

DECODING DATA

A Complete Guide
to Business
Intelligence

Amara Atif, MA Qureshi, Yang Yang,
Meena Jha, Florence Mwagwabi &
Marcella Papini

EDITED BY AMARA ATIF



Decoding Data: a Complete Guide to Business Intelligence

DECODING DATA: A COMPLETE GUIDE TO BUSINESS INTELLIGENCE

A Complete Guide to Business Intelligence

AMARA ATIF; M.A. QURESHI; YANG YANG; MEENA JHA; FLORENCE
MWAGWABI; AND MARCELLA PAPINI

UTS ePress
Sydney



Decoding Data: a Complete Guide to Business Intelligence Copyright © 2026 by The authors is licensed under a [Creative Commons Attribution-NonCommercial 4.0 International License](https://creativecommons.org/licenses/by-nc/4.0/), except where otherwise noted.

This book was published by UTS ePress via the Council of Australasian University Librarians Open Educational Resources Collective. The online version is available at <https://utsepress.lib.uts.edu.au>

Disclaimer

Note that corporate logos and branding are specifically excluded from the [Creative Commons Attribution Noncommercial Licence](https://creativecommons.org/licenses/by-nc/4.0/) of this work, and may not be reproduced under any circumstances without the express written permission of the copyright holders.

The views and opinions expressed in this book are those of the authors and do not necessarily reflect the official policy or position of any agency or organisation. The information provided in this book is for general educational purposes only and is not intended to be a substitute for professional advice. While every effort has been made to ensure the accuracy and completeness of the information contained herein, the authors and publishers assume no responsibility for errors, omissions, or contrary interpretation of the subject matter. The reader is advised to consult with appropriate professionals before making any decisions based on the content of this book. The authors and publishers shall not be liable for any loss or damage caused or alleged to be caused directly or indirectly by the information or ideas contained in this book.

Copyright

Decoding Data: A Complete Guide to Business Intelligence by Amara Atif, M.A. Qureshi, Yang Yang, Meena Jha, Florence Mwagwabi and Marcella Papini is licensed under a Creative Commons Attribution Noncommercial 4.0 Licence by the authors.

Recommended citation: Atif, A., Qureshi, M.A., Yang, Y., Jha, M., Mwagwabi, F. & Papini, M. (2026). *Decoding Data: A Complete Guide to Business Intelligence*. UTS ePress. <https://utsepress.lib.uts.edu.au/>.

Recommended attribution: *Decoding Data: A Complete Guide to Business Intelligence* by Amara Atif, MA Qureshi, Yang Yang, Marcella Papini, Florence Mwagwabi & Meena Jha is licensed under a Creative Commons Attribution Noncommercial 4.0 Licence by UTS ePress.

The book cover was created by UTS ePress using Canva and is licensed under a [Creative Commons Attribution 4.0 International Licence](https://creativecommons.org/licenses/by-nc/4.0/). The cover photo of a light bulb is by [Marek Piwnicki](https://unsplash.com/photos/1jZ8v8v8v8v) on [Unsplash](https://unsplash.com/) and licenced for reuse under the [Unsplash licence](https://unsplash.com/).

CONTENTS

Acknowledgement of Country	vii
About this book	1
About the authors	3
Acknowledgements	vi
Foreword by Karen Whelan and Paul Kennedy	vii
Foreword by Jose M Merigo	viii

Main Body

1. Data to Knowledge Continuum	10
Amara Atif	
2. Rethinking Business Decisions: The Synergy of Decision Support System, Analytics, Business Intelligence, Artificial Intelligence, and Data Science	52
Amara Atif and M.A. Qureshi	
3. Architecture of Business Intelligence	82
Amara Atif	
4. Statistical Modelling for Business Analytics	102
Yang Yang	
5. Data/Information Visualisation	138
Florence Mwagwabi	
6. Information Dashboards	213
Amara Atif	
7. Predictive Analytics	234
Marcella Papini	
8. Prescriptive Analytics	278
Meena Jha	
9. Bringing It All Together	302
<i>A Summary of Key Concepts in Business Intelligence</i>	
Amara Atif	

Review Statement	310
Accessibility Statement	312
Versioning History	314
Feedback and corrections	315

ACKNOWLEDGEMENT OF COUNTRY

The authors pay their deepest respect to the Indigenous Elders and Communities upon whose lands we all live, work and learn.

The University of Technology Sydney recognises the Gadigal people of the Eora Nation, the Boorooberongal people of the Dharug Nation, the Bidiagal people and the Gamaygal people, upon whose ancestral lands our university stands. We would also like to pay respect to the Elders both past and present, acknowledging them as the traditional custodians of knowledge for these lands.

Torrens University Australia acknowledges the Australian Aboriginal and Torres Strait Islander peoples as the first inhabitants of the nation and the traditional custodians of the lands where we live, learn and work. Torrens University Australia respectfully acknowledges their Ancestors and Elders, past and present.

The University of Newcastle acknowledges the traditional custodians of the lands within our footprint areas: Awabakal, Darkinjung, Biripai, Worimi, Wonnarua, and Eora Nations. We also pay respect to the wisdom of our Elders past and present.

CQUniversity acknowledges the Traditional Owners and Custodians of the land on which work and learn. We recognise their deep and ongoing connection to this land, its seas, skies, and waterways, and the role they continue to play in caring for Country. We also honour their right to self-determination, acknowledging the importance of their guidance and leadership in shaping our futures. We pay our respects to Elders past, present, and emerging, honouring their wisdom, leadership, and the cultural traditions they uphold. May their strength and connection to the land guide us.

Murdoch University is situated on the lands of the Whadjuk and Binjareb Noongar people. We pay our respects to their enduring and dynamic culture and the leadership of Noongar elders past and present. The boodjar (country) on which Murdoch University is located has, for thousands of years, been a place of learning. We at Murdoch University are proud to continue this long tradition.

ABOUT THIS BOOK

Welcome to *Decoding Data: A Complete Guide to Business Intelligence*!

How this project came to life

This textbook was created by Dr Amara Atif, who wanted to make learning about business intelligence more open, practical, and accessible. Drawing on her passion for open educational practice, she developed this resource to be freely available to her students – and to anyone interested in understanding how data is used to support real business decisions. Dr Atif has also brought together a team of expert contributors to guide you through key ideas, tools, and case studies that bring business intelligence to life.

Is this textbook right for you?

Audience

This book is designed primarily for Australian university students studying business intelligence (BI), data analytics, or related fields. However, it is also useful for professionals who want to build or refresh their understanding of business intelligence in practice.

Key features

One of the strengths of this textbook is its focus on Australian case studies and business scenarios. These examples help you see how concepts apply in familiar, real-world contexts, rather than just in theory.

Most chapters also include interactive activities at the end, giving you a chance to reflect on what you've learned and check your understanding as you go. You're encouraged to use the references and suggested readings to deepen your knowledge and explore topics that interest you further.

How the chapters connect

The eight chapters in this text form a single, integrated system. See the structure and concepts covered in each:

Table 9.1: Chapter focus and core questions answered

Chapter	Title	Core question answered	Key concepts
Chapter 1	Data to Knowledge Continuum	What is data and how do we prepare it?	DIKW hierarchy, data taxonomy, data quality metrics, pre-processing
Chapter 2	Rethinking Business Decisions: The Synergy of Decision Support System, Analytics, Business Intelligence, Artificial Intelligence, and Data Science	How do organisations use data to make better decisions?	DSS evolution, drivers of change, analytics/BI/AI/data science synergy
Chapter 3	Architecture of Business Intelligence	How does the infrastructure move data to decision-makers?	BI components, OLTP vs OLAP, ETL, data warehouses, architecture frameworks
Chapter 4	Statistical Modelling for Business Analytics	How do we analyse data rigorously?	Data types, distributions, correlation, statistical reasoning
Chapter 5	Data/Information Visualisation	How do we represent data meaningfully?	Visualisation history, design principles, chart taxonomy, Gestalt principles, data storytelling
Chapter 6	Information Dashboards	How do we communicate insights effectively?	Info visualisation vs visual analytics, report types, dashboard design, data storytelling
Chapter 7	Predictive Analytics	What will happen?	Supervised/unsupervised learning, data mining, text mining, web & social analytics
Chapter 8	Prescriptive Analytics	What should we do about it?	Optimisation, simulation, decision theory, heuristics, ethical considerations

For a longer summary of each chapter, jump to [Chapter 9: Bringing it all together](#).

About currency of content

This open textbook reflects knowledge and practice at the time it was written (2024-2026). Some information may change over time. Readers are encouraged to check current research, policies, and professional guidance to ensure they are using the most up-to-date information.

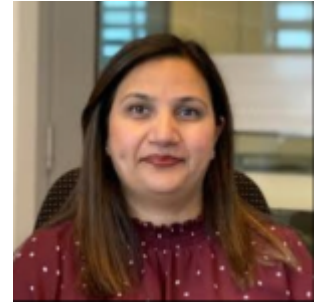
Get in touch

If you have any questions or comments, please contact the Editor Amara.Atif@uts.edu.au or the Publisher utsepress@uts.edu.au.

ABOUT THE AUTHORS

Amara Atif (Lead author, editor, book project manager)

Dr Amara Atif is a Senior Lecturer in the School of Computer Science, Faculty of Engineering and IT (FEIT), at the University of Technology Sydney (UTS), Australia. Amara is an expert in educational technology, with a comprehensive skill set that includes research and development in the design of technology-enhanced learning environments, human-computer interaction, artificial intelligence in education, and learning analytics/educational data mining. At UTS, she developed the undergraduate subject “Business Intelligence”, integrating culturally informed Indigenous knowledge systems through the FEIT Indigenous Graduate Attribute. She has received multiple UTS Vice Chancellor’s Learning and Teaching Awards, including recognition for data-informed teaching practices and inclusive assessment. Passionate about the accessibility and sustainability of educational resources, Amara is an advocate for Open Educational Resources (OER). *Decoding Data* continues this mission, bringing rigorous Business Intelligence concepts to learners through an open, practitioner-oriented resource grounded in Australian business contexts.



M.A. Qureshi

Dr. M.A. Qureshi is an expert in Software Modelling, Meta-modelling, and Modelling Languages, holding a PhD in Software Engineering. With years of combined experience in both the IT industry and academia, he has worked with leading organisations to implement innovative BI solutions that drive data-driven decision-making, optimise business processes, and enhance organisational performance.



As an academic, he employs constructive and effective teaching methods that foster a collaborative learning environment. He takes a futuristic approach to learning, engaging students through both theory and practical application.

Additionally, as a certified business analyst (ECBA-IIBA), he integrates his unique blend of industry experience and academic insight to provide a comprehensive, real-world perspective on Business Intelligence. This enables him to make complex concepts accessible to both beginners and seasoned professionals.

Yang Yang

Dr Yang Yang is a lecturer at the University of Newcastle, Australia, with expertise in demography forecasting, climate data analysis, and functional time series modelling. He holds a PhD in Statistics from ANU and has published extensively in top-tier journals.



Meena Jha

With a degree in Electronics and Communication Engineering and a Master and Ph.D. in the area of Computer Science, Associate Professor Meena Jha is a researcher and educator in the field of computer science and IT in the School of Engineering and Technology (SET), at Central Queensland University (CQUniversity) in Sydney, Australia. Her research interests include in particular Big Data and Artificial Intelligence, Expert Systems, Business Intelligence, Real-Time Analytics, Legacy System Modernization, the Internet of Things (IoT), and Academic Integrity in education. She has developed courses on Information Systems Analysis and Business Intelligence using Big Data and Data Science. She is also working in the area of learning analytics, which is an emerging field and uses sophisticated analytic tools to improve learning and education. It draws from and is closely tied to, a series of other fields of study including business intelligence, web analytics, academic analytics, educational data mining, and action analytics. She is passionate about women in STEM and working on how to improve and enhance women's power in the STEM field.



Florence Mwagwabi

Dr Florence Mwagwabi is a Senior Lecturer in Information Technology and the Academic Chair of Transnational Education (IT) at Murdoch University, Singapore. She has taught Business Intelligence application development for more than nine years and has led curriculum design that integrates real-world datasets, BI tools and current industry trends. Her teaching spans databases and systems analysis, data communications, business intelligence, machine learning foundations and information security across undergraduate and postgraduate programs. Her research in this area includes the use of SDG data in analytics education, the impact of real-world projects on BI skill development and the design of data literacy programs.



Her broader research spans cybersecurity, AI applications in learning, user security behaviour and ICT use across different population groups. Her work has been featured in national media and supported by competitive research funding. She is a Fellow of the Higher Education Academy (FHEA), a member of IEEE and the Australian Computer Society (ACS) and a recipient of Murdoch University's Citation for Excellence in Enhancing Student Learning.

Marcella Papini

Dr Marcella Papini is an expert in mathematical optimisation with a strong commitment to advancing the United Nations Sustainable Development Goals (SDGs). She actively aligns her research, teaching, and engagement efforts with the SDGs, using mathematics as a tool to support evidence-based decision-making for a more sustainable and equitable world.



Marcella has led several publicly and privately funded research projects focused on improving the management of operations across diverse domains—optimising performance not only in terms of economic costs but also environmental and social impacts. Her work has been supported by prestigious institutions such as Australian universities, Transport for NSW, and the Brazilian National Research Council (CNPq). Her research has resulted in high-quality publications in leading international journals, and she regularly collaborates with academic partners around the globe. She brings a truly international perspective to her work, having studied, taught, and conducted research in Brazil, Australia, and Germany.

ACKNOWLEDGEMENTS

We would like to thank the following for their support of this project.

Publishing support

Thank you to UTS, UTS Library and UTS ePress for their support in getting this book from ideation to publication!

- UTS Library: Anna Troiano and Patrick Tooth for copyright advice
- UTS ePress: Marie-Louise Taylor for copyediting and proofreading; Helen Chan for book project management
- UTS School of Computer Science: Luke Mathieson for support of this publication

A special thank you to the Council of Australasian University Librarians (CAUL) for their support in getting this book published. The [CAUL OER Collective](#) provided extensive advice, guidance and publishing infrastructure that helped us create a book that we can all be proud of.

Design

The authors would like to acknowledge and thank Khadeejah Atif who contributed to the design of figures used in Chapter 1 and Chapter 2.

Author Acknowledgements

Dr Florence Mwagwabi

First I would like to thank Dr Amara Atif for initiating this much needed project and for inviting us to contribute to this open educational textbook in Business Intelligence with a unique focus on Australian data cases. I would also like to acknowledge my fellow authors without whom this project would not be possible and their shared commitment to create this open educational resource.

I would like to thank UTS, UTS Library and UTS ePress for their support in moving this book from ideation to publication. I am grateful to Helen Chan from UTS ePress for guidance on copyright and publication processes and to the Editor Dr Amara Atif for insightful editorial direction. I also thank the anonymous reviewers whose feedback strengthened the quality of this work.

Finally, I would like to thank Murdoch University Singapore for supporting me toward this book.

FOREWORD BY KAREN WHELAN AND PAUL KENNEDY

Business intelligence (BI) turns raw data into actionable insights for decision-making. It is vital for contemporary organisations to make data-informed and data-driven decisions for a multitude of processes including enhancing productivity, maintaining a competitive advantage and identifying new opportunities, just to name a few.

Now more than ever before, we swim in an ocean of data. Organisations need professionals trained in business intelligence to make sense of this data and to identify actionable insights for domain experts.

This new Open Educational Resource, “Decoding Data: A Complete Guide to Business Intelligence”, provides a thorough grounding in the tools and techniques, and the art and science of business intelligence. It has been carefully developed by a team of academics across Australia based on their experiences in teaching BI in multiple institutions. It covers all the key topics in the area, supported by exercises, interactive tools and real-world case studies. Notably, the case studies and examples derive from the Australian context and business environment, a useful addition for Australian students and practitioners.

Educators in business intelligence and related fields will find this a valuable resource for their classrooms. Students will readily engage with this clear and practical introduction as they make their way in their business intelligence journey.

**Professor Karen Whelan, Associate Dean (Teaching and Learning) & Professor Paul Kennedy,
Head, School of Computer Science**

Faculty of Engineering and Information Technology, University of Technology Sydney

FOREWORD BY JOSE M MERIGO

Decoding Data: A Complete Guide to Business Intelligence, edited by Amara Atif (University of Technology Sydney), is a comprehensive resource in the field of business intelligence covering key topics for the analysis of information and decision-making in business and management contexts. The book equips readers with a practical set of methods and techniques to analyse a wide range of situations that inform research and business decisions.

A standout feature of this resource is its grounding in Australian case studies and business scenarios, offering localised insights that are highly relevant to students and professionals operating in the Australian context. Written in accessible, student-friendly language, each chapter builds foundational knowledge progressively, supported by key takeaways that reinforce core concepts and help readers consolidate their learning. This book is an invaluable resource for any student or practitioner interested in business intelligence and the related concepts underpinning efficient, data-informed decision-making. It is particularly well-suited to those new to the field, offering a clear and approachable introduction to a research and practice area that has grown significantly in recent years.

The book starts with an emphasis on basic concepts that sustain the methods developed in business intelligence. Particularly, it considers four fundamental issues: data, information, knowledge and wisdom. In this context, it is important to clarify that data is the first step in the process of collecting observations, while information is the second step when the data has been correctly organised. Knowledge is the subsequent step where the information is studied with the aim of reaching a deeper understanding of the specific problem considered. Finally, wisdom is the process of making intelligent decisions based on the data and information available and the accumulated knowledge.

The second chapter focuses on business decision making methods considering the modern theories that have been developing during the last years around analytics, business intelligence, artificial intelligence and data science. The third chapter studies the architecture of business intelligence to provide a deeper understanding of the processes needed for performing an adequate assessment of the problems considered. The subsequent chapters examine different fundamental issues needed for building a complete analysis of a business intelligence scenario.

With the publication of this book, the authors reach a significant milestone in their academic career providing a very complete overview of the field of business intelligence that will be of great usefulness for the future generations of decision makers and researchers. I have enjoyed reading the book and I strongly recommend it to anyone interested in business intelligence.

Finally, I would like to thank Amara Atif and the rest of the contributing authors, for having invited me to provide the preface of their book. I wish you a lot of success with this book and I am very sure it will be of great interest for the current and future generations of students and decision makers.

Sincerely,

Professor José M. Merigó
School of Computer Science
Faculty of Engineering and Information Technology
University of Technology Sydney
Australia
Sydney, 28 November 2025

1. DATA TO KNOWLEDGE CONTINUUM

Amara Atif

Learning Objectives (LO)

By the end of this chapter, you will be able to:

LO1 Define data and distinguish it from information, knowledge and wisdom while exploring the progression from raw data to actionable insights.

LO2 Recognise the role of data in supporting business decisions across strategic, tactical and operational levels.

LO3 Describe the taxonomy of data, including its classification and types.

LO4 Understand the concept of analytics-ready data and the importance of data quality metrics for accurate and effective analysis.

LO5 Explore the process of data pre-processing, detailing its sub-tasks and methods with real-world examples from various industries.

1 Foundations of data: Building the basics for analytics and business intelligence

This chapter provides an overview of the fundamentals of data as a basis for analytics and business intelligence. First, it defines data and distinguishes it from information, knowledge and wisdom by describing the process that starts with raw data and ends with actionable insight. Then, it covers the role that data plays in supporting business decisions and how it impacts the firm at the strategic, tactical and operational levels. The chapter introduces a simple taxonomy of data by discussing data classification and types – this will help in building a clear understanding of its structure and usage. Further, the chapter discusses analytics-ready data, with a major focus on data quality metrics, something that becomes essential to ensure correct and effective analysis. It then goes on to cover data pre-processing, its most important sub-tasks and methods, using practical examples for various industries to illustrate the essential role it can play in representing data ready for meaningful insights.

Understanding these concepts is important because data forms the foundation of all analytics

and business intelligence processes. Without a clear grasp of how to structure, assess, and prepare data effectively, organisations risk making decisions based on incomplete, low-quality, or misleading insights – something no business can afford in today’s data-driven world.

1.1 The nature of data

1.1.1 Data

The term **data (singular: datum)** refers to **a collection of raw, unprocessed facts, measurements or observations.**

Data can take many forms, including numbers, text, symbols, images, audio, video, and more. In its original form, data is unorganised and lacks context, making it difficult to interpret without further processing. Data represents the most basic level of abstraction, from which higher levels, such as *information*, *knowledge* and *wisdom*, are developed.

Data is classified based on *how it is organised and stored* (structured, semi-structured or unstructured), the *types of information it represents* (quantitative or qualitative), its *purpose* (primary or secondary) and *origin/source* (internal, external or generated data). For more on how data is classified, see Section 1.3 on the taxonomy of data.

1.1.2 From data to information

Information is data that has been **processed, organised or structured** to **provide meaning and context**, making it more understandable and relevant to specific purposes.

Unlike raw data, information serves a particular goal, often answering specific questions or helping with decision-making.

For example, a report summarising monthly sales totals, a demographic profile of customers or a chart illustrating website traffic over time takes raw data and organises it into a meaningful format. This organised data can now provide insights that address business questions and support objectives.

In business, converting raw data into information allows organisations to identify trends, make comparisons and gain insights that guide tactical and operational decisions.

1.1.3 From information to knowledge

Knowledge is created **by analysing and synthesising information to identify patterns, relationships or insights** that contribute to a deeper understanding.

It connects various pieces of information, enabling interpretation based on experience, expertise or reasoning. Knowledge goes beyond basic facts, providing meaningful insights that often lead to actionable recommendations for strategic decisions.

For example, recognising that a particular demographic consistently purchases a specific product or identifying a seasonal trend in sales data illustrates how knowledge can be used to understand customer behaviour patterns. This understanding informs strategic decisions, such as adjusting marketing efforts or planning inventory to meet customer needs.

In daily life, knowledge might be applied when someone notices that they feel more energised after a good night's sleep and chooses to adjust their bedtime to improve daily energy levels.

1.1.4 From knowledge to wisdom

Wisdom represents the **highest level in the progression from data to decision-making**. It is the ability to make sound decisions and judgements based on ***accumulated knowledge and experience***.

Wisdom involves applying knowledge with insight, ethics and a long-term perspective, carefully weighing the potential impacts of decisions. It balances short-term benefits with long-term sustainability, guiding actions that are both effective and responsible.

For example, a business leader might use historical knowledge and ethical considerations to develop a sustainable strategy that respects both profitability and environmental impact. This could involve implementing policies that foster customer trust or making decisions that balance profit with societal well-being.

In daily life, wisdom is reflected in choices such as investing time in developing a lifelong skill rather than seeking a short-term reward. Wisdom is about knowing not only “what” to do, but also “why” and “how” to make decisions that lead to positive, long-term outcomes.

1.1.5 The Data-Information-Knowledge-Wisdom (DIKW) Hierarchy

This **progression from data to wisdom** is commonly referred to as the **DIKW hierarchy** (Ackoff, 1989). Over the years, the DIKW theory has evolved and has been applied across various disciplines. In the field of business intelligence, it highlights the transformation of raw data into meaningful insights, which ultimately support wise and informed decision-making. Data serves as the foundational building block – once processed, it becomes information. When this information is analysed and synthesised, it evolves into knowledge. Finally, when knowledge is applied with ethical and long-term considerations, it becomes wisdom. Each stage in this hierarchy adds value, turning raw, unorganised facts into powerful tools for insightful and responsible decision-making that drives business success.

PYRAMID OF DIKW THEORY IN BUSINESS INTELLIGENCE

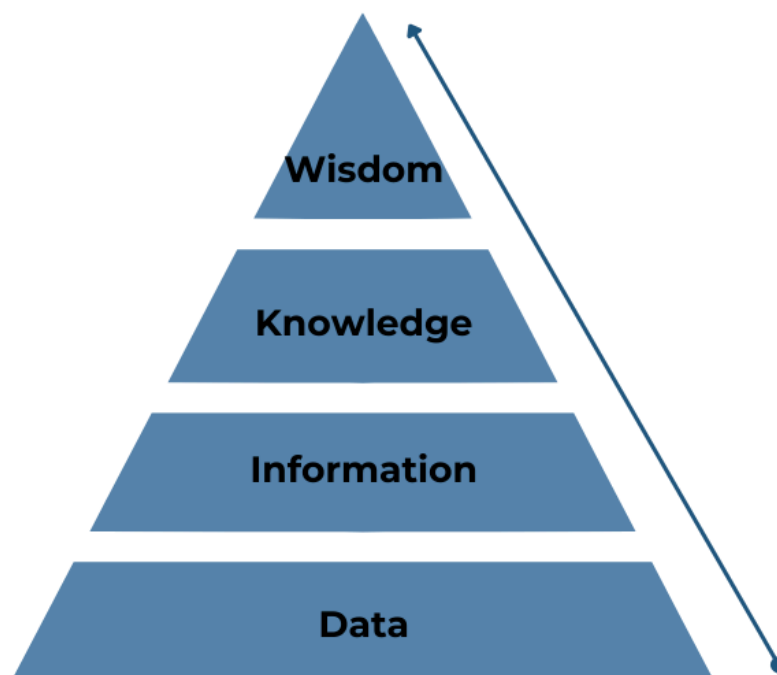


Figure 1.1: Classic DIKW Pyramid

Alternative Text for Figure 1.1

The Classic DIKW pyramid shows the upwards progression of 'Data' at the bottom of the pyramid, to the next level which is 'Information', followed by 'Knowledge' and finally 'Wisdom'.

1.1.6 Role of data, information, knowledge and wisdom in business intelligence

Below are two business scenarios from different industries that demonstrate how data, information, knowledge and wisdom are used in the analytics process to enhance decision-making. **Scenario 1** describes how improving customer experience in retail is essential for building loyalty and satisfaction through personalised services and streamlined operations.

Scenario 1: Enhancing customer experience in a retail chain

A large retail chain wants to improve its in-store and online customer experience to boost customer satisfaction and loyalty. They decide to use analytics to understand customer behaviour, identify improvement areas and make strategic adjustments.

- **Data:** The retail chain collects a wide range of raw data, including transaction records, customer demographics, website click-through rates and social media engagement. They also gather customer feedback from surveys and in-store feedback forms. This data forms the foundation, providing detailed, unprocessed facts about customer behaviours and interactions.
- **Information:** The analytics team processes and organises this data into meaningful insights. They create reports summarising key metrics, such as popular products, peak shopping times and regional purchasing patterns. Additionally, they analyse survey responses to identify trends in customer preferences, such as a demand for eco-friendly products or faster checkout processes. This organised information helps them understand what customers are buying, when, and what preferences or pain points they have.
- **Knowledge:** By analysing the information, the retail chain gains knowledge about customer behaviour patterns and preferences. They observe, for example, that customers in urban areas prefer sustainable product options and that online shoppers abandon carts more frequently when shipping costs are high. This knowledge allows them to make data-driven decisions, such as adjusting product offerings in specific locations or creating promotions for online shoppers to reduce cart abandonment.
- **Wisdom:** The leadership team uses this knowledge to make strategic, ethically sound decisions that align with the company's values of sustainability and customer satisfaction. They decide to introduce a loyalty program that rewards customers for eco-friendly purchases, balancing short-term profit with long-term brand loyalty and

environmental impact. Additionally, they opt to subsidise shipping costs in regions with high cart abandonment rates, prioritising customer experience over immediate revenue gains. These decisions reflect wisdom by considering not only profitability but also long-term customer trust and brand reputation.

In this scenario, **data** provides the foundational facts, **information** organises these facts into trends and patterns, **knowledge** enables actionable insights about customer behaviour, and **wisdom** guides strategic decisions that align with both business goals and ethical considerations. Together, these elements allow the retail chain to improve customer experience in a way that is both effective and aligned with their brand values.

Scenario 2 describes how enhancing patient care in hospitals involves leveraging advanced technologies and efficient workflows to ensure better outcomes and a compassionate healthcare environment.

Scenario 2: Improving patient care in a hospital

A hospital wants to enhance patient outcomes and streamline its operations to deliver faster, more efficient care. To achieve this, it uses analytics to understand patient needs, optimise resource allocation and guide strategic decisions.

- **Data:** The hospital collects a wide range of raw data, including patient demographics, medical histories, vital signs, lab test results and feedback from post-treatment surveys. Operational data, such as wait times in the emergency room, bed occupancy rates and staff schedules, are also gathered. This raw data serves as the foundation, capturing detailed and unprocessed information about both patient care and hospital operations.
- **Information:** The analytics team organises and processes this data to provide insights that are more actionable. They create reports that highlight patterns, such as the average wait times for different departments, peak admission times and common health issues in certain demographics. They also analyse feedback from patients to identify areas for improvement, such as communication during treatment or responsiveness of medical staff. This organised information provides a clearer picture of hospital performance and patient needs.
- **Knowledge:** By analysing the information, the hospital gains knowledge about operational inefficiencies and patient care patterns. For example, the team discovers that wait times are highest in the emergency room during certain hours and that certain patient demographics are more likely to require readmission. This knowledge enables

hospital administrators to make informed decisions, such as allocating more staff during peak hours in the emergency room or developing post-treatment plans for high-risk demographics to reduce readmissions.

- **Wisdom:** Hospital leadership uses this knowledge to make ethical, strategic decisions that balance patient care with operational efficiency. Recognising the importance of equitable care, they decide to increase staffing and resources during peak hours to reduce emergency room wait times, even if it means higher short-term costs. They also implement a patient follow-up program for high-risk groups to provide proactive care and reduce readmissions, prioritising patient well-being over immediate financial gains. These decisions reflect wisdom, as they consider both the hospital's mission to provide quality care and the long-term benefits of improved patient satisfaction and health outcomes.

In this scenario, **data** provides the raw details about patients and operations, **information** organises this data to reveal patterns, **knowledge** offers actionable insights for targeted improvements, and **wisdom** guides strategic decisions that align with the hospital's values and long-term objectives. Together, these elements enable the hospital to enhance patient care, optimise resources and fulfil its mission responsibly and sustainably.

1.2 Overview of data's role in business decisions

1.2.1 Importance of clean, structured and analytics-ready data

Data is powerful only when it is prepared effectively for analysis. Raw data, while valuable, often contains noise, inconsistencies and missing information that can mislead/skew the results if not addressed. Clean, structured and analytics-ready data is essential for accurate analyses, informed decisions and meaningful business insights. Without proper data preparation, businesses risk making decisions based on incorrect or incomplete information, which can result in less effective strategies and missed opportunities.

1.2.2 Data as the foundation for business intelligence

For any organisation to leverage business intelligence tools and analytics effectively, it must first establish a robust foundation of analytics-ready data. Business intelligence systems rely on data to generate visualisations, reports and dashboards that provide managers and executives with a quick, accurate snapshot of performance.

1.2.3 The shift to data-driven decision-making

Traditionally, business decisions were often made based on intuition, past experiences or anecdotal

evidence. However, with the rise of big data and analytics, there has been a shift towards data-driven decision-making, where objective insights derived from data guide strategic choices. This approach allows organisations to minimise biases, improve accuracy and make more reliable predictions about the future.

Data-driven decision-making is now widely adopted across various sectors, allowing companies to harness their data to understand customer preferences, identify trends, optimise operations and drive innovation. By incorporating data insights into the decision-making process, businesses are better equipped to respond to market changes and consumer demands with agility and confidence.

1.2.4 How data drives business insights

Figure 1.2 illustrates how data analytics drives business success across key areas in the retail industry, including trend identification, future outcome prediction, process optimisation and customer experience enhancement.

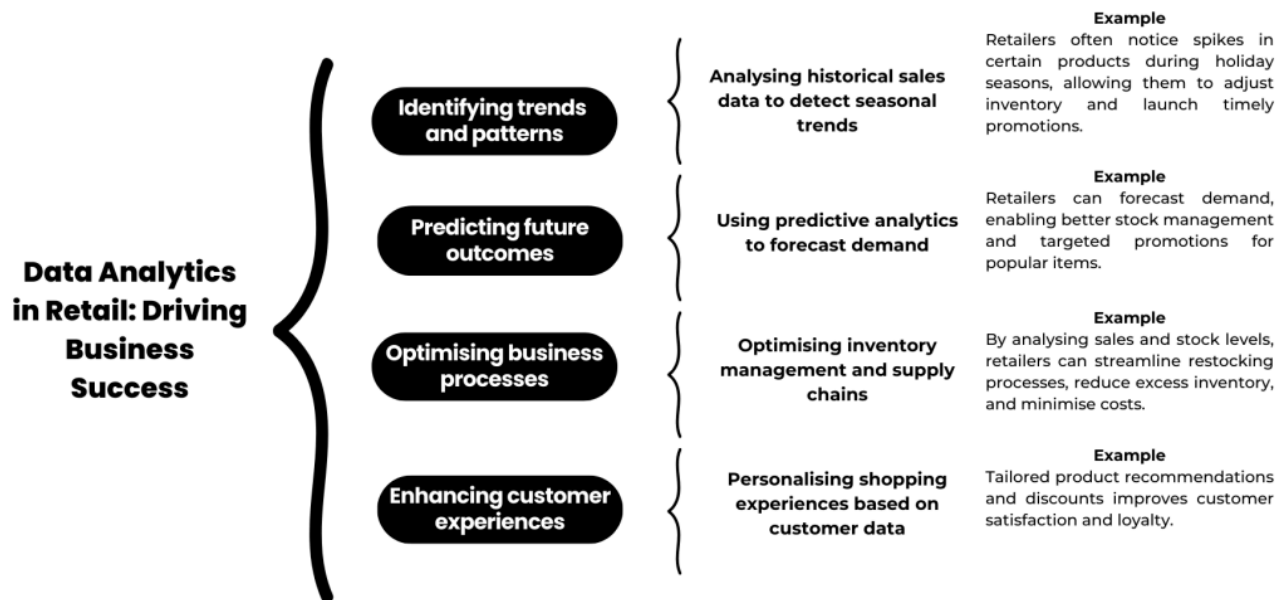


Figure 1.2: Data analytics in retail

Alternative Text for Figure 1.2

Identifying trends and patterns: Data analytics helps businesses detect patterns in consumer behaviour, operational efficiency and market dynamics. For example, analysing historical sales data may reveal seasonal trends, helping companies adjust their inventory and marketing strategies accordingly.

Predicting future outcomes: Predictive analytics uses data from the past and present to forecast future trends, risks and opportunities. For instance, by analysing data on customer

purchasing habits, businesses can predict which products are likely to be popular in the future, enabling proactive stock management and targeted marketing.

Optimising business processes: Data analytics supports process optimisation by identifying inefficiencies, bottlenecks or areas for improvement. In manufacturing, data from production lines can help reduce downtime and enhance quality control, while in customer service, analytics can streamline workflows, reducing response times and improving satisfaction.

Enhancing customer experiences: Data allows businesses to gain a deep understanding of customer preferences, behaviours and needs. Through data analytics, companies can personalise marketing efforts, recommend products based on customer history and improve customer service, all of which enhance the overall customer experience.

1.2.5 The role of data in driving competitive advantage

Organisations that effectively use data gain a significant competitive edge. By turning raw data into actionable insights, businesses can anticipate trends, adapt to market changes and make quicker, better-informed decisions than their competitors. In Australia, companies across various industries are using data analytics to drive success and innovation, focusing on improving customer experiences, streamlining operations (making business processes more efficient by reducing unnecessary steps, cutting down on waste and improving workflow) and developing better products.

Examples of data-driven competitive strategies in Australian industries

The examples below from Australian industries demonstrate the strategic value of data in enhancing customer personalisation, operational efficiency and product development. By integrating data analytics into core operations, Australian companies are better positioned to adapt, innovate and lead within their markets.

Customer personalisation: Australian retailers, such as Woolworths and Coles, use customer data from loyalty programs to recommend products based on previous purchases, tailor promotions and improve customer satisfaction. Woolworths, for instance, uses its *Everyday Rewards* program to analyse shopping habits and offer personalised discounts, boosting both engagement and sales. Similarly, Coles uses its *Flybuys* program to provide targeted offers that align with customer preferences, enhancing loyalty and driving sales growth.

Operational efficiency: Australian logistics company Toll Group uses data analytics to

streamline delivery routes, reduce fuel consumption and improve delivery times. By analysing delivery data, Toll can optimise its supply chain logistics, minimising operational costs while enhancing customer satisfaction through timely deliveries.

Product development: Australian telecommunications giant Telstra uses customer usage data to understand product needs, identify service improvements and introduce new features that align with customer preferences. By analysing patterns in network use and customer feedback, Telstra can improve its offerings, refine its service plans and stay competitive in a dynamic market.

1.2.6 The link between data quality and decision quality

The quality of decisions is only as good as the quality of data driving them. Inaccurate, incomplete or outdated data can lead to misguided decisions, whereas high-quality data allows reliable and actionable insights. Organisations must establish rigorous data quality standards and continuously monitor and improve their data sources to support high-stakes decision-making.

Table 1.1 demonstrates the impacts of poor data quality.

Table 1.1: Impacts of poor data quality

Type of impact	Impacts
Incorrect assumptions	Decisions based on faulty data can lead to misinterpretations of customer behaviours, misalignment with market demands or improper allocation of resources.
Lost opportunities	Poor data quality can hide potential opportunities, such as emerging customer segments or untapped market needs.
Risk exposure	In industries such as finance or healthcare, low-quality data can increase risks, leading to compliance issues or compromised customer safety.

A hypothetical business scenario: GreenBite Organics' product launch

Context

A large retail chain, **GreenBite Organics**, plans to launch a new product line of organic snacks based on market research and sales data from its existing product portfolio. The success of this

initiative heavily depends on the quality of the data used to make decisions about product selection, pricing and marketing strategies.

Scenario description

- **Incorrect assumptions:** The company relies on old sales data that shows a high demand for organic snacks in a specific region. However, if they had used updated data, they would have seen that demand for these snacks was actually decreasing. As a result, the company misunderstands what customers want and spends too much on the product for that region, ending up with unsold stock and losing money.
- **Lost opportunities:** Because the company's data was not well-integrated across departments, it missed noticing a new customer group, i.e., health-conscious young professionals in cities. This group was buying organic products in small amounts, but the trend was hidden due to inconsistent data formats and was overlooked during analysis. As a result, the company did not create targeted marketing campaigns for this promising group, missing a chance to grow in a valuable market.
- **Risk exposure:** Mistakes in the product labelling caused by inaccurate data result in wrongly claiming that one of the snacks is allergen-free. In the health-conscious food industry, this puts the company at risk of breaking regulations and facing customer lawsuits. These issues harm the company's reputation and lead to costly penalties.

Outcome

This scenario highlights how poor data quality can lead to incorrect decisions, missed opportunities for growth and increased risks. To address these challenges, GreenBite Organics adopts robust data quality standards, focusing on timely updates, accurate analysis and consistency across departments. By implementing these measures, the company ensures more reliable decision-making and builds a foundation for sustained business success.

1.3 The taxonomy of data

In this section, we explore the fundamental classification of data, including *how it is organised and stored* (structured, semi-structured or unstructured), the *types of information it represents* (quantitative or qualitative), its *purpose* (primary or secondary) and *origin/source* (internal, external or generated data). We also discuss *data granularity*, which refers to the level of detail within the data. Understanding these perspectives is important for effective data management and analysis, as they play a key role in supporting business operations, generating insights and driving analytics.

In addition to these classifications, data can also be categorised by:

- **Levels of measurement:** Data can be categorised by levels of measurement – nominal, ordinal, interval and ratio. Each level adds a layer of understanding to data, influencing how data values can

be interpreted and analysed. These levels apply to both quantitative and qualitative data and will be explored in more detail in Chapter 4: Statistical Modelling for Business Analytics.

- **Usage and role in analytics:** Data can be categorised as descriptive, predictive or prescriptive, each with unique functions and applications in business intelligence and analytics. These types of data will be discussed in more detail in Chapter 4: Statistical Modelling for Business Analytics, Chapter 7: Predictive Analytics and Chapter 8: Prescriptive Analytics. By classifying data based on its role in analytics, organisations can focus their analysis to achieve specific objectives, whether it's understanding historical trends, forecasting future events or determining optimal strategies. This approach enables businesses to make more proactive, competitive and effective decisions.

Big data and its role in the taxonomy of data

Big data is not a separate type of data in the traditional taxonomy but rather **a category or characteristic** that encompasses all types of data. It serves as an **umbrella term** that includes structured, semi-structured, unstructured, quantitative, qualitative, internal, external, primary, secondary or generated, all of which are managed and analysed at scale.

In the context of a data taxonomy, big data is better understood as a **characteristic or context** that describes data environments with specific attributes: **high volume** (large-scale datasets), **high variety** (diverse types of data), **high velocity** (rapid data generation and processing), and often **lower veracity** (uncertainty or inconsistency in data quality). These characteristics distinguish big data by its **complexity and scale** rather than by its format or origin.

Instead of being a distinct type of data, big data integrates various data structures and formats, offering a **comprehensive landscape for advanced analytics and business intelligence**. It highlights the challenges and opportunities of managing and analysing vast, diverse datasets in real-time or near real-time scenarios.

1.3.1 Types of data in a business context

In business, data comes in various forms, each playing an important role in helping companies make informed decisions. Data can also vary in its level of granularity or the degree of detail it captures, which affects its usability in analysis.

1.3.1.1 Structured and quantitative data

Quantitative data, being numerical and measurable, is often stored as structured data. Examples include sales figures, customer counts and revenue records, which are typically organised into rows and

columns within relational databases. Structured data formats (e.g., spreadsheets and databases) facilitate easy calculation of key performance indicators (KPIs), trend tracking and statistical analysis, aligning closely with the goals of quantitative data.

Granularity further defines the level of detail in structured quantitative data. High granularity provides specific and detailed information, such as individual transaction records that include data like purchase time, items bought and customer details. This level of detail is useful for identifying precise patterns and trends. In contrast, low granularity involves aggregated data, such as monthly sales totals, which offer a broader and less detailed view. High-granularity data is ideal for deep analysis, while low-granularity data is better suited for observing overall trends and making strategic decisions. The appropriate level of granularity depends on the purpose of the analysis and the type of insights required.

1.3.1.2 Unstructured, semi-structured and qualitative data

Qualitative data, being descriptive and often non-numeric, frequently appears in semi-structured or unstructured formats. Examples for unstructured formats include images (both human- and machine-generated), video files, audio files, product reviews, social media posts, and messages sent by Short Messaging Service (SMS) or online platforms. Semi-structured formats, such as emails, combine structured metadata (e.g., headers like date, language, and recipient's email address) with unstructured data in the email body, which contains the actual message. Additionally, qualitative data, such as customer feedback may be stored in semi-structured formats, e.g., JSON (JavaScript Object Notation) or XML (Extensible Markup Language) files.

These types of data require specialised methods, such as text mining and natural language processing (NLP), to extract meaningful insights about customer preferences, opinions and areas where processes can be improved. High-granularity unstructured data, such as individual customer comments, provides rich detail but may require extensive processing, whereas summaries or tags offer a more manageable, lower-granularity view.

1.3.1.3 Integration of data types for decision-making

Quantitative and qualitative data often overlap and complement each other, providing different perspectives that support effective decision-making. Quantitative data can exist in semi-structured or unstructured formats, such as numerical information in web analytics logs (e.g., session times or page views) or metadata in image and video files (e.g., timestamps or geolocation). While most structured data is numerical (quantitative), certain qualitative data can also be structured when categorised or coded systematically. For example, qualitative data may be embedded in structured formats, such as survey responses that combine predefined categorical answers with open-ended feedback.

When integrated, these data types create a more complete picture. Quantitative data provides measurable insights, enabling businesses to track performance and make objective comparisons, while qualitative data adds context by explaining the “why” behind the numbers. Together, they allow businesses to make informed, evidence-based decisions that are both data-driven and enriched with deeper insights.

Table 1.2 illustrates specific examples of quantitative and qualitative data across different business functions, including customer and marketing, sales and revenue, and operations and production. This table provides a clearer understanding of how various data types can support business operations, enable performance tracking and contribute to strategic decision-making.

Table 1.2: Examples of data types and how they are used in business

Types of Data in Business	Quantitative Data Examples	Qualitative Data Examples
Customer and marketing	- Number of purchases made per customer - Click-through rates on email campaigns - Average customer satisfaction scores from surveys - Monthly social media engagement metrics (likes, shares, retweets) - Cost per acquisition (CPA) for marketing campaigns	- Comments or reviews from customers about their experience - Preferences for product features (e.g., prefers eco-friendly packaging) - Feedback on advertisements (e.g., “ad felt relatable” or “ad was too long”) - Brand sentiment (positive, neutral or negative) - Reasons for customer complaints (e.g., “delayed delivery”, “damaged product”)
Sales and revenue	- Daily sales revenue for each product category - Number of units sold per store - Average transaction value (ATV) - Discounts offered and redeemed - Year-over-year revenue growth percentages	- Reasons for low sales in certain regions (e.g., competition, seasonality) - Categories of most returned items (e.g., electronics, clothing) - Payment methods preferred by customers (e.g., credit card, cash, digital wallet) - Descriptions of best-selling products (e.g., “popular for holidays”, “limited edition”) - Sales team feedback on customer trends
Operations and production	- Production cycle times (e.g., time to manufacture a single product) - Machine uptime percentages - Energy consumption in kilowatt-hours (kWh) - Defect rates (e.g., number of defective items per batch) - Inventory turnover rates	- Machine maintenance status (e.g., needs repair, fully operational) - Types of defects identified in production (e.g., cracked items, mislabelling) - Supplier performance descriptions (e.g., on-time deliveries, frequent delays) - Workflow bottlenecks identified by staff (e.g., packaging delays, short staffing) - Employee feedback on production schedules
Employee and workforce	- Hours worked per week by employees - Number of employees in each department - Monthly employee absenteeism rates - Average training hours completed per employee - Annual employee turnover percentage	- Employee opinions on workplace culture (e.g., “supportive”, “stressful”) - Feedback from exit interviews about reasons for leaving - Descriptions of skill gaps in the workforce (e.g., need more IT training) - Performance appraisal comments (e.g., “team player”, “needs improvement in time management”) - Employee preferences for flexible work options

Web and digital analytics	• Daily website visits • Conversion rates (percentage of visitors who make a purchase) • Average time spent on a webpage • Bounce rates (percentage of visitors who leave without interacting) • Downloads of digital assets (e.g., eBooks, reports)	• User feedback on website design (e.g., “confusing navigation”, “visually appealing”) • Reasons for unsubscribing from a newsletter (e.g., “too frequent”, “irrelevant content”) • Keywords users search for most often on the site • Types of digital assets users prefer to download (e.g., “templates”, “how-to guides”) • Feedback on mobile app features (e.g., “intuitive design”, “needs better speed”)
Finance and compliance	• Total expenses per quarter • Tax payments made annually • Number of financial transactions per day • Ratio of debts to equity • Compliance scores from audits	• Descriptions of financial risks (e.g., currency fluctuation, market volatility) • Auditor comments on areas of improvement (e.g., better expense tracking) • Reasons for compliance failures (e.g., incomplete documentation) • Client feedback on payment processes (e.g., “too complicated”, “user-friendly”) • Stakeholder opinions on budgeting priorities (e.g., “focus on innovation”, “reduce costs”)

1.3.1.4 Data purpose and origin/sources

Data used in business comes from various sources, broadly categorised as primary internal, external or generated data, and secondary internal, external or generated data.

Primary data is collected firsthand by an organisation for a specific purpose. It is original and obtained directly from the source through surveys, experiments, observations or direct measurements. Primary data is often more reliable for specific analysis but can be time-consuming and costly to gather. Primary data can be internal (collected within the organisation, typically structured), external (collected from sources outside the organisation, structured or semi-structured) or generated (created through interactions or automated systems, structured, semi-structured or unstructured).

Secondary data refers to data collected by a third party, often for general purposes, which can be reused in other contexts. Secondary data is often more accessible and affordable than primary data but may be less tailored to specific analytical needs. Using secondary data saves time and resources by reusing existing data to gain insights. Like primary data, secondary data can also be internal (structured or semi-structured), external (typically semi-structured or unstructured) or generated (can be structured, semi-structured or unstructured).

Internal data is collected and generated within an organisation’s own systems, processes or operations. It originates from the organisation’s activities and is typically exclusive to the organisation.

External data is acquired from sources outside the organisation. This data provides insights into external

factors that impact the business, such as market conditions, industry trends and customer behaviour outside direct interactions with the organisation.

Generated data is created as a result of interactions, activities or usage patterns, and can be internal or external. Generated data may come from various sources, including user interactions, sensor readings or automated systems. It can be structured, semi-structured or unstructured and offers unique insights into behaviours and performance.

The data classification matrix in Table 1.3 illustrates these categories and their relationships, making it easy to understand how these data types overlap and interact.

Table 1.3: Data classification matrix

Aspect	Primary Internal Data	Primary External Data
Definition	Data collected within the organisation	Data collected by the organisation from external sources
Use	Used for operational decisions, performance monitoring and process improvements	Used to understand customer needs, market trends and external factors for strategic planning
Examples	- Sales transaction records from POS systems - Employee performance data - Production data from internal machines Generated data: Website clickstream data, sensor data from IoT (Internet of Things) devices within company facilities, customer service interaction logs	- Customer survey responses - Observational data on customer behaviour outside stores - Competitor analysis conducted by the organisation Generated data: Data from market research conducted directly by the organisation
	Secondary Internal Data	Secondary External Data
Definition	Previously collected internal data repurposed (reused)	Data collected by other organisations or entities
Use	Useful for trend analysis, long-term planning and optimising existing processes	Provides broader market insights and benchmarking and supports strategic decision-making without primary collection
Examples	- Archived sales data repurposed for forecasting - Historical customer feedback analysed for trends - Production data reused for process optimisation Generated data: Internal social media engagement data repurposed for sentiment analysis	- Industry reports from research firms - Public demographic data (e.g., census) - Social media analytics from third-party platforms Generated data: IoT sensor data from public infrastructure (e.g., smart cities), social media engagement metrics from third-party platforms, web traffic data from external ad platforms

1.3.2 Data as a strategic asset

Data is often referred to as the “new oil” because it is essential to driving competitive advantage. Data-driven companies use insights derived from data to make informed decisions, improve products, optimise operations and personalise customer experiences. Businesses depend on data to guide strategic, tactical and operational decisions, using it to reduce uncertainty and provide evidence-based support for selecting the best course of action.

The following real-world example illustrates how Australian retailers treat data as a strategic asset to enhance their competitiveness and customer satisfaction.

Scenario: Australian retailers using data as a strategic asset

In Australia's retail industry, leading companies such as Woolworths, Coles, IGA and Aldi all leverage data as a strategic asset to enhance competitiveness and improve customer satisfaction. These retailers collect extensive data from loyalty programs, online shopping behaviours and in-store purchases, enabling them to track customer preferences, seasonal trends and regional demand patterns. By analysing this data, they can make informed decisions at all levels – from strategic to operational.

At the **strategic level**, Australian retailers use data insights to identify and respond to market trends, such as the growing demand for organic products and sustainable packaging. For example, Woolworths and Coles have expanded their organic product ranges and environmentally friendly options, positioning themselves as leaders in sustainable retail.

At the **tactical level**, data helps retailers plan regional promotions tailored to customer preferences. For instance, during the holiday season, stores in warmer areas might promote outdoor products and summer foods, while stores in cooler regions stock up on winter essentials. This targeted approach allows retailers to connect more effectively with local customers.

At the **operational level**, data on customer transactions helps these retailers optimise inventory management, ensuring popular items are consistently available while minimising overstock. By using predictive models, they can anticipate demand surges for specific products and adjust stock levels, accordingly, reducing waste and improving the customer experience.

Through these data-driven practices, Australian retailers like Woolworths, Coles, IGA and Aldi enhance customer personalisation, streamline operations and adapt to shifting market trends, giving them a competitive edge. This example demonstrates how data, when treated as a strategic asset, empowers these companies to make evidence-based decisions that drive growth, foster customer loyalty and maintain their market positions in Australia's retail sector.

1.4 Metrics for analytics-ready data

In this section, we explore the essential metrics that define analytics-ready data, highlighting the qualities that make data reliable, meaningful and suitable for analysis. Analytics-ready data is crucial for organisations that aim to derive insights effectively and efficiently, as it minimises the risk of errors and maximises the value extracted from data.

1.4.1 Definition of analytics-ready data

What are analytics-ready data? Analytics-ready data refers to data that has been prepared to meet the requirements of analysis. This means the data is clean, consistent, complete and formatted in a way that allows for easy interpretation. To achieve this state, the data undergoes pre-processing steps, such as cleaning, transformation and validation, to align it with the specific goals of the analysis.

Why is analytics-ready data important? Having analytics-ready data saves time, reduces errors and enables quicker insights by providing a solid foundation for analysis. Without proper data quality measures in place, analysts risk incomplete results, inaccuracies or misinterpretations, potentially leading to costly business decisions based on flawed data.

1.4.2 Data quality metrics

Data quality is essential to ensure the accuracy, reliability and value of analytics. High-quality data supports effective decision-making, minimises errors and enhances the overall impact of analysis. Below are the key metrics for analytics-ready data, each explained with examples. Here is a simple poem for you that was generated using Copilot (Microsoft, 2024).

Reliable and **complete**, data stands tall,
Consistent and **accurate**, the key to it all.
Timely and **valid**, with **uniqueness** in play,
Accessible and **rich** to brighten the day.
Diverse and **traceable**, it tells the tale,
Interoperable, **scalable**, it will never fail.
Relevance and **context** give it its might,
Granularity and **integrity** keep it tight.
Secure and **private**, it passes the test,
With **interpretability**, it's simply the best!

1.4.2.1 Data reliability

Reliability refers to the consistency and dependability of data over time and across processes or systems, ensuring unbiased and accurate information regardless of when or where it is collected or used.

Impact on analytics: Unreliable or inaccurate data can lead to misleading insights and flawed decisions.

Examples:

1. A hospital depends on dependable readings from medical devices (e.g., heart rate monitors) to ensure accurate diagnosis and treatment. If the data fluctuates or becomes unreliable due to sensor malfunctions, it can compromise patient safety.
2. A retailer analysing weekly sales data needs consistency in how sales are recorded across different stores and systems. If some stores log discounts differently or omit promotions, the data becomes unreliable for decision-making.

1.4.2.2 Data consistency

Consistency means that data values remain uniform (identical and accurate) across datasets and systems.

Impact on analytics: Inconsistent data creates confusion and reduces trust in analytics results.

Example: If a customer's address is updated in a customer relationship management (CRM) system, the same updated address should appear in the billing system and any associated databases. If the CRM shows the new address while the billing system still shows the old one, the data is inconsistent, which could lead to billing errors or delivery issues.

1.4.2.3 Data completeness

Completeness refers to the extent to which all required data fields, including the metadata that describes the characteristics and context of the data, are available and properly filled.

Impact on analytics: Incomplete/missing data can lead to gaps in analysis and poor decision-making, as critical insights may be overlooked or misinterpreted due to missing information.

Examples:

1. A retail business must ensure all customer profiles include essential details such as name, email and purchase history to personalise marketing campaigns effectively.
2. A hospital requires complete patient medical records, including allergies and medication history, to avoid treatment errors and improve patient safety.
3. Metadata for a dataset of images might include details such as file size, resolution, timestamps and descriptive tags, which provide additional context and improve the ability to retrieve, analyse and interpret the data.

1.4.2.4 Data content accuracy

Accuracy ensures that data values are correct and reflect the true state of the object or event they represent.

Impact on analytics: Inaccurate data can lead to false conclusions, impacting business strategies and operational processes.

Examples:

1. An airline reservation system must accurately capture flight schedules and passenger bookings to avoid overbooking.
2. Sensor data from IoT devices used in agriculture must correctly measure soil moisture levels to optimise irrigation systems.

1.4.2.5 Data timeliness and temporal relevance

Timeliness (also referred to as data currency) ensures that data is up-to-date and available at the moment when it is needed for decision-making.

Temporal relevance ensures that data corresponds to the appropriate time frame for the specific analysis or decision-making process.

Impact on analytics: Outdated data makes insights irrelevant and can lead to missed opportunities. Using data from an inappropriate time period can result in incorrect conclusions, as it may no longer reflect the current or relevant conditions. Together, these metrics ensure decisions are based on both the most current data and data that is contextually appropriate for the situation.

Examples:

1. A logistics company needs real-time tracking data to monitor delivery statuses and optimise routes.
2. A retail business should rely on recent sales data to plan seasonal promotions, rather than outdated figures from prior years.

1.4.2.6 Data validity

Validity ensures that data follows predefined formats, types or business rules.

Impact on analytics: Invalid data can disrupt processes and analytics workflows.

Examples:

1. A payroll system requires dates in a valid format (e.g., MM/DD/YYYY) to process employee salaries correctly.
2. Customer phone numbers in a CRM system must match a specific format (e.g., country code included) to enable seamless communication.

1.4.2.7 Data Uniqueness

Uniqueness means that no duplicate records exist within a dataset.

Impact on analytics: Duplicate data can inflate metrics, skew results and increase storage costs.

Examples:

1. An email marketing platform must ensure that each email address in its database is unique to avoid sending multiple messages to the same recipient.
2. A university admissions system must ensure that no duplicate student records are created during the application process.

1.4.2.8 Data accessibility

Data accessibility refers to the ease with which authorised users can locate, retrieve and use data when needed for analysis or decision-making.

Impact on analytics: Poor accessibility can delay operations, hinder decision-making and reduce the effectiveness of analytics.

Examples:

1. A doctor accessing a patient's electronic health records (EHR) during an emergency needs quick and seamless access to critical medical data, such as allergies and previous treatments, to make life-saving decisions.
2. A logistics company relies on real-time accessibility of tracking data to monitor shipments, reroute deliveries or notify customers of delays.

1.4.2.9 Data usability

Data usability is the ease with which data can be accessed, understood and utilised by end-users.

Impact on analytics: Complex or poorly organised data may hinder its use in analytics, reducing the effectiveness of insights.

Example: A dashboard that provides easy access to interactive data visualisations ensures users can quickly explore trends and make decisions.

1.4.2.10 Data integrity

Data integrity ensures that relationships between data elements are maintained correctly.

Impact on analytics: Data with broken relationships can result in incomplete or misleading analyses.

Examples:

1. In a relational database, foreign key relationships between order details and customer information must remain intact for accurate reporting.
2. A supply chain management system must ensure that product IDs in the inventory table match those in the supplier table.

1.4.2.11 Data richness

Data richness refers to the depth, breadth and diversity of data available for analysis.

Impact on analytics: Lack of rich data restricts the ability to gain meaningful insights and can result in inefficient strategies or actions.

Examples:

1. A hospital collects rich data about patients, including demographics, medical history, lab results, prescriptions and lifestyle habits to create personalised treatment plans and predictive models for patient care.
2. An e-commerce platform gathers data on customer browsing behaviour, purchase history, product reviews and payment preferences to create personalised recommendations.

1.4.2.12 Data diversity

Data diversity refers to the variety of data types and sources, ensuring a multi-dimensional view.

Impact on analytics: Incorporating diverse datasets (structured, unstructured, quantitative, qualitative) provides a richer, more comprehensive perspective.

Example: Combining social media sentiment (qualitative) with sales data (quantitative) enables a company to align product popularity with customer feedback.

1.4.2.13 Data interpretability

Data interpretability is the extent to which the data can be analysed and understood in a meaningful way.

Impact on analytics: Data that is too complex or poorly presented can hinder actionable insights.

Example: In machine learning models, interpretable datasets allow analysts to understand how certain variables influence predictions.

1.4.2.14 Data relevance

Data relevance refers to measures of whether the data is meaningful and applicable to the specific analysis or business question.

Impact on analytics: Irrelevant data can clutter analysis, leading to noise that distracts from actionable insights.

Example: In marketing, data on customer purchasing behaviours is relevant, but unrelated demographic data (e.g., age of non-target groups) may be less useful.

1.4.2.15 Data contextuality

Data contextuality refers to whether the data captures the context surrounding an event or transaction, enhancing its interpretability.

Impact on analytics: Data without context may lead to inaccurate assumptions.

Example: Knowing that sales spiked in a specific region is valuable, but understanding that it occurred during a festival (context) provides actionable insights for future marketing campaigns.

1.4.2.16 Data granularity

Data granularity refers to the level of detail in the data. High granularity means the data is very detailed, while low granularity provides a summarised view.

Impact on analytics: The level of detail impacts the depth of analysis; high granularity enables micro-level insights, while low granularity is better for high-level trends.

Example: In customer analytics, transaction-level data provides granular insights into individual behaviour, while monthly summaries show macro trends.

1.4.2.17 Data traceability

Data traceability is the ability to trace data back to its source and track its transformations during processing.

Impact on analytics: Traceability ensures transparency, builds trust in the analysis and enables verification of results.

Example: In financial audits, having a clear record of data origins and modifications is critical to validate outcomes.

1.4.2.18 Data interoperability

Data interoperability refers to the ability of data to integrate seamlessly across different platforms, systems and tools.

Impact on analytics: Non-interoperable data creates silos, making it harder to conduct a comprehensive analysis.

Example: A supply chain system that integrates data from logistics providers, inventory systems and customer orders enables a unified view of operations.

1.4.2.19 Data scalability

Data scalability ensures that the data infrastructure can handle growth in data volume, variety and complexity without degrading performance.

Impact on analytics: Analytics systems should be able to process increasing amounts of data as organisations grow.

Example: An IoT platform should scale to accommodate additional sensors as a factory expands operations.

1.4.2.20 Data security and data privacy

Data security refers to the measures and technologies used to protect data from unauthorised access, breaches, theft or corruption.

Data privacy refers to the rights of individuals to control how their personal data is collected, stored and shared in compliance with ethical and legal standards.

Impact on analytics: Weak data security and privacy protocols can result in financial losses, reputational damage and legal consequences.

Examples:

1. Banks and financial institutions must secure customer financial records and transaction data to prevent fraud or hacking attempts. Similarly, hospitals use encryption and access controls to safeguard patient medical records, ensuring only authorised personnel can view sensitive information.
2. Social media platforms, e.g., Facebook, must ensure that user data is shared only with consent, particularly for advertising purposes. Additionally, a retailer collecting customer email addresses and purchase history must comply with privacy regulations by allowing customers to opt out of marketing communications.

1.5 Data pre-processing

Data pre-processing is the process of preparing raw data for analysis by applying techniques such as cleaning, transformation, reduction and integration. It ensures that data is in an optimal state for generating reliable and accurate insights. Proper data pre-processing is essential for analytics as it enhances data quality, minimises errors and improves the performance of analytical models.

Pre-processing is a prerequisite for analysis, focusing on data preparation rather than generating insights. By cleaning, transforming, integrating and formatting the data, pre-processing ensures high-quality, well-structured and analytics-ready data, which forms the foundation for meaningful and reliable analysis.

1.5.1 Importance of pre-processing for analytics

Raw data is often messy, incomplete, inconsistent and unsuitable for direct analysis. It may contain errors, outliers and redundancies that can distort analytics and lead to inaccurate or misleading results. Effective pre-processing addresses these issues by cleaning and organising the data, making it cleaner, more consistent and ready for meaningful analysis. Without proper pre-processing, the accuracy and reliability of analytics outcomes are significantly compromised.

1.5.2 Steps in data pre-processing

Data pre-processing is a structured process that involves several sequential steps to prepare raw data for analysis. Each step is designed to address specific issues within the data, ensuring that it is clean, accurate and in a usable format for generating meaningful insights. These steps often include sub-tasks and methods tailored to meet the unique requirements of the dataset or analytical task. By following these steps, businesses can ensure that their data is of high quality and ready for reliable analysis.

Figure 1.3 describes the steps in basic data pre-processing.

Steps in data pre-processing

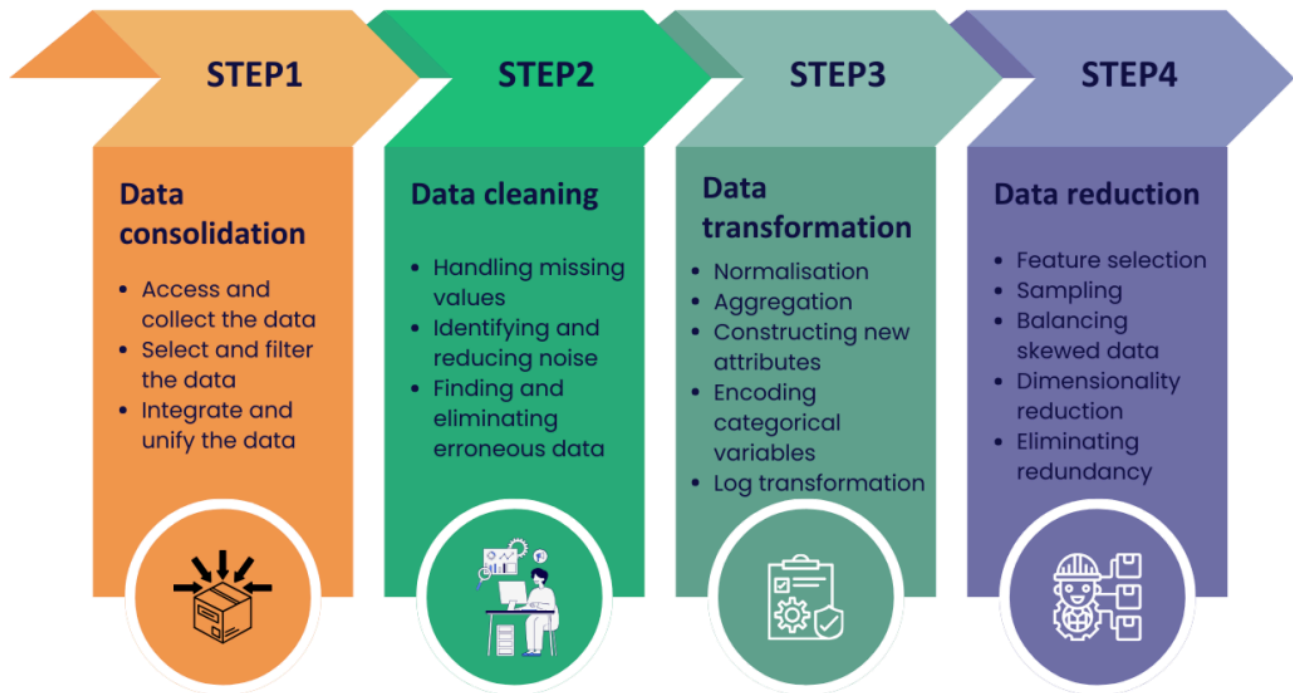


Figure 1.3: Steps in data pre-processing

Alternative Text for Figure 1.3

Steps in Data Pre-Processing:

Step 1: Data Consolidation

- Access and collect data.
- Select and filter the data.
- Integrate and unify the data.

Step 2: Data Cleaning

- Handle missing values.
- Identify and reduce noise.
- Find and eliminate erroneous data.

Step 3: Data Transformation

- Perform normalisation.
- Aggregate data.

- Construct new attributes.
- Encode categorical variables.
- Apply log transformation.

Step 4: Data Reduction

- Select features.
- Sample data.
- Balance skewed data.
- Perform dimensionality reduction.
- Eliminate redundancy.

Each step is visually represented as a distinct arrow, highlighting the sequential nature of the process. Icons at the bottom of each arrow symbolise tasks within the respective steps.

1.5.2.1 Data consolidation

In a business setting, **data consolidation is the process of combining information collected from different systems or tools into one central place.** This is especially important because businesses often use multiple systems to manage different aspects of their operations. When these systems operate separately, the data they hold can become siloed, meaning it is scattered across different platforms and not easily accessible as a whole. This makes it harder for businesses to see the “big picture” or draw meaningful conclusions. By consolidating this data into one system, businesses can create a unified view of their operations, customers and performance. This unified data is easier to analyse, enabling better decision-making and more efficient business processes.

Table 1.4 provides an overview of the sub-tasks in data consolidation, along with an example for each.

Table 1.4: Sub-tasks of data consolidation

Sub-task	Description	Example
Access and collect the data	This step involves locating and retrieving raw data from various sources such as databases, cloud storage, web services or even physical records.	A retail company collects sales data from its point-of-sale (POS) system, inventory data from its warehouse system and customer feedback from social media platforms.
Select and filter the data	After collecting data, this step involves identifying and selecting only the relevant data for the analysis. Irrelevant or unnecessary data is removed to ensure the focus remains on actionable insights.	A company analysing sales performance may filter out fields such as “employee birthdays” or “holiday schedules” as they are irrelevant to the sales analysis.
Integrate and unify the data	This step aligns data from different sources to create a consistent and unified dataset. It resolves discrepancies such as mismatched formats, different naming conventions or varying levels of granularity.	A financial institution consolidates transaction data from multiple branches, ensuring that date formats and account names follow a uniform standard.

Table 1.5 provides an overview of the methods involved in data consolidation, followed by a business scenario illustrating how these methods can be applied in practice.

Table 1.5: Methods of data consolidation

Method Name	Description
SQL (Structured Query Language) queries	SQL is widely used to extract, transform and merge data from relational databases.
Software agents and automation tools	Automation tools, e.g., robotic process automation (RPA) or software agents, can collect and consolidate data from multiple sources, including APIs or web services.
Domain expertise and statistical tools	Experts and tools are often needed to determine which data is relevant and how to align it. This includes cleaning inconsistencies during consolidation.
Ontology-driven data mapping	Ontology-driven methods help map data fields from different sources into a unified schema. This is particularly useful when merging datasets with different structures.

Business scenario: Enhancing customer insights with consolidation methods

A **retail company** aims to improve its understanding of customer behaviour by integrating

data from multiple systems, including a CRM database, an online sales platform, social media reviews and in-store transaction records. By using various consolidation methods, the company ensures that the data is consistent, comprehensive and ready for analysis.

- **SQL queries for data integration**

The company uses SQL queries to join customer information from the CRM database with purchase histories stored in the sales database. For example, they query customer IDs in both databases to link customer demographics (age, gender, location) with their recent purchases.

Result: This allows the company to identify purchasing patterns, such as which age group prefers specific product categories, enabling targeted marketing campaigns.

- **Software agents for automation**

The company implements an automated agent to collect product reviews from multiple online platforms, such as their website, social media channels and third-party review sites. The agent consolidates this data into a central database for further analysis.

Result: By analysing the consolidated reviews, the company identifies common customer concerns, such as delayed deliveries or packaging quality, and prioritises improvements.

- **Domain expertise and statistical tools**

Using statistical tools, domain experts examine the merged datasets to resolve inconsistencies, such as duplicate customer records or outliers in transaction amounts. For instance, they detect and correct a data entry error where a product price was mistakenly recorded as \$10,000 instead of \$100.

Result: This ensures that the final dataset is clean and accurate, providing reliable insights for decision-making.

- **Ontology-driven data mapping for unified structure**

The company faces a challenge when merging data from in-store transactions and online sales, as these systems use different terminologies (e.g., “item number” in one system vs. “product ID” in another). By applying ontology-driven data mapping, the company aligns these fields into a unified structure.

Result: This enables seamless integration, allowing the company to analyse sales performance across both online and offline channels holistically.

1.5.2.2 Data cleaning

Data cleaning is the process of identifying and correcting errors, inconsistencies and missing pieces of information in a dataset to improve its quality and reliability. This step in data pre-

processing ensures that data is accurate, complete and suitable for analysis. Without cleaning, the data might contain mistakes or gaps that can lead to incorrect results and poor decision-making.

In practice, raw data collected from various sources often contains issues such as missing values, duplicate entries, incorrect formats and outliers, which must be resolved before meaningful analysis can take place.

Table 1.6 provides an overview of the sub-tasks in data cleaning, along with an example for each.

Table 1.6: Sub-tasks of data cleaning

Sub-task	Description	Example
Handling missing values	Missing values occur when data points are not recorded or are unavailable. Handling missing values ensures the dataset remains usable.	Filling in missing ages in a customer dataset using the average age of the existing entries.
Identifying and reducing noise	Noise refers to irrelevant or meaningless data that can obscure useful insights. Reducing noise helps clarify patterns and relationships within the data.	Smoothing extreme spikes in stock price data that result from anomalies in the recording process.
Finding and eliminating erroneous data	Erroneous data includes invalid, incorrect or inconsistent values in the dataset. Identifying and correcting such errors ensures accuracy.	Correcting product prices listed as negative due to a data entry error.

Table 1.7 provides an overview of the methods involved in data cleaning, followed by a business scenario illustrating how these methods can be applied in practice.

Table 1.7: Methods of data cleaning

Method Name	Description
Imputation for missing values	Replace missing data with estimated values based on statistical techniques or machine learning models.
Outlier detection and treatment	Detect outliers using statistical methods (e.g., standard deviation, interquartile range) or clustering algorithms. Treat them by removing, capping or adjusting their values.
Data de-duplication	Identify and remove duplicate entries in the dataset to ensure uniqueness.
Error correction tools	Use automated tools or manual review processes to correct invalid or inconsistent data values.
Smoothing techniques	Address noisy data points (due to general fluctuations, which may or may not include outliers) by replacing them with averages (of surrounding records) or applying regression techniques.

Note: Outlier detection can be part of a smoothing process. For instance, you might first detect and remove outliers and then apply smoothing to clean the dataset further.

Business scenario: Preparing healthcare data for patient outcome analysis

A hospital wants to analyse patient data to identify patterns in treatment outcomes and optimise resource allocation. However, the raw data collected from different sources, such as electronic health records (EHR), lab reports and wearable health devices, contains several issues, such as missing values, outliers, duplicates and inconsistencies. The hospital applies various data cleaning and preparation methods to create a reliable dataset for analysis.

- **Imputation for missing values**

Some patient records have missing values for critical measurements, such as blood pressure or glucose levels, during routine check-ups. The hospital uses mean imputation to fill in the gaps by calculating the average value for these parameters based on other patients of similar age and medical history.

Result: The completed dataset ensures that missing values do not compromise the accuracy of health trend analysis.

- **Outlier detection and treatment**

While reviewing wearable device data, the hospital identifies abnormally high heart rate readings for a few patients that were far outside the normal range. Using statistical methods like the interquartile range (IQR), these outliers are flagged. Upon review, some outliers are corrected (e.g., a heart rate of 600 beats per minute due to a device malfunction), while others (valid but extreme cases) are retained for further investigation.

Result: The hospital prevents incorrect conclusions caused by faulty device readings while maintaining valid data for rare cases.

- **Data de-duplication**

The patient database contains duplicate entries caused by multiple visits, where patients used slightly different identifiers (e.g., one visit uses their full name, another their initials). Using a data de-duplication algorithm, the hospital matches records based on attributes like patient ID, date of birth and contact information to merge duplicates.

Result: A clean, unique patient dataset enables accurate tracking of treatment histories and outcomes.

- **Error correction tools**

Some lab results have inconsistencies in how measurements are recorded, e.g., cholesterol levels in one system are recorded in mg/dL, while another system uses mmol/L. The hospital applies an error correction script to standardise all units to mg/dL.

Result: Standardised units ensure that comparisons across lab results are accurate and meaningful for medical analysis.

- **Smoothing techniques**

Wearable health devices report daily step counts for patients recovering from surgery. These reports show erratic fluctuations due to device errors or irregular recording intervals. The hospital applies a moving average smoothing technique to replace noisy values with smoothed data that represents overall activity trends.

Result: The smoothed data allows the hospital to monitor patient recovery progress and identify when additional interventions may be needed.

1.5.2.3 Data transformation

Data transformation is the process of converting raw data into a format suitable for analysis. It involves modifying the structure, format or values of data to meet specific requirements of the analytics

task. Transformation ensures consistency and standardisation, allowing better compatibility with analytical tools and techniques. By applying transformation techniques, data becomes easier to interpret, analyse and apply in tasks like machine learning, statistical studies or business reporting.

Table 1.8 provides an overview of the sub-tasks in data transformation, along with an example for each.

Table 1.8: Sub-tasks of data transformation

Sub-task	Description	Example
Normalisation	Adjusting the scale of numerical data to a common range, often between 0 and 1 or -1 and $+1$. This eliminates biases caused by differences in measurement units or magnitudes.	Scaling sales revenue (measured in millions) and customer ratings (measured on a scale of 1 to 5) to the same range for consistent comparison.
Discretisation or aggregation	Converting continuous numerical data into discrete categories (discretisation) or summarising multiple data points into a single value (aggregation).	Converting customer ages into ranges, e.g., “20–30” or aggregating daily sales data into monthly totals.
Constructing new attributes	Deriving new variables or features from existing ones by applying mathematical, logical or statistical operations.	Creating a “customer loyalty score” by combining purchase frequency, average transaction value and account age.
Encoding categorical variables	- Converting categorical data into numerical format so that it can be analysed by algorithms that require numeric inputs. - Representing categorical variables as binary vectors.	- Encoding “Male” as 0 and “Female” as 1 in a dataset for a predictive model. - Representing “Red”, “Blue” and “Green” as $[1, 0, 0]$, $[0, 1, 0]$ and $[0, 0, 1]$, respectively.
Log transformation	Applying a logarithmic function to data to reduce skewness or compress large ranges of values.	Transforming sales figures in millions using logarithms to reduce large disparities.

Table 1.9 provides an overview of the methods involved in data transformation, followed by a business scenario illustrating how these methods can be applied in practice.

Table 1.9: Methods of data transformation

Method Name	Description
Scaling techniques	Adjusts data values to fit within a standard range or distribution.
Binning or discretisation	Groups numerical data into intervals or bins.
Mathematical transformations	Applies functions like logarithms, square roots or exponentials to data values.
Feature engineering	Creates new variables by combining or modifying existing ones.
Text transformation	Converts unstructured text data into structured formats for analysis.
Dimensionality reduction	Reduces the number of features in a dataset while retaining critical information.

Business scenario: Enhancing supply chain analytics for a manufacturing company

A manufacturing company wants to optimise its supply chain operations to reduce costs, improve delivery times and forecast demand more effectively. However, the raw data from various sources, such as supplier records, inventory logs and shipment tracking systems, requires significant pre-processing to be suitable for analysis. The company applies various data transformation methods to prepare the data for insights.

• **Scaling techniques**

The company’s inventory data includes a wide range of values, such as the number of units (ranging from 1 to 100,000) and shipment costs (ranging from \$0.01 to \$1,000). To ensure all features are comparable, the company applies Min-Max Scaling to normalise these variables between 0 and 1.

Result: Scaled data ensures that no single variable dominates the analysis due to differences in magnitude, improving the performance of predictive models for inventory optimisation.

• **Binning or discretisation**

The company’s demand forecast dataset includes the age of customers ordering products, which ranges widely from teenagers to seniors. To simplify the analysis, the company groups ages into bins, such as “Under 18”, “18–35” and “35+”.

Result: The binned data allows the company to analyse demand patterns for different age groups and target its marketing efforts accordingly.

- **Mathematical transformations**

The company's shipment cost data is highly skewed, with a small number of unusually high costs distorting the averages. To address this, the company applies a logarithmic transformation to reduce the skewness and make the data more suitable for statistical modelling.

Result: Transformed data provides more meaningful insights into average shipment costs and highlights areas where cost optimisation is needed.

- **Feature engineering**

The company created a new feature, "inventory turnover rate", by dividing the total number of units sold by the average inventory level. This variable helps measure how efficiently inventory is managed across different product categories.

Result: The new feature enables the company to identify slow-moving products and adjust its inventory strategy to reduce holding costs.

- **Text transformation**

Supplier feedback forms contain unstructured text data describing delays, quality issues and other concerns. The company applies text transformation techniques such as TF-IDF (Term Frequency-Inverse Document Frequency) to extract keywords and convert the text into numerical representations.

Result: The structured text data enables the company to identify common supplier issues and prioritise improvements in supplier relationships.

- **Dimensionality reduction**

The dataset includes over 50 variables related to shipment tracking, such as delivery times, route distances, vehicle types and fuel costs. Many of these features are highly correlated. The company uses Principal Component Analysis (PCA) to reduce the dataset to a smaller set of uncorrelated components while retaining the majority of the original information.

Result: Dimensionality reduction simplifies the dataset, making it easier to build machine learning models that predict delivery delays or cost overruns.

1.5.2.4 Data reduction

Data reduction is the process of making large datasets smaller and simpler without losing important information. The goal is to reduce the size, complexity or number of variables in the data while keeping its key insights and usefulness intact. Data reduction helps improve the speed and efficiency

of analysis, lowers storage costs and removes unnecessary “noise” from the data, making it easier to focus on meaningful insights.

Large datasets often include extra or irrelevant information that can overwhelm analysis tools and waste resources. Data reduction ensures that only the most important parts of the data are kept, helping analysts focus on the variables and records that truly matter. This makes the analytics process faster, more efficient and easier to manage.

Table 1.10 provides an overview of the sub-tasks in data reduction, along with an example for each.

Table 1.10: Sub-tasks of data reduction

Sub-task	Description	Example
Reducing the number of attributes (feature selection)	Identifying and retaining only the most relevant variables (attributes) for analysis.	In predicting customer churn, variables such as “purchase frequency” and “customer satisfaction score” might be more relevant than “store location”.
Reducing the number of records (sampling)	Selecting a representative subset of the dataset while discarding redundant or unnecessary records.	A bank may analyse a sample of transactions rather than the entire dataset to detect fraud trends.
Balancing skewed data	Adjusting datasets with imbalanced classes to ensure fair representation of all categories.	In fraud detection, oversampling fraudulent transactions ensures the model recognises minority-class patterns effectively.
Dimensionality reduction	Reducing the number of features or dimensions while retaining the dataset’s essential characteristics.	Combining highly correlated features (e.g., “height” and “weight”) into a single dimension using Principal Component Analysis (PCA).
Eliminating redundancy	Removing duplicate or irrelevant records and attributes.	A CRM system might eliminate duplicate customer profiles to streamline marketing efforts.

Table 1.11 provides an overview of the methods involved in data reduction, followed by a business scenario illustrating how these methods can be applied in practice.

Table 1.11: Methods of data reduction

Method Name	Description
Feature selection techniques	Retains only the most informative variables by using methods such as correlation analysis (identifying and removing highly correlated variables) and chi-square tests (selecting features based on their statistical significance).
Sampling	Extracts a subset of the data using techniques such as random sampling (selecting a random subset of data points) and stratified sampling (ensuring that the sample represents key characteristics of the population).
Dimensionality reduction techniques	These techniques reduce the number of features through transformation or combination, such as Principal Component Analysis (PCA), which combines features into fewer components.
Aggregation	Summarises data by grouping or combining records.
Resampling for balancing skewed data	Adjusts the distribution of data classes using techniques such as oversampling (duplicating data points from underrepresented classes) and undersampling (removing data points from overrepresented classes).
Clustering	Groups similar records together, allowing representative records (centroids) to stand in for the full dataset.

Business scenario: Improving academic performance insights in the education sector

An education analytics company aims to analyse student performance data across various schools to identify factors influencing academic success. The data collected includes demographic information, grades, attendance records, extracurricular involvement and socioeconomic indicators. However, due to the large and complex nature of the dataset, the company applies several data preparation methods to streamline the analysis process.

- **Feature selection techniques**

The dataset contains variables such as “parental income”, “postcode”, “number of siblings” and “study hours”. Using correlation analysis and chi-square tests, the company identifies that “study hours”, “attendance rate” and “parental education level” are the most informative features. Less relevant features, such as “postcode”, are excluded.

Result: By focusing on the most important variables, the analysis becomes more efficient and yields clearer insights into factors that impact student performance.

- **Sampling**

The company has data for 1 million students but wants to build a predictive model to identify students at risk of failing. To speed up processing, the company applies stratified sampling to

ensure that the sample includes a balanced representation of high-performing, average and low-performing students.

Result: The smaller but representative sample allows for faster model development without compromising the quality of insights.

- **Dimensionality reduction techniques**

The dataset includes over 100 variables related to student activities, such as extracurricular participation, hours spent on homework and subject preferences. Using Principal Component Analysis (PCA), the company combines these variables into a smaller set of components, such as “academic engagement” and “extracurricular involvement”.

Result: The reduced dataset simplifies the analysis while retaining the key patterns and relationships needed to understand student success.

- **Aggregation**

The attendance records are collected daily for each student, creating a large volume of data. The company aggregates this data into monthly attendance rates for each student, summarising the information without losing its relevance.

Result: The aggregated dataset makes it easier to identify long-term attendance trends and their correlation with academic performance.

- **Resampling for balancing skewed data**

The company notices that the dataset is heavily skewed, with only 5 per cent of students categorised as “at risk of failing”. To balance the dataset, it uses oversampling to duplicate records of at-risk students and undersampling to reduce the number of high-performing student records.

Result: The balanced dataset prevents bias in the predictive model, ensuring that the system accurately identifies students at risk.

- **Clustering**

The company uses clustering techniques to group students based on their performance patterns, study habits and extracurricular participation. For example, one cluster might represent high-performing students with strong academic engagement, while another might represent struggling students with poor attendance.

Result: Clustering helps the company develop tailored intervention strategies, such as tutoring programs for struggling students or advanced coursework for high achievers.

1.6 Test your knowledge

Drag the words into the correct boxes.

1) (in singular form) refers to a collection of facts usually obtained as the result of experiments, observations, transactions, or experiences.

2)

may consist of numbers, letters, words, images, and/or voice recordings.

3) Data is the contextualisation of

set in context.

4)

is a tedious process of converting raw data into an analytic ready state.

5)

is a structured representation of the sub-groups/sub-types of data.

Data pre-processing

Data

Datum

Data taxonomy

information



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=95#h5p-1>

An alternative source for this activity: <https://uts.h5p.com/content/1292858503633700099>

References

- Ackoff, R. L. (1989). From data to wisdom. *Journal of applied systems analysis*, 16(1), 3-9.
- Microsoft. (2024). Copilot [Large language model]. <https://copilot.microsoft.com/>

Additional Readings for Australian case studies

- Agile Analytics. (n.d.). *Tourism Australia Azure Power BI DAX solution case study*. Retrieved January 24, 2025, from <https://www.agile-analytics.com.au/stories/tourism-australia-azure-power-bi-dax-solution-case-study/>
- BizData. (n.d.). *Case studies*. Retrieved January 24, 2025, from <https://www.bizdata.com.au/case-studies>
- CSIRO. (n.d.). *Case studies*. Retrieved January 24, 2025, from <https://research.csiro.au/ads/case-studies/>
- Dear Watson Consulting. (n.d.). *Power BI case studies*. Retrieved January 24, 2025, from <https://www.dearwatson.net.au/power-bi-case-studies/>
- DVT Group. (n.d.). *Case studies*. Retrieved January 24, 2025, from <https://dvtgroup.com.au/case-studies/>
- Expose Data. (n.d.). *Case studies*. Retrieved January 24, 2025, from <https://exposedata.com.au/case-studies/>

[studies/](#)

- NetSuite. (n.d.). *Business intelligence examples*. Retrieved January 24, 2025, from <https://www.netsuite.com.au/portal/au/resource/articles/business-strategy/business-intelligence-examples.shtml>
- SAP Concur. (n.d.). *Case study*. Retrieved January 24, 2025, from <https://www.concur.com.au/casestudy>
- Telstra Purple. (n.d.). *Case studies*. Retrieved January 24, 2025, from <https://purple.telstra.com/insights/case-studies>
- WingArc Australia. (n.d.). *Case studies*. Retrieved January 24, 2025, from <https://wingarc.com.au/customers/case-studies/>

Acknowledgements

Thank you to Khadeejah Atif who contributed to the design of images used in this chapter (Figure 1.1, Figure 1.2 and Figure 1.3)

About the author



Dr Amara Atif
UNIVERSITY OF TECHNOLOGY SYDNEY
<https://profiles.uts.edu.au/Amara.Atif>

Dr Amara Atif is a Senior Lecturer in the School of Computer Science, Faculty of Engineering and IT (FEIT), at the University of Technology Sydney (UTS), Australia. Amara is an expert in educational technology, with a comprehensive skill set that includes research and development in the design of technology-enhanced learning environments, human-computer interaction, artificial intelligence in education, and learning analytics/educational data mining.

At UTS, she developed the undergraduate subject “Business Intelligence”, integrating culturally informed Indigenous knowledge systems through the FEIT Indigenous Graduate Attribute. She has received multiple UTS Vice Chancellor’s Learning and Teaching Awards, including recognition for data-informed teaching practices and inclusive assessment. Passionate about the accessibility and sustainability of educational resources, Amara is an advocate for Open Educational Resources (OER). *Decoding Data* continues this mission, bringing rigorous Business Intelligence concepts to learners through an open, practitioner-oriented resource grounded in Australian business contexts.

2. RETHINKING BUSINESS DECISIONS: THE SYNERGY OF DECISION SUPPORT SYSTEM, ANALYTICS, BUSINESS INTELLIGENCE, ARTIFICIAL INTELLIGENCE, AND DATA SCIENCE

Amara Atif and M.A. Qureshi

Learning Objectives (LO)

By the end of this chapter, you will be able to:

- LO1** Understand the history, development, and transformation of decision support systems over time.
- LO2** Define decision support systems, explain its key features, and recognise its role in modern business environments impacted by technology, globalisation, and data growth.
- LO3** Learn about the frameworks used to design decision support systems and understand how these frameworks guide decision-making processes.
- LO4** Understand the five stages of the decision-making process and how decision support systems support each stage.
- LO5** Identify the key characteristics, capabilities, and components of decision support systems, and examine how these elements work together to enhance decision-making.
- LO6** Understand the key concepts and fundamentals of analytics, business intelligence, artificial intelligence, and data science, and explore the roles these disciplines play in supporting decision-making.

2 Transforming decision-making: The role of decision support systems in the modern business landscape

In today's rapidly evolving business landscape, the ability to make informed and timely decisions can determine an organisation's success or failure. As businesses face unprecedented challenges such as rapid technological advancements, globalisation, the digital economy, consumer behaviour shifts, and the explosion of data generation, these factors are redefining how businesses operate and compete. These changes are creating both challenges and opportunities, necessitating a transformation in how decisions are made. Traditional decision-making approaches, which often rely on intuition or historical data alone, are increasingly inadequate to address the speed, complexity, and scalability of modern challenges.

This chapter is essential to start with because it introduces the readers to **Decision Support Systems (DSS)** – *a foundation for integrating data, technology, and analytics to enhance decision-making*. It highlights how the evolution of DSS into modern technologies such as Analytics, Business Intelligence (BI), Artificial Intelligence (AI), and Data Science equips decision-makers with tools to navigate complexity, automate processes, and gain competitive advantages. Understanding these systems and their frameworks is key to unlocking better insights, making faster decisions, and driving innovation.

2.1 The evolution, definitions, and frameworks of decision support systems (DSS)

The development of DSS can be seen as a natural progression in the evolution of organisational information systems. This evolution began with Electronic Data Processing (EDP), which focused on automating routine, structured tasks, and progressed through Management Information Systems (MIS), which aimed to generate structured reports for managerial decision-making. The emergence of DSS represents the next step in this trajectory, addressing the need for systems capable of handling complex, unstructured problems.

Defining decision support systems

Decision support systems are characterised as interactive, computer-based systems designed to assist decision-makers in solving unstructured or semi-structured problems by utilising data and models.

Unlike EDP and MIS, which focus on predefined tasks and structured reports, DSS offers flexibility and

interactivity, enabling users to explore data dynamically and simulate various scenarios to arrive at informed decisions.

Expanding perspectives on decision support systems

While traditionally defined as tools for aiding decision-making through data and models, some authors and researchers have broadened the scope of DSS to include **any system that contributes to decision-making**. This broader perspective implies that virtually all systems, except for basic transaction processing, could be considered a DSS. However, this overly inclusive definition risks diluting the concept, making it challenging to distinguish DSS from other systems or to understand their specific value and technical requirements.

Redefining decision support systems: The case of Navigation Apps

For example, a navigation app such as Google Maps illustrates the broader perspective on what could be considered a DSS. Such an app collects real-time data (traffic conditions, road closures), utilises models (shortest route algorithms, estimated arrival times), provides interactive options (alternative routes, live traffic updates), and assists users in making decisions (which route to take based on time, traffic, or preferences). From a broader viewpoint, this app “supports decisions” by helping users choose the best travel route, making it a potential DSS.

However, if the definition of DSS is applied too broadly, systems like navigation apps or even weather apps could fall under the DSS umbrella, even though they may lack the depth of functionality or technical requirements traditionally associated with DSS, such as advanced “what-if” analysis or scenario modelling.

Frameworks for understanding decision support systems

A **framework** is a structured and organised approach to developing or understanding a concept, system, or process, providing a blueprint or guidelines to systematically analyse, design, or implement solutions. It emphasises structure by defining interconnected components, offers guidance through principles and practices, and remains flexible and scalable to adapt to varying contexts or project sizes. Frameworks are widely used across disciplines, from technology to management, to simplify complex problems and ensure consistency in methodology.

In the context of DSS, frameworks play a critical role in organising thoughts and enhancing decision-making by breaking down tasks into manageable parts. Researchers often develop DSS frameworks tailored to their specific projects, ensuring that the systems address the unique requirements of the problems being studied. A good starting point for understanding DSS frameworks is the classic model developed by Gorry and Scott-Morton's "A framework for management information systems" (1971), which uses a 3×3 matrix to categorise decisions based on problem structure (ranging from structured to unstructured) and organisational level (from operational to strategic). This foundational framework highlights the necessity of integrating data, models, and user interactivity to effectively support decision-making processes.

Challenges in defining decision support systems

The term ***decision support system*** holds intuitive appeal, as any system that aids a decision might seem to qualify. Yet, restrictive definitions fail to fully capture the diverse functionalities of DSS, while overly broad definitions lack practical utility. Additionally, people from varying professional backgrounds and organisational context's view DSS differently, further complicating the development of a unified definition. This ambiguity highlights the importance of foundational works, such as Alter's "A taxonomy of decision support systems" (1977), to gain clarity on the core principles and design approaches of DSS.

2.2 Changing nature of business environments and evolving needs for decision support and analytics

Modern businesses are experiencing significant changes driven by rapid technological advancements, global interconnectedness, and increasing volumes of data. To thrive in this fast-changing environment, businesses are turning to DSS powered by data-driven technologies. These tools help businesses make quicker and smarter decisions. Below are the key drivers of change in business environments, with examples from Australian businesses included to illustrate their impact.

2.2.1 Rapid technological advancements

Emerging technologies such as the AI, machine learning, Internet of Things (IoT), and cloud computing are transforming how businesses operate. These innovations enable the collection and analysis of vast amounts of data in real-time, providing valuable insights for decision-making.

Example

Australian retailers, such as Woolworths and Myer, utilise e-commerce platforms alongside data collected from mobile apps, social media interactions, and online purchase histories to analyse customer behaviour. By understanding preferences and shopping patterns, they tailor

marketing campaigns, personalise offers, and enhance customer engagement to maintain a competitive edge in the market.

2.2.2 Globalisation

Globalisation has increased competition by allowing businesses to operate across borders, exposing them to diverse markets, regulations, and cultural nuances. This has amplified the need for agile decision-making tools.

Example

Australian wine exporters, such as those from the Barossa Valley (South Australia) and Margaret River (Western Australia) regions, face intense competition from wine-producing countries such as Chile and South Africa in critical international markets, including China. To remain competitive, these companies analyse global market trends, monitor pricing strategies, and adapt their offerings to align with regional consumer preferences, such as flavour profiles and packaging styles, in real-time.

2.2.3 Massive data generation

The explosion of digital channels, IoT devices, and customer touchpoints generates massive amounts of structured and unstructured data. Managing and extracting actionable insights from this data is essential for businesses to remain competitive.

Example

Retailers, such as major Australian chains like Coles, Woolworths, Kmart and Bunnings, collect data from in-store sensors and facial recognition technology that track customer movements, online transactions that reveal purchasing patterns, and social media interactions that highlight consumer preferences and sentiment. These data are analysed to refine product offerings, optimise inventory, and create targeted marketing campaigns that cater to specific customer segments, enhancing overall business effectiveness.

2.2.4 Consumer expectations

Modern consumers demand personalised experiences, rapid service, and ethical business practices. Businesses that fail to meet these expectations risk losing customer loyalty.

Example

An Australian retailer such as The Iconic uses predictive analytics to analyse customer browsing and purchase history, enabling it to offer tailored product recommendations and personalised discounts. This targeted approach enhances customer satisfaction and loyalty, giving the retailer a competitive advantage over rivals who rely on generic, one-size-fits-all marketing strategies.

2.2.5 Uncertainty and volatility

Unpredictable events, such as the COVID-19 pandemic and geopolitical tensions, have introduced unprecedented levels of uncertainty and volatility. These disruptions highlight the importance of robust and flexible decision-making systems.

Example

During the pandemic, Australian businesses such as Woolworths in retail and Toll Group in logistics adopted digital solutions to address challenges like supply chain disruptions and shifting consumer behaviours. Woolworths enhanced its online shopping platform and implemented real-time inventory tracking, while Toll Group leveraged digital logistics tools to optimise delivery routes and ensure timely shipments, enabling both to maintain operations and meet evolving customer demands.

2.3 The transformation of decision support systems over time

Over decades, DSS has transformed from static tools limited to structured data into sophisticated systems powered by AI and real-time processing. This evolution of DSS reflects the growing complexity of the business world and the increasing demand for agility, efficiency, and accuracy.

For Australian businesses, this journey highlights the importance of innovation in maintaining competitiveness. Whether it's a small grocer relying on basic systems or a multinational retailer using AI to predict consumer behaviour, DSS continues to shape the way decisions are made.

Early DSS: The beginning of data-driven decisions

The story of DSS begins in the 1960s and 1970s when businesses started using basic systems to assist decision-making. These early DSS were simple, focusing on structured data stored in databases and generating static reports. For instance, a retailer might have used these systems to track inventory levels manually and generate monthly stock reports. While groundbreaking at the time, these systems lacked the ability to handle dynamic environments or real-time changes.

Example: A local Australian grocer using a basic system to calculate stock levels for staples such as bread and milk manually and generate monthly stock and sales reports.

Executive Information Systems (EIS): Putting insights in the hands of leaders

By the 1980s, the rise of EIS brought DSS into the boardrooms of senior executives. These systems provided high-level summaries and dashboards that allowed decision-makers to quickly grasp business performance. Unlike their predecessors, EIS featured customisable visualisations and drill-down capabilities, giving leaders a clearer picture of their operations.

Example: An Australian manufacturing company, such as BlueScope Steel or Boral, might use an EIS to monitor key metrics related to its production processes and financial health. For instance, if the company notices that a particular factory consistently has higher costs due to overtime caused by machine breakdowns, the EIS might recommend redirecting funds to upgrade or repair machinery. Executives can then adjust their strategy to reduce costs and increase efficiency, ultimately improving profitability.

Business Intelligence (BI): Interactive insights for better decisions

In the 1990s, BI tools revolutionised DSS by introducing interactive capabilities. BI allowed businesses to integrate data from multiple sources and generate detailed analyses. For the

first time, decision-makers could go beyond static reports and explore data dynamically through ad-hoc querying and visualisations.

Example: A major Australian retailer such as Myer or David Jones using BI to analyse seasonal sales trends across multiple store locations, identifying which regions required restocking or promotional efforts.

Analytics and big data: Harnessing the power of vast data sources

The early 2000s marked the era of analytics and big data, where DSS grew smarter and more powerful. Businesses started having access to advanced statistical models and machine learning methods to process vast amounts of both structured and unstructured data. This shift enabled predictive and prescriptive analytics, allowing organisations to anticipate trends and make proactive decisions.

Example: Australian logistics companies such as Toll Group using predictive analytics to optimise delivery routes by factoring in real-time traffic, weather conditions, and customer locations.

Automation and AI: The rise of intelligent decision systems

The integration of automation and AI into DSS represents the latest stage in their evolution. These advanced systems have the capabilities to analyse vast amounts of data in real-time, dynamically adapt to changing conditions, and even autonomously execute decisions. By automating routine and repetitive tasks, these intelligent systems allow human resources to focus on more strategic and creative activities, ultimately improving efficiency and decision-making speed.

Example: Woolworths leverages AI-driven DSS to manage inventory in real-time, ensuring shelves are stocked with high-demand items during peak periods, such as holiday seasons or unexpected surges. For instance, during the COVID-19 pandemic, the system anticipated spikes in demand for essential items like toilet paper and hand sanitiser, enabling proactive restocking to meet customer needs.

2.4 Decision-making process

Decision-making is a structured and systematic process that enables organisations to address challenges or seize opportunities effectively. This process comprises five key stages, each critical to achieving informed and successful outcomes. Figure 2.1 provides a visual representation of these stages in the decision-making process.

Stage 1: Intelligence

The first stage involves identifying and defining the problem or recognising a potential opportunity. DSS supports this stage by providing tools to collect and consolidate data from various sources, such as market trends, operational metrics, and customer feedback, into a unified platform. Advanced analytics within DSS can detect patterns, identify anomalies, and flag potential opportunities automatically.

Stage 2: Design

In this stage, potential solutions are developed by brainstorming and exploring different options. DSS helps by providing tools to simulate scenarios and predict outcomes of various strategies. This allows decision-makers to test how feasible each solution is and understand the risks and impacts.

Stage 3: Choice

After identifying alternatives, the next step is to evaluate their strengths and weaknesses and choose the best option based on factors such as cost, risk, and potential benefits. DSS supports this process by providing tools such as decision matrices¹, optimisation algorithms², and visualisations. These tools simplify complex comparisons and make the decision-making process more objective and efficient.

Stage 4: Implementation

This stage focuses on putting the chosen solution into action. For implementation to succeed, teams must work together effectively, progress needs to be monitored closely, and objectives must be met as planned. DSS supports this process by providing dashboards and real-time monitoring tools to track progress. It also offers alerts and recommendations if the project deviates from the plan, helping to address issues quickly and keep the implementation on track.

Stage 5: Evaluation

-
1. Decision matrices are tables that help compare different options against key criteria. For example, when choosing a supplier, a company might use a matrix to rate suppliers on cost, delivery speed, and quality, assigning scores to each and selecting the highest-rated option.
 2. Optimisation algorithms are mathematical methods used to find the best solution from a set of options. Some commonly used optimisation algorithms are Linear Programming (LP), genetic algorithms etc. For instance, a logistics company might use an algorithm to determine the most efficient delivery routes that minimise fuel costs and travel time.

The final stage assesses how well the decision, and its implementation worked. This includes comparing actual results to the original goals, identifying successes, and finding areas for improvement. Lessons from this stage help refine future decisions. DSS supports this process by creating performance reports, highlighting what went well, and pointing out where results fell short. The system also uses feedback from these assessments to improve future decision-making strategies, enabling continuous improvement.

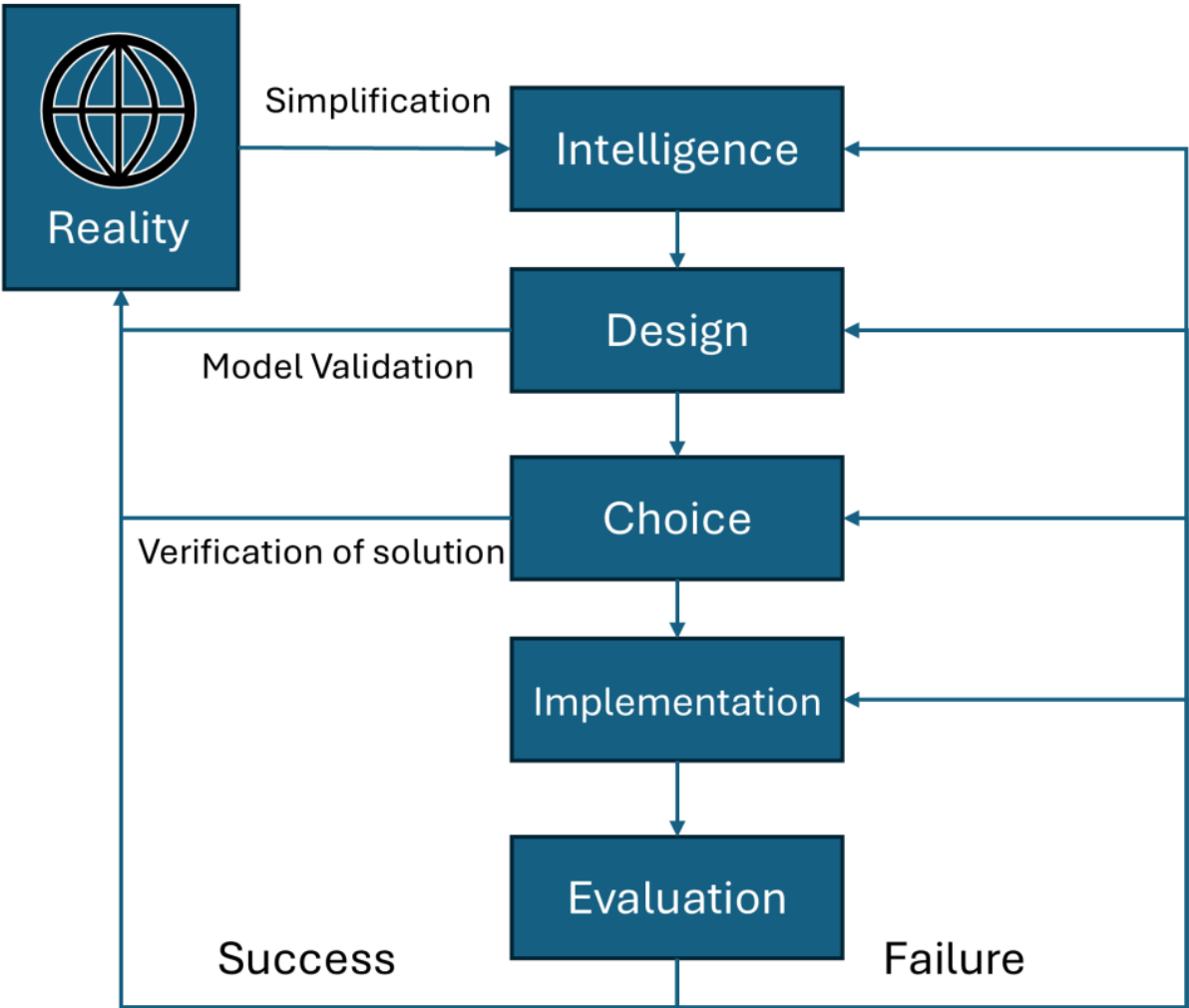


Figure 2.1 Decision-making process

Alternative Text for Figure 2.1

A flowchart of the decision-making process, illustrating five stages: Intelligence, Design, Choice, Implementation, and Evaluation, with feedback loops for validation and verification leading to success or failure.

A common misconception

The decision-making process is the same as the BI development life cycle. The answer is No – they are distinct concepts, although they can overlap in the context of decision support.

The decision-making process refers to the steps individuals or organisations take to identify problems, evaluate options, and choose the best solution. On the other hand, the BI development life cycle focuses on creating and deploying a BI system – a technical and procedural approach to building the infrastructure needed for data analysis and reporting.

While the BI development life cycle supports decision-making by providing the tools and systems needed for data-driven decisions, they serve different purposes. The BI life cycle will be discussed in detail in Chapter 9.

Example business scenario: Implementing a new product line in an Australian retail chain

An Australian retail chain, **AussieMart**, is considering introducing a new line of eco-friendly household cleaning products. This scenario describes how the decision-making process and its five stages help the company address this opportunity effectively.

Stage 1: Intelligence

The management team identifies a growing demand for eco-friendly products among Australian consumers, driven by sustainability trends and environmental awareness. DSS supports this stage by integrating and analysing data from multiple sources:

- **Market trends:** DSS tools consolidate external reports, highlighting a 25% year-on-year growth in the eco-friendly product segment.
- **Operational metrics:** DSS dashboards analyse historical sales data and detect declining demand for non-eco-friendly products.
- **Customer feedback:** Sentiment analysis through DSS tools reveals that 70% of AussieMart's customers are willing to pay more for sustainable options.

Through these capabilities, the DSS not only helps gather data but also provides actionable

insights, such as pinpointing customer preferences and emerging market trends, which allow *AussieMart* to clearly define the opportunity to launch a new product line. This ensures the organisation is well-prepared to move into the design stage.

Stage 2: Design

The team explores different strategies for launching the product line such as partnering with a local supplier for a co-branded product, importing products from a well-known international brand, and developing an in-house eco-friendly brand.

Using their DSS, they simulate scenarios for each option:

- **Scenario modelling:** Predicting potential costs, revenue, and market share for each strategy.
- **Risk analysis:** Evaluating risks like supply chain disruptions, initial investment, and brand alignment.

This stage helps the team understand the feasibility, risks, and impacts of each option, ensuring they are well-prepared to evaluate alternatives.

Stage 3: Choice

The team uses DSS tools to objectively compare the three alternatives:

- **Decision matrices:** A table evaluates each option based on key criteria such as cost, customer appeal, and time to market. Scores indicate that partnering with a local supplier is the most balanced choice.
- **Optimisation algorithms:** These help identify the optimal pricing strategy for maximising revenue while staying competitive.
- **Visualisations:** Charts and heatmaps show the projected profitability of each option over three years.

Based on the insights, *AussieMart* decides to partner with a trusted local supplier to co-brand an eco-friendly cleaning product line.

Stage 4: Implementation

To launch the new product line, the team coordinates across departments:

- Procurement works with the local supplier to finalise product specifications.
- Marketing designs a campaign emphasising sustainability and local partnership.
- Operations ensures distribution to stores and online channels.

The DSS provides:

- Dashboards to track the production, distribution, and marketing activities in real-time.
- Alerts to notify the team of delays in packaging or supply chain issues.
- Recommendations suggesting adjustments to distribution if certain regions show higher demand.

The coordinated efforts ensure that the product is launched on schedule and meets customer expectations.

Stage 5: Evaluation

After the product launch, the team evaluates its performance:

- **Sales metrics:** Reports indicate a 20% increase in sales in the household cleaning category.
- **Customer feedback:** Positive reviews highlight the product's eco-friendly benefits and affordability.
- **Profitability analysis:** The DSS compares actual revenue with projections, showing the product exceeded expectations by 10%.

The evaluation identifies areas for improvement, such as optimising supply chain logistics to reduce costs further. Lessons from this launch are incorporated into future product strategies, ensuring continuous improvement.

By following the structured decision-making process and utilising DSS at every stage, *AussieMart* successfully capitalises on a market opportunity. DSS tools play an integral role in supporting all stages – from identifying the opportunity in the intelligence stage to evaluating the success of the product launch in the evaluation stage. This approach ensures data-driven decisions, continuous learning, and long-term growth.

2.5 Decision support system characteristics, capabilities, and components

2.5.1 Characteristics and capabilities of decision support systems

The diverse nature of DSS stems from the lack of consensus on its exact definition, leading to varying interpretations of its standard characteristics and capabilities. To better understand this diversity, the key characteristics and capabilities of DSS can be grouped into four broad categories. As depicted in Figure 2.2, these categories encompass essential functionality, user support, process enhancement, data and tools, and technical versatility. The characteristics and capabilities highlighted are drawn from the works of Alter (1977), Sprague (1980), Sharda (2018), and others, and include:

1. **Core functionality and adaptability:** DSS is designed to handle not just structured problems but also semi-structured and unstructured problems, making it highly versatile. Its flexible nature allows it to adapt to a wide range of decision-making contexts, providing effective solutions for diverse challenges. This makes DSS equally valuable for structured tasks, like inventory management, and more ambiguous, unstructured challenges, such as long-term strategic planning or market entry

analysis.

2. **User support and interactivity:** DSS provides comprehensive support to individuals, teams, and managers, emphasising ease of use and adaptability to user needs. Its interactive interfaces allow users to manipulate data, test scenarios, and explore outcomes in real-time, fostering better understanding and collaboration.
3. **Process enhancement and data integration:** DSS enhances decision-making effectiveness by supporting all stages of the decision process, from problem identification to evaluation. Its data-driven approach integrates structured (e.g., sales data) and unstructured data (e.g., customer feedback), empowering users with deep insights.
4. **Analytical tools and technical versatility:** DSS offers advanced analytical tools and models, including decision matrices, optimisation algorithms, and visualisations, to simplify complex comparisons. Its technical versatility ensures it can function as a standalone system, integrate with existing platforms, or operate as a web-based solution.

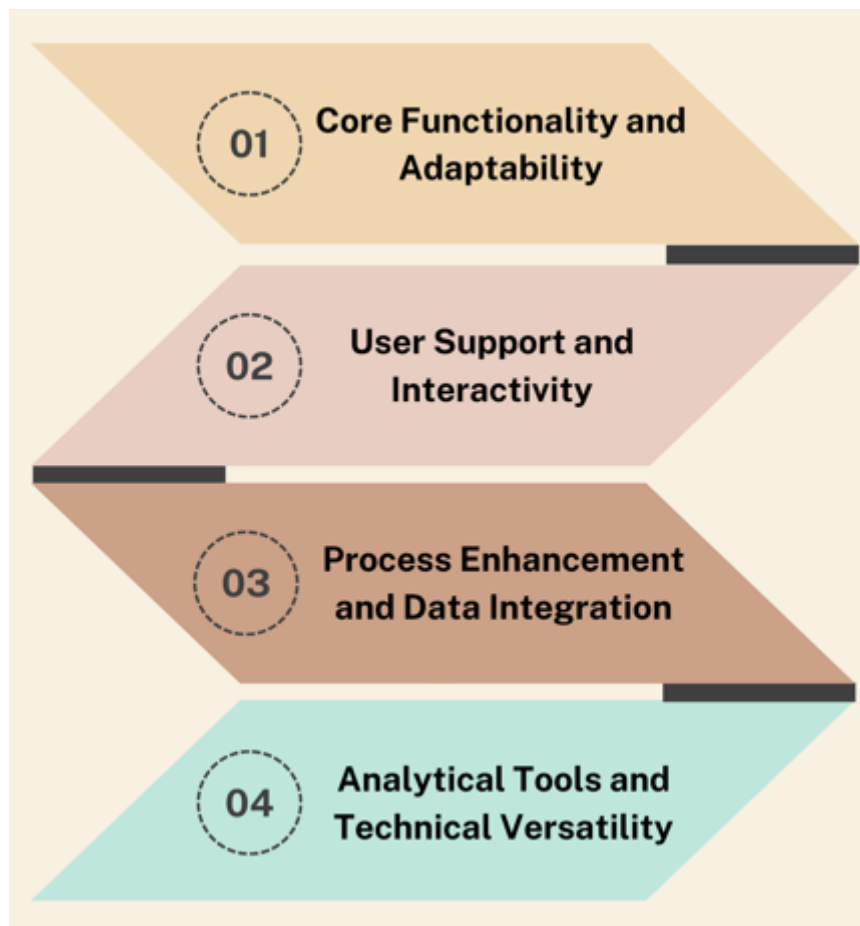


Figure 2.2: Characteristics and capabilities of decision support systems

Alternative Text for Figure 2.2

A diagram representing the four primary characteristics of Decision Support Systems:

1. Core Functionality and Adaptability,
2. User Support and Interactivity,
3. Process Enhancement and Data Integration, and
4. Analytical Tools and Textual Versatility, displayed in a stepwise format.

Business scenario: DSS in supply chain management for an Australian retail industry

To illustrate the characteristics and capabilities of DSS, consider the scenario of a large Australian retail company, such as a supermarket chain managing its supply chain operations. Retailers in Australia often face semi-structured and unstructured problems, such as predicting demand fluctuations during public holidays (e.g., Australia Day), optimising inventory across remote or regional stores, and coordinating with local and international suppliers. Table 2.1 below outlines how a DSS designed for supply chain management demonstrates its characteristics and capabilities (Column 1), applies to real-world scenarios (Column 2), and serves specific applications (Column 3):

Table 2.1: Key DSS characteristics and capabilities of retail supply chain scenario in Australia

Characteristic/ Capability	Scenario	Application
Core functionality and adaptability	The company uses DSS to address the semi-structured problem of restocking inventory during peak seasons. The system predicts demand using historical sales data, current trends, and external factors like holidays or economic shifts.	DSS supports diverse decision-making styles, from analytical (data-driven forecasting) to heuristic approaches (experience-based decisions). It enables interdependent decision-making by integrating with logistics planning, ensuring seamless coordination between inventory levels and supply chain operations.
User support and interactivity	Managers at all organisational levels from warehouse operators to supply chain executives use DSS to coordinate inventory and supplier activities effectively.	The DSS provides a user-friendly interface that allows non-technical users to generate reports, adjust restocking parameters, and analyse outcomes. It supports team collaboration by enabling multiple users to work simultaneously, share insights, and make collective decisions in real-time.
Process enhancement and data integration	During unexpected disruptions, such as a delayed shipment, the DSS helps the supply chain team analyse alternative suppliers or rerouting options in real-time.	The application of DSS supports all stages of the decision-making process in a seamless and integrated manner. During the intelligence stage, the system identifies disruptions and assesses their impact on operations. In the design stage, DSS helps explore alternative solutions, such as evaluating different suppliers or rerouting options. Moving to the choice stage, the system evaluates these alternatives by analysing costs, timelines, and reliability to determine the most suitable option. Finally, in the implementation stage, DSS facilitates the execution of the selected solution, ensuring that decisions are carried out efficiently and effectively. This comprehensive approach allows the team to manage tight deadlines while maintaining full control of the decision-making process.
Analytical tools and technical versatility	The company integrates DSS with its sales and supplier database to provide comprehensive insights into historical sales, supplier reliability, and transportation costs.	The DSS leverages advanced analytical tools like predictive models to generate actionable insights, such as optimal inventory levels or prioritising reliable suppliers. Its technical versatility ensures it can function as part of a larger Enterprise Resource Planning (ERP) system while providing web-based access for regional managers, allowing flexibility for both centralised and remote operations.

2.5.2 Components of decision support systems

Figure 2.3 illustrates the key components of a DSS, which integrates various elements such as data,

models, knowledge, and interactivity to support managers in making informed decisions. Below is a brief description of these components and their essential roles in facilitating effective decision-making.

01 Data management

Data management serves as the foundation of the DSS, responsible for collecting, storing, and organising data from various sources. These sources include internal systems, such as operational data from ERP or Point of Sale (POS) systems, legacy systems (older databases or applications still in use), and external sources, such as web services, ¹ (APIs), or public databases.

02 Model management

Model management provides the analytical tools necessary for decision-making. It includes statistical models (e.g., regression analysis, time series analysis, clustering), optimisation techniques, and simulation tools (e.g., spreadsheet simulations, Monte Carlo simulation) that allow users to explore different scenarios and conduct “what-if” analyses.

03 External models

External models enhance the system by integrating specialised software, industry-specific algorithms, or AI models that might reside outside the core DSS. This enables access to industry-specific expertise or advanced computational capabilities.

1. An API is a set of rules and protocols that allows different software applications to communicate with each other. It acts as an intermediary, enabling one system to access the functionality or data of another system in a structured and secure way. APIs are commonly used in web development, mobile apps, cloud computing, and various other fields to integrate services, retrieve data, or automate processes.

04 Knowledge-based subsystems

Knowledge-based subsystems add another layer of intelligence to the DSS by leveraging organisational knowledge and domain expertise. This includes best practices, rules, heuristics, and accumulated expertise, often supported by AI or expert systems. These subsystems ensure that decisions are contextually informed and aligned with organisational goals.

05 User interface for end-user/decisionmaker

The user interface acts as the bridge between the system and the end-user or decision-maker (e.g., manager, business analyst). It offers an interactive platform, such as dashboards, where users can input preferences or assumptions, explore insights, and visualise reports, trends, and analysis results. It ensures the system remains user-friendly, facilitating intuitive interaction and delivering actionable insights in an accessible and comprehensible format.

06 Organisational knowledge base

The organisational knowledge base supports the entire system by storing critical information such as historical data, policies, and guidelines. This provides essential context for decision-making.

07 Interaction with external systems

The DSS interacts seamlessly with other computer-based systems, such as databases, customer relationship management (CRM) systems, or supply chain systems. It also connects to external networks, such as the internet or intranets, to gather real-time data, including global market data or competitor analysis. Together, these interactions ensure a comprehensive and effective decision-making process.

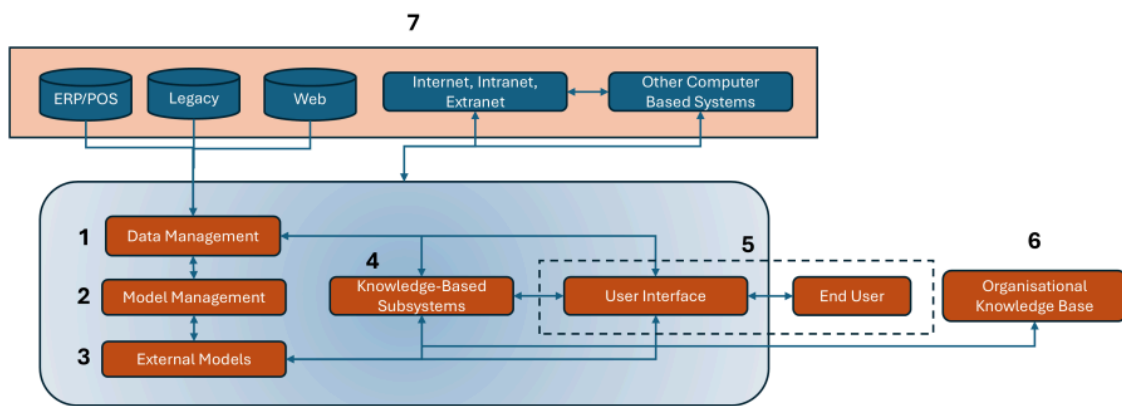


Figure 2.3: Components of a decision support system

Alternative Text for Figure 2.3

A detailed architecture diagram of a Decision Support System (DSS), featuring interconnected components: Data Management, Model Management, External Models, Knowledge-Based Subsystems, User Interface, End User, and Organisational Knowledge Base. These components interact with external systems such as ERP/POS, legacy systems, and web-based solutions.

Example: A Bushfire management DSS

Australia frequently faces bushfires and managing them effectively requires data-driven decisions. A Bushfire Management DSS used by emergency services such as the NSW Rural Fire Service integrates various components to help decision-makers assess risks, allocate resources, and protect communities.

How it works in practice?

During a bushfire, the system gathers weather and satellite data, runs predictive fire spread models, and integrates knowledge from past bushfires. Emergency managers interact with the DSS through an intuitive dashboard, visualise the predicted fire behaviour, and make critical decisions, such as allocating firefighting resources and planning evacuations.

Components of the Bushfire Management DSS

Components of the Bushfire Management DSS are critical in enabling effective decision-making during emergencies by integrating real-time data, predictive models, and user-friendly interfaces to assist emergency managers. Table 2.2 includes the components and practical applications of a Bushfire Management Decision Support System.

Table 2.2: Components and practical applications of a Bushfire Management Decision Support System

Component	Source	Function	Example in use
Data management	Real-time data from satellites, weather stations, drones, and ground sensors	Collects and stores data on temperature, wind speed, humidity, and vegetation dryness	Data from the Bureau of Meteorology (BoM) is integrated into the system to predict fire spread
Model management	Simulation and prediction models	Uses fire spread algorithms and weather forecast models to simulate bushfire behaviour	A fire simulation model predicts how a fire will spread based on wind patterns and vegetation
External models	Advanced fire prediction models or climate models developed by external institutions such as CSIRO (Australia's national science agency)	Provides detailed forecasts and long-term fire risk assessments	A climate model predicts areas of high fire risk for the upcoming season
Knowledge-based subsystems	Historical data on bushfire patterns, expert knowledge from firefighters, and lessons learned from past incidents	Adds context to the models by incorporating expert advice, past fire management strategies, and rules for resource allocation	Knowledge from previous large bushfires helps prioritise which areas to evacuate first
User interface for end-user/decision-maker	Dashboards and visualisations accessible by fire management officers and emergency coordinators	Provides an interactive platform for users to input data, explore scenarios, and visualise outputs	Emergency managers use a dashboard to evaluate predicted fire paths, resource needs, and evacuation timelines
Organisational knowledge base	Policies, emergency protocols, and historical incident records stored within the system	Supplies additional context and supports compliance with national fire management guidelines	Accessing past incident records to refine evacuation plans and strategies
Interaction with external systems	Integration with external agencies like the BoM, local councils, and fire services	Ensures collaboration and access to external data for comprehensive decision-making	Synchronisation with local government systems to issue evacuation orders and community alerts in real-time

2.6 Analytics vs Business Intelligence vs Artificial Intelligence vs Data Science

2.6.1 Overview and historical context

2.6.1.1 Analytics

According to the Institute for Operations Research and the Management Sciences (INFORMS), **analytics** is defined as **the scientific process of transforming data into insight for making better decisions**.

This definition emphasises the systematic approach of utilising data analysis to derive actionable insights that enhance decision-making processes. Analytics has its roots in statistics and mathematics, which date back centuries. Formal methods such as regression analysis began emerging in the late 1800s. By the mid-1900s, analytics evolved further with the introduction of computational tools and statistical software. It is the foundational concept that laid the groundwork for more specialised fields. Figure 2.4 illustrates the categories of business analytics, highlighting its evolution and applications in modern decision-making.

INFORMS further categorises analytics into three distinct types:

- **Descriptive (or reporting) analytics:** Provides insight into past events by analysing historical data.
- **Predictive analytics:** Offers foresight into future events by identifying patterns and trends.
- **Prescriptive analytics:** Recommends actionable strategies to achieve desired outcomes.

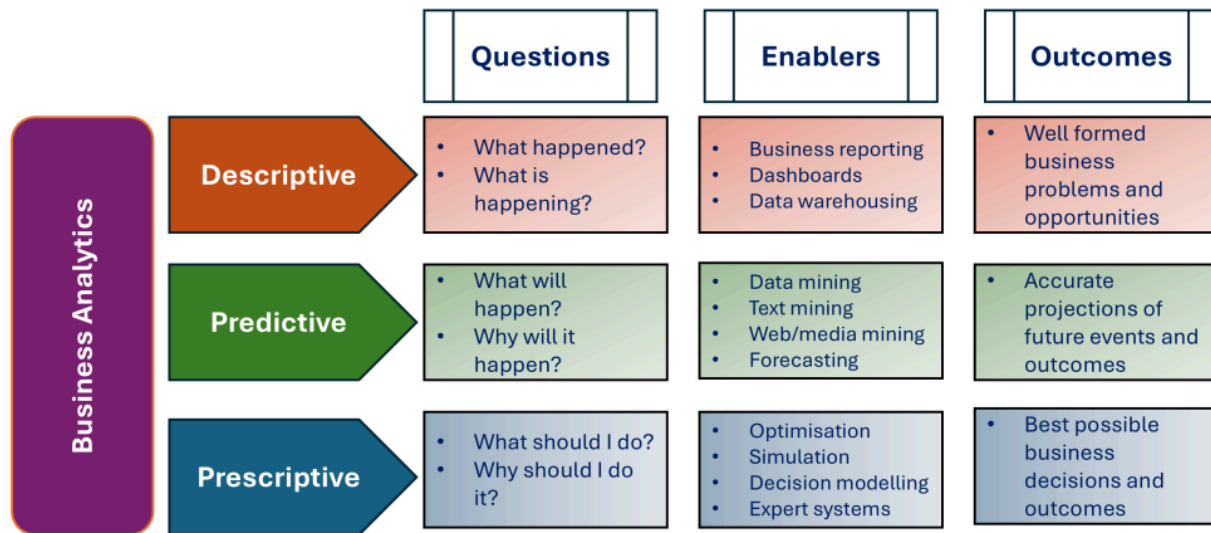


Figure 2.4: Categories of business analytics

Alternative Text for Figure 2.4

A layered framework of business analytics types: Descriptive, Predictive, and Prescriptive analytics. Each layer highlights its associated questions (e.g., What happened?), enabling tools (e.g., dashboards, data mining), and outcomes (e.g., actionable insights or decision optimisation).

2.6.1.2 Business Intelligence

Business Intelligence (BI) refers to **the technologies, processes, and practices used to collect, analyse, and present business data to help organisations make informed decisions**. It involves gathering data from various sources, transforming it into actionable insights, and presenting it through visualisations, reports, and dashboards.

Key elements of BI include data mining, querying, reporting, and online analytical processing (OLAP), which help to improve efficiency, optimise processes, and support decision-making at all levels of an organisation. The term “business intelligence” was first used by Richard Millar Devens in his 1865 book, *Cyclopædia of Commercial and Business Anecdotes*, to describe how banker Sir Henry Furnese profited by receiving and acting upon information about his environment before his competitors. In 1958, IBM researcher Hans Peter Luhn further developed the concept by describing a business intelligence system

as an “automatic system... developed to disseminate information to the various sections of any industrial, scientific, or government organisation”. The 1980s and 1990s saw the development of DSS and the emergence of BI tools from companies such as SAP and Oracle, focusing on creating dashboards and reporting mechanisms for improved decision-making. Today, BI continues to evolve, building on analytics to support structured business decision-making processes and enabling organisations to gain competitive advantages through data-driven insights.

2.6.1.3 Artificial Intelligence

Artificial Intelligence (AI) refers to **the simulation of human intelligence in machines that are programmed to perform tasks that typically require human cognition.**

These tasks can include learning, reasoning, problem-solving, perception, natural language understanding, and decision-making.

Artificial Intelligence began as a theoretical field in the 1940s and 1950s. Early milestones include Alan Turing’s 1950 paper, “Computing Machinery and Intelligence”, which introduced the concept of machines capable of simulating human intelligence and proposed the Turing Test as a criterion for machine intelligence. The term AI was officially coined at the 1956 Dartmouth Conference. This event is considered the birth of AI as a distinct field of study. Progress in AI was initially slow, facing periods of reduced funding and interest known as “AI winters”. However, advancements in computing technology during the late 20th century, particularly in the 1980s and 1990s, revitalised the field. The development of machine learning algorithms increased computational power, and the availability of large datasets contributed to this revival. In recent years, AI has experienced significant growth, with advancements in deep learning, natural language processing, and robotics, leading to applications across various industries, including healthcare, finance, and transportation.

2.6.1.4 Data Science

Data Science is an **interdisciplinary field** that focuses on **extracting knowledge and insights from typically large data sets to solve problems across various application domains.** It encompasses preparing data for analysis, formulating data science problems, analysing data, developing data-driven solutions, and presenting findings to inform high-level decisions.

The term *Data Science* has been attributed to various origins. In 1974, Peter Naur used the term in his

book “Concise Survey of Computer Methods”, defining it as “the usefulness of data and data processes derives from their application in building and handling models of reality”. The field became more distinct in the early 2000s, driven by the big data revolution and the increasing need to analyse and interpret vast amounts of data. Data Science combines elements of statistics, AI, and computer science, focusing on extracting insights and building predictive models. It integrates aspects of analytics, BI, and AI to address modern data challenges, offering a comprehensive framework for working with complex datasets. The evolution of Data Science has been marked by significant milestones, including the development of machine learning algorithms, advancements in data processing technologies, and the emergence of data visualisation techniques. These developments have enabled Data Science to become a critical component in decision-making processes across various industries, from healthcare to finance.

Figure 2.5 shows the timeline summary of the evolution of key concepts in data-driven decision-making, including Analytics (1880–1940), Artificial Intelligence (1940–1970), Data Science (1970–1980), and Decision Support Systems (1980–1990), highlighting their major milestones and contributions to modern practices.

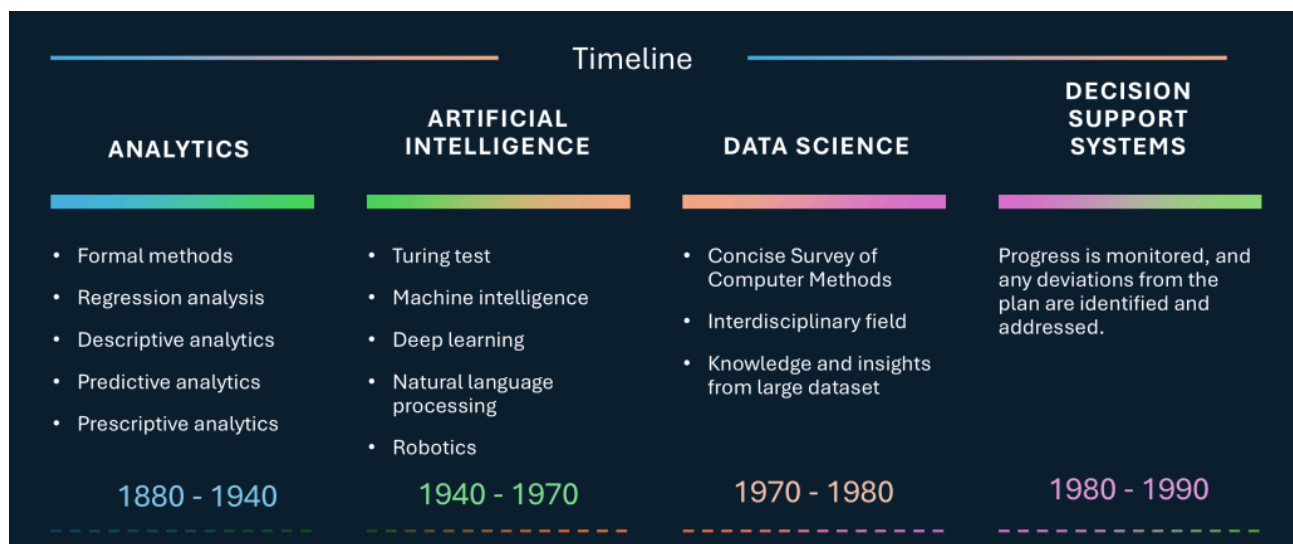


Figure 2.5: Timeline summary

Alternative Text for Figure 2.5

A timeline summarising the evolution of data-driven technologies, spanning from Analytics (1880–1940) with regression and descriptive methods, Artificial Intelligence (1940–1970) with advancements in machine learning and robotics, Data Science (1970–1980) focusing on insights from large datasets, to Decision Support Systems (1980–1990) monitoring and addressing deviations in plans.

2.6.2 Primary purpose and focus

Table 2.3 outlines the primary purpose and focus of each field, along with the key questions they address. This comparison provides a clear understanding of how analytics, BI, AI, and data science differ in their objectives and the types of insights they deliver.

Table 2.3: Primary purpose and focus of analytics, BI, AI, and data science

Field	Primary Purpose	Key questions addressed
Analytics	Interpreting past and present data to identify trends and patterns.	What happened? Why did it happen?
BI	Organising and visualising data for operational and strategic decisions.	What is happening now? What should we focus on?
AI	Automating tasks and mimicking human intelligence.	How can tasks be automated or decisions optimised?
Data Science	Solving complex problems and building predictive models.	What patterns exist? What will happen next?

2.6.3 Key features and tools

Analytics involves tools such as **spreadsheets**, **SQL**, **PostgreSQL**, **Mongo DB**, and **statistical software** e.g., SAS or R, which help derive descriptive and diagnostic insights. A notable Australian example is the analysis of retail sales trends by Woolworths Group to identify growth areas and customer preferences using historical transaction data.

Business Intelligence utilises **dashboards**, **Extract-Transform-Load (ETL) processes**, and **BI software** such as Qlik, Tableau and Power BI. For instance, the Australian financial services company ANZ Bank uses BI to create real-time performance dashboards, enabling managers to monitor key operational metrics and make data-driven decisions effectively.

Artificial Intelligence employs **machine learning frameworks** such as **TensorFlow** and **PyTorch** to create automated systems and predictions. An Australian application includes the use of AI in mining by companies such as Rio Tinto, which deploy autonomous vehicles and predictive maintenance systems to enhance efficiency and safety in mining operations. Additionally, AI-powered chatbots are increasingly used by Australian companies such as Qantas to provide 24/7 customer support, answering queries about flights and bookings efficiently. AI-driven recommendation engines, like those used by e-commerce platforms in Australia, personalise product suggestions based on user behaviour to enhance customer satisfaction and drive sales.

Data Science relies on **programming languages** such as **Python** and **R**, alongside **big data platforms** such as **Hadoop** and **Spark**, to develop predictive models and actionable insights. In Australia, CSIRO's Data61 uses data science to create models that predict bushfire risks, combining environmental data and advanced analytics to support emergency planning and response. Another example is the use of data science by telecommunications companies such as Telstra, which builds predictive models to identify potential customer churn. By analysing customer behaviour patterns, Telstra can proactively address concerns and offer tailored services to improve customer retention.

2.6.4 Role in decision support

Table 2.4 compares how analytics, BI, AI, and data science contribute to decision-making processes. It highlights the level of automation in each field and the extent to which they directly or indirectly support human decisions, providing clarity on their unique roles in guiding or automating decisions.

Table 2.4: Role in decision support

Field	Decision support role	Level of automation
Analytics	Direct support through trend and pattern analysis	Mostly human-driven
BI	Direct support via dashboards and real-time insights	Semi-automated dashboards
AI	Indirect support; often automates decisions rather than supporting them	Highly automated
Data science	Indirect support by providing models and patterns	Mix of automation and human-driven exploration

2.7 Test your knowledge

Drag the words into the correct boxes.

1) DSS is highly versatile because it adapts to both and problems.

2) Business Intelligence aims to collect, analyse, and present

data for decision-making.

3) The revival of AI in the late 20th century was driven by advancements in

technology, learning algorithms, and data .

4) Data Science focuses on extracting

and from data to solve problems.

5) Data Science integrates

, , and computer science to address modern data challenges.

structured

insights

AI

computing

knowledge

machine

business

statistics

informed

unstructured

availability



The interactive version of this H5P content is available at:
<https://oercollective.caul.edu.au/decoding-data/?p=221#h5p-2>

An alternative source for this activity: <https://uts.h5p.com/content/1292858504922397299>

References and additional readings

- Alter, S. (1977). A taxonomy of decision support systems. *Sloan Management Review* (pre-1986), 19(1), 39.
- Foote, K. D. (2021, October 16). *A brief history of data science*. DATAVERSITY. Retrieved from <https://www.dataversity.net/brief-history-data-science/>
- Gorry, G. A., & Scott Morton, M. S. (1971). A framework for management information systems.
- Guimaraes, T., Igarria, M., & Lu, M. T. (1992). The determinants of DSS success: An integrated model. *Decision Sciences*, 23(2), 409-430.
- Sharda, R., Delen, D., and Turban, E. *Business Intelligence, Analytics and Data Science: A Managerial Perspective*, Fourth Edition, Global Edition, Pearson Education Limited, 2018.
- Sprague Jr, R. H. (1980). A framework for the development of decision support systems. *MIS quarterly*, 1-26. Retrieved from:
<https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=618e96a432e22e2fe353274b45d397b71a7b9ec8>

Acknowledgements

Thank you to Khadeejah Atif who contributed to the design of Figure 2.2 used in this chapter.

About the authors



Dr Amara Atif

UNIVERSITY OF TECHNOLOGY SYDNEY

<https://profiles.uts.edu.au/Amara.Atif>

Dr Amara Atif is a Senior Lecturer in the School of Computer Science, Faculty of Engineering and IT (FEIT), at the University of Technology Sydney (UTS), Australia. Amara is an expert in educational technology, with a comprehensive skill set that includes research and development in the design of technology-enhanced learning environments, human-computer interaction, artificial intelligence in education, and learning analytics/educational data mining.

At UTS, she developed the undergraduate subject “Business Intelligence”, integrating culturally informed Indigenous knowledge systems through the FEIT Indigenous Graduate Attribute. She has received multiple UTS Vice Chancellor’s Learning and Teaching Awards, including recognition for data-informed teaching practices and inclusive assessment. Passionate about the accessibility and sustainability of educational resources, Amara is an advocate for Open Educational Resources (OER). *Decoding Data* continues this mission, bringing rigorous Business Intelligence concepts to learners through an open, practitioner-oriented resource grounded in Australian business contexts.



Dr Muhammad Atif Qureshi

TORRENS UNIVERSITY AUSTRALIA

<https://www.torrens.edu.au/>

Dr. M.A. Qureshi is an expert in Software Modelling, Meta-modelling, and Modelling Languages, holding a PhD in Software Engineering. With years of combined experience in both the IT industry and academia, he has worked with leading organisations to implement innovative BI solutions that drive data-driven decision-making, optimise business processes, and enhance organisational performance.

As an academic, he employs constructive and effective teaching methods that foster a collaborative learning

environment. He takes a futuristic approach to learning, engaging students through both theory and practical application. Additionally, as a certified business analyst (ECBA-IIBA), he integrates his unique blend of industry experience and academic insight to provide a comprehensive, real-world perspective on Business Intelligence. This enables him to make complex concepts accessible to both beginners and seasoned professionals.

3. ARCHITECTURE OF BUSINESS INTELLIGENCE

Amara Atif

Learning Objectives (LO)

By the end of this chapter, you will be able to:

LO1 Understand how business intelligence (BI) architecture supports decision-making by transforming raw data into actionable insights.

LO2 Identify the main components and layers of BI architecture.

LO3 Understand the difference between OLTP (online transactional systems) and OLAP (online analytical systems) and their roles in BI architecture.

LO4 Compare and contrast various BI architecture frameworks, understanding their advantages and challenges.

LO5 Explore real-world examples of BI architecture implementation to understand how organisations successfully leverage BI systems for business success.

3 Overview of business intelligence architecture

BI architecture refers to the framework that manages the flow of data through the system to provide valuable insights. It involves various components such as data sources, Extract-Transform-Load (ETL) processes, data storage, data analytics tools, reporting systems, and more. A well-designed BI architecture ensures that data (i) flows efficiently through the system, (ii) is processed correctly, and (iii) is made available to decision-makers in a usable format.

3.1 Key concepts in BI architecture

3.1.1 What is BI architecture?

BI architecture is a **blueprint** that ***outlines the structure of BI systems and their components***. It defines how data is collected from different sources, processed, stored, analysed and presented to users. The BI architecture ensures that all parts of the BI system interact seamlessly, providing a cohesive and efficient process for generating business insights.

3.1.2 The role of BI architecture in modern business

The role of BI architecture is to support organisations in converting raw data into strategic insights. This involves the integration of various technologies and processes that help collect data from different sources, process it, and then deliver it to the appropriate stakeholders. A strong BI architecture is essential for the smooth flow of information and helps organisations make quick, data-backed decisions.

3.1.3 Components of BI architecture

As mentioned in the overview, the architecture of BI systems consists of several essential components, including data sources, ETL processes, storage systems, analytics tools and reporting systems. Each component plays a vital role in transforming raw data into actionable insights for decision-making. To fully understand how BI systems work, it is important to comprehend the underlying data processing systems that power these components. The way data is processed, stored and accessed directly influences how organisations utilise their BI systems to generate valuable insights and make informed decisions.

Example scenario: A retail store using BI architecture

A retail chain, **ShopSmart**, operates physical stores and an online platform. The company aims to optimise inventory management and make data-driven business decisions.

Data collection: ShopSmart collects raw data from multiple sources:

- point-of-sale (POS) systems (sales transactions in physical stores)

- e-commerce platforms (customer orders)
- social media (customer feedback).

Data processing (ETL): The collected data is processed through an ETL pipeline:

- **Extract** – data is pulled from all sources (sales, online, social media).
- **Transform** – the data is cleaned, standardised and aggregated for applying business logic (e.g., sales totals by region, customer sentiment).
- **Load** – transformed data is stored in a central storage system such as a data warehouse for structured analysis and a data lake for unstructured data like social media reviews.

Data analysis: With BI tools such as Tableau or Power BI, ShopSmart can analyse the processed data. For example, the analysis reveals that certain products are selling fast in specific regions and that social media sentiment around new promotions is positive.

Reporting and dashboards: Real-time dashboards provide managers¹ with updated data on sales, inventory levels and customer feedback, allowing for quick decisions on restocking or adjusting marketing strategies.

Delivering insights: These insights are delivered to key stakeholders (e.g., inventory managers and marketing teams) who use the information to make informed decisions about stock levels and promotions.

In this scenario, ShopSmart's BI architecture integrates various technologies and processes to convert raw data into valuable, strategic insights. Data is collected from multiple sources, processed for analysis and delivered to the appropriate stakeholders, enabling the company to make quick, data-backed decisions that optimise its operations.

3.2 OLTP vs. OLAP

To ensure the data in BI systems is useful and actionable, organisations need to manage two fundamental types of data processing that form the backbone of BI systems.

3.2.1 OLTP (OnLine Transaction Processing)

OLTP systems are designed for managing and processing everyday business transactions that keep an organisation running. These systems are designed for fast data entry and retrieval, ensuring real-time processing of transactional data.

1. Here the managers referred to are those who oversee various operational and strategic areas of the business. Specially, these might include inventory managers, marketing managers, sales managers or operations managers.

Transaction and transactional data

A **transaction** is an event or activity that involves the exchange or transfer of goods, services or information and **transactional data** refers to the information captured during the execution of a transaction. It represents the details of an event or interaction and is typically recorded and processed by various systems (e.g., databases, accounting software, POS systems). This data is used to track the actions taken and the outcomes of those actions, often forming the basis of operational business processes.

For example, in a sales transaction, the data might include customer details (name, contact information), product information (product ID, price, quantity), transaction date and time, payment method (credit card, cash, online payment), sales tax applied and total amount paid.

Table 3.1 provides an overview of some key characteristics of OLTP.

Table 3.1: Characteristics of OLTP

Characteristic	Description
High Transaction Volume	OLTP systems handle frequent transactions, such as sales orders, inventory updates or financial transactions.
Real-time Processing	Transactions are processed immediately, allowing businesses to update and access critical information in real time.
Data Structure	OLTP systems are designed to manage normalised, relational databases, which focus on maintaining data integrity and reducing redundancy.

Examples

Relational DBMS (e.g., Microsoft SQL Server, Oracle Database, MySQL, PostgreSQL), ERP systems (e.g., SAP ERP, Microsoft Dynamics 365, Oracle ERP), CRM systems (e.g., Salesforce, HubSpot CRM), POST Systems (e.g., Square² POS, Toast³ POS), and e-commerce platforms.

3.2.2 OLAP (OnLine Analytical Processing)

OLAP systems focus on analysing large complex datasets over time to generate insights that help guide long-term strategy and decision-making. They allow users to query large datasets, often involving multi-dimensional analysis, to understand trends, patterns and business performance. OLAP is used primarily for decision support and strategic planning.

Table 3.2 provides an overview of some key characteristics of OLAP.

Table 3.2: Characteristics of OLAP

Characteristic	Description
Historical Data	OLAP systems work with aggregated and historical data rather than transactional or real-time data.
Multi-dimensional Analysis	Data in OLAP systems is typically stored in multi-dimensional formats such as OLAP cubes, allowing users to slice and dice data along various dimensions (e.g., time, geography and product categories).
Complex Queries	OLAP queries are typically more complex, involving large datasets and requiring significant computational resources to generate reports and dashboards.

Examples

Data warehouse tools (e.g., Amazon Redshift, Google BigQuery, Snowflake), business intelligence and analytics tools (e.g., Tableau, Microsoft Power BI, QlikView, IBM Cognos Analytics), OLAP Cube tools (Microsoft SQL Server Analysis Services – SSAS, Oracle OLAP), and Excel with OLAP Add-ins (e.g., Excel with Power Pivot).

Now that we have reviewed both, can you highlight the key differences between

-
- 2. Square POS is a versatile, cloud-based system suitable for various industries, including retail, food and beverage, and professional services.
 - 3. Toast POS is specifically tailored for the restaurant industry. It provides comprehensive solutions for order management, payment processing, and customer engagement. Toast POS integrates with various hardware components, such as handheld devices for tableside ordering and kitchen display systems, to streamline restaurant operations.

OLTP and OLAP?

3.2.3 How do OLTP and OLAP work together in BI architecture?

In a typical BI system, OLTP systems provide the raw transactional data, which is then extracted, transformed and loaded (ETL) into OLAP systems for deeper analysis. While OLTP systems ensure the smooth operation of business processes, OLAP systems help with reporting, trend analysis and decision-making by providing a historical view of business data.

Scenario in the health industry: OLTP and OLAP systems in a hospital

Imagine a **hospital** that uses both OLTP and OLAP systems to manage and analyse patient data. Here is how these systems work together in this healthcare environment:

OLTP system – real-time transactional data

In the hospital, OLTP systems are responsible for handling real-time transactional data. These systems record every action that happens in the hospital's day-to-day operations. For example:

- **Patient registration:** When a new patient arrives at the hospital, the registration system records their personal details (e.g., name, age, medical history, insurance).
- **Appointment scheduling:** When a patient books or reschedules an appointment with a doctor, the OLTP system updates the hospital's calendar, records the time slot and confirms availability.
- **Billing and payment:** As medical services are rendered (e.g., tests, treatments, surgeries), the billing system processes the charges, generates invoices and tracks payments in real time.

The key here is that OLTP systems handle high-frequency, operational transactions that need to be processed immediately and accurately to ensure smooth daily operations of the hospital.

OLAP system – in-depth analysis and decision-making

Once the OLTP systems capture the transactional data, it is extracted, transformed and loaded (ETL) into OLAP systems for deeper, multi-dimensional analysis. For example:

- **Trend analysis:** OLAP systems allow the hospital to analyse patient outcomes over time, such as identifying trends in the recovery rates for certain procedures or diseases. By looking at historical data (e.g., treatment results over the past five years), hospital administrators can identify which treatments are most effective.

- **Resource allocation:** OLAP systems aggregate patient admission data, medical staff workloads and equipment usage to identify patterns in hospital resource utilisation. For example, they can reveal peak times for patient admissions, allowing the hospital to allocate resources (e.g., beds, staff) more efficiently.
- **Cost management:** Using historical financial data, OLAP tools can generate detailed reports on hospital spending, identifying trends in operational costs, patient billing and insurance payments. This helps the finance department make informed decisions about budgeting and cost control.

The OLAP systems provide strategic insights by analysing large volumes of aggregated and historical data, helping hospital administrators and decision-makers make long-term decisions on resource allocation, patient care strategies and operational improvements.

How do OLTP and OLAP work together in a hospital scenario?

In this healthcare setting, OLTP systems ensure the smooth running of day-to-day tasks such as patient registration, billing and appointment scheduling. These systems provide real-time transactional data that is crucial for the hospital to function efficiently.

Meanwhile, OLAP systems enable healthcare administrators to analyse historical data over months or years to identify patterns, trends and insights that support strategic decision-making. For example, they might identify seasonal spikes in flu cases, forecast the need for additional staff or resources, or track the effectiveness of specific treatments over time.

3.3 BI architecture components and layers

The components and layers of BI architecture are often divided into tiers to organise and structure the BI system effectively. This tiered approach helps break down the architecture into manageable sections, ensuring that each component serves a specific function in the process of collecting, storing, analysing and presenting data. Typically, these tiers are referred to as three-tier architecture, four-tier architecture or even five-tier architecture, depending on the complexity and needs of the organisation.

In the context of this book, our focus will be limited to the three-tier architecture.

3.3.1 Three-tier BI architecture

This is the most common and simplest form of BI architecture, and it consists of three primary tiers:

3.3.1.1 Tier 1: Data source layer (data collection layer)

This tier is where raw data is generated and stored. It includes transactional data (e.g., from OLTP systems) and other types of data (e.g., external data feeds, social media data, unstructured data).

Components:

- Operational databases – these include **OLTP systems** that manage transactional data (e.g., sales transactions, customer interactions).
- External data sources – this includes **APIs**, third-party data and external datasets that are integrated into the system for a broader view of data.
- Flat files – includes **CSV**, **Excel** and other structured files used for data transfer or manual uploads.
- Unstructured data – data from sources such as **social media**, **text data** and other formats that don't have a predefined structure.
- IoT devices and sensors – devices that capture **real-time data** (e.g., environmental sensors, patient monitoring devices) that feed into the system for immediate analysis.

3.3.1.2 Tier 2: Data integration, storage and modelling layer (ETL, data warehouse and data modelling)

This tier handles the process of extracting, transforming and loading (ETL) data from various sources into a centralised storage system, which could be a data warehouse, data lake, or both. It also includes data modelling to ensure that the data is structured within the storage system in a way that facilitates easy analysis and efficient querying and reporting.

Components:

- ETL tools – these tools (e.g., **Talend**, **Informatica**, **Apache Nifi**) handle the process of extracting data from multiple sources, transforming it (cleaning, standardising and enriching) and loading it into storage systems.
- Data warehouse – a system for storing structured data that is ready for analysis, typically based on relational databases (e.g., **Amazon Redshift**, **Google BigQuery**, **Microsoft SQL Server**).
- Data lake – a storage system designed to handle unstructured or semi-structured data, often cloud-based (e.g., **Hadoop**, **AWS S3**, **Azure Data Lake**, **Apache Spark**).
- Data marts – subsets or smaller, specialised versions of a data warehouse, designed for specific departmental or business unit (e.g., finance, marketing or sales) needs.
- For data modelling – common data models include *star schema* and *snowflake schema*.
- Star schema – a simplified model where a central fact table is connected to multiple dimension tables, facilitating fast queries and reporting.
- Snowflake schema – an extension of the star schema that normalises the dimension tables, reducing redundancy and enhancing data integrity.

3.3.1.3 Tier 3: Analytics and presentation layer (BI tools and dashboards)

This tier is where data is analysed and presented. BI tools are used to generate reports, perform data analysis and provide visualisations. This layer allows users to interact with the data and access actionable insights.

It bridges the gap between raw data stored in previous layers and business decision-makers who need clear, understandable insights.

Components:

- BI tools – these tools allow users to perform ad-hoc queries, generate reports and explore data visually. Examples include **Tableau**, **Power BI** and **QlikView/Qlik Sense**, which offer powerful data exploration and reporting capabilities.
- Reporting tools – these tools are specifically designed to create formal reports, often with advanced features for scheduling, distribution and formatting. Examples include **IBM Cognos** and **SAP BusinessObjects**.
- Data visualisation – data is presented in the form of *charts*, *graphs* and *dashboards*, allowing users to visually interpret trends, patterns and insights quickly.
- OLAP tools – tools such as **Microsoft SSAS** and **Oracle OLAP** allow for multi-dimensional data analysis, where users can slice and dice data across various dimensions (e.g., time, geography, product) for deeper insights.
- Advanced analytics – this includes tools and techniques such as *predictive analytics* and *machine learning models* that help identify trends, forecast future outcomes and make data-driven predictions.

Misconceptions

01 Microsoft Excel as an ETL tool

Excel itself is not traditionally considered a fully-fledged **ETL (Extract, Transform, Load)** tool, but it can be used for simple ETL processes, especially on a smaller scale. Excel's built-in features allow users to extract, transform and load data, but it lacks the advanced automation and scalability of dedicated ETL tools. Excel is not built for handling very large datasets (millions of rows), which can cause performance issues. While you can set up some automation with **VBA (Visual Basic for Applications)**, it's not as robust or scalable as dedicated ETL tools. ***Excel can serve as a useful, albeit limited, ETL tool.***

02 Microsoft Excel as a data warehouse or tool for creating a data warehouse

While **Excel** is an excellent tool for data analysis, calculations and reporting, it **cannot** be considered a **data warehouse**, nor is it typically used to **create a data warehouse** in a traditional sense.

Excel is not designed to handle the massive data volumes that a typical data warehouse manages. Excel files have row limits (e.g., 1,048,576 rows in Office 365 Excel 2024), and

performance reduces significantly when dealing with large datasets. Excel lacks the ability to automatically integrate and consolidate data from multiple sources in real-time or on a scheduled basis, which is a fundamental characteristic of a data warehouse. **ETL processes** are usually required to populate a data warehouse, something Excel cannot handle efficiently without manual intervention.

Excel is not built for concurrent access by multiple users and is not optimised for collaborative work on a large scale, which is often required for data warehouse systems. Data warehouses are optimised for running complex queries across large datasets efficiently, which Excel is not designed for. Running complex queries on a large dataset in Excel can lead to poor performance or crashes.

While Excel is **not suitable as a data warehouse** or **for creating a data warehouse**, it can still play a role in conjunction with a data warehouse. After data is loaded into a data warehouse, Excel can be used as a **front-end tool for data analysis, creating reports and visualisations**. Excel can connect to data warehouses using tools like **Power Query** or **Power Pivot**. For small projects or teams with limited data needs, Excel can be used to store and manage data for analysis, but this is not the same as a full data warehouse solution. Data extracted from a data warehouse can be exported into Excel for detailed analysis, reporting or ad-hoc querying.

03 Excel as a data mart: a limited use case

Having said that, **Excel can be used as a data mart** in specific scenarios, but it has many limitations compared to traditional data mart solutions. While Excel is not typically used as a fully-fledged **data mart** for large-scale enterprise environments, it can serve as a **simple, lightweight data mart** for smaller or less complex needs.

3.4 BI architecture frameworks

The architecture framework for BI provides various approaches to structuring and implementing BI systems. These frameworks help organisations determine the best strategies for managing data, performing analysis and making informed decisions. In this section, we will explore key BI architecture approaches.

3.4.1 Top-down vs. bottom-up approach

The top-down and bottom-up approaches are two strategies for implementing BI architecture, specifically focusing on how to organise and structure the data storage and analysis layers (such as data warehouses and data marts). Both approaches involve creating a system that consolidates data and makes it available for reporting and decision-making, but they do so in different ways.

3.4.1.1 Top-down approach

The top-down approach begins by first building a central data warehouse that acts as a single, comprehensive repository for all organisational data. The data warehouse is populated with data from various operational systems (e.g., sales, finance, marketing) and external sources. The idea is to create a unified source of truth where data from across the organisation is integrated and cleaned. Once the data warehouse is established, data marts and reporting systems are built around it.

Table 3.3 lists the advantages and challenges of the top-down approach.

Table 3.3: Advantages and challenges of the top-down approach

Advantages	Challenges
Centralised data management ensures consistency across the organisation.	The upfront effort to build a central data warehouse can be complex and time-consuming.
Reporting and analytics are based on a unified data source, reducing duplication and errors.	Requires significant planning and a clear understanding of data integration needs.
Scalable for long-term, large-scale data analysis needs.	

3.4.1.2 Bottom-up approach

The bottom-up approach, in contrast, starts by building small, departmental data marts rather than a centralised data warehouse. These data marts are designed to serve specific business functions (e.g., finance, sales or HR), and they store data tailored to the needs of the department. Once these individual data marts are established and working, they are gradually integrated into a larger, enterprise-wide data warehouse. This approach allows for quicker wins because each department can start using its data mart immediately, and they can later consolidate the data as part of a broader system.

Table 3.4 lists the advantages and challenges of the bottom-up approach.

Table 3.4: Advantages and challenges of the bottom-up approach

Advantages	Challenges
Quicker implementation since departments can work independently and don't have to wait for a centralised warehouse.	Potential for inconsistency between departments, as each data mart may have different data models.
Business units can directly control and tailor their data to meet specific needs.	As the system grows, integrating data marts into a central warehouse can be complex and require significant rework.

3.4.2 Real-time BI vs. batch processing

In BI systems, real-time BI and batch processing represent two different methods for managing how and when data is updated, processed and analysed.

3.4.2.1 Real-time BI

Real-time BI focuses on continuous, live data feeds. It involves processing and analysing data as it becomes available. This means that business intelligence reports, dashboards and insights are updated in real time (or near real time), providing organisations with the most up-to-date information. For instance, in a retail environment, real-time BI could track customer purchases as they happen, enabling immediate inventory adjustments or personalised offers.

Table 3.5 lists the advantages and challenges of the real-time BI approach.

Table 3.5: Advantages and challenges of real-time BI

Advantages	Challenges
Provides up-to-the-minute insights that can lead to quick decision-making.	Complex to implement and maintain, as it requires handling continuous data streams.
Essential for environments where timely information is critical, such as stock trading, e-commerce or customer service.	Often needs specialised tools and infrastructure (e.g., message brokers, streaming data platforms such as Apache Kafka or AWS Kinesis).

3.4.2.2 Batch processing

Batch processing involves collecting and processing data in large batches at regular intervals (e.g., daily, weekly or monthly). The data is typically extracted from various sources, transformed and loaded into a data warehouse or reporting system, often during off-peak hours to minimise system load. This approach is common for historical analysis or reporting purposes, where immediate updates are not required.

Table 3.6 lists the advantages and challenges of the batch processing approach.

Table 3.6: Advantages and challenges of batch processing

Advantages	Challenges
Simpler to implement than real-time BI, as data can be processed in bulk at scheduled intervals.	Does not provide the most up-to-date information, which can be a drawback for real-time decision-making.
Less resource-intensive during the processing phase since updates are less frequent.	Delays in data processing can lead to stale insights, especially in fast-paced industries.

3.4.3 Cloud BI architecture

Cloud BI architecture refers to the use of cloud-based technologies to manage, store and analyse data for business intelligence. Traditional BI often relied on on-premises infrastructure (refers to IT hardware and software systems that are physically located within an organisation’s facilities or data centres, rather than being hosted remotely or in the cloud), but cloud BI solutions have grown in popularity due to their flexibility, scalability and cost-effectiveness.

3.4.3.1 Cloud-based BI solutions

Cloud BI platforms allow organisations to store and analyse large amounts of data on remote servers (the “cloud”) rather than maintaining physical data centres and hardware. These platforms offer various tools and services for data integration, reporting and analytics. Popular examples of cloud BI solutions include:

- **Google BigQuery** – a serverless data warehouse for handling large-scale data analytics
- **Amazon Web Services (AWS)** – a suite of cloud computing services that includes analytics platforms such as Amazon Redshift (data warehouse) and Amazon QuickSight (BI and analytics service)
- **Microsoft Azure** – offers a variety of cloud analytics services, including Azure Synapse Analytics for data integration and Power BI for visualisation and reporting.

Table 3.7 lists the advantages and challenges of the cloud BI architecture.

Table 3.7: Advantages and challenges of cloud BI architecture

Advantages	Challenges
Cloud solutions can easily scale up or down based on the needs of the organisation, accommodating fluctuating data volumes.	Storing sensitive data in the cloud raises concerns about data privacy and security, though cloud providers often offer robust security features.
Pay-as-you-go pricing models allow organisations to only pay for the storage and computing resources they actually use.	Cloud-based BI relies on consistent internet access, and connectivity issues can hinder performance or accessibility.
Cloud platforms support a wide range of tools, including data storage, analysis and visualisation, and can be integrated with various other cloud services.	
Since the platform is hosted in the cloud, authorised users can access the data and tools from anywhere with an internet connection.	

3.5 Case studies and real-world applications of BI architecture

Several Australian companies/institutions have successfully utilised BI solutions to drive business success.

3.5.1 Australian examples of BI architecture in action

All these examples are from [Telstra Purple](#), which was launched by Telstra in 2019. Telstra Purple is the largest Australian-owned technology services business, bringing together Telstra Enterprise's business technology services capabilities and a number of its recently acquired companies focused on outcome-based, transformative tech solutions.

3.5.1.1 InfraBuild's procurement transformation

This case study exemplifies how leveraging modern digital solutions can transform procurement processes, leading to significant efficiency gains and improved operational effectiveness.

The problem

InfraBuild, Australia's only fully vertically integrated sustainable steel manufacturing and recycling business, faced challenges in its device procurement process. Employees experienced a cumbersome and time-consuming ordering system, leading to inefficiencies and a lack of visibility into device inventory and procurement status. To address these issues, InfraBuild partnered with Telstra Purple to develop an automated procurement platform.

The solution

The solution, built on Microsoft's Power Platform, offers a user-friendly interface where employees can select and order devices from a centralised catalogue. This integration streamlined the ordering process, reducing order processing time from two weeks to approximately four hours and enabling faster device deployment. The platform also provided real-time visibility into orders, invoices and device inventory, enhancing control over approved devices and preventing unauthorised purchases.

The results

The automated procurement system not only improved operational efficiency but also enhanced the onboarding experience for new employees by ensuring timely access to necessary devices and support services. Encouraged by these benefits, InfraBuild plans to expand the platform's use to include mobile

services, telecommunications, software procurement and printers, aiming to create a comprehensive procurement solution that maintains consistent control and visibility across various services.

3.5.1.2 Victoria University's digital transformation

This case study illustrates how adopting advanced digital infrastructure and smart campus solutions can enhance connectivity, security and operational efficiency, creating a seamless and secure environment for both students and staff at Victoria University.

The problem

Victoria University (VU), established in 1916, is a public institution offering both higher education and Technical and Further Education (TAFE) to over 40,000 students across multiple campuses in Melbourne, Sydney and Brisbane. As part of a multimillion-dollar project to consolidate five locations in Melbourne's CBD, VU developed the state-of-the-art VU City Tower at its City Campus.

To enhance its digital infrastructure, VU partnered with Telstra Purple and Cisco to implement a connected, digitised and smart campus. The objectives included automating IT troubleshooting, delivering next-generation network services, improving network resilience and security, maximising student engagement, and supporting a range of digital services for students, staff and industry partners.

The solution

The solution involved deploying Cisco's Software-Defined Access (SDA) network architecture, which utilises Software-Defined Networking (SDN) with physical network devices. This approach enabled policy-based automation from edge to cloud, including automated user and device policies, end-to-end segmentation, and zero-trust security across all users and devices. The implementation simplified IT operations by automating processes like device provisioning, issue identification and policy updates, resulting in better security, transparency and visibility across the campus.

The results

The successful deployment of Cisco SDN led by Telstra Purple improved and simplified VU's IT operations, offering better device security and a consistent user experience across the City Campus and from remote locations. It also reduced network management complexity and enhanced cyber threat management and visibility. The project was completed within budget and has been recognised as a significant achievement in creating a digitally enabled campus environment.

3.5.1.3 Jemena's IoT-driven gas metering transformation

This case study demonstrates how leveraging IoT-enabled smart monitoring and data analytics can transform operational efficiency and customer service, providing real-time insights and enhancing the management of gas metering systems at Jemena.

The problem

Jemena, a leading Australian utilities provider, faced challenges with its outdated gas metering systems, which impacted operational efficiency and customer service. The existing 3G and 2G meters did not provide real-time data collection or remote monitoring capabilities, limiting the company's ability to accurately track and manage gas usage across its network.

The solution

To address these challenges, Jemena partnered with Telstra Purple and Nucleus3 to develop an Internet of Things (IoT)-enabled platform. This solution involved replacing the old gas meters with advanced IoT-enabled devices, which transmit data through Telstra's Narrowband IoT network. The project began with a successful proof-of-concept deployment of 50 Eden Worth modems in high-rise locations, including areas with poor network coverage. The solution uses the Cumulocity connected device platform for centralised management, offering dashboards and analytics tools for real-time visibility of gas usage and network health.

The results

The IoT solution enabled Jemena to achieve better billing accuracy and operational efficiency. With real-time data, the company could remotely monitor gas usage, improving customer service and providing insights into consumption patterns. The project has expanded from an initial 50 units to 600, with plans to deploy around 5,000 units, covering 200,000 customers. In the future, Jemena plans to utilise this IoT infrastructure for additional applications, such as pressure monitoring and gas leakage detection, further enhancing its digital capabilities and service offerings.

3.5.1.4 University of Tasmania's campus navigation

This case study illustrates how implementing innovative digital solutions like UniMaps and

UniNav can transform the campus navigation experience, enhancing accessibility, improving user engagement and streamlining operations at the University of Tasmania.

The problem

The University of Tasmania (UTAS), with its multiple campuses across the state, faced significant challenges in helping students, staff and visitors navigate its expansive and often complex campus infrastructure. Despite offering induction sessions and support from frontline staff, these measures were not enough to improve the overall campus navigation experience. Students and visitors had difficulty finding buildings, classrooms and public transport information, resulting in frustration and inefficiencies, particularly for those new to the campus.

The solution

To address these challenges, UTAS partnered with Telstra Purple to develop an innovative digital campus navigation tool. The project began with a discovery phase, involving workshops with students, staff and other stakeholders to identify pain points and develop effective solutions. The primary objective was to provide an accessible, user-friendly digital platform that would allow students to navigate the campus easily, improve accessibility, enhance spatial data and integrate information such as timetables and transport schedules.

The solution involved creating two interconnected tools:

- **UniMaps:** A digital wayfinding platform that allows users to explore the campus from their mobile, tablet or desktop devices. This platform was integrated into the UTAS website and the UniApp mobile app.
- **UniNav:** A solution embedded into physical **QR code signage** around the campus, enabling students and visitors to scan codes and access navigation tools, making the entire campus experience more interactive and accessible.

The results

The implementation of **UniMaps** and **UniNav** greatly enhanced the campus navigation experience for students, staff and visitors. With **UniMaps**, users can now easily explore the campus and access key information on their devices, whether on mobile, tablet or desktop. **UniNav** connected the physical campus to the digital experience through strategically placed QR codes, further simplifying navigation.

The tools were seamlessly integrated into UTAS's public website, the UniApp mobile app, and physical signage, ensuring easy access to campus information and real-time updates. This transformation not only improved the overall user experience but also boosted operational efficiency, as it streamlined campus

services and made vital information more accessible. The project has contributed to a more connected, modern and accessible campus, enhancing student satisfaction and engagement.

3.5.1.5 Queensland Police Service's data-driven decisions

This case study illustrates how leveraging data-driven solutions, such as the Mobile Business Intelligence dashboard, can enhance resource allocation, improve operational efficiency and enable informed decision-making at the Queensland Police Service.

The problem

The Queensland Police Service (QPS) faced challenges in effectively managing and monitoring the usage of over 8,000 iPads deployed statewide. Limited visibility into device utilisation hindered their ability to maximise return on investment and ensure officers had the necessary tools for efficient operations. Disparate data sources provided fragmented insights, making it difficult to obtain a comprehensive view of device performance and application uptake.

The solution

To address these issues, QPS partnered with Telstra Purple to develop a Mobile Business Intelligence (MBI) dashboard powered by Microsoft Power BI. This data visualisation tool consolidated multiple data sources, offering actionable insights into device usage, application performance and user engagement. The dashboard was designed to be intuitive and user-friendly, enabling Officers-In-Charge (OIC) and Mobile Capability Centre (MCC) personnel to monitor and manage device performance effectively.

The results

The implementation of the MBI dashboard provided the QPS with enhanced visibility into device and application usage, offering several key benefits. By identifying underutilised devices, QPS was able to reallocate resources to areas with higher demand, optimising resource distribution. Additionally, monitoring application usage and device performance helped reduce costs associated with managed services and device maintenance. The dashboard also ensured that officers had access to the most effective applications and tools, thereby improving their operational efficiency. Furthermore, the dashboard empowered leadership to make data-driven decisions by providing real-time insights, ultimately enhancing the overall effectiveness of QPS operations.

3.6 Test your knowledge

Drag the words into the correct boxes.

1) Drag each component or technology into the correct layer of the BI architecture.

Data Source Layer: , .

Data Integration and Storage Layer: , , .

Analytics and Presentation Layer: , , .

2) A company has decided to implement a BI system. The first step is to create a central

to consolidate data from all departments. Afterwards, are created for specific departments like sales and marketing, feeding into the data warehouse for company-wide analysis. This describes the approach because it starts with a centralised structure.

3) Scenario: A hospital monitoring and recording patient vital signs in

.

4) Scenario: A company updating its financial reports at the end of each business day using

.

OLAP Tools

Data Warehouse

Power BI

top-down

Data Lake

data warehouse

batch processing

ETL Tools

real-time

Operational Databases (OLTP)

Tableau

SQL Database

data marts



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=269#h5p-3>

An alternative source for this activity: <https://uts.h5p.com/content/1292858505401511929>

Additional readings for Australian case studies

- Agile Analytics. (n.d.). *Tourism Australia: Azure Data Platform, Synapse, Power BI case study*. Agile Analytics. Retrieved May 28, 2024, from <https://www.agile-analytics.com.au/stories/tourism-australia-azure-power-bi-dax-solution-case-study/>
- BizData. (n.d.). *Customer case studies*. BizData. Retrieved June 2, 2024, from <https://www.bizdata.com.au/case-studies>
- CSIRO. (n.d.). *Case studies – Analytics and Decision Sciences*. CSIRO. Retrieved June 2, 2024, from <https://research.csiro.au/ads/case-studies/>
- Dear Watson Consulting. (n.d.). *Power BI case studies*. Dear Watson Consulting. Retrieved June 2, 2024, from <https://www.dearwatson.net.au/power-bi-case-studies/>
- dVT Group. (n.d.). *Case studies*. dVT Group. Retrieved June 2, 2024, from <https://dvtgroup.com.au/case-studies/>
- Expose Data. (n.d.). *Case studies*. Expose Data. Retrieved June 2, 2024, from <https://exposedata.com.au/case-studies/>
- NetSuite. (n.d.). *18 real-world examples of business intelligence*. NetSuite. Retrieved June 2, 2024, from <https://www.netsuite.com.au/portal/au/resource/articles/business-strategy/business->

[intelligence-examples.shtml](#)

- SAP Concur Australia. (n.d.). *Case studies*. SAP Concur Australia. Retrieved June 2, 2024, from <https://www.concur.com.au/casestudy>
- Telstra Purple. (n.d.). *Case studies*. Telstra Purple. Retrieved June 2, 2024, from <https://purple.telstra.com/insights/case-studies>
- WingArc Australia. (n.d.). *Customer case studies*. WingArc Australia. Retrieved June 2, 2024, from <https://wingarc.com.au/customers/case-studies/>

About the author



Dr Amara Atif

UNIVERSITY OF TECHNOLOGY SYDNEY

<https://profiles.uts.edu.au/Amara.Atif>

Dr Amara Atif is a Senior Lecturer in the School of Computer Science, Faculty of Engineering and IT (FEIT), at the University of Technology Sydney (UTS), Australia. Amara is an expert in educational technology, with a comprehensive skill set that includes research and development in the design of technology-enhanced learning environments, human-computer interaction, artificial intelligence in education, and learning analytics/educational data mining.

At UTS, she developed the undergraduate subject “Business Intelligence”, integrating culturally informed Indigenous knowledge systems through the FEIT Indigenous Graduate Attribute. She has received multiple UTS Vice Chancellor’s Learning and Teaching Awards, including recognition for data-informed teaching practices and inclusive assessment. Passionate about the accessibility and sustainability of educational resources, Amara is an advocate for Open Educational Resources (OER). *Decoding Data* continues this mission, bringing rigorous Business Intelligence concepts to learners through an open, practitioner-oriented resource grounded in Australian business contexts.

4. STATISTICAL MODELLING FOR BUSINESS ANALYTICS

Yang Yang

Learning Objectives (LO)

By the end of this chapter, you will be able to:

LO1 Know the distinction between cases and variables in datasets.

LO2 Distinguish different types of data including nominal, ordinal, continuous and discrete variables.

LO3 Recognise that different graphs and statistics are effective for visualising and describing the distributions of different types of data.

LO4 Understand that common discrete and continuous distributions are useful for modelling various real-world phenomena.

LO5 Measure the strength of the linear relationship between two continuous variables by the correlation coefficients.

4 Statistics in business

Knowledge of statistics is essential for gaining insights into a company's business operations, customer behaviour, market dynamics and the industry in which it competes. Business statistics is a systematic process of collecting data, conducting analysis, interpreting results and presenting visualisations. Utilising various statistical methods and techniques, management teams can make informed decisions to help companies remain competitive and adaptable in an increasingly data-driven business landscape. To better understand these fundamental concepts, let's explore an example that illustrates their real-world application.

This chapter focuses on foundational statistical methods relevant to business analytics. More advanced topics, such as non-parametric tests and time series analysis, are excluded to maintain accessibility but are suggested for further study. While advanced inference procedures are valuable in certain applications, they are beyond the scope of this introductory text. In addition, Bayesian approaches are increasingly popular

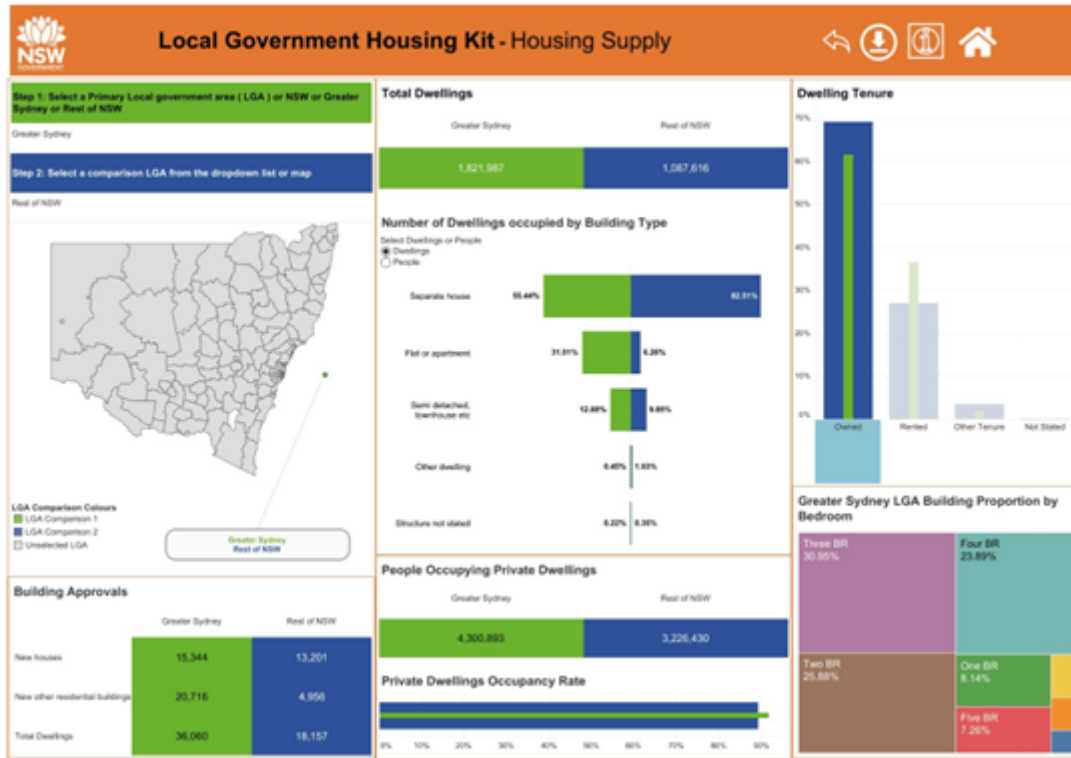
in data-driven decision-making, especially in modern analytics. This chapter introduces foundational frequentist methods. Readers interested in these topics are referred to more advanced texts and the additional resources listed at the end of this chapter.

Statistical tools play a crucial role across different industries and business applications. This chapter utilises real estate scenarios to demonstrate concepts such as random variables, probability distributions, and regression analysis. These real-world examples aim to showcase the practical impact and effectiveness of statistics. Let us begin with the following introduction example.

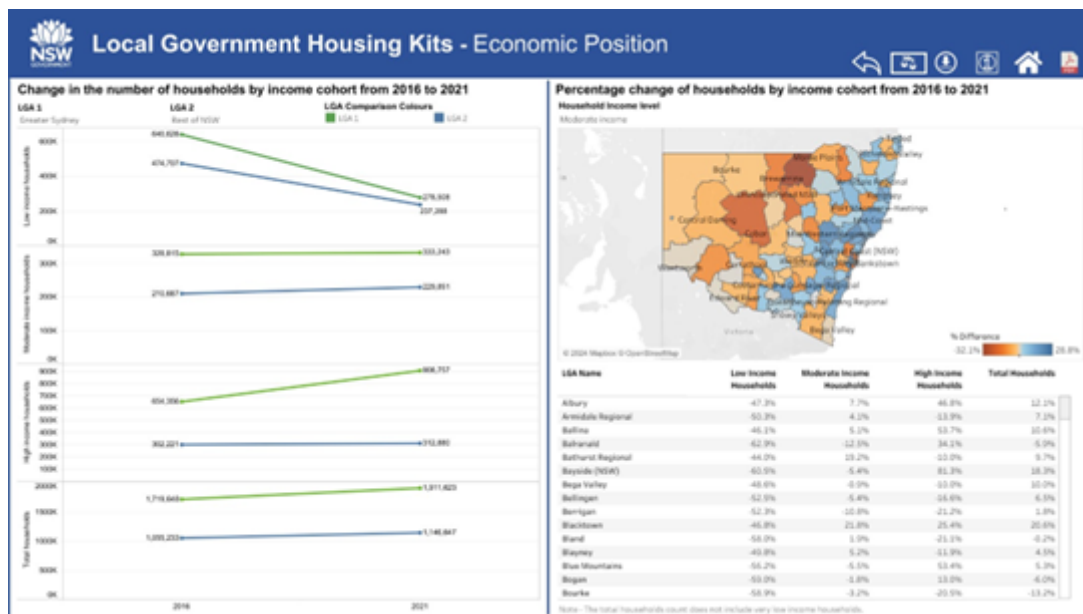
Leveraging statistics for data-driven decisions in Sydney's rental market

Consider the scenario of a **real estate agent navigating Sydney's rental market**. With various factors influencing property demand such as **dwelling types, location, and economic conditions** making informed business decisions requires analysing **complex data patterns**. This is where **statistics and statistical analysis tools** play a crucial role.

By examining **Figure 4.1**, we can see how **housing supply, market trends, and economic factors** shape rental pricing and customer demand. This [Local Government Housing Kit](#) published by DCJ Statistics (2024), provides valuable insights into **housing supply, market distributions, and the economic status of customers in the Greater Sydney local government area (LGA) and the rest of NSW**. However, effectively interpreting this data **requires a solid grasp of statistical principles**. Let's examine this example to understand how statistical concepts help make sense of complex data.



(a) Housing supply in Greater Sydney LGA and the rest of NSW



(b) Economic positions in Greater Sydney LGA and the rest of NSW

Figure 4.1: Housing supply and the economic positions in Greater Sydney LGA and the rest of NSW. The data and visualisations are downloaded from the Local Government Housing Kit provided by the NSW Government.

Alternative Text for Figure 4.1 (a) (b)

This image contains two sets of data visualisations comparing housing supply and economic positions in Greater Sydney Local Government Areas (LGAs) and the rest of New South Wales (NSW). The visualisations are sourced from the **Local Government Housing Kit** provided by the **NSW Government**.

- **Left (Figure a: Housing Supply)**: Displays total dwellings, dwelling types, building approvals, and occupancy rates for Greater Sydney and the rest of NSW. Includes a **map of NSW** with selectable LGAs, **bar charts** showing dwelling tenure types (owned, rented, and others) and illustrating the distribution of dwellings by the number of bedrooms.
- **Right (Figure b: Economic Position)**: Presents economic trends from **2016 to 2021**, highlighting changes in household income groups. Includes a **line graph** showing income cohort trends, a **heat map** depicting the percentage change in household incomes by LGA, and a **data table** summarising these changes.

This chapter will introduce theories and applications of statistics, focusing on their applications in analysing real-world business datasets.

4.1 Data and statistics

Data is the collection of observations or measurements that can provide quantitative and qualitative information to assist decision-making. It is the fundamental input in statistics to produce meaningful statistical studies. Data is the foundation for providing evidence to assist management in making informed decisions.

4.1.1 Types of data

There are two broad types of data in the business environment: the *numerical* data providing quantitative information and the *categorical* data presenting qualitative responses.

Numerical data is most often either continuous in nature or can only take certain discrete values. The *discrete* data is used for counting occurrences of separate events and hence contains distinct integer values. The *continuous* data can take any value within a range, yielding an infinite number of possible measurements. Continuous data is measured, not counted, and can represent any value, including decimals and fractions, depending on the precision of the measurement tool.

Categorical data is used to classify individuals, items or events based on shared characteristics. Depending

on whether a natural order or ranking exists in the measurement levels, categorical data can be further broken down into nominal and ordinal. The ***nominal*** data are labels or categories without any inherent order. The frequencies of observations that fall into each labelled category are typically analysed to provide information about the spread of the data. The ***ordinal*** data often consists of rankings or ratings of objects based on attributes. The frequencies of data in categories, as well as the ordering of individual observations, are all important for ordinal data.

Figure 4.2 illustrates the four types of data covered in this chapter.

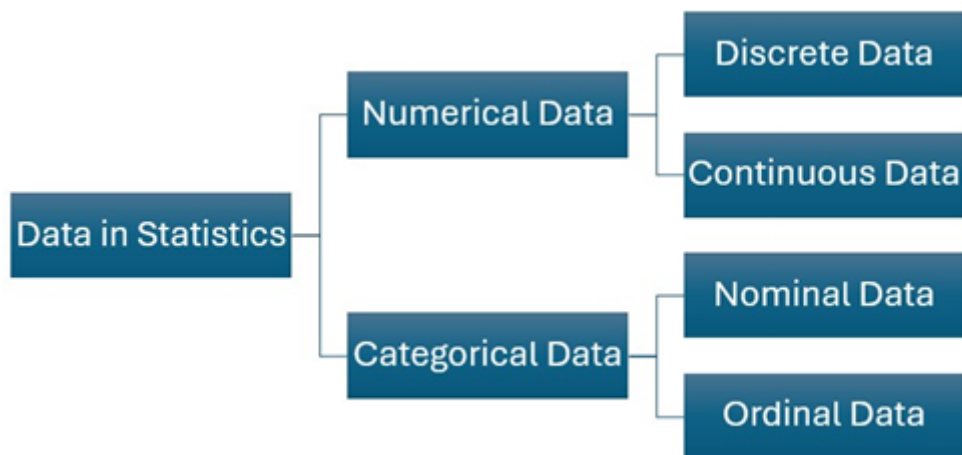


Figure 4.2: The types of data in statistics

Alternative Text for Figure 4.2

A hierarchical diagram illustrating the types of data in statistics. The diagram starts with “Data in Statistics” and branches into two main categories: **Numerical Data** and **Categorical Data**.

- **Numerical Data** further divides into **Discrete Data** and **Continuous Data**.
- **Categorical Data** branches into **Nominal Data** and **Ordinal Data**.

Understanding data types from Figure 4.1 – Sydney’s Rental Market

Using the dataset described in **Figure 4.1**, we illustrate the four data types as follows:

- **Discrete data**

Number of bedrooms in rental houses is a discrete variable whose values are integer counts starting from one. Fractions do not make sense in this case. Residential houses typically do not have designs with “zero bedrooms”.

- **Continuous data**

Living space is a continuous variable that is usually reported in square metres (m^2). Living space can be measured to any level of precision depending on the measurement tools used, e.g., 68.1 m^2 and 68.105 m^2 .

- **Nominal data**

Building type is a nominal variable consisting of several categories, e.g., apartment, detached house, townhouse, unit, etc.

- **Ordinal data**

Property condition is an ordinal variable describing a ranked order of property quality, with possible categories like “Poor”, “Fair”, “Good” and “Excellent”. Note that the variable does not quantify how much better one level is compared to another.

4.1.2 Descriptive statistics

Real-world data analysis often begins by computing statistics for the collected data. Before delving into the computation of statistics, it is worthwhile to distinguish between **cases** and **variables**.

A **case** is a subject or individual from whom data has been collected. For example, a 3-bedroom apartment in Sydney’s CBD can be considered a case in a dataset collected by real estate agents.

In contrast,

A **variable** is any characteristic measured on the case. Following the discussion of real estate data, a variable can be defined as the monthly rent for all 3-bedroom apartments in Sydney.

It is easily noticed that real-world data often has significant variations: not all 3-bedroom apartments in Sydney are rented out at the same price! In statistics, data is considered a mixture of “signal and noise” or “patterns and variations”.

Descriptive statistics is a key branch of statistics that focuses on summarising and describing the main features of a dataset. By studying descriptive statistics, one can gain insights into the characteristics of the current dataset. Common descriptive statistics include **measures of central tendency (such as mode, mean and median)** and **measures of dispersion (such as range, variance and standard deviation)**.

Consider a variable X consisting of n observations. The common measures of **central tendency** include:

- The sample **mean**, denoted by $\bar{x}$, is the average of the n observations:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$$

where the lower case x_i denotes the i^{th} observation of the random variable X .

Throughout this chapter, the uppercase notations will be reserved for random variables, whereas their lowercase counterparts are used for actual observations.

Mean values can be computed for continuous or discrete numerical variables. However, this measure of central tendency can be easily affected by outliers, which are observations that are considerably larger or smaller than the rest of the data points in the sample.

- The **median** is the middle observation (the average of two middle observations if n is an even number) when the data are arranged in increasing order. Medians are used with ordinal, continuous or discrete data. An advantage of the median measure is being immune to outliers or data skewness.
- The **mode** is the most common observation of a variable. Mode is often used with nominal data and discrete numerical data. However, the mode should not be used as the only measure of central tendency for a given dataset.

Commonly used **measures of dispersion** include:

- The **range** is the difference between the highest and lowest observations:

$$\text{Range} = \text{Maximum} - \text{Minimum}$$

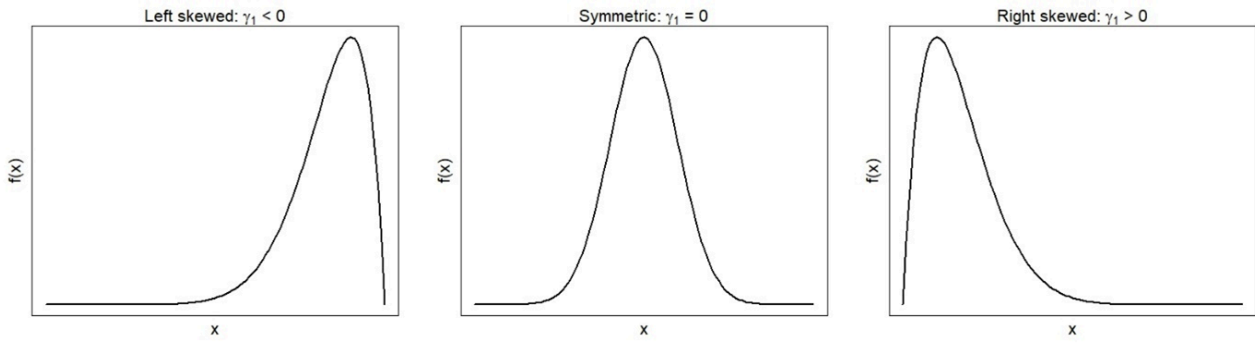


Figure 4.3: The coefficient of skewness is dimensionless, which means that X and cX have the same value of γ_1 , where c is a constant

- The sample **variance**, denoted by s^2 , measures squared differences from the mean:

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

- The sample **standard deviation**, denoted by s , is the square root of the variance:

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}}$$

Both variance and standard deviation measure how spread out the observations in a given dataset are.

- The **quartiles** are the values that divide a collection of sorted numerical data into quarters. The difference between the 3rd quartile (Q3) and the 1st quartile (Q1) is denoted as the *interquartile range (IQR)*:

$$\text{IQR} = Q3 - Q1$$

- The **coefficient of skewness (γ_1)** measures asymmetry in the distribution, with acceptable values falling between -3 and 3 . Distributions with positive coefficients of skewness (a.k.a. positively skewed distributions) have longer tails on the right side. Mean and median values will be greater than the mode. In contrast, negative coefficients of skewness correspond to negatively skewed distributions with longer left tails. The mean and median of negatively skewed distributions are less than the mode.
- The **coefficient of kurtosis (γ_2)** is a measure that quantifies a probability distribution's tailedness. Its definition ensures that for the Normal distribution (to be introduced in Section 4.2), the coefficient of kurtosis is zero. The coefficient of kurtosis is often used to identify outliers in the distribution: kurtosis greater than 3 indicates a leptokurtic distribution with potential large outliers.

Which central tendency measure should be used to describe a dataset?

- Using the mean is appropriate when data is not skewed or affected by outliers.
- Use the median when the data contains outliers or is ordinal in nature.
- Use the mode when the data is nominal.
- The **best approach** is to use all three together to capture and represent the central tendency of the data set from three perspectives.

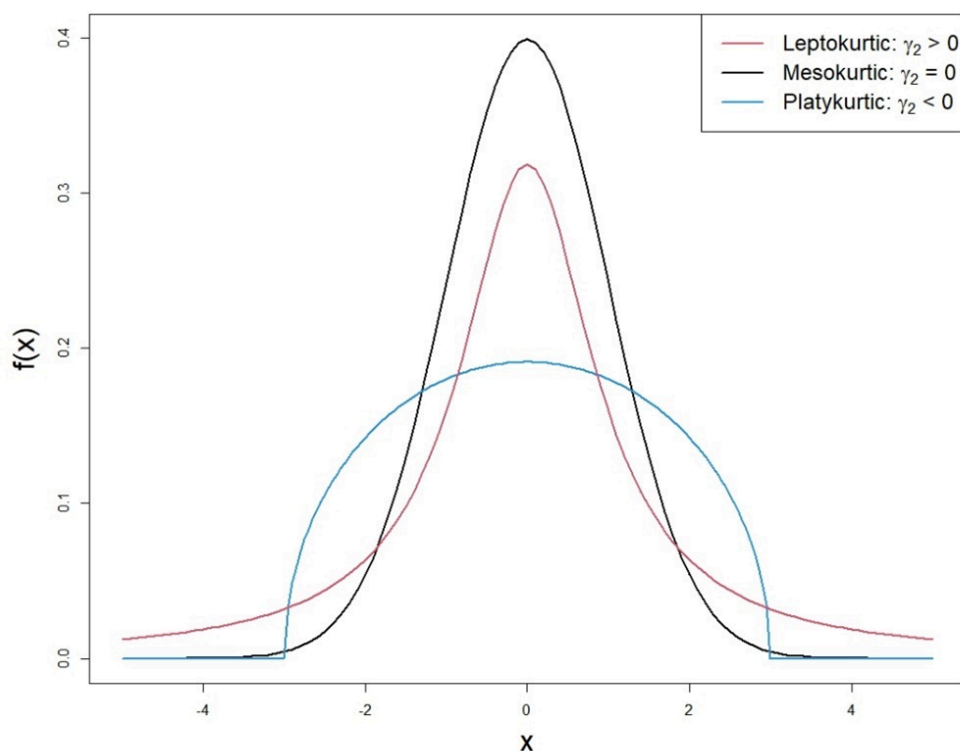


Figure 4.4: Kurtosis greater than 3 (less than 3) indicate the distribution of interest is peaked/tall/thin (flat/short/broad)

Alternative Text for Figure 4.4

A line graph illustrating three types of distributions based on kurtosis, which describes the shape of a probability distribution in terms of how peaked or flat it is. The graph includes three curves:

- **Leptokurtic (Tall and Thin)** – Represented by the black curve, this distribution has a high peak and long tails, indicating a kurtosis value greater than three.
- **Mesokurtic (Normal Shape)** – Shown by the red curve, this distribution has a moderate peak and tails, representing the normal distribution with a kurtosis value of three.
- **Platykurtic (Flat and Broad)** – Depicted by the blue curve, this distribution has a low peak and short tails, indicating a kurtosis value less than three.

The x-axis represents the data values, while the y-axis represents the probability density function. A legend in the top right corner labels the three types of kurtosis. The figure highlights how kurtosis affects the shape of a distribution.

What are the pros and cons of variance (standard deviation) over IQR as a dispersion measure?

- Standard deviation is more commonly used because it uses all of the data. However, it is sensitive to outliers.
- IQR is a more reliable measure of spread for skewed data or data with extreme values because it is resistant to outliers.

Which combination of measures should we report?

Shape	Measure of centre	Measure of spread
Symmetrical	Mean	Standard deviation
Skewed/outliers	Mean and Median	SD and IQR

Graphical representations (such as histograms, bar charts, box plots, scatter plots, etc.) are also considered important tools for describing datasets. Different types of data require specific graphical tools to effectively highlight their features. Recall the introduction example of real estate agents investigating rental properties in Sydney. Given the city's vast size and diverse range of properties, capturing their unique features through text and numbers alone is highly challenging. To address this, various graphical tools can be used to visualise the characteristics of residential properties, as demonstrated in the following example.

Exploring Sydney's property market: Insights from web-scraped data

A dataset about residential property information in Sydney was collected by Alex Lau in 2021 through web scraping from the Australian leading digital property portal [Domain.com.au](https://www.domain.com.au). The dataset covers 11,160 property sale transactions in the Greater Sydney LGA between 2016 and 2021. There are eight variables that describe the characteristics of properties, such as sale price, number of bedrooms, number of bathrooms and property size in square metres. Additionally, there are eight variables related to the suburb where the property is located, including the suburb's population and median income.

The complete dataset can be downloaded from Kaggle (2021). This example focuses on apartments, houses and townhouses with less than 10 bedrooms and 1000 m² property size and uses the appropriate graphical tools to visualise the following variables.

- **Number of bedrooms** as a discrete variable with a small number of observed values can be visualised by a bar chart, as shown in Figure 4.5. It can be seen that the plot has a longer tail on the right side, indicating a right-skewed distribution. Properties with 4 bedrooms are the most common in the dataset, hence mode = 4 for the “number of bedrooms” variable. Median = 5 for the 9 discrete values considered. The mean value is located between the two highest bars (i.e., 3-bedroom and 4-bedroom bars) and turns out to equal 3.69.

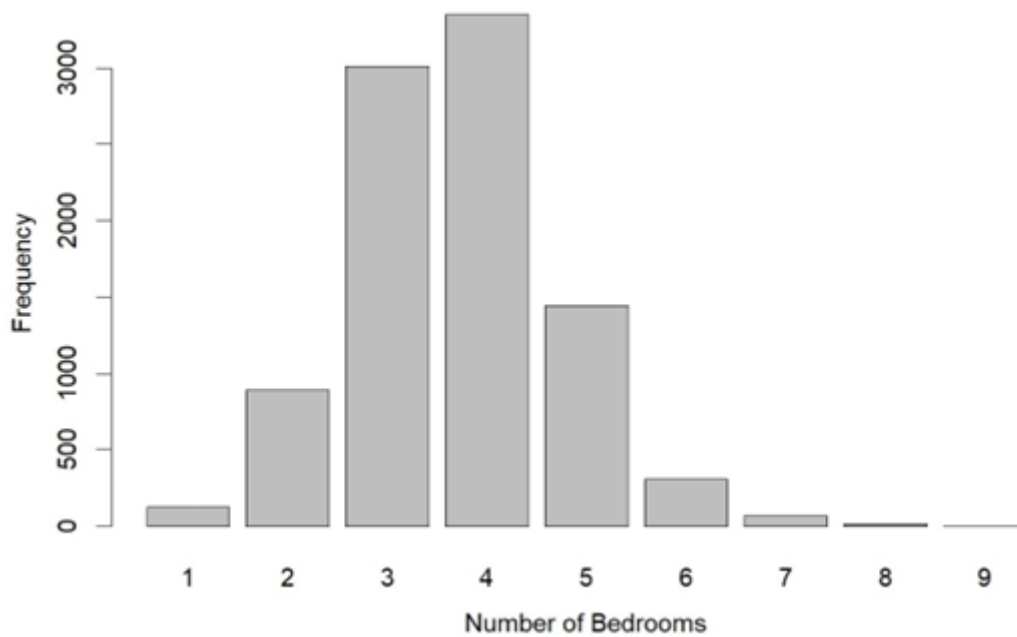


Figure 4.5: Number of bedrooms of the selected residential properties in Sydney

Alternative Text for Figure 4.5

A bar chart showing the distribution of the number of bedrooms in selected residential properties in Sydney. The x-axis represents the **number of bedrooms**, ranging from 1 to 9, while the y-axis represents **frequency**, indicating how many properties have each bedroom count.

- The most common properties have **4 bedrooms**, followed by **3-bedroom properties**.
- Properties with **5 bedrooms** are less frequent but still notable.
- Fewer properties have **1, 2, 6, 7, 8, or 9 bedrooms**, with **8 and 9-bedroom properties being the least common**.

The chart highlights the typical bedroom distribution in Sydney's residential properties, showing that **3- and 4-bedroom homes are the most prevalent**.

- **Property size** is a continuous variable that can be best visualised by a histogram or a box plot, as shown in Figure 4.6. The distribution of property size is roughly symmetrical with a mean size of 551 m² and a median size of 575 m². There are two outliers in the left tail of the distribution as indicated by the “circles” in the box plot.

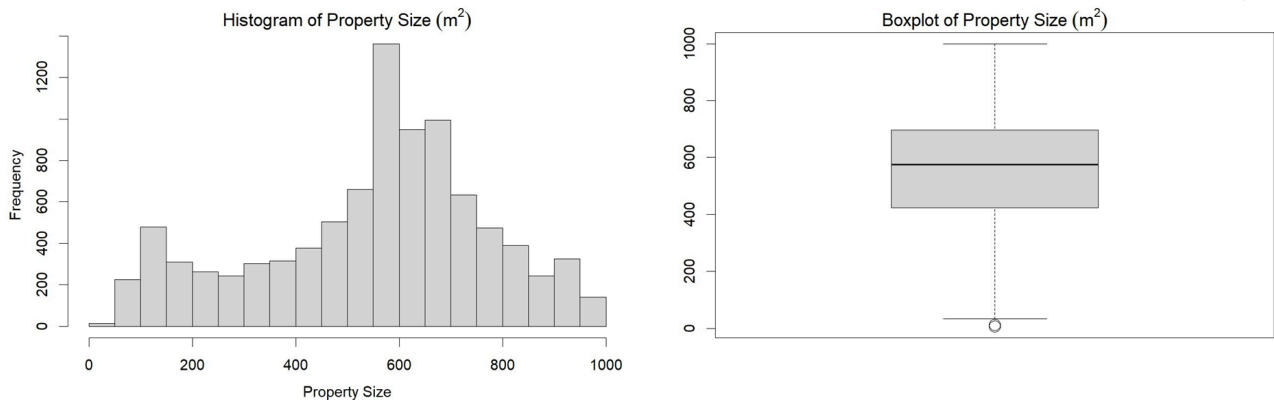


Figure 4.6: Property size of the selected residential properties in Sydney

Alternative Text for Figure 4.6

A figure consisting of two graphs representing the distribution of property sizes for selected residential properties in Sydney.

- **Left: Histogram of Property Size** – A bar chart displaying the frequency distribution of property sizes. The x-axis represents property size in square meters, while the y-axis represents frequency. The histogram shows that most properties have sizes clustered around 600 square meters, with a spread ranging from smaller to larger properties.
- **Right: Boxplot of Property Size** – A boxplot summarising the spread and central tendency of property sizes. The box represents the interquartile range (middle 50% of the data), the horizontal line inside the box indicates the median, and the whiskers extend to the minimum and maximum values within a defined range. A few outliers are visible below the lower whisker, suggesting some properties have significantly smaller sizes than the majority.

These visualisations provide insight into the distribution and variability of residential property sizes in Sydney.

The following diagram 4.7 illustrates how outliers beyond the upper limit (UT) and lower limit (LT) are determined in the box plot. This “1.5 IQR rule” was developed by statistician John W. Tukey in 1969.

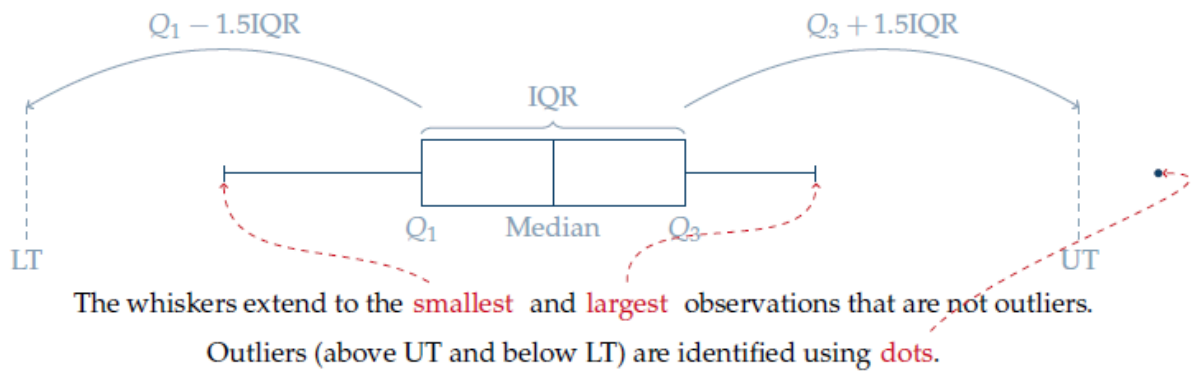


Figure 4.7: Illustration of a Boxplot and Its Components

- **Property type** is a nominal variable that can be visualised by a Pareto chart as shown in Figure 4.8. A Pareto chart combines a bar chart arranged in a decreasing frequency with a line plot tracking the cumulative percentage of the categories from left to right.

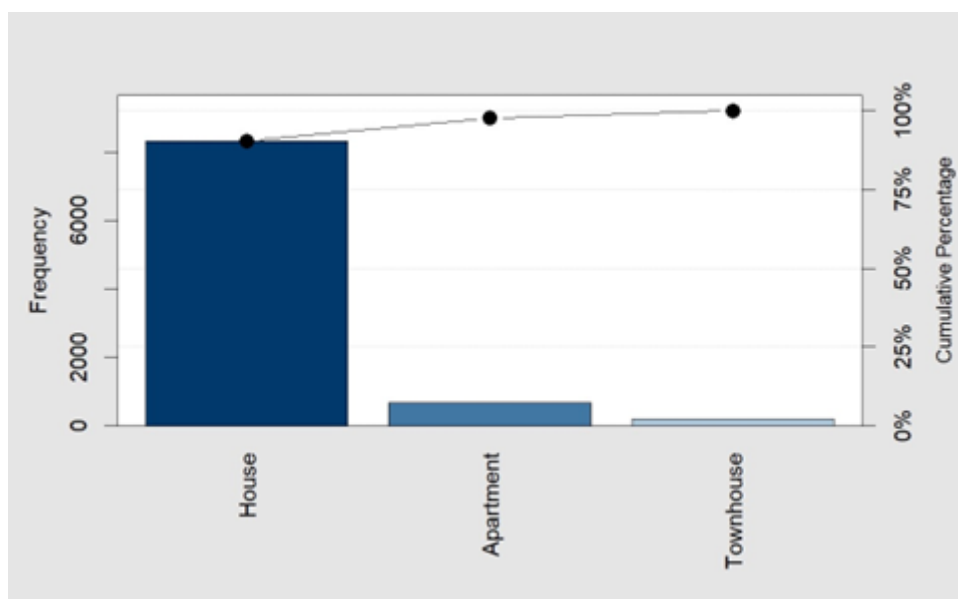


Figure 4.8: Property type of the selected residential properties in Sydney

Alternative Text for Figure 4.8

A **Pareto chart** displaying the distribution of property types for selected residential properties in Sydney. The **x-axis** represents three property types: **House, Apartment, and Townhouse**. The **left y-axis** indicates the **frequency** of each property type, while the **right y-axis** shows the **cumulative percentage**.

- **Houses** have the highest frequency, significantly outnumbering other property types.
- **Apartments** appear in much smaller numbers.
- **Townhouses** have the lowest frequency among the three property types.
- A **cumulative percentage line** with black dots is plotted above the bars, showing that houses make up the majority of residential properties in the dataset.

This visualisation highlights the predominant property type in Sydney, emphasising the dominance of houses over apartments and townhouses.

Real-world data often include missing values. For instance, property transaction records may lack sale prices due to confidentiality agreements between buyers and sellers. Depending on the context and the extent of missing data, the most common approaches are either excluding incomplete cases or imputing missing values using estimates such as the mean or median. Advanced analytics tools such as Python or R provide libraries for automating these processes, such as *scikit-learn* for machine learning or *pandas* for data cleaning. Details of missing value handling techniques are omitted to avoid overwhelming the readers.

4.1.3 Inferential statistics

Inferential statistics provide evidence for making generalisations about the entire population based on sampled data. The “**inference**” refers to extending findings from the collected datasets to the population from which the samples were drawn.

In the real world, studying every individual in a population is often impractical or impossible due to time and budget constraints in sample collection. Inferential statistics help users to draw conclusions and make predictions about populations by studying a small sample of representative data.

More details about the techniques used for making statistical inferences will be introduced in Section 4.3.

Table 4.1 summarises the main differences between descriptive and inferential statistics, highlighting their respective purposes, representations and commonly used techniques.

Table 4.1: Key differences between descriptive and inferential statistics

	Descriptive Statistics	Inferential Statistics
Purpose	Summarises and describes characteristics of a specific dataset	Generalises findings to make inferences about the entire population
Representation	Represents characteristics within the current dataset	Represents characteristics of the entire population
Techniques	Measures of central tendency and spread, graphical tools	Confidence intervals and regression analysis

4.2 Common probability distributions

Real-world data has variations, with sources of variation including the inherent differences in the objects being observed and measurement errors. Because of the uncertainty, data is collected in practice by following certain systematic processes or rigorous experiments. A random variable is a numerical value associated with each outcome of a process or experiment. Referring to the real estate agent example, would descriptive statistics and graphical tools be sufficient to handle all properties and accommodate every existing and potential customer?

Estimating living space: The role of measurement precision in real estate

Referring to the previous real estate dataset, the living space of a rental property can be considered a random variable that both agents and customers find significant. However, even the most experienced agents cannot accurately determine a house's total living space by visual inspection. Laser measurement tools¹ need to be used, and proper procedures should be followed to ensure precise results. It is nearly impossible to find two houses with identical living areas when the values are measured to two decimal points. A statistical tool that allows for generalising findings from a specific dataset to a larger population would be highly valuable in this case. It enables real estate agents to draw meaningful insights about rental properties in Sydney despite the city's vast size and diverse housing market.

The large volume of data in practice means observations cannot be effectively analysed on an individual basis. What's more, inferential statistics aims to generalise findings from the current dataset to the entire

1. Laser measurement tools are electronic devices that use laser beams to accurately measure distances, areas, and volumes. These tools are commonly used in real estate, construction, engineering, and interior design for precise measurements.

population with an infinite number of unobserved objects. Statisticians apply probability theory to analyse random variables and utilise probability distributions to standardise their findings. A probability distribution is a statistical function that describes all the possible values and likelihoods that a random variable can take within a given range. Commonly used probability distributions for discrete and continuous variables are introduced in this section.

4.2.1 Discrete probability distributions

The **Binomial** and **Poisson** distributions are the two commonly used probability distributions for dealing with discrete random variables that count the occurrences of events. Many other discrete probability distributions are extensions of these two distributions.

4.2.1.1 Binomial distribution

Consider a sequence of n independent trials, each resulting in either success or failure with fixed probabilities. Let X denote the count of “successes” in this experiment and p denote the probability of success. Then X follows a Binomial distribution denoted by $X \sim B(n, p)$, for which the probability mass function is:

$$f(x) = \binom{n}{x} p^x (1-p)^{n-x}, \quad x = 0, 1, \dots, n$$

where $\binom{n}{x} = \frac{n!}{x!(n-x)!}$ is a Binomial coefficient. Key features of the Binomial distribution include:

- There is a fixed number n of trials.
- Each of the n trials is independent.
- Each trial can be classified as a “success” or a “failure”.
- The probability of success, p , remains constant across all trials.

The mean and variance of the Binomially distributed X are given by:

$$E(X) = np \quad \text{and} \quad Var(X) = np(1-p)$$

Note that the $E(X)$ and $Var(X)$ are used to represent the population mean and variance of X , respectively. These notations for population parameters will be used throughout this section. The population mean and variance reflect the “true” characteristics of the entire population and do not depend on any specific dataset. In contrast, the sample mean ($\bar{x}$) and sample variance (s^2) introduced in Section 4.1 only describe features of a given dataset. The following examples illustrate the difference.

Coin flip

A fair coin is flipped 10 times. Define X as the number of heads observed in 10 flips. What is the chance of observing $P(X=2)$, i.e., the probability of observing 2 heads?

$X \sim B(10, 0.5)$ because it satisfies the following conditions of the Binomial setting:

- There is a fixed number of trials ($n=10$). The trials are independent.
- The result of one flip does not influence the result of another.
- A coin flip results in two possible outcomes: heads or tails. The heads are classified as successes in this experiment.
- The probability of success is constant for each flip at $p=0.5$ since a fair coin is flipped.

The probability of observing 2 heads in 10 flips is

$$\begin{aligned} P(X = 2) &= \binom{10}{2} (0.5)^2 (0.5)^{(10-2)} \\ &= 45 \times 0.5^2 \times 0.5^8 \\ &= 0.0439. \end{aligned}$$

Figure 4.9 illustrates the probability mass function for this experiment. The mean and variance for the $X \sim B(10, 0.5)$ can be computed as

$$\begin{aligned} E(X) &= 10 \times 0.5 = 5, \\ \text{Var}(X) &= 10 \times 0.5 \times (1 - 0.5) = 2.5 \end{aligned}$$

where the $E(X) = 5$ means that the expected number of heads in a fixed number of 10 flips of a fair coin is 5. This means one would expect to observe an average of 5 heads over many repeated experiments under the same conditions.

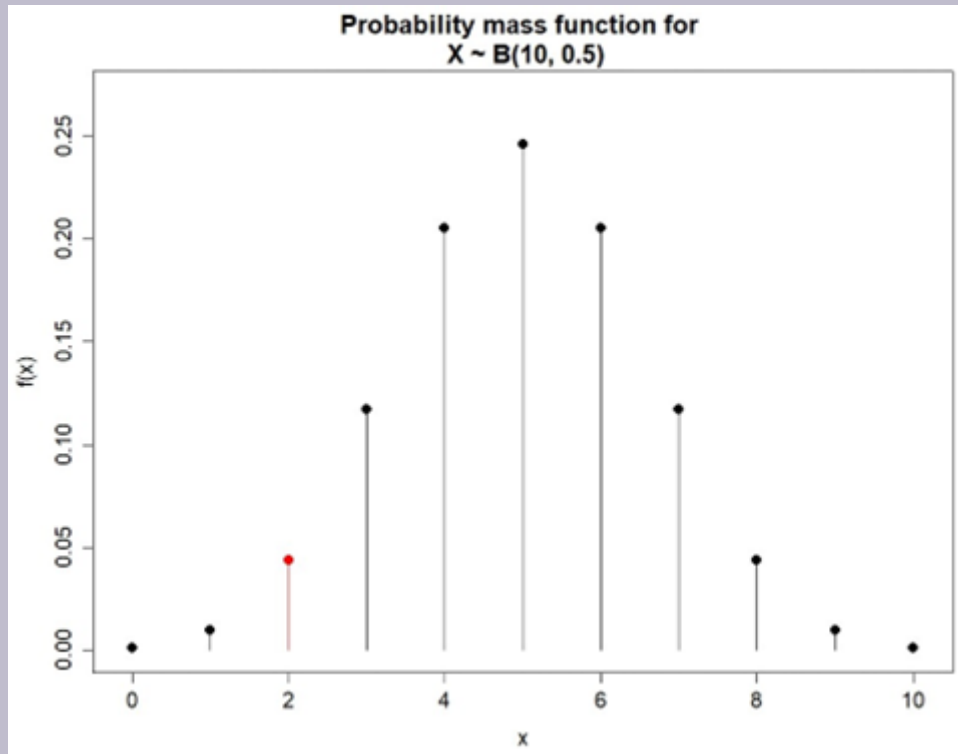


Figure 4.9: Probability mass function for $X \sim B(10, 0.5)$. The red colour indicates the probability of observing 2 heads in 10 flips

4.2.1.2 Poisson distribution

Consider events that occur independently at a fixed average rate $\lambda > 0$. Let X be the count of the number of events over a fixed time interval. The random variable X then follows a Poisson distribution, $X \sim P(\lambda)$, for which the probability mass function is:

$$P(X = x) = \frac{e^{-\lambda} \lambda^x}{x!}, \quad x = 0, 1, 2, \dots$$

Where $e \approx 2.7183$ is the Euler number. The Poisson distribution is named after the French mathematician Simeon-Denis Poisson and is useful for modelling a wide range of real-world phenomena. Some examples include:

- The number of customer service calls at a call centre in the next hour.
- The number of claims received by an insurance company in a month.
- The number of patients arriving in an emergency room over the period of an hour.
- The number of server failures or technical issues reported per week by an IT company.

Key features of the Poisson distribution include:

- The events occur independently.
- The average rate of events occurring is fixed at λ .
- The events are observed over a fixed interval.

The mean and variance of $X \sim P(\lambda)$ are given by:

$$E(X) = \lambda \quad \text{and} \quad \text{Var}(X) = \lambda.$$

Predicting customer visits: A probability approach in real estate

Suppose a real estate agent preparing a house open for inspection believes that, on average, 3 customers arrive every 10 minutes. What is the probability that in the next 30 minutes, 10 customers will arrive?

Let X be the number of customers arriving in a 30-minute period. X follows the Poisson distribution with $\lambda = 3 \times 3 = 9$ customers per 30 minutes.

$$\begin{aligned} P(X = 10) &= \frac{9^{10} \times e^{-9}}{10!} \\ &= 0.1186 \end{aligned}$$

Therefore, there is about an 11.86% chance that 10 customers will arrive within the next 30 minutes. Figure 4.10 illustrates the probability mass function for this example.

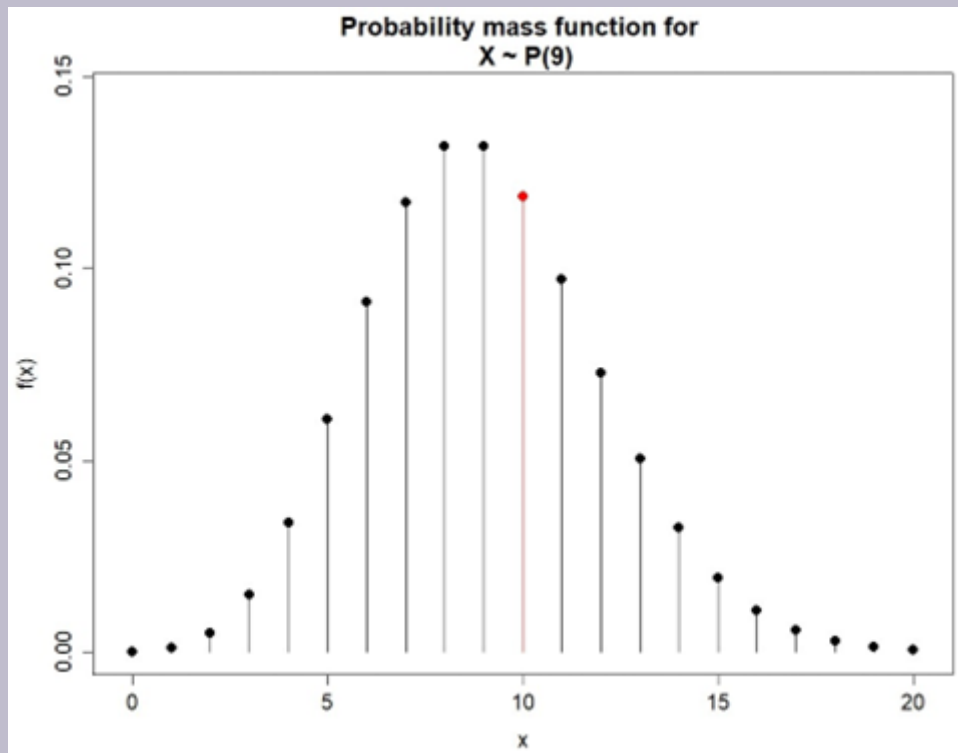


Figure 4.10: Probability mass function for $X \sim P(9)$. The red colour indicates the probability of 10 customers arriving in the next 30 minutes

There are other discrete distributions that are linked to the Binomial and Poisson distributions introduced above. Here are some examples:

- In analogy to the Binomial distribution, a *Geometric* distribution arises when the number of failures is counted before observing the first success.
- A *Negative Binomial* distribution arises when the Geometric distribution is extended to observe the r^{th} success.
- A sum of a number of independent and identically distributed Poisson random variables yields *Compound Poisson* distribution.

Choosing the appropriate probability distribution depends on the context of the data and the problem being solved.

4.2.2 Continuous probability distributions

When data arises because of measurements or observations of continuous processes, continuous distributions are needed to model such variables. Two most popular continuous probability distributions are introduced in this section.

4.2.2.1 Exponential distribution

Consider events that occur over a period of time where the number of occurrences following a Poisson distribution have a fixed rate λ . Let X denote the length of time until the next event occurs. Then the random variable X follows an Exponential distribution with parameter λ , denoted by $X \sim \text{Exp}(\lambda)$. The probability density function (PDF) of X is given by:

$$f(x) = \lambda e^{-\lambda x}, \quad x > 0$$

The mean and variance of $X \sim \text{Exp}(\lambda)$ are given by

$$E(X) = \frac{1}{\lambda} \quad \text{and} \quad \text{Var}(x) = \frac{1}{\lambda^2}$$

Modelling customer arrivals: Poisson distribution in retail operations

Suppose that an average of 3 customers every 10 minutes arrive at a shop in accordance with a Poisson distribution. What is the probability of there being a 5-minute break (i.e., no customers arriving) so the shop owner can have a cup of tea? Given that it has been 5 minutes since the last customer arrived, what is the probability of there being a further 5-minute break?

Let X be the time until the next customer arrives. Let Y denote the number of customers that arrive within a x -minute period. According to the question, Y is a Poisson distributed random variable with a rate parameter $\lambda = \frac{x}{60} \times 18 = 0.3x$. It can be seen that:

$$\begin{aligned} F(x) &= P(X \leq x) = 1 - P(X > x) \\ &= 1 - P(0 \text{ customers arriving within time } x) \\ &= 1 - P(Y = 0) \\ &= 1 - e^{-0.3x} \end{aligned}$$

where $F(x)$ is the cumulative density function of X . Taking the first derivative of $F(x)$ yields the probability density function of x given by:

$$f(x) = F'(x) = 0.3e^{-0.3x}$$

This is an exponential distribution with $\lambda=0.3$. The probability of no customers arriving in the next 5 minutes is:

$$\begin{aligned}
 P(X \geq 5) &= 1 - P(X < 5) \\
 &= 1 - \int_0^5 0.3e^{-0.3x} dx \\
 &= 0.2231
 \end{aligned}$$

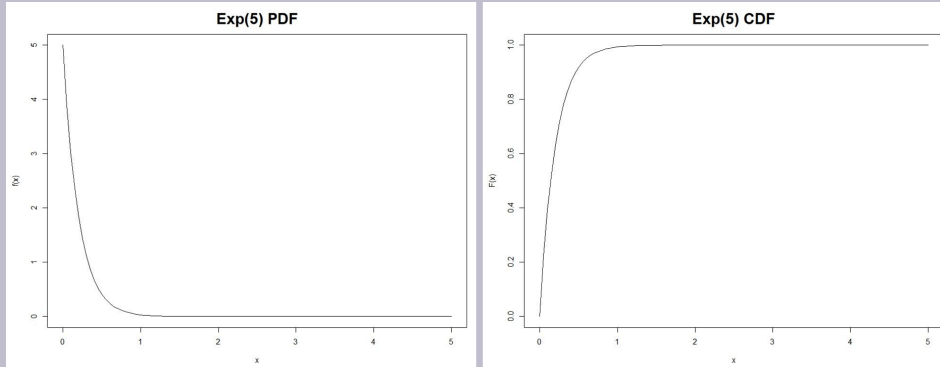


Figure 4.11:
Probability
density function
and cumulative
density function
of $X \sim \text{Exp}(5)$

Thus, there is about a 22% probability of at least a 5-minute break. Given that it has been 5 minutes since the last customer arrived, the probability of there being a further 5-minute break is given by:

$$\begin{aligned}
 P(X \geq 5 + 5 | X \geq 5) &= \frac{P(X \geq 10 \cap X \geq 5)}{P(X \geq 5)} \\
 &= \frac{P(X \geq 10)}{P(X \geq 5)} \\
 &= \frac{0.3e^{-0.3 \times 10}}{0.3e^{-0.3 \times 5}} \\
 &= e^{-0.3 \times 5} \\
 &= P(X \geq 5) \\
 &= 0.2231
 \end{aligned}$$

The fact that the shop owner has already waited 5 minutes for a customer does not change the probability that there will be another 5 minutes without a customer.

The memoryless property demonstrated in this example is a key feature of the Exponential distribution. That is, the probability of an event occurring in the next time interval does not depend on how much time has already passed, i.e.,

$$P(X > t + s | X > x) = P(X > t)$$

The Exponential distribution has a wide range of applications in business:

- Modelling customer service processes, such as the time between customer arrivals or the service time at a counter.
- Estimating the time until failure of machines, equipment or systems.
- Modelling the time between restocking.
- Analysing the time between events in risk scenarios, such as a borrower defaults on a loan.
- Estimating the time until a response to an advertisement or marketing campaign is received.

4.2.2.2 Normal distribution

The most common continuous distribution is the Normal distribution, often referred to as the Gaussian distribution in honour of mathematician Carl Friedrich Gauss. Let X be a random variable that is Normally distributed with mean μ and variance σ^2 , i.e., $X \sim N(\mu, \sigma^2)$. The probability density function of X is given by:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp \left[-\frac{(x - \mu)^2}{2\sigma^2} \right], \quad -\infty < x < \infty$$

The mean μ controls the location of the curve along the X -axis, whereas the variance σ^2 controls the spread of the distribution about the mean, as illustrated in Figure 4.12.

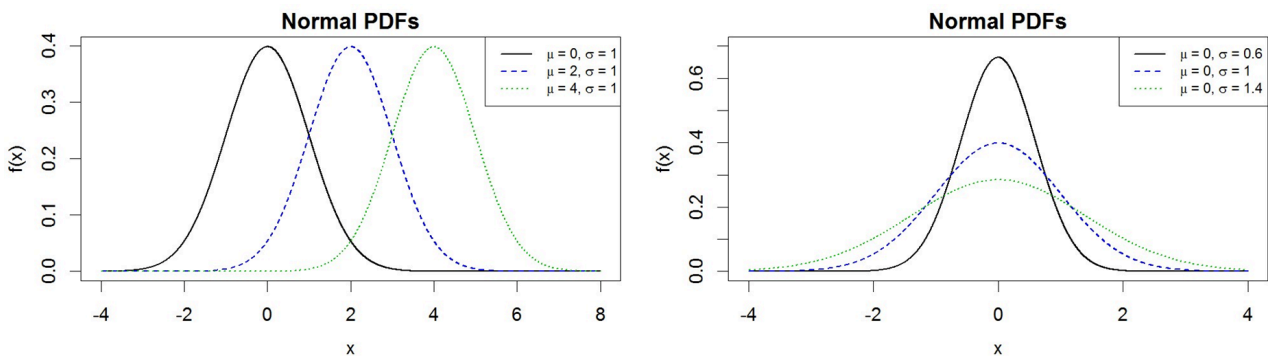


Figure 4.12: Normal distributions with different mean and variance parameters

When $\mu=0$ and $\sigma^2=1$, X becomes the Standard Normal distribution that is often denoted by $Z \sim N(0, 1)$. The Standard Normal distribution is useful for computing probabilities using the standardisation of Normally distributed variable $X \sim N(\mu, \sigma)$ given by:

$$Z = \frac{X - \mu}{\sigma}$$

Assessing chocolate weight variability using normal distribution

A chocolate maker produces a certain type of chocolate with a nominal weight of 21 grams. Assume that the distribution of weights is actually $N(21.5, 0.16)$. Let X denote the weight of a single chocolate selected randomly from the production process. What is the probability that this selected chocolate weighs less than 21 grams, i.e., $P(X < 21)$? It is known that $X \sim N(21.5, 0.16)$ according to the description. Using the standardisation of Normally distributed variables, the required probability of a randomly selected chocolate weighing less than nominal is given by:

$$\begin{aligned} P(X < 21) &= P\left(\frac{x - \mu}{\sigma} < \frac{21 - 21.5}{\sqrt{0.16}}\right) \\ &= P(Z < -1.25) \\ &= \int_{-\infty}^{-1.25} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \\ &= 0.3958 \end{aligned}$$

Therefore, there is about a 39.58% chance that a randomly selected chocolate weighs less than 21 grams, as illustrated in Figure 4.13.

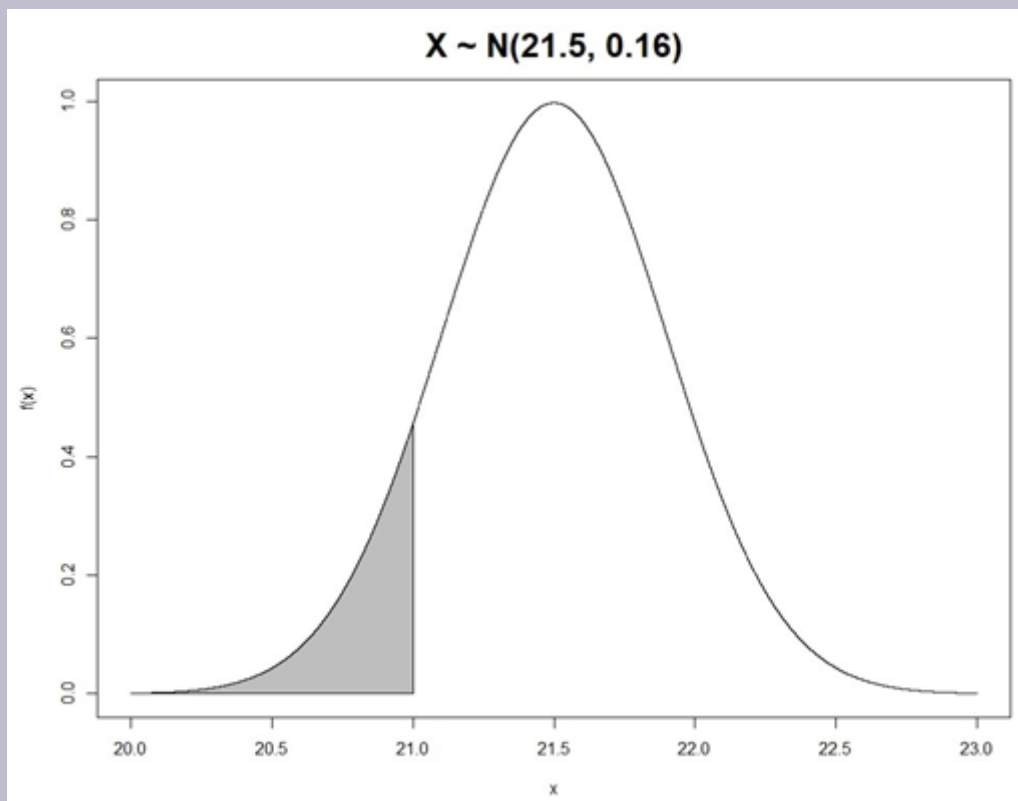


Figure 4.13: Probability of a randomly selected chocolate weighing less than 21 grams

Other commonly used continuous probability distributions include:

- The *Uniform* distribution describes a situation where all outcomes within a given range are equally likely. This distribution is commonly used when there is no reason to favour one outcome over another or when data is distributed evenly.
- The *Gamma* distribution is often used for modelling skewed data that describe the time until a certain number of events occur. It is a generalisation of the exponential distribution, making it versatile in various contexts, such as reliability analysis, queuing systems and financial modelling.

4.3 Statistical inference

Data collected in the real world is always subject to random variation. Repeating a 10-trial coin-flipping experiment multiple times with the same coin will likely yield different numbers of heads. When estimating the parameters for a probability distribution, there is always a question to answer:

Which dataset contains the most accurate information about the population?

Statistical inference, as a process of analysing data with uncertainty and making conclusions about the parameters of the entire population, is the answer to this question.

Statistical inference provides a probable range for the true values of a population parameter after accounting for sample-to-sample variations (e.g., sample size, variation within the sample, differences across samples, etc.) of real-world datasets.

This section will introduce the most popular statistical inference techniques: **confidence interval** and **regression analysis**.

4.3.1 Confidence interval

Confidence intervals provide a range of values within which a population parameter of interest (e.g., mean or variance) is expected to be located. They are crucial for making informed decisions in business and scientific research. Suppose a population parameter θ is estimated based on a random sample $x = \{X_1, X_2, \dots, X_n\}$. It is possible to determine two statistics, $a(x)$ and $b(x)$, based on the sample information such that:

$$P(a(x) < \theta < b(x)) = 100(1-\alpha)\% = C\%$$

where $0 < \alpha < 1$ is the significance level parameter. The interval $[a(x), b(x)]$ specifies a probable range

for the population parameter θ at a $C\%$ confidence level, which is called a $C\%$ confidence interval for the population parameter θ .

The confidence interval reflects the variability in the data since both limits (i.e., $a(x)$ and $b(x)$) are computed based on all observations x . The confidence interval does not mean that the population parameter θ will be contained within the range $[a(x), b(x)]$. Instead, if many random samples of size n were selected, then $100(1-\alpha)\%$ of them should contain θ . Alternatively, one can interpret the confidence interval as being $100(1-\alpha)\%$ confident that the population parameter θ lies somewhere within the range $[a(x), b(x)]$.

The computation of confidence intervals relies on the probability distribution of a statistic (i.e., sampling distribution of statistics) estimating the population parameter of interest. Therefore, it is necessary to choose the appropriate computation equations for different datasets. When the sample size is sufficiently large, the following Central Limit Theorem applies to simplify the computation.

Theorem 1. For a random sample of size n , $X_1, X_2, \dots, X_n$ from X with finite mean μ and variance σ^2 ,

$$Z = \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{d} N(0, 1) \text{ as } n \rightarrow \infty.$$

Equivalently,

$$\bar{X} \xrightarrow{d} N\left(\mu, \frac{\sigma^2}{n}\right) \text{ as } n \rightarrow \infty.$$

The Central Limit Theorem specifies that, as $n \rightarrow \infty$, the sampling distribution of $\bar{X}$ follows a Normal distribution irrespective of the sampling distribution of the variable X . In practice, this theorem works reasonably well if n is greater than 30.

With the help of the Central Limit Theorem, a general $(1-\alpha)\%$ confidence interval for the population mean parameter μ can be computed as:

$$\bar{X} - z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{\text{th}}$ quantile of the Standard Normal distribution. To use this general confidence interval equation, the sample size n and the population variance σ^2 (or standard deviation σ) must be known.

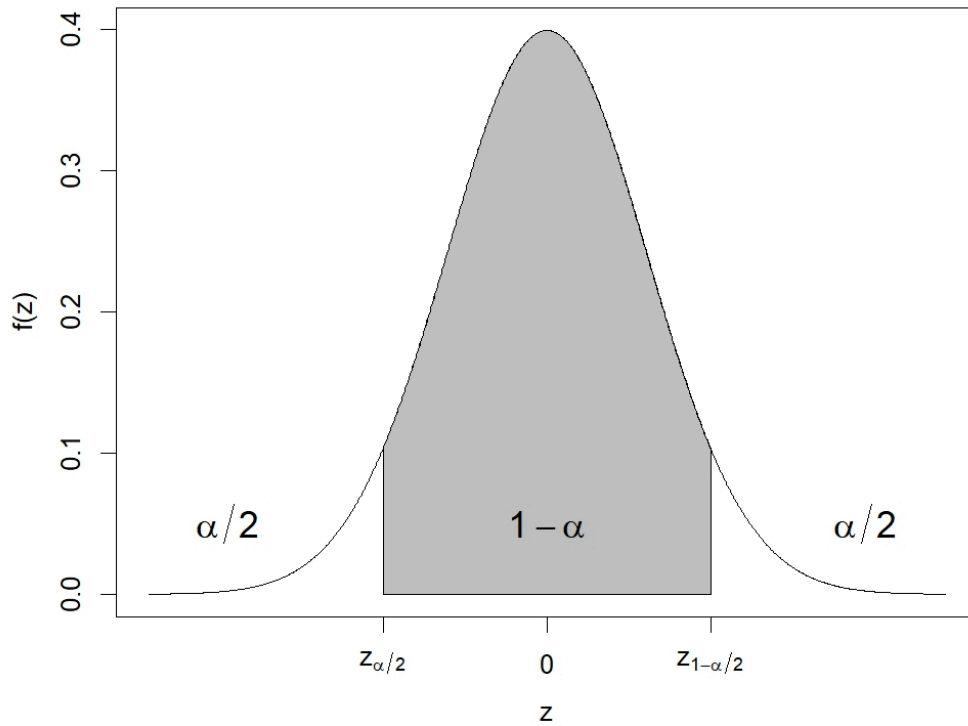


Figure 4.14: The $(1-\alpha)\%$ confidence interval uses the $z_{1-\alpha/2}$ quantile of the Standard Normal distribution.

The following derivations explain the computation of this confidence interval.

$$\begin{aligned}
 P\left(z_{\frac{\alpha}{2}} < Z < z_{1-\frac{\alpha}{2}}\right) &= 1 - \alpha \\
 P\left(z_{\frac{\alpha}{2}} < \frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < z_{1-\frac{\alpha}{2}}\right) &= 1 - \alpha \\
 P\left(z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \bar{X} - \mu < z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\
 P\left(-\bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\
 P\left(\bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} > \mu > \bar{X} - z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha \\
 P\left(\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}\right) &= 1 - \alpha
 \end{aligned}$$

It can be seen that the computed confidence interval corresponds to the central $(1-\alpha)\%$ area in the Standard

Normal distribution, as illustrated in Figure 4.14. Different values of α will lead to different $z_{1-\alpha/2}$ values, and thus, the choice of α will reflect the width of the interval.

Estimating customer satisfaction: A 95% confidence interval analysis

A company surveyed 100 customers about a new product. The average rating was $\bar{x} = 8.1$ (out of 10). It is assumed the population standard deviation $\sigma = 0.6$ for all ratings. Find a 95% confidence interval for the mean customer rating score.

For a 95% confidence interval, $\alpha = (1-0.95) = 0.05$ and $z_{1-\alpha/2} = z_{0.975} = 1.96$ using a statistical table or computation software. The 95% confidence interval for the mean customer rating score can be computed as:

$$\begin{aligned}\bar{x} \pm z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} &= 8.1 \pm 1.96 \times \frac{0.6}{\sqrt{100}} \\ &= 8.1 \pm 0.1176 \\ &= [7.9824, 8.2176]\end{aligned}$$

Therefore, the company is 95% confident that the true mean rating score for this product lies between 7.9824 and 8.2176.

In practice, it is difficult to get a good estimate of the population parameter σ . In this case, a t -distribution is used to compute the confidence interval for the population mean as:

$$\bar{X} - t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}} < \mu < \bar{X} + t_{n-1, 1-\alpha/2} \frac{s}{\sqrt{n}}$$

where s is the sample standard deviation introduced in Section 4.1 and $t_{n-1, 1-\alpha/2}$ is the $1-\alpha/2$ quantile value from the Student's t -distribution with $n-1$ degrees of freedom, as illustrated in Figure 4.15.

The t -distribution was introduced by English statistician William Sealy Gosset in 1908. At the time, Gosset worked for Guinness Breweries in England, which prevented their workers from disclosing confidential information after an employee was found to have leaked trade secrets of the brewery. For this reason, Gosset, who considered himself a student of statistics, published his work under the name “Student”. The t -distribution is a continuous distribution whose probability density function is symmetrical and analogous to that of the Normal distribution.

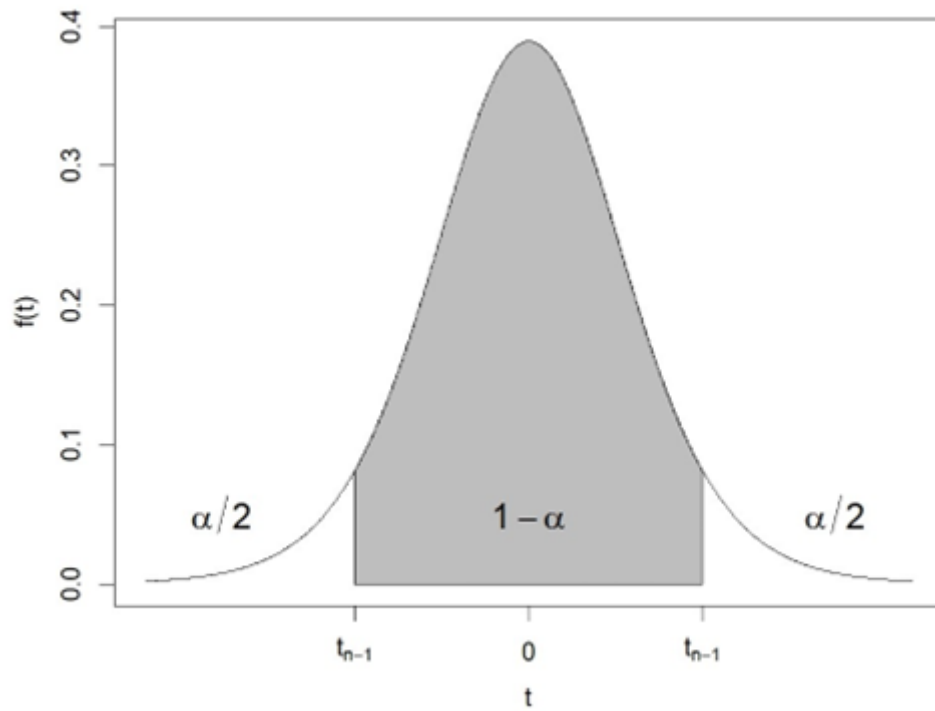


Figure 4.15: The $(1-\alpha)\%$ confidence interval uses the $t_{n-1,1-\alpha/2}$ quantile of the t-distribution.

Estimating weekly sales: A confidence interval approach for inventory planning

A company that manufactures technology gadgets wants to estimate the average number of units sold to improve inventory management and production planning. Sales data was collected over a period of eight weeks. The average number of units sold per week in this sample is $\overline{x} = 160$ units with a standard deviation of $s=20$ units. Find a 95% confidence interval for the average number of units sold per week over the entire lifecycle of this product.

The population standard deviation of sales is not given. The t-distribution should be used to compute the confidence interval as:

$$\begin{aligned}\overline{x} \pm t_{n-1,1-\frac{\alpha}{2}} \frac{s}{\sqrt{n}} &= 160 \pm 2.3646 \times \frac{20}{\sqrt{8}} \\ &= 160 \pm 16.7204 \\ &= [143.2796, 176.7204]\end{aligned}$$

The $t_{n-1,1-\alpha/2} = t_{7,0.975} = 2.3646$ in the above computation is the 97.5 percentile value from a t-distribution with 7 degrees of freedom. This confidence interval means that based on the data

from an 8-week sample, the company is 95% confident that the true average number of units sold per week over the entire lifecycle of this product lies between 143.2796 and 176.7204 units.

4.3.2 Regression analysis

Real-world datasets most likely contain two or more variables that are not completely independent. The correlation coefficient is often used to measure the strength of the linear relationship between two numerical variables X and Y :

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

where s_x and s_y are sample standard deviation for x and y variables, respectively.

The correlation coefficient has no unit and falls in the interval $-1 \leq r \leq 1$. A positive correlation coefficient (i.e., $r > 0$) indicates that X and Y increase together, whereas a negative value represents the situation when X increases, Y tends to decrease. The larger the absolute value of the correlation coefficient, the stronger the linear relationship between X and Y . $r = 1$ (or $r = -1$) is a perfect positive (or negative) correlation, which means the observations lie exactly on a straight line without any scatter. Figure 4.16 shows simulated datasets in which the variables X and Y exhibit linear relationships with varying strengths. The *scatter plots* in the figure display the predictor variable X and response variable Y on the horizontal and vertical axes, respectively.

Most computational software can calculate correlation coefficients with just a single click of a button. However, it is strongly recommended that the correlation coefficient be used together with a scatter plot. This is to avoid incorrect interpretations of a zero correlation coefficient. When $r = 0$, there is no linear relationship between the two variables. However, this does not rule out other kinds of relationships, e.g., a quadratic relationship. The last scatter plot in Figure 4.16 illustrates this situation.

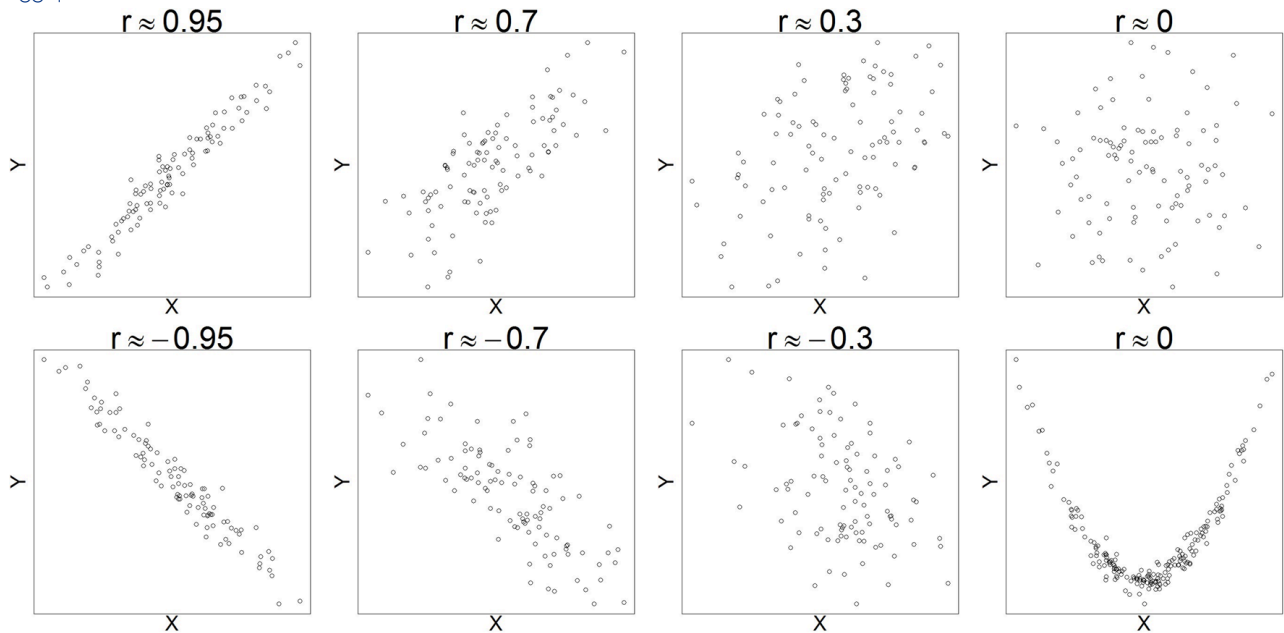


Figure 4.16: Correlation coefficients and scatter plots

After establishing the correlation between two variables X and Y , it is often necessary to investigate how one variable affects another. Linear regression models use mathematical equations to quantify how changes in the explanatory (input) variable X affect the response (output) variable Y , in the following structure:

$$\text{response} = \text{trend} + \text{scatter}$$

The *trend* discovered from the data is the linear relationship between the two variables. Mathematically, a simple linear regression model can be expressed as:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

where β_0 and β_1 are fixed coefficients, and the ϵ_i is a modelling error term representing the random scatters of data observations around the trend line. The slope parameter β_1 indicates how much Y increases with each unit increase in X . The following example illustrates the simple linear regression.

Galton's discovery: Regression to mediocrity and the origins of regression analysis

Sir Francis Galton was the first to formally study the relationship between the heights of parents and children to gain insights into what degree height is an inherited characteristic. He noted that the heights of the children tended to be more moderate than the heights of their parents in the work by Galton (1886).

For example, if the parents were very tall, the children tended to be tall but shorter than

their parents. If parents were very short, the children tended to be short but taller than their parents were. Galton referred to this phenomenon as “regression to mediocrity”, which led to the development of regression analysis. Figure 4.17 shows the data of the heights of fathers and sons studied by Pearson and Lee (1903). The left panel suggests that the sons’ heights tend to increase as the fathers’ heights increase. For example, the blue-coloured points have average heights of 161 cm and 169 cm for fathers and sons, respectively. In contrast, the red-coloured points show the mean heights for fathers and sons are 183 cm and 180 cm, respectively. Using a straight line to connect the intersections of the mean heights for fathers and sons across the entire dataset captures the general trend of the data, as shown in the right panel of Figure 4.17.

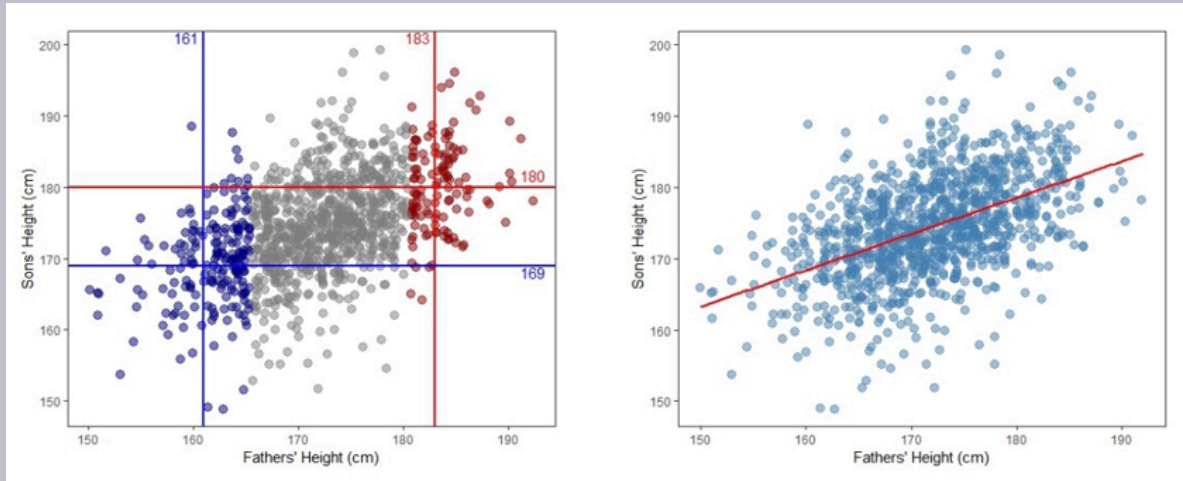


Figure 4.17: Heights of fathers and sons studied by Pearson and Lee (1903)

Regression analysis is useful for explaining the general pattern of real-world datasets (hypothesis testing) and making predictions (forecasting) for new observations from the same population. Recall the the real estate example introduced at the start of this chapter (*Exploring Sydney’s property market: Insights from web-scraped data*). It is possible to apply regression techniques to build a linear model for housing prices based on the dataset introduced in the example. The following demonstration illustrates this approach.

The influence of property size and bedrooms on real estate pricing

It is well-known that property size is an important variable for determining property prices, especially for apartments. In practice, builders of apartment properties often use the number of bedrooms to attract specific customer groups: compact apartments with one or two bedrooms

are designed for young adults and retired people, whereas large properties with more than three bedrooms are designed for families with children.

Figure 4.18 presents apartment properties with less than 500 m² property size selected from the Kaggle (Kaggle 2021) dataset. The sale prices of the selected properties are shown in a natural logarithm scale on the vertical axis. The property sizes of the apartments are indicated on the horizontal axis. Two separate regression lines are fitted to the apartments depending on the design (i.e., compact vs. large apartments). Similar slopes for the two fitted straight lines suggest that a 1 square metre increase in property size is likely to cause comparable price increases for compact and large apartment properties. The cyan line above the red line indicates that apartments with three or more bedrooms are more expensive than those with one or two bedrooms of the same size. This difference is likely due to the higher construction costs associated with building more bedrooms.

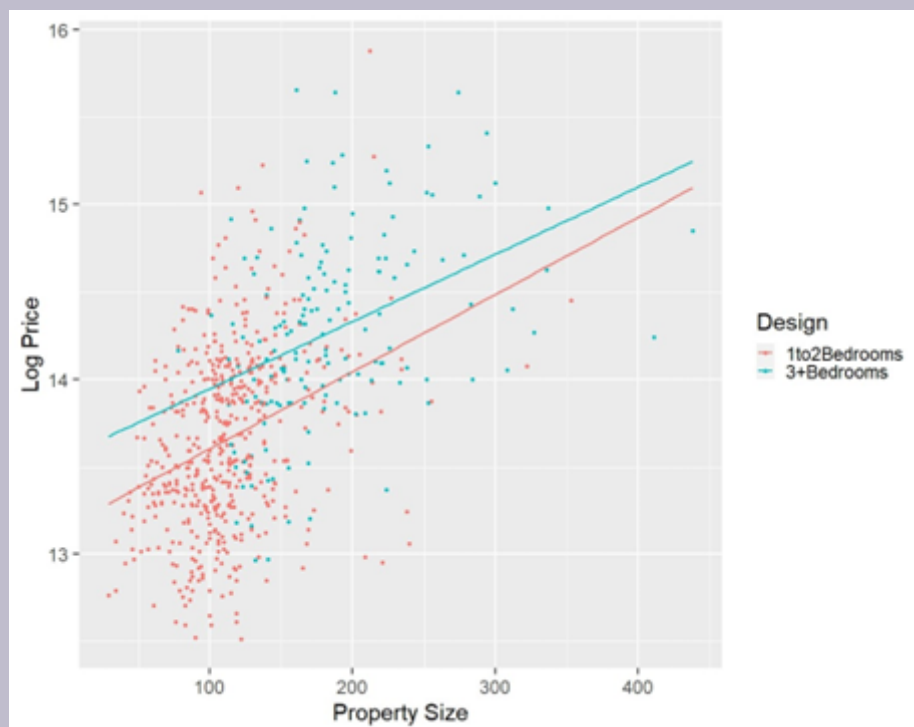


Figure 4.18: A regression analysis to examine the relationship between apartment prices and property size

4.4 Statistical tools in practice

Advancements in computational software have greatly streamlined statistical analysis and data visualization processes. Modern tools like Microsoft Excel (Microsoft Corporation, 2018), Python (van Rossum & Drake, 2009), and R (R Core Team, 2021) offer built-in functions for various statistical analyses. In this chapter, most visualization and computation examples are created using R, with the source code available at <https://github.com/YangANU/Chapter-4-R-Code>

4.5 Test your knowledge

Drag the words into the correct boxes.

1) The rating of a movie on a scale of 1-5 is an example of a(n) variable.

2) For a dataset with a right-skewed distribution, the most appropriate measures of location are

and .

3) The average lifespan of a random sample of car batteries is a

.

4) A dataset with a right-tailed distribution is said to be

to the right.

5) A correlation coefficient of -0.8 indicates a strong

between two variables.

skewed

negative correlation

median

mean

ordinal

statistic



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=391#h5p-5>

An alternative source for this activity: <https://uts.h5p.com/content/1292858506000764799>

References

- DCJ Statistics (2024). *Local Government Housing Kit – Census 2021*. [Tableau Dashboard]. Retrieved from <https://public.tableau.com/app/profile/dcj.statistics/viz/LocalGovernmentHousingKit-Census2021/Summary#1>.
- Galton, F. (1886), Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15, 246–263.
- Kaggle. (2021). *Sydney house prices: Greater Sydney Property Sample from 2016– 2021*. Retrieved from <https://www.kaggle.com/datasets/alexlau203/sydney-house-prices?resource=download>.
- Pearson, K., & Lee, A. (1903). On the laws of inheritance in man: Inheritance of physical characters. *Biometrika*, 2(4), 357–462. <http://dx.doi.org/10.1093/biomet/2.4.357>

Additional readings

- **Introduction to statistics and data visualisation**
 - Simple Learning Pro. (n.d.). *Introduction to statistics and data visualisation* [Video]. YouTube. <https://www.youtube.com/watch?v=MXaJ7sa7q-8>
 - Simple Learning Pro. (n.d.). *Statistics and data visualisation explained* [Video]. YouTube. <https://www.youtube.com/watch?v=uHRqkGXX55I>
 - Simple Learning Pro. (n.d.). *Data visualisation basics* [Video]. YouTube.

<https://www.youtube.com/watch?v=mk8tOD0t8M0>

- **Distribution and central tendency**

- Simple Learning Pro. (n.d.). *Understanding distribution and central tendency* [Video]. YouTube. <https://www.youtube.com/watch?v=sK0RY-Qkug4>
- Simple Learning Pro. (n.d.). *Exploring central tendency in statistics* [Video]. YouTube. https://www.youtube.com/watch?v=vcbMinm_1Q8

- **Central limit theorem**

- 3Blue1Brown. (n.d.). *The Central Limit Theorem explained visually* [Video]. YouTube. https://www.youtube.com/watch?v=zeJD6dqJ5lo&ab_channel=3Blue1Brown

About the author



Yang Yang

UNIVERSITY OF NEWCASTLE

<https://www.newcastle.edu.au/profile/yang-yang10>

Yang Yang is a lecturer at the University of Newcastle, Australia, with expertise in demography forecasting, climate data analysis, and functional time series modelling. He holds a PhD in Statistics from ANU and has published extensively in top-tier journals.

5. DATA/INFORMATION VISUALISATION

Florence Mwagwabi

Learning Objectives (LO)

By the end of this chapter, you will be able to:

- LO1** Compare traditional and contemporary visualisation techniques.
- LO2** Evaluate the role of visualisation in BI and apply fundamental design principles.
- LO3** Design and create various visualisations using the data visualisation taxonomy.
- LO4** Select appropriate chart types for different contexts and communication goals.
- LO5** Create persuasive data stories to communicate insights.
- LO6** Identify and apply Gestalt principles and preattentive attributes to make visualisations impactful.

5 Introduction and evolution of data visualisation

This chapter explores data visualisation and its application in business intelligence (BI). You will learn to create visuals that reveal insights and inform better decisions, learn to apply design principles and discover how to avoid common mistakes that cause misinterpretation. You will also learn how to choose appropriate charts and graphs from basic to specialised visualisations. This chapter covers a range of visualisation techniques and shows you how to create interactive visuals that tell stories with data.

5.1 The evolution of data visualisation

Data visualisation goes back as early as the 10th century (Friendly, 2008). Even then, the intention was to make complex information easy to understand. This section reviews key historical moments in data visualisation, from early graphs to modern interactive dashboards.

5.1.1 Historical overview

Key moments in data visualisation history

- **Pre-17th century:** One of the earliest known examples is a 10th century graph showing how planets move over time.
- **1600s:** Michael Florent van Langren created what many regard as the first known statistical graph to compare longitude estimates between Toledo and Rome (Tufte, 2001).
- **1700s:** William Playfair introduced several foundational chart types, including line graphs, bar charts and pie charts to represent economic data. One notable example is the time-series chart shown in **Figure 5.1**. The chart plots Denmark and Norway's imports and exports from 1700 to 1780 and shows fluctuations between deficits and surpluses in England's trade balance.

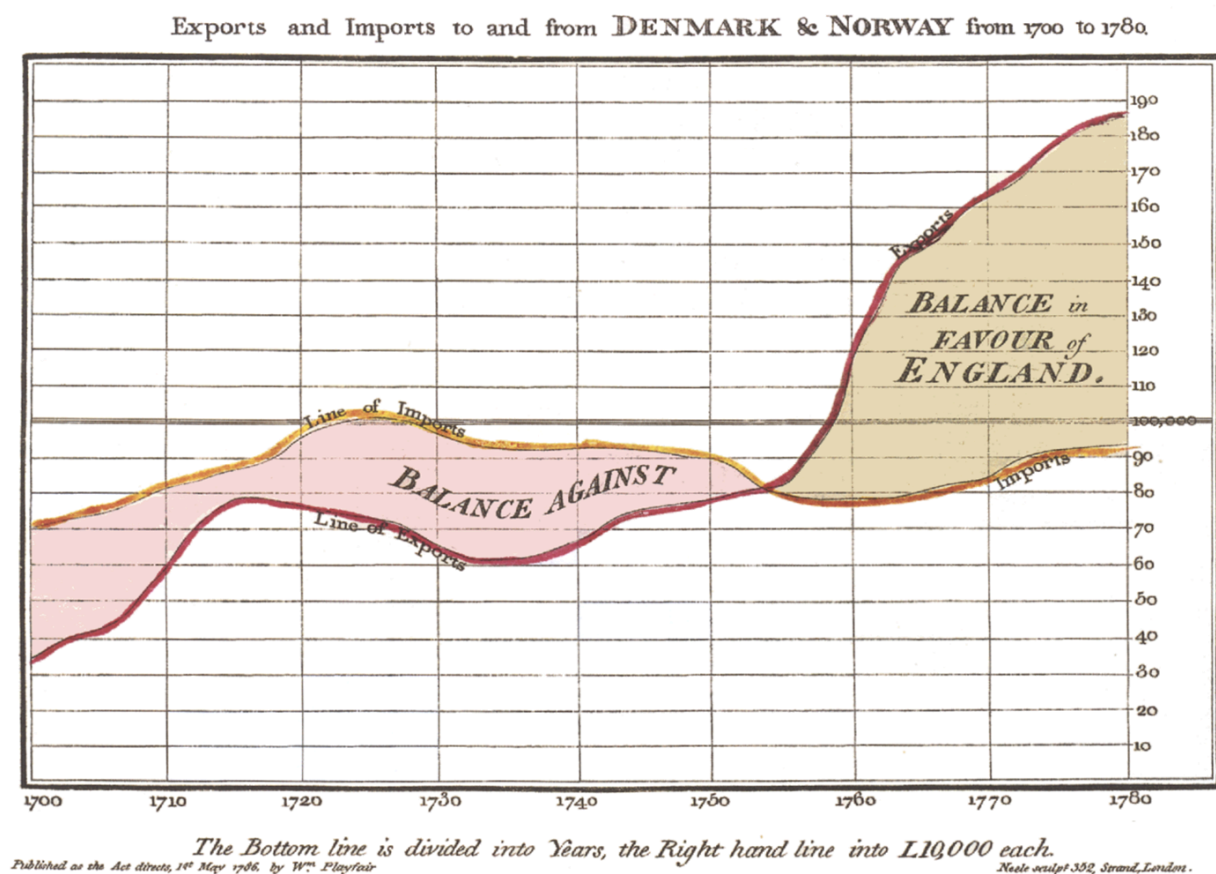


Figure 5.1: One of William Playfair's line charts in his Commercial and Political Atlas (1786). This work is out of copyright and in the public domain. Image via George Mason University Press (https://publishing.gmu.edu/wp-content/uploads/2017/11/Playfair_LineChart3_CommercialandPoliticalAtlas_1786_balanceoftrade.png).

Alternative Text for Figure 5.1

Historical line chart titled 'Exports and Imports to and from Denmark & Norway from 1700 to 1780.'

- Two lines show England's exports (red) and imports (yellow) over time.
- Shaded areas indicate trade balance: a negative balance against England in the early 1700s, shifting to a strong positive balance in England's favor after about 1760 as exports rise sharply above imports.
- X-axis shows years 1700–1780; right y-axis shows values in £10,000 units.

- **1800s:** Charles Dupin introduces shaded (choropleth) maps to display social data, such as illiteracy rates across France.
- **1869:** Figure 5.2 presents Charles Minard's innovative map, later praised in *The Visual Display of Quantitative Information* (1983). The map visually depicts Napoleon's disastrous Russian campaign, correlating the dramatic loss of troops with falling temperatures to illustrate the campaign's catastrophic outcome.

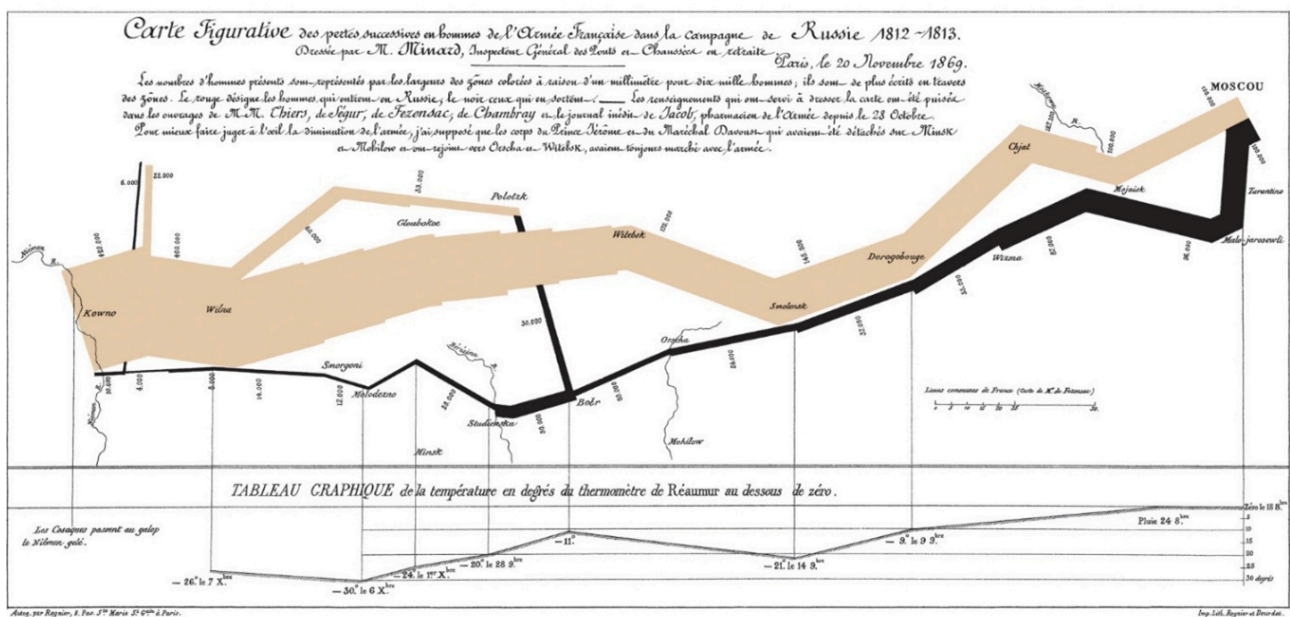


Figure 5.2: Charles Minard's map of Napoleon's Russian campaign (1869). This work is in the public domain.

Alternative Text for Figure 5.2

Charles Minard's 1869 flow map of Napoleon's 1812–1813 Russian campaign, showing the

army's route to and retreat from Moscow.

The width of the band represents troop numbers shrinking dramatically, while a lower chart tracks falling temperatures during the retreat, highlighting the campaign's devastating losses.

- **1853 (Crimean War):** The British Army asks Florence Nightingale to report on hospital sanitation. When officials ignore her report, she creates rose diagrams that show more soldiers die from disease than from wounds. These powerful visuals convince the government to act (Brasseur, 2005). Her work demonstrates how data visualisation can influence social and political reform. We revisit her work in [Section 5.4.2.2](#).
- **Late 1800s:** The French Ministry of Public Works publishes the *Albums de Statistique Graphique*, which contain detailed statistical graphs. Around the same time, Francis Galton develops correlation and regression techniques and uses visual plots to identify these patterns.

The modern era: From spreadsheets to AI

Things changed with computers. Before, it was all manual charts and graphs. Then, in the late 1980s, computers let us store and handle more data, paving the way for modern BI.

- **1980s (spreadsheets, but static):** Excel and similar programs let users organise data and draw basic charts, yet the visuals remained static and users could not interact with them (Friendly, 2008).
- **2000s (interactive dashboards):** The introduction of self-service BI platforms such as Tableau and Power BI start to appear and allow non-technical users to build interactive dashboards and explore data on their own (Theoris, 2024).
- **2020s:** Modern BI tools incorporate AI that suggests visualisations and performs automated pattern detection.
- **The 2020s (AI and AR/VR):** BI integrates AI for automated pattern detection and helps visual selection. Augmented and virtual reality technologies provide new dimensions for BI. They allow users to explore data in immersive three-dimensional environments. For example, AR and VR transform static charts and graphs into interactive experiences. This process improves a user's ability to gain insights and recognise patterns. Full immersion allows a user to walk through a supply chain model or manipulate a 3D financial market representation (Olshannikova et al., 2015).
- **Geospatial revolution:** Organisations began to place data on maps. For example, UPS used geospatial analytics to view traffic and weather data with the aim of finding the best delivery routes (Song et al., 2023). Map-based visuals transformed logistics and urban planning.

How BI visualisation improves decision-making

The evolution of BI has transformed how businesses make decisions. Organisations once relied on slow

reports and static charts. Today, real-time data, interactive dashboards and instant insights enable quick action. This agility allows governments to track a virus during a pandemic (Reeves, et al., 2021), [banks](#) to identify fraud patterns and marketing teams to review campaign performance without a data scientist. ***The goal is to deliver the right information to the right people at the right moment.***

Data visualisation makes BI accessible to non-technical staff. Modern tools allow users to build interactive dashboards. OLAP, a multi-dimensional technique, helps organisations explore data from many angles. This chapter discusses modern data visualisations, from basic graphs to advanced specialised visuals.

5.1.2 Introducing the Airbnb data (Australian case study)

This chapter uses **Australian Airbnb data** to illustrate business intelligence concepts and visualisations. The dataset contains information on reservation patterns, prices and regional performance. We use this real data to demonstrate how BI tools extract insights and to highlight the role of visualisation in informing decisions. Data from this textbook are sourced from Inside Airbnb [<http://insideairbnb.com>].

Data structure and dimensions

The Airbnb dataset has many different categories that fit with BI structures.

- **Listings:** Location, room type, minimum nights and reviews.
- **Pricing:** Daily rates, seasonal prices and regional comparisons.
- **Availability:** Data showing when listings are booked or available.
- **Reviews:** Customer feedback used to create BI metrics.

Summary of the AirBnB data source, copyright and usage information

- **Source Name:** Inside Airbnb
- **Website (URL):** <http://insideairbnb.com>, data source <https://insideairbnb.com/get-the-data/>
- **Brief description:** Inside Airbnb is an independent, non-commercial project that scrapes and provides publicly available data from the Airbnb website. It compiles datasets for numerous cities worldwide, including the Australian regions we've been analysing (Sydney, Melbourne, etc.).
- **Content of the data:** The datasets typically include *listings.csv* (detailed information about properties, hosts, location, price, room type, etc.) and *calendar.csv* (daily availability and pricing for each listing), which matches the data we are using.
- **License:** The data from Inside Airbnb is licensed under a [Creative Commons Attribution 4.0 International \(CC BY 4.0\) license](#). This means that you can copy and redistribute the material in any medium or format, transform and build upon the material for any purpose, even commercially.
- **Usage requirement:** Give **appropriate credit** (attribution) to Inside Airbnb as the source.
- **Attribution.** Inside Airbnb data should be attributed and cited appropriately.

5.2 The role of data visualisation in business

intelligence

5.2.1 Why visualisation matters in BI

Today, organisations generate and collect large volumes of data. The real value of data depends on an organisation's ability to interpret the data and turn it into actionable insights. Data visualisation enables this translation, a critical role in BI. These visual insights support informed organisational decisions. The goal of visualisation is not for aesthetics but to inform (Cairo, 2016).

5.2.2 What is the difference between data visualisation and information visualisation?

Let's start by distinguishing **data visualisation** and **information visualisation**.

Data visualisation is the graphical representation of **raw, structured numerical data**. Its **scope** includes quantitative datasets such as revenue figures, sales data, and time-series data.

The purpose of data visualisation is to help users identify patterns, trends, comparisons, and insights by presenting numerical data in a visual form. By converting numbers into charts and graphs, data visualisation makes it easier to interpret large datasets and supports analytical thinking.

The application of data visualisation is commonly seen in BI tools, where numerical data points are transformed into visuals such as bar charts, line charts, and dashboards. These visuals allow users to explore data, monitor performance, and make data-informed decisions. For example, the bar charts and line charts used in this chapter enable users to visually compare revenue across regions and identify revenue trends over time.

Information visualisation is the broader practice of visually representing **abstract, complex, or non-numerical information**. Its **scope** includes unstructured or semi-structured data such as relationships, processes, systems, and conceptual information.

The purpose of information visualisation is to enhance understanding of complexity, reveal relationships, and support sense-making rather than precise numerical comparison. These

visualisations help users understand how elements are connected and how systems or processes function as a whole.

The application of information visualisation includes visual forms such as concept maps, network diagrams, and flow charts. These are often used to support decision-making in areas such as strategy, process improvement, and systems analysis. For example, a network diagram mapping the connections between Airbnb properties and their booking patterns can reveal relationships and clusters that a simple chart cannot show, helping users uncover insights within complex datasets.

Key Takeaways

- Data visualisation **is a subset** of information visualisation.
- Both aim to make data easier to understand.
- The key difference is in their scope, purpose and application.

5.2.2.1 Why is data visualisation important in BI?

Data visualisation helps us make sense of data and use it to make better decisions. Whether the task is to identify a trend, understand a relationship or communicate with stakeholders, data visualisation plays a critical role in BI. ***It transforms raw data into meaningful insights for informed action.*** Within the BI life cycle, visualisation supports data discovery, models relationships and presents insights in reports.

- **Data discovery:** Visualisations identify patterns and outliers not visible in spreadsheets. They highlight valuable details hidden in raw data. For example, the ordered bar chart (**Figure 5.3**) shows Sydney generates more revenue than other regions. This critical insight might otherwise remain hidden among raw numbers. The heatmap of occupancy rates (**Figure 5.15**) uses dark green shades to make Melbourne's high occupancy immediately visible.
- **Modelling relationships:** Sometimes you must understand how different variables relate to each other. Visual tools such as relationship diagrams or scatter plots highlight correlations and dependencies. The scatter plot (**Figure 5.8**) is a great example. It illustrates the link between average price and occupancy rates across states and how this link varies over time. The plot allows a quick analysis of whether higher prices lead to higher or lower occupancy or if no distinct relationship

exists.

- **Reports and dashboards:** Dashboards summarise complex data for stakeholders to support quick, informed decisions. Interactive charts organise multiple graphs in a single location and make it easy to explore and interact with related data with more efficiency.

5.2.2.2 What is the purpose of visualisation in the context of BI?

Data visualisation communicates insights. However, the effectiveness of visualisation depends on the audience's needs and understanding. Before you create a visual, you must define why you present the data, who needs it if the information remains relevant and accurate. Below are four primary purposes of visualisation in BI.

- **Analysis:** To examine data and identify trends, outliers or unusual patterns. For example, the occupancy heatmap (**Figure 5.15**) identifies regions with sustained high or low occupancy throughout the year.
- **Communication:** To share insights and findings with others. A well-designed visual, such as the treemap that shows revenue by room type and state (**Figure 5.20**), delivers key takeaways without needing detailed explanations.
- **Monitoring:** To track key metrics over time. A real-time dashboard with a line chart (**Figure 5.7**) shows revenue trends. This visual highlights whether revenue increases, decreases or remains steady. Another good example is the use of KPI donut charts to track key metrics in real time.
- **Planning:** To support decisions about future actions. Heatmaps that illustrate occupancy and revenue by month (**Figure 5.15** and **Figure 5.16**) help Airbnb hosts plan price strategies and promotions. For instance, hosts may offer discounts during low occupancy months to attract more reservations. Visualisation supports this process with forecasts of revenue or occupancy rates, which use line charts as a start point for future projections.

5.2.2.3 The role of modern BI tools

Modern BI tools such as Power BI and Tableau help users create visualisations to support decisions. These tools allow users to create interactive dashboards, reports, charts and graphs without technical expertise. Their user-friendly interfaces make it easy to connect to a variety of data sources, from simple spreadsheets to complex databases.

Users can build visualisations through intuitive drag-and-drop actions. For example, Airbnb data can be visualised on a map to highlight popular rental locations, or a line chart can be used to show reservation trends over time. Modern BI tools also support collaboration by making it easy to share dashboards and insights with others.

In addition, tools such as Power BI and Tableau offer built-in forecasting and analytical features that help users plan for future scenarios. As a result, modern BI tools have become essential for organisations seeking to maximise the value of their data and support data-informed decision-making.

5.3 Choose the right charts with Andrew Abela's data visualisation taxonomy

Choosing the correct chart type supports the effective communication of insights. Different charts and graphs serve different purposes. For example, bar charts summarise categorical data, scatter plots illustrate relationships and heatmaps reveal patterns. Knowledge of when and how to use these tools allows you to create meaningful and actionable visualisations.

This section introduces several chart types, from basic options like bar charts, line charts and pie charts to specialised visualisations such as heatmaps, treemaps and geospatial maps. You will learn best practices to select and use these charts, with practical examples from Airbnb data that covers price trends and regional occupancy rates. This knowledge supports your ability to justify your chart choices and to align them with your data context and BI objectives.

Andrew Abela's data visualisation taxonomy (Abela, 2013)

Being able to select the right chart type is critical. Andrew Abela's taxonomy provides a practical framework that categorises visualisations based on what you want your audience to understand. Sometimes referred to as "Chart Chooser", this framework categorises visuals based on their communication purpose and helps practitioners match the right visual to the story they want to tell with their data.

Abela first presented this taxonomy in his book *Advanced Presentations by Design* (2013) and based it on decades of experience at Harvard Business School.

The taxonomy focuses on the fundamental question, *What would you like to show?* and has wide adoption today as a standard reference in academia and industry.

The charts introduced in this section follow Abela's taxonomy, organised into four main categories: ***comparison***, ***relationship***, ***distribution***, and ***composition***.

5.3.1 Comparison charts (among items or over time)

Do you need to compare items side by side or track trends over time? Comparison charts help users compare values between different items or across time periods. According to Abela's data visualisation taxonomy (Abela 2013), comparison charts highlight differences or similarities between items or periods. The taxonomy assigns bar charts for categorical comparisons and line or area charts for time-based comparisons. This process guides users to focus on differences rather than absolute values alone.

To select comparison charts, first you must decide whether your comparison involves items or periods.

- **Comparing items**
 - **Few items:** Use simple bar or column charts.
 - **Many items:** Consider dot plots or sorted bar charts.
- **Comparing periods**
 - **Few periods:** Choose clustered or stacked column/area charts.
 - **Many periods:** Use stacked area charts or small multiples.
 - **Data cyclicity:** Decide if your data involves cyclic patterns, for which you should use radar or cyclic line charts or non-cyclic trends, which require standard line charts.

5.3.1.1 Comparison of two or more items, using basic charts

Do you need to compare categories? Bar charts represent one of the most common and effective chart types in Andrew Abela's data visualisation taxonomy (Abela 2013). You encounter bar charts regularly because they provide an immediate visual comparison of categorical data. Column bars, for instance, allow you to spot differences between values immediately.

When to use bar charts? Use bar charts to compare discrete categories like regions. For example, a bar chart can compare total revenue for Airbnb listings across Australian regions or identify the room type that generates the highest revenue. While bar charts can show data over time, line charts are more appropriate for that purpose.

How to use bar charts? Keep bar charts simple. Each bar represents one category and its length or height represents its value. Do not overcrowd bar charts with too many categories.

Best practices for Bar Charts

- Keep the bars the same width.
- The bars should have even spacing.
- Use concise and clear labels on bars and axes.
- Always start the value axis at zero to prevent data distortion. Bar charts require a zero baseline, as bar length is a visual representation of the data values (Tufte, 2001), otherwise the baseline proportions will be distorted.

Bar Chart Examples from Airbnb data

Bar Chart Example 1 –Total revenue by region

Figure 5.3 shows the total revenue for each region. You can see Sydney generates the highest revenue at over \$284 million, while Barossa Valley generates the least, around \$8.8 million. This bar chart allows a quick comparison across all regions.

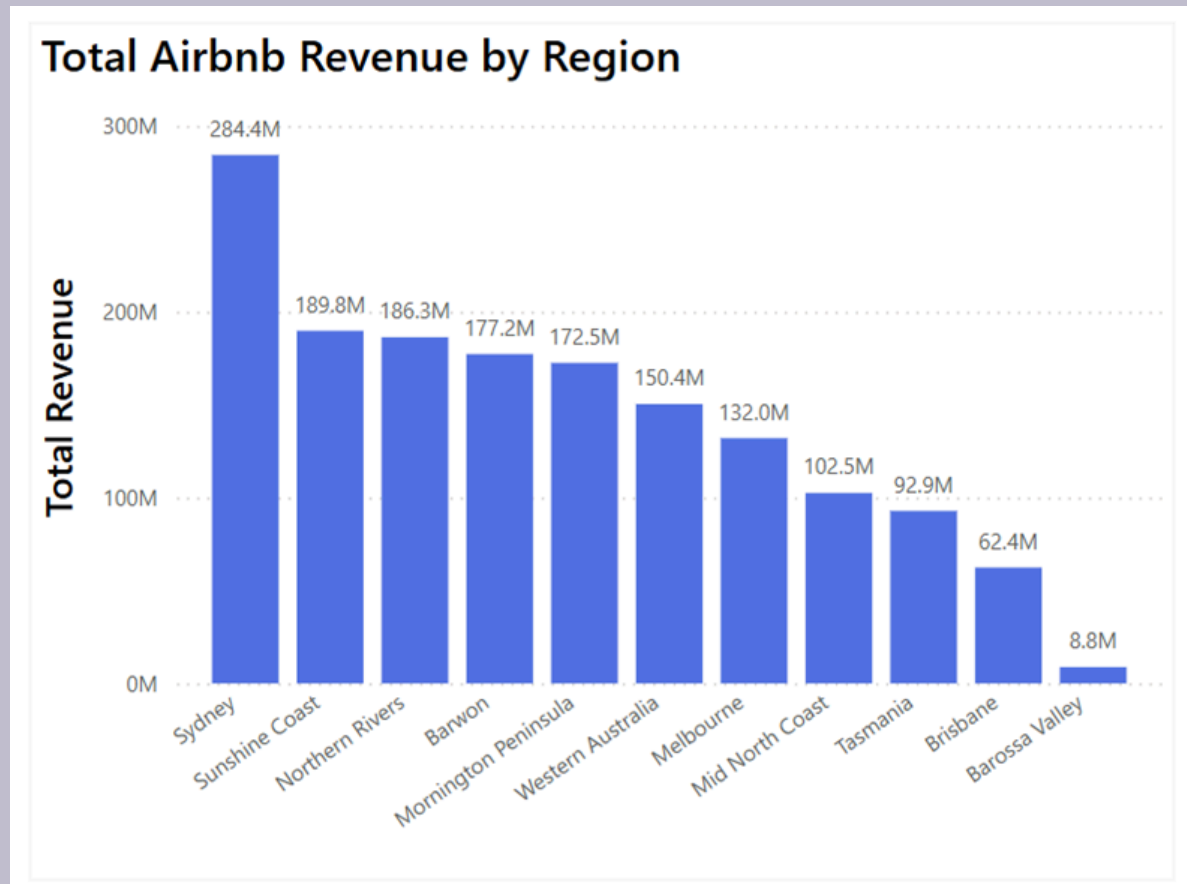


Figure 5.3: Total Airbnb revenue by region. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.3

Bar chart titled 'Total Airbnb Revenue by Region' showing Sydney with the highest revenue (about \$284M), followed by Sunshine Coast and Northern Rivers, with revenues steadily decreasing across regions to Barossa Valley at about \$9M.

Bar Chart Example 2 –Total revenue by region (grouped)

A bar chart can appear cluttered if it has too many categories, as seen in **Figure 5.3**.

Figure 5.4 a better solution because it groups smaller regions into a single category called “Other Regions”. This solution simplifies the comparison and draws attention to major regions like Sydney and Brisbane & Sunshine Coast.

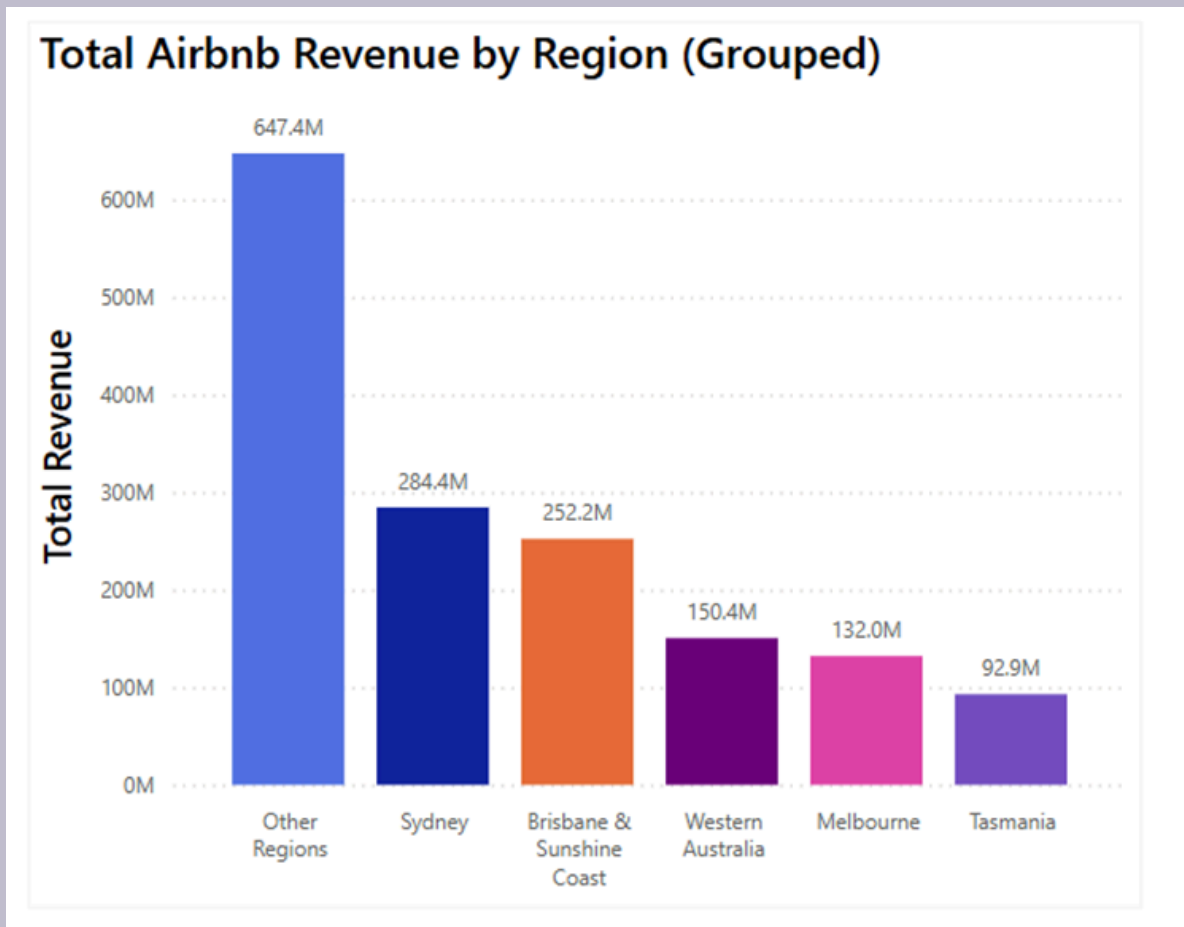


Figure 5.4: Total Airbnb revenue by region (grouped, less cluttered). Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.4

Grouped bar chart titled ‘Total Airbnb Revenue by Region (Grouped)’ showing ‘Other Regions’ with the highest total revenue (about \$647M), followed by Sydney (\$284M), Brisbane & Sunshine Coast (\$252M), Western Australia (\$150M), Melbourne (\$132M), and Tasmania (\$93M).

Bar Chart Example 3 – Which room type brings in the most revenue?

Figure 5.5 shows revenue by room type. The graph reveals that the “Entire home/apt”

category generates substantially more revenue, over \$1.3 billion, compared to other types. Private rooms follow with approximately \$233 million, while hotel and shared rooms contribute much less. Different bar chart formats allow us to explore the data from different perspectives. Note strategic use of colour to highlight key insights, a topic we discuss further in [Section 5.5.3.2](#).

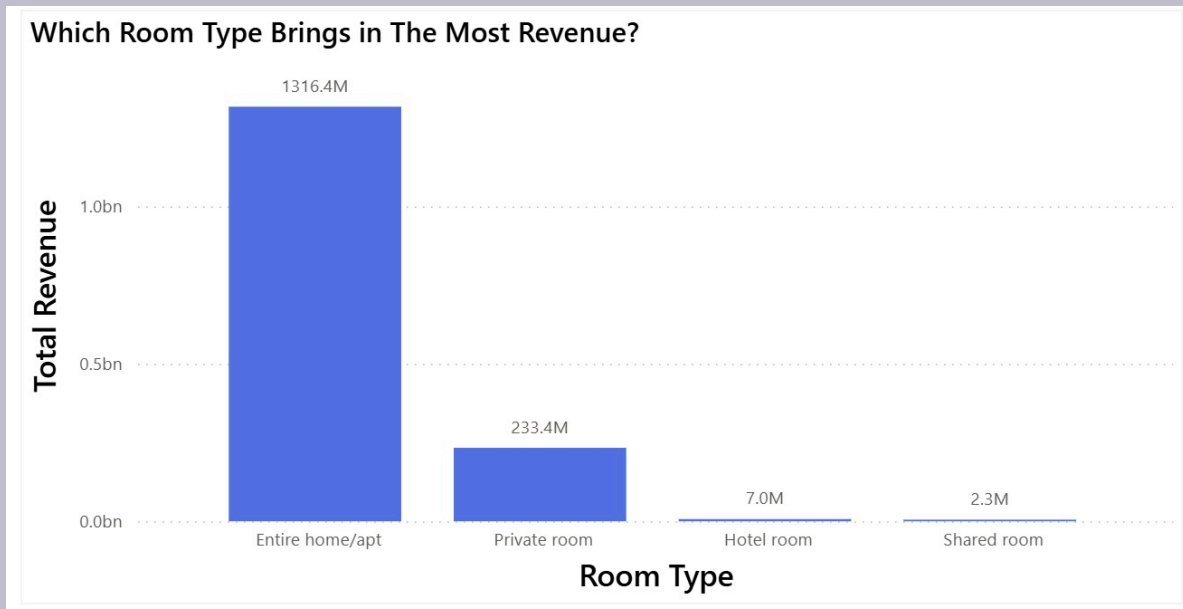


Figure 5.5: Airbnb revenue by room type. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.5

Bar chart titled 'Which Room Type Brings the Most Revenue?' showing entire homes/apartments generating the highest revenue (about \$1.3B), far exceeding private rooms (\$233M), hotel rooms (\$7M), and shared rooms (\$2M).

Key insights

- **Figure 5.3:** Sydney generates the highest revenue, significantly more than all other regions.
- **Figure 5.4:** The “Other Regions” category, combining multiple smaller regions, contributes to total revenue. Perhaps further analysis is needed?
- **Figure 5.5:** Entire home/apt listings generate more revenue than all other room types combined.

Additional questions to explore

1. **Sydney vs Barossa Valley:** Sydney performs well whereas Barossa Valley struggles. Further analysis should clarify why: does it relate to listing counts, pricing, occupancy or a combination of factors?
2. **Other Regions:** Investigate specific regions within “Other Regions” to identify strong performers. Examine for any hidden insights.
3. **Entire home/apt focus:** Since “Entire home/apt” listings dominate revenue, further analysis should identify regions generating the highest revenue for this category. Also, explore how prices and occupancy rates affect revenue outcomes.

5.3.1.2 Comparison over time, using line charts (over many periods)

Do you need to visualise how values change over days, months or years? To observe trends across periods, Abela’s data visualisation taxonomy (Abela 2013) recommends line charts. Line charts illustrate continuous changes over intervals such as days, months or years.

When to use line charts? Following Abela’s taxonomy, select line charts for the comparison of data points connected in a sequence over consistent intervals, usually time-based data. Common examples include reservation trends, price fluctuations or website traffic. Ensure the intervals remain consistent, such as daily, monthly or yearly periods. Lines must only connect data points from intervals that are adjacent.

How to use line charts? Plot each data point on the chart, with the time along the horizontal x-axis and the measured value along the vertical y-axis. Draw lines to connect adjacent data points to illustrate overall trends, such as a rise, a fall or a stable pattern.

Best practices for Line Charts

- Use a solid line to make the trend apparent. Avoid dashed or dotted lines unless you show something specific like a forecast or a goal.
- Do not plot more than 3–4 lines on a single chart as it gets messy. If you need to

show more, consider the use of small multiples, a topic we discuss later.

- Label the axes so the chart's purpose is easy to understand.
- Start the value axis at zero if it makes sense for the data. If all your data points are high, you might use a different start point to show the fluctuations with more detail. Use caution, as this can create a false impression if done without care (Yau, 2024). Always add a note if you do not start at zero.

Line Chart Examples from Airbnb data

Figure 5.6 tries to show revenue over time for several key states on one chart. The result is hard to follow because the lines overlap. The overlap of lines is a good example of what not to do when you have too many categories. Note the y-axis does not start at zero. This choice lets us see the fluctuations with more detail, but it can create a false impression, so you must take care.

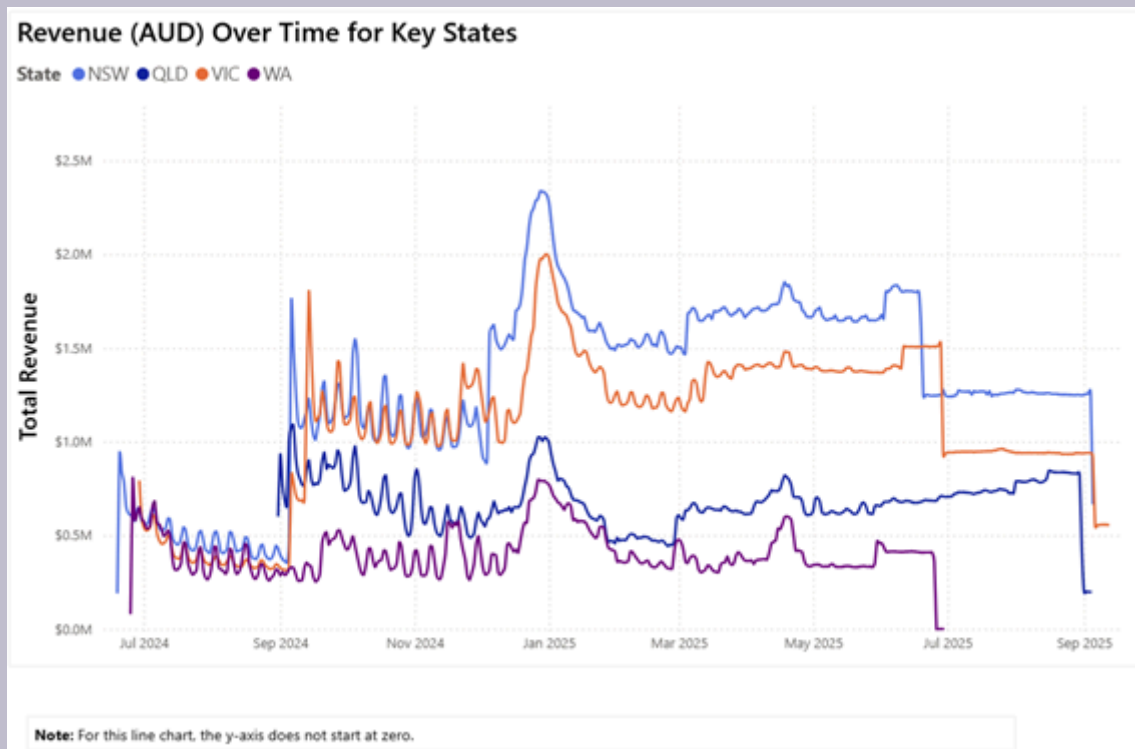


Figure 5.6: Revenue over time for key states (follows the 3-4 liners best practice rule).
Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.6

Line chart titled 'Revenue (AUD) Over Time for Key States' showing daily Airbnb revenue trends from mid-2024 to late-2025 for NSW, QLD, VIC, and WA.

NSW consistently records the highest revenue with a peak in early 2025, followed by VIC, while QLD and WA show lower and more variable revenue levels.

The y-axis does not start at zero.

Remember **Figure 5.6** presented a cluttered line chart with overlapping revenue data for several states.

Figure 5.7 fixes this issue with the use of small multiples, a feature in Power BI. Small multiples are a series of smaller line charts, with one chart for each state: New South Wales, Queensland, Victoria and Western Australia. Each chart displays the same metric, monthly revenue over time, for its respective state.

This format makes it simple to compare trends. New South Wales shows the highest and most volatile revenue, with significant peaks around January and October. Queensland's revenue is

more stable, while Victoria shows some variability, though less extreme than New South Wales. Western Australia's revenue follows a different pattern. The y-axis in **Figure 5.7** starts at zero, which allows for an accurate comparison of overall revenue levels between states.

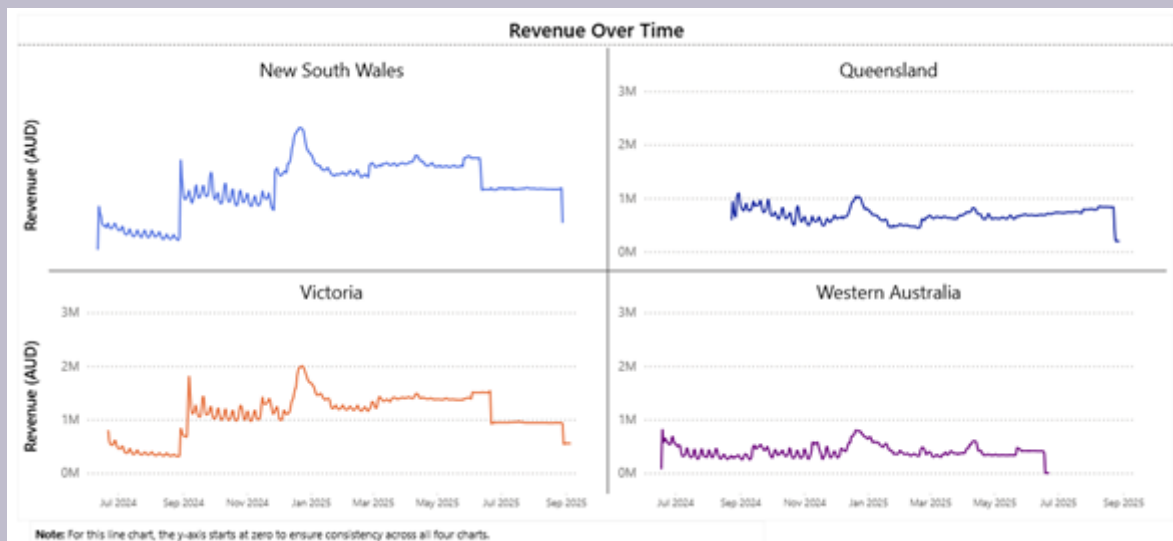


Figure 5.7: Revenue over time by state (one-line charts for each state). Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.7

Four-panel line chart titled 'Revenue Over Time' showing Airbnb revenue trends (AUD) from mid-2024 to late-2025 for New South Wales, Queensland, Victoria, and Western Australia. New South Wales and Victoria show higher and more volatile revenues with peaks in early 2025, while Queensland and Western Australia remain lower and steadier. All charts share a y-axis starting at zero for consistency.

Key insights

- We will discuss information overload and clutter, including the Gestalt principles, in **Section 5.5.3.1**. The *small multiples* approach in **Figure 5.7** is effective due to these

principles.

- The charts share a uniform appearance (axes and line style), which relates to the principle of similarity.
- Because the charts are grouped together, an application of proximity, the viewer can compare them with ease. This design allows a person to focus on one state at a time and compare it to others without the confusion of crossed lines.

Additional questions to explore

- **NSW deep dive:** What causes the revenue spikes in January and October? Can this success be replicated elsewhere?
- **Queensland consistency:** How can revenue be increased while stability is maintained?
- **Victoria vs. NSW:** What drives the difference in performance between these two states?
- **Western Australia's pattern:** Why is its pattern different? Do the line gaps indicate zero revenue due to data issues or seasonality?

5.3.2 Relationship charts

Do you want to explore how variables interrelate? Relationship charts illustrate connections or correlations between two or more numerical variables. According to Andrew Abela's data visualisation taxonomy (Abela 2013), these charts highlight relationships to show if variables connect and the strength of that connection. Industries such as healthcare, finance and technology often use relationship charts, like scatter plots, to identify correlations and inform strategic decisions.

Abela's taxonomy classifies scatter plots for analyses that involve two variables and bubble charts for scenarios that involve three variables. These charts identify patterns, clusters and outliers.

How to select relationship charts (Abela 2013)

- **Two variables:** Use a 2D scatter plot.
- **Three variables:** Add size or colour encoding to a scatter plot. This addition creates a bubble chart.
- **More than three variables:** Choose small multiples or interactive filters rather than complex 3D charts.

5.3.2.1 Reveal relationships with scatter plots

Do you need to identify if changes in one metric align with another? Scatter plots offer a visual representation of the relationship between two numerical variables. One variable appears on the horizontal x-axis, another on the vertical y-axis and each dot represents a single data point.

When to use scatter plots? Select scatter plots to identify potential relationships between two numerical variables. For example, in our Airbnb dataset, a scatter plot might explore the relationship between average price per listing and occupancy rate.

How to use scatter plots? Plot one variable on the x-axis and the other on the y-axis. Each dot represents a data point, for example, a state's data in a specific month. Observe the pattern that results. Do dots form a line, curve, cluster together or scatter without a pattern? The dot pattern provides insight into the relationship.

Best practices for Scatter Plots

- Use clear labels for the x and y axes so it is easy to understand what variables are being compared.
- If you have a lot of data points, consider using different colours or shapes to represent various categories (like other regions in the Airbnb data).
- Add a trend line if it helps to show the relationship more, *but only if it makes sense for the data*. Don't force a trend line if there isn't a clear one.

Scatter Plot Example from Airbnb data

Figure 5.8 is a scatter plot that shows the relationship between average price and occupancy for Airbnb listings, segmented by state and month. Each bubble on this chart represents a state in a specific month. A slider at the bottom allows the user to change the month and observe how the data relationships change. The bubble's position indicates the average daily rate on the horizontal x-axis and the occupancy rate on the vertical y-axis. The bubble's size corresponds to the total revenue for that state and month, while colour distinguishes the different states.

The chart helps an analyst identify the connection between price and occupancy and how that relationship changes throughout the year. The ideal outcome is to see bubbles in the top right quadrant, an indication of high occupancy at high prices. A user can hover over any bubble to see the exact details for that state and month: average price, occupancy rate and total revenue.

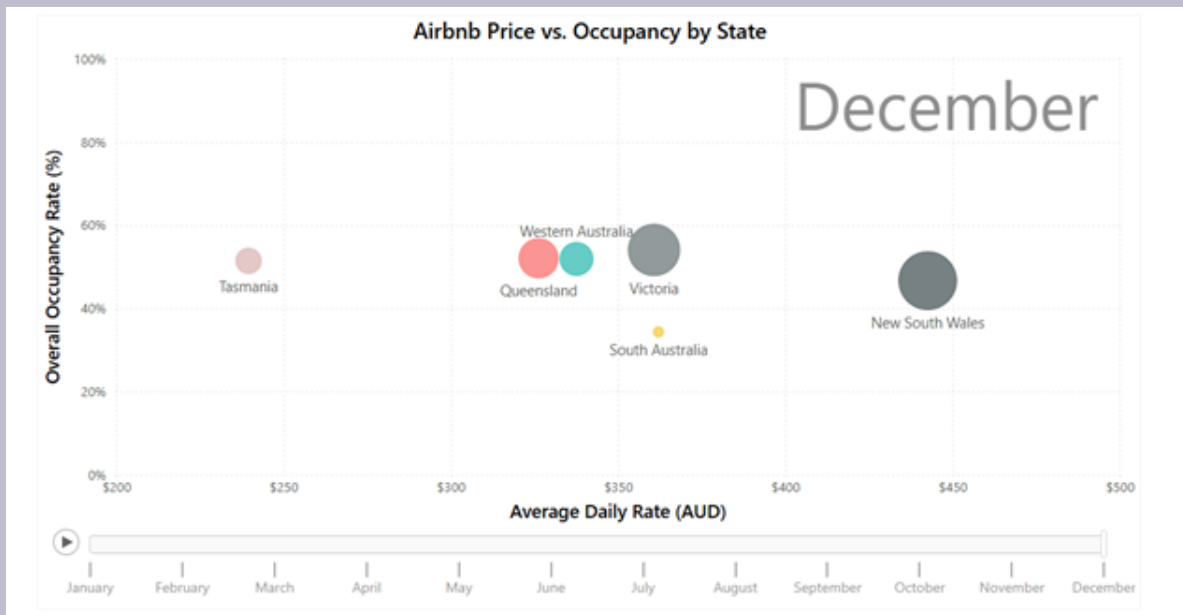


Figure 5.8: Scatter plot showing the relationship between price and occupancy. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.8

Bubble chart titled 'Airbnb Price vs. Occupancy by State' for December, showing average daily rate on the x-axis and overall occupancy rate on the y-axis.

New South Wales has the highest prices and mid-range occupancy, Victoria and Western Australia show moderate prices with higher occupancy, Queensland and Tasmania cluster at lower prices, and bubble sizes indicate relative listing volume by state.

Key insights

- Just looking at December reveals that different states have distinct price and occupancy combinations. No single pattern applies to all states.
- The relationship between price and occupancy is not static; it shifts from month to month. An analysis over time reveals these changes.

Additional questions to explore

1. **Investigate individual states:** Investigate states with different patterns, such as NSW and Tasmania, with the use of time-series line charts.
2. **Compare to external factors:** Determine if these patterns correlate with known events, holidays or seasonal trends in each state.
3. **Consider other variables:** Investigate the influence of other variables, such as the number of reviews or the room type, on price and occupancy. This may require the integration of more data for a complete picture.

5.3.3 Distribution charts

Do you need to understand how your data values spread or how frequently they occur?

Distribution charts show how data points spread across a range of values. They show frequency, variability and patterns within the data. Organisations often use distribution charts for quality control, risk management and market research to identify data patterns and anomalies.

According to Andrew Abela's data visualisation taxonomy (Abela 2013), distribution charts visualise how data points distribute across a range. For single variable distributions, Abela recommends histograms, box plots or density curves. For bivariate (two variable) distributions, Abela suggests scatter plots, 2D density plots or contour plots.

Decision flow for distribution charts (Abela 2013)

- **Single variable:** Use histograms for many data points. For fewer data points, select box plots or dot plots.
- **Multiple variables:**
 - **Two variables:** Select scatter or 2D density plots.
 - **Three variables or more:** Consider contour plots or small multiple histograms rather than complex 3D visuals.

5.3.4 Composition charts

Do you need to illustrate how parts come together to form a whole? Composition charts show proportions and how individual components contribute to the overall total. According to Andrew Abela's data visualisation taxonomy (Abela 2013), composition charts represent part to whole relationships, either at a single time point or across multiple periods.

Abela's taxonomy identifies static composition charts, which show one moment in time, as pie charts, stacked bar charts and waterfall charts. Dynamic composition charts, which show changes over time,

include stacked area charts or small multiples. Composition charts appear with frequency in financial reports, budgets and resource allocation plans across many areas.

How to select composition charts (Abela 2013)

- **Static (single time point)**
 - **Few categories:** Use a pie chart.
 - **Multiple categories:** Select a stacked bar chart.
 - **Sequential components:** Choose a waterfall chart.
- **Dynamic (over time)**
 - **Few periods:** Consider clustered or stacked column charts or 100% stacked area charts.
 - **Many periods:** Select stacked area charts or small multiples.

5.3.4.1 Displaying the composition of static data with pie charts

Do you need to quickly see how a total divides into category shares? Pie charts show parts of a whole, usually categorical data, and each slice represents a proportion of the total. According to Andrew Abela's data visualisation taxonomy (Abela 2013), pie charts illustrate static composition with skill, especially for a small number of categories.

When to use pie charts? Abela's taxonomy recommends pie charts to illustrate how different parts contribute to a single total. An example is a chart that shows the percentage share of revenue by region. Pie charts work well with fewer categories but become unclear and complicated with many. When category numbers increase, alternative visuals like bar charts or treemaps become more suitable.

How to use pie charts? Ensure each slice is large enough to see and labelled and that all slices together total 100%. Arrange slices from largest to smallest, starting from the 12 o'clock position, to aid readability.

Best practices for Pie Charts

- Limit slices to six or fewer categories.
- Use bar charts or treemaps (discussed later in this chapter) for seven or more categories to avoid clutter.
- Label slices, indicating the category and its percentage.
- Avoid 3D effects (Few, 2007).

Pie Chart Examples from Airbnb data

Let's see how pie charts work (and *don't* work) with our Airbnb data. **Figure 5.9** shows the revenue breakdown for *all* our regions. But, as you can see, it's overloaded. There are too many small slices, making it hard to compare them, and the labels are cluttered. A bar chart would be *much* better for this (like we saw in **Figure 5.3**).

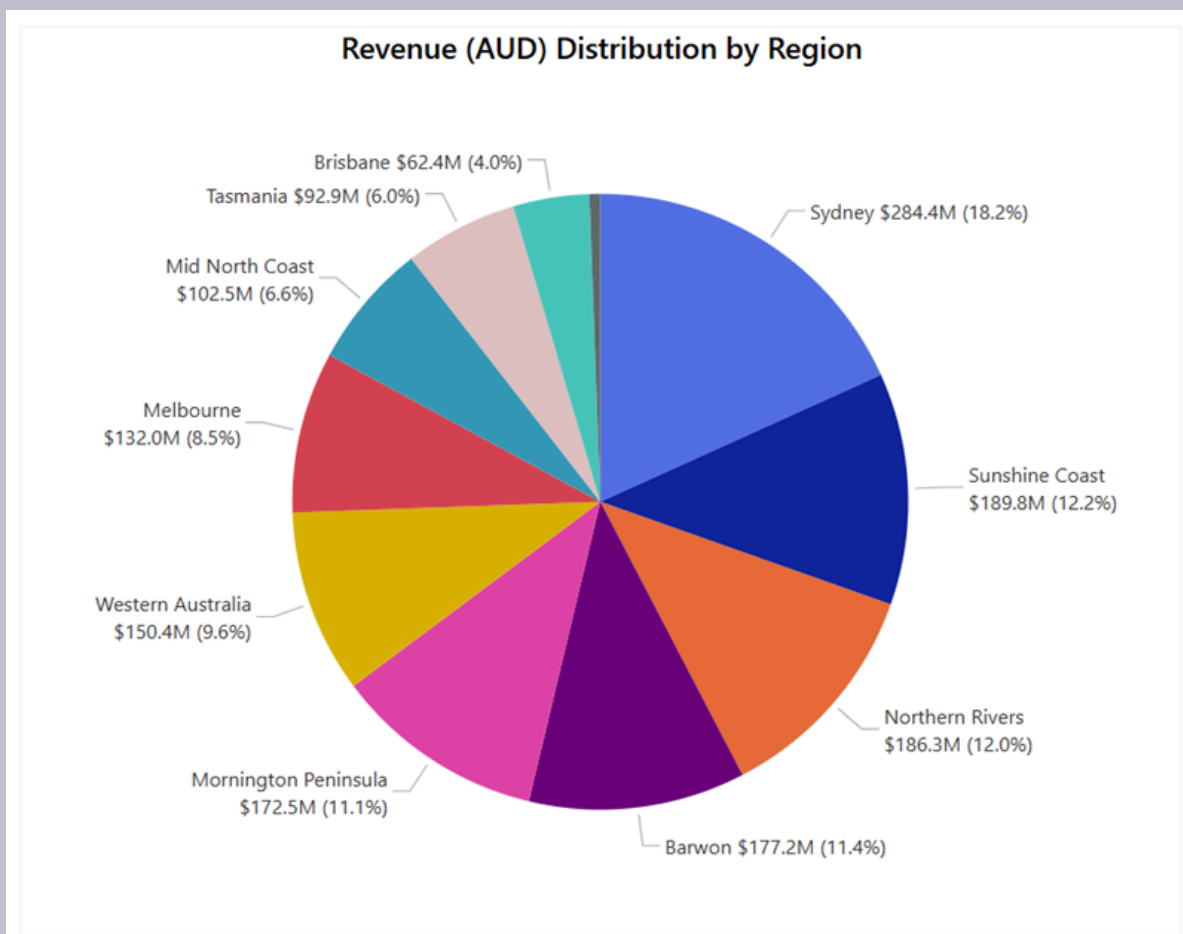


Figure 5.9: Overall revenue distribution by region (overloaded pie chart). Source: <http://insideairbnb.com>

Alternative Text for Figure 5.9

Pie chart titled 'Revenue (AUD) Distribution by Region' showing Sydney as the largest share at about \$284M (18%), followed by Sunshine Coast, Northern Rivers, Barwon, and Mornington

Peninsula, with smaller shares from Western Australia, Melbourne, Mid North Coast, Tasmania, and Brisbane.

The pie chart in **Figure 5.9** shows a common problem: it is overloaded with 11 categories. This large number of slices makes it difficult to distinguish between the different regions, especially the smaller ones, and creates a cluttered chart. A bar chart would offer a better comparison of revenue across all regions. However, to improve the pie chart, one could group the smallest regions into an “Other” category. **Figure 5.10** shows this method of reducing clutter to improve readability.

Improving our pie chart

We saw in the pie chart shown in **Figure 5.9** that too many slices can be a problem. It was hard to compare the different regions, especially the smaller ones, and the labels were all over the place. Cluttered slices made it tough to get a clear picture of revenue share. Now, let’s look at a better way to present this data, following the best practices we discussed earlier.

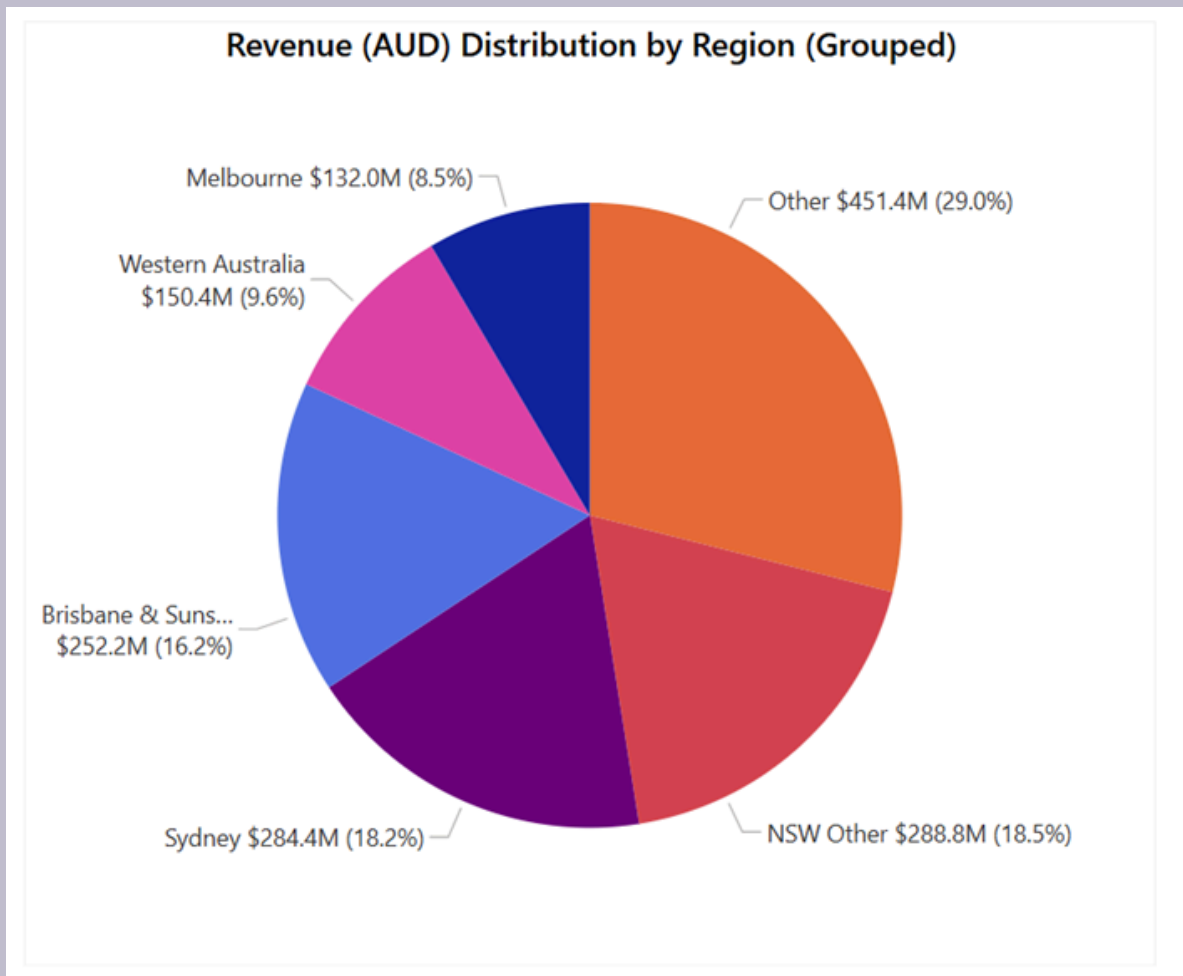


Figure 5.10: Overall revenue distribution by region (less cluttered pie chart).

Alternative Text for Figure 5.10

Pie chart titled 'Revenue (AUD) Distribution by Region (Grouped)' showing 'Other' regions as the largest share at about \$451M (29%), followed by NSW Other (\$289M), Sydney (\$284M), Brisbane & Sunshine Coast (\$252M), Western Australia (\$150M), and Melbourne (\$132M).

Figure 5.10 is the *same* data, but we've grouped some smaller regions into an "Other" category and combined some into "NSW Other". See how much clearer that is? Now you can compare the main contributors.

Key insights

- This pie chart in **Figure 5.9** gives us a general idea. Too many slices makes comparing smaller regions difficult.
- **Figure 5.10** is the *same* data, but we've grouped some smaller regions. The "Other" category represents a significant portion of total revenue (29.0%). "NSW Other" also represents a significant portion (18.5%). Sydney (18.2%) and Brisbane & Sunshine Coast (16.2%) are the largest contributors to revenue.

Additional questions to explore

1. Drill into Other and NSW Other.
2. Use a bar chart or treemap for many categories. As you can see, a treemap (**Figure 5.19**) is a good alternative to a pie chart with many categories.

5.3.4.2 Display the composition of raw data with tables

Do you need exact figures rather than visual summaries? While charts allow a rapid perception of patterns, sometimes exact details are essential. According to Andrew Abela's data visualisation taxonomy (Abela 2013), tables display raw numbers and precise values. Analysts often prefer tables when exact data points matter more than visual form (Kolokolov & Zelensky 2024).

When to use tables? Abela's taxonomy recommends tables when charts cannot present detailed, exact information with clarity or when audiences require precise data lookups. For example, detailed breakdowns of Airbnb listings by region and room type, along with key metrics, often appear with more clarity as tables. Additionally, tables support visuals and provide the exact numbers behind graphs.

How to use tables? Organise tables simply, using defined rows and columns. Each row represents a single item (e.g., region or listing) and each column represents a distinct variable (e.g. price or occupancy). Clearly label rows and columns with consistent formatting throughout.

Best practices to use Tables (Few, 2012)

- Keep tables simple and uncluttered. Avoid excessive information.
- Use clear and concise labels for rows and columns.

- Ensure consistent formatting (e.g., number of decimal places, units).
- Use gridlines sparingly. Too many lines make the table cluttered.
- Emphasise important data points with bold or shading, but don't overdo it.

Table Examples from Airbnb data

Figure 5.11 shows a basic table with details for region, room type, listing counts, average price, occupancy rate and total revenue. Such a table is valuable for a lookup of a specific number, like the average price of a private room in Barossa Valley. It also allows a side-by-side comparison of all metrics for different regions. However, the volume of data makes it difficult to see trends or a quick overview. The table serves as the raw material for other visualisations that can tell the story more visually.

Airbnb Listing Details by Region and Room Type

State	Region	Room Type	Listing Count	Average Price	Occupancy Rate	Total Revenue
Tasmania	Tasmania	Shared room	9	\$43.375	0.59	\$70.23K
New South Wales	Northern Rivers	Shared room	6	\$60.75	0.23	\$24.58K
Queensland	Brisbane	Shared room	13	\$75.5	0.55	\$88.24K
Victoria	Melbourne	Shared room	273	\$79.2424	0.84	\$799.38K
Western Australia	Western Australia	Shared room	48	\$85.0434	0.67	\$211.11K
New South Wales	Northern Rivers	Hotel room	3	\$101.7191	0.27	\$39.17K
New South Wales	Mid North Coast	Shared room	3	\$128.3333	0.43	\$85.90K
Victoria	Melbourne	Private room	6527	\$128.4341	0.76	\$25,301.70K
Western Australia	Western Australia	Private room	2300	\$129.7057	0.51	\$10,020.02K
Victoria	Barwon	Hotel room	16	\$134.1667	0.48	\$238.80K
Queensland	Brisbane	Private room	1233	\$137.1773	0.52	\$10,984.53K
Tasmania	Tasmania	Private room	717	\$152.633	0.47	\$6,336.84K
Tasmania	Tasmania	Hotel room	89	\$171.1081	0.45	\$1,103.93K
Victoria	Barwon	Private room	689	\$178.9565	0.46	\$6,591.76K
New South Wales	Northern Rivers	Private room	573	\$191.4107	0.45	\$7,916.96K
New South Wales	Sydney	Shared room	87	\$191.4691	0.27	\$195.70K
Western Australia	Western Australia	Hotel room	49	\$197.2253	0.22	\$375.31K
New South Wales	Mid North Coast	Private room	332	\$209.8634	0.40	\$6,719.33K
Victoria	Melbourne	Entire home/apt	18229	\$229.5332	0.71	\$105,834.55K
Queensland	Brisbane	Entire home/apt	3620	\$230.7321	0.54	\$51,196.09K
Queensland	Sunshine Coast	Private room	460	\$238.7962	0.56	\$6,625.58K
South Australia	Barossa Valley	Private room	35	\$250.7059	0.39	\$513.42K
Tasmania	Tasmania	Entire home/apt	5415	\$253.1829	0.48	\$85,360.15K
South Australia	Barossa Valley	Hotel room	1	\$270	1.00	\$98.55K
Queensland	Brisbane	Hotel room	5	\$285.8	0.30	\$108.61K
Victoria	Mornington Peninsula	Private room	93	\$301.4167	0.47	\$2,588.39K
New South Wales	Mid North Coast	Hotel room	10	\$320.5	0.40	\$436.65K
New South Wales	Sydney	Hotel room	78	\$341.4583	0.17	\$526.35K
New South Wales	Mid North Coast	Entire home/apt	4489	\$341.5771	0.39	\$95,233.93K
Total			94724	\$373.3255	0.49	\$1,559,159.90K

Figure 5.11: Table showing Airbnb listings by region and room type. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.11

Table titled 'Airbnb Listing Details by Region and Room Type' listing Australian states and regions with room type, listing count, average price, occupancy rate, and total revenue. Entire homes/apartments dominate revenue, with Melbourne, Sydney, and Tasmania generating the highest totals. A summary row shows 94,724 listings, an average price of about \$373, 49% occupancy, and total revenue of about \$1.56B.

Figure 5.12 presents some of that same data in a visual format. The matrix is a table that groups data in rows and columns. The table uses colour-coded icons: a green checkmark for high occupancy, a yellow exclamation mark for moderate occupancy and a red X for low occupancy. The icons make it simple to look across the rows to see how occupancy changes throughout the year. For example, Melbourne shows high occupancy for most of the year, while

Barossa Valley shows largely low occupancy. This type of table is effective to identify seasonal trends and to compare regional performance each month.

Region	January	February	March	April	May	June	July	August	September	October	November	December
Barossa Valley	37.00%	34.70%	35.55%	44.76%	38.07%	41.16%	43.52%	34.00%	30.46%	35.88%	32.50%	34.35%
Barwon	61.54%	53.14%	55.16%	60.11%	59.12%	59.18%	41.74%	32.02%	31.67%	36.64%	39.84%	46.13%
Brisbane	44.75%	42.89%	55.75%	58.08%	56.72%	60.65%	61.39%	60.15%	62.92%	45.71%	38.30%	48.34%
Melbourne	71.52%	68.07%	73.61%	72.64%	71.81%	74.79%	75.39%	74.76%	75.91%	71.73%	68.83%	71.92%
Mid North Coast	47.72%	29.17%	36.05%	39.99%	35.27%	40.07%	41.93%	40.41%	45.32%	41.01%	29.79%	45.90%
Mornington Peninsula	45.96%	30.30%	35.27%	39.13%	37.91%	44.16%	47.58%	47.29%	44.81%	29.02%	31.64%	41.45%
Northern Rivers	49.10%	43.94%	46.12%	50.14%	48.25%	49.98%	38.17%	31.43%	32.20%	36.50%	36.34%	44.97%
Sunshine Coast	48.96%	35.54%	47.57%	50.79%	47.19%	51.80%	54.91%	58.19%	61.16%	48.93%	41.49%	54.19%
Sydney	44.01%	40.00%	50.63%	50.10%	48.53%	53.85%	54.48%	54.34%	53.23%	46.24%	44.36%	49.05%
Tasmania	48.33%	38.87%	48.46%	47.31%	46.28%	52.28%	52.67%	52.65%	51.58%	44.60%	41.23%	51.43%
Western Australia	56.85%	46.53%	46.28%	51.16%	47.75%	51.53%	53.99%	43.12%	44.43%	45.61%	44.10%	51.89%

Figure 5.12: Visual matrix (table with visualisations). Source: <http://insideairbnb.com>

Alternative Text for Figure 5.12

Table titled 'Airbnb Regions: Occupancy Trends' showing monthly occupancy rates from January to December for Australian regions including Sydney, Melbourne, Brisbane, and Tasmania. Occupancy generally peaks in mid-year, with Melbourne showing consistently high rates across most months, while regional areas display greater seasonal variation.

Table legend for Figure 5.12

- **Green checkmark (✓):** High occupancy (70%–100%)
- **Yellow exclamation mark (!):** Medium occupancy (50%–69%)
- **Red X:** Low occupancy (0%–49%)

Key insights

- Melbourne maintains a high occupancy rate throughout the year, as indicated by the frequency of green checkmarks.
- Most regions show different seasonal patterns with variations in occupancy at different

times of the year.

- Barossa Valley shows a pattern of low occupancy rates across most months.

Additional questions to explore

This matrix gives us an excellent overview of *when* each region's occupancy rates are high or low. Now, we need to dig into the *why*. Here are some next steps.

1. **Investigate seasonal drivers:** What events, holidays or weather patterns might cause the peaks and troughs in occupancy for each region?
2. **Barossa Valley focus:** Why does Barossa Valley show persistent low performance? Is the cause a lack of awareness, price issues, listing quality or something else?
3. **Compare with revenue:** How do these occupancy trends relate to revenue? A comparison with the revenue heatmap **Figure 5.16** (shows that some regions with high occupancy do not generate the most revenue. For example, Melbourne has high occupancy most of the year, but its revenue is not as high as NSW, which suggests a potential to increase prices.

5.3.4.3 Display composition of changes over times with area charts

Do you need to visualise both volume and trends over time? Area charts, closely related to line charts, show changes over a period and show both trends and overall quantities. Area charts show part to whole compositions and trends (Abela 2008). They fill the area between the line and the horizontal axis. Area charts can appear in two formats: stacked, which emphasises composition; or unstacked, which focuses on individual trends.

When to use area chart? Use area charts to display changes over time with focus on total volumes and the contribution of different parts to a whole.

- **Stacked area chart (composition):** Shows the total volume and each part's contribution over time. For example, a chart of total revenue broken down by region (**Figure 5.14**) or total reservations segmented by room type (**Figure 5.13**).
- **Unstacked area chart (composition):** Compares trends of multiple categories without emphasising the combined total. The chart is effective in showing how individual categories change, even when areas overlap.

How to use area charts? Plot data points like a line chart. Place time intervals on the horizontal x-axis and measured values on the vertical y-axis. Fill the area beneath the line with colour. For stacked area charts, select distinct yet complementary colours to ensure readability.

Best practices for Area Charts

- Use transparent colours to differentiate overlapping areas.
- Limit the number of categories displayed to avoid visual clutter.
- Clearly label axes to help readers interpret the data accurately.

Area Chart Examples from Airbnb data

We can use area charts to show different aspects of the Airbnb data. **Figure 5.13** is a stacked area chart that shows total Airbnb revenue over time, segmented by room type. The stacked, coloured areas allow a user to see both the overall revenue trend and the contribution of each room type. The chart shows that the “Entire home/apt” category comprises the largest portion of annual revenue.

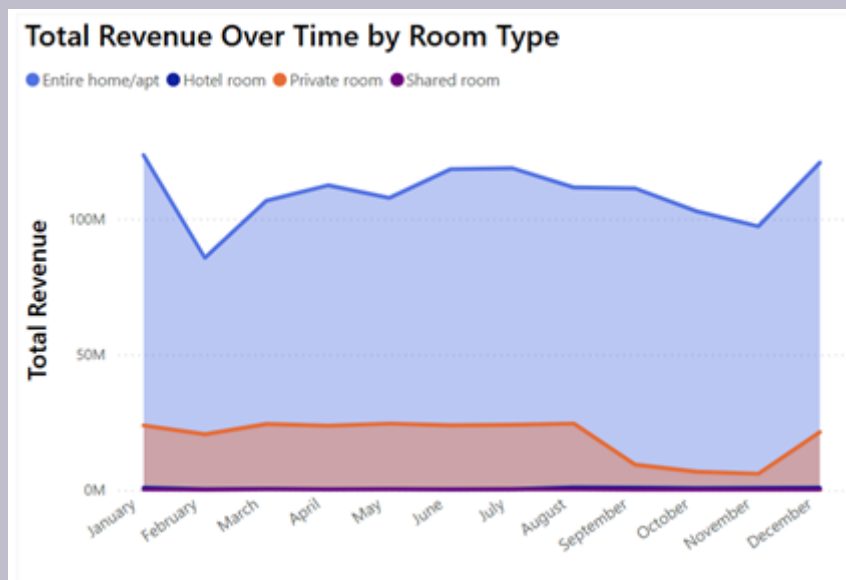


Figure 5.13: Area chart revenue over time by room type. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.13

Area chart titled 'Total Revenue Over Time by Room Type' showing monthly Airbnb revenue by room type. Entire homes/apartments generate the highest revenue throughout the year, peaking mid-year and again in December, while private rooms contribute a smaller share and hotel and shared rooms remain minimal.

Figure 5.14 is another stacked area chart that segments total revenue by region instead of room type and highlights which regions contribute the most to total revenue over time. For example, New South Wales and Victoria are major contributors throughout the year.

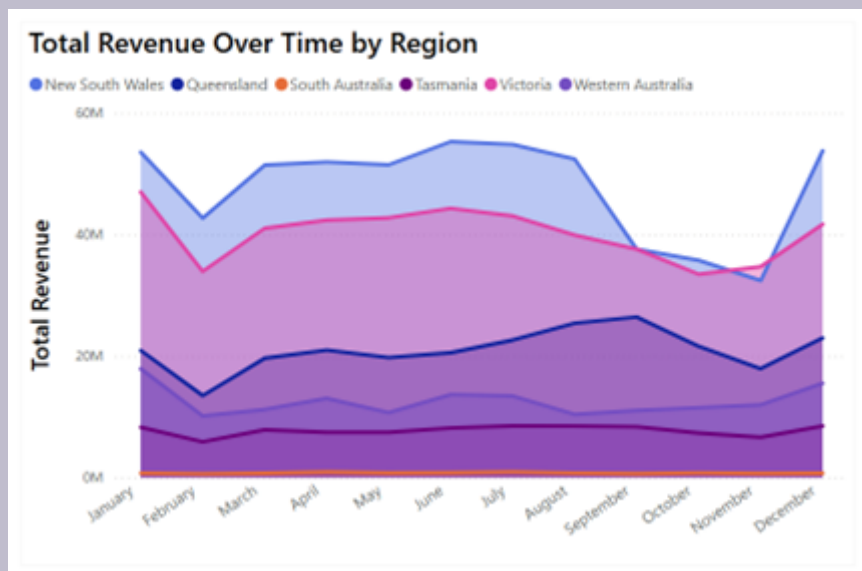


Figure 5.14: Area chart revenue over time by state. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.14

Stacked area chart titled 'Total Revenue Over Time by Region' showing monthly Airbnb revenue for Australian states. New South Wales and Victoria contribute the largest shares throughout the year, peaking mid-year and in December, while Queensland, Western Australia, Tasmania, and South Australia show lower but steady revenue levels.

Key insights

- No matter how you slice it, the “Entire home/apt” category generates the most revenue (**Figure 5.13**).
- New South Wales and Victoria are the largest contributors to total revenue and both show seasonal effects (**Figure 5.14**).
- Both charts reveal that revenue fluctuates throughout the year, with clear peaks and dips. **Figure 5.14** shows that most regions have their highest revenue at the start and end of the year, with a drop in the middle.

Additional questions to explore

These area charts give us a good overview, but they also raise some questions:

1. **Regional differences?** What factors cause the large revenue fluctuations in NSW? Why are NSW and Victoria such strong performers? Can other regions adopt their strategies?
2. **Region-specific room type breakdown:** What is the breakdown of revenue by room type for each region? Why is the “Entire home/apt” category so dominant? Are there opportunities for growth in other room types?
3. **Combine with occupancy:** Do high revenue months also have high occupancy rates or are the prices just higher? A comparison of these area charts with an occupancy heatmap, which we perform in **Section 5.3.5.1**, can answer this question.

5.3.5 Specialised visualisations

Do you want to explore advanced visualisations beyond basic charts? Specialised visualisations include heatmaps, treemaps, geospatial maps, key influencer charts and more. This section introduces several specialised visualisations and shows how each aligns with Andrew Abela’s data visualisation taxonomy (Abela 2008).

5.3.5.1 Revealing distributions using heatmaps (matrix visuals)

Do you need to identify intensity patterns across two dimensions? Heatmaps show how one measure, represented by colour intensity, distributes across two categorical axes like region and month. A heatmap functions as a two-dimensional frequency or intensity visualisation, like a histogram but for two categorical dimensions instead of one. Colour gradients indicate high or low values, which highlights patterns (Abela 2008).

When to use heatmaps? Use heatmaps to observe how a measure changes across two categorical variables.

Examples include an analysis of which website sections receive the most clicks or identify busy times for customer service. For Airbnb data, heatmaps show occupancy rates and revenue variations across regions and months.

How to use heatmaps? Select two categorical variables to compare, such as region and month. Apply a colour gradient to represent the third variable like occupancy rate or revenue. Darker shades usually represent higher values, while lighter shades represent lower values.

Best practices for Heatmaps

- Choose a simple, intuitive colour scheme. Diverging schemes (e.g., light orange to white to dark green) or sequential schemes (e.g., light blue to dark blue) generally provide clarity.
- To avoid confusion, don't use too many colours.
- Provide a clear legend to explain colour meanings.

Heatmap Examples from Airbnb data

Figure 5.15 uses a heatmap to show the occupancy rate for each region on a monthly basis. Darker green shades indicate higher occupancy, white indicates a medium rate and orange indicates lower occupancy. This format allows a user to scan the chart and see how occupancy changes throughout the year. For example, Melbourne shows high occupancy for most of the year, while Barossa Valley shows low occupancy, which warrants further investigation.

Occupancy Rate by Region and Month

Region	January	February	March	April	May	June	July	August	September	October	November	December
Barossa Valley	37.0%	34.7%	35.6%	44.8%	38.1%	41.2%	43.5%	34.0%	30.5%	35.9%	32.5%	34.3%
Barwon	61.5%	53.1%	55.2%	60.1%	59.1%	59.2%	41.7%	32.0%	31.7%	36.6%	39.8%	46.1%
Brisbane	44.7%	42.9%	55.7%	58.1%	56.7%	60.6%	61.4%	60.1%	62.9%	45.7%	38.3%	48.3%
Melbourne	71.5%	68.1%	73.6%	72.6%	71.8%	74.8%	75.4%	74.8%	75.9%	71.7%	68.8%	71.9%
Mid North Coast	47.7%	29.2%	36.0%	40.0%	35.3%	40.1%	41.9%	40.4%	45.3%	41.0%	29.8%	45.9%
Mornington Peninsula	46.0%	30.3%	35.3%	39.1%	37.9%	44.2%	47.6%	47.3%	44.8%	29.0%	31.6%	41.5%
Northern Rivers	49.1%	43.9%	46.1%	50.1%	48.3%	50.0%	38.2%	31.4%	32.2%	36.5%	36.3%	45.0%
Sunshine Coast	49.0%	35.5%	47.6%	50.8%	47.2%	51.8%	54.9%	58.2%	61.2%	48.9%	41.5%	54.2%
Sydney	44.0%	40.0%	50.6%	50.1%	48.5%	53.8%	54.5%	54.3%	53.2%	46.2%	44.4%	49.0%
Tasmania	48.3%	38.9%	48.5%	47.3%	46.3%	52.3%	52.7%	52.7%	51.6%	44.6%	41.2%	51.4%
Western Australia	56.9%	46.5%	46.3%	51.2%	47.7%	51.5%	54.0%	43.1%	44.4%	45.6%	44.1%	51.9%

Figure 5.15: Occupancy rate by region and month (heatmap/matrix). Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.15

Heatmap table titled 'Occupancy Rate by Region and Month' showing monthly Airbnb occupancy percentages across Australian regions. Melbourne and Barwon display consistently high occupancy, especially mid-year, while regions such as Barossa Valley and Mid North Coast show lower and more seasonal rates.

Revenue by Region and Month

Region	January	February	March	April	May	June	July	August	September	October	November	December
Barossa Valley	\$13,707,538	\$10,004,556	\$13,749,164	\$14,408,252	\$14,089,808	\$15,362,728	\$16,221,916	\$15,818,634	\$15,644,504	\$12,473,518	\$10,586,434	\$14,044,094
Barwon	\$711,471	\$590,800	\$700,616	\$894,956	\$741,347	\$789,131	\$925,289	\$716,552	\$611,471	\$753,957	\$651,794	\$706,859
Brisbane	\$19,192,621	\$14,842,163	\$16,816,860	\$17,456,408	\$17,667,204	\$17,229,979	\$13,654,732	\$10,817,936	\$10,347,776	\$11,798,008	\$12,602,877	\$14,812,187
Melbourne	\$4,491,587	\$3,776,746	\$5,730,946	\$5,928,471	\$5,924,986	\$6,010,140	\$6,308,977	\$6,181,917	\$5,974,403	\$4,332,345	\$3,257,731	\$4,459,211
Mid North Coast	\$10,872,632	\$9,400,893	\$11,528,181	\$11,017,542	\$11,209,036	\$11,391,669	\$11,920,477	\$11,773,169	\$11,131,480	\$10,714,073	\$10,037,886	\$11,019,962
Mornington Peninsula	\$10,616,676	\$5,798,297	\$7,762,262	\$8,293,947	\$7,567,741	\$8,534,394	\$9,111,727	\$8,859,709	\$9,686,620	\$9,153,185	\$6,627,646	\$10,463,606
Northern Rivers	\$16,945,622	\$9,672,025	\$12,675,435	\$13,884,006	\$13,864,193	\$15,684,226	\$17,480,075	\$17,319,070	\$16,115,252	\$10,944,307	\$12,057,584	\$15,875,720
Sunshine Coast	\$16,935,154	\$13,665,534	\$15,880,996	\$17,112,269	\$16,850,262	\$17,961,256	\$15,818,540	\$13,446,969	\$12,423,719	\$14,637,623	\$14,455,116	\$17,145,999
Sydney	\$16,403,419	\$9,710,974	\$13,912,294	\$15,034,729	\$13,826,605	\$14,532,229	\$16,303,649	\$19,203,407	\$20,438,096	\$17,256,161	\$14,659,643	\$18,504,441
Tasmania	\$26,011,598	\$23,226,198	\$27,708,430	\$26,540,383	\$27,052,145	\$28,760,207	\$28,504,508	\$8,485,874	\$15,480,611	\$11,968,007	\$11,334,032	\$26,178,867
Western Australia	\$8,264,603.32	\$5,877,106.48	\$7,862,701.16	\$7,453,989	\$7,449,378	\$8,166,327	\$8,504,508	\$8,485,874	\$8,354,407.26	\$7,323,919.56	\$6,631,757.02	\$8,496,579.96
Western Australia	\$17,911,752	\$10,125,848	\$11,195,975	\$13,019,327	\$10,668,615	\$13,639,034	\$13,419,779	\$10,381,688	\$11,028,978	\$11,498,401	\$11,955,170	\$15,514,750

Figure 5.16: Total revenue by region and month (heatmap/matrix). Source: <http://insideairbnb.com>

Alternative Text for Figure 5.16

Heatmap table titled 'Revenue by Region and Month' displaying monthly Airbnb revenue amounts for Australian regions. Sydney shows the highest revenue across all months, followed by Melbourne and Sunshine Coast, with darker green cells indicating higher revenue.

and red cells indicating lower revenue, and clear seasonal increases toward mid-year and December.

Figure 5.16 is a heatmap similar to **Figure 5.15**, but it shows total revenue. Darker blue shades indicate higher revenue, white indicates a medium level and lighter blues indicate lower revenue. This visual allows an analyst to identify when and where the most revenue is generated. For example, the chart shows that New South Wales has significant revenue spikes in January and October. A comparison with the occupancy heatmap **Figure 5.15** can determine if high revenue months are a result of high occupancy, high prices or both.

Key insights

- Melbourne shows a high occupancy rate for most of the year (**Figure 5.15**).
- Many regions, such as Brisbane and Sunshine Coast, show increased activity from September to January (**Figure 5.15**).
- Occupancy rates tend to decline around July and August for most regions (**Figure 5.15**).
- NSW shows very high revenue, with peaks in January and October, though its occupancy is not always the highest (**Figure 5.16**).
- Melbourne's revenue is more consistent throughout the year and does not show the same extreme peaks (**Figure 5.16**).
- Each region has a unique revenue pattern (**Figure 5.16**).

Additional questions to explore

The **Figure 5.15** heatmap provides a good overview of high or low occupancy rates. The next step is to determine the cause of these patterns.

1. **Seasonal strategies:** Regions with large seasonal fluctuations may require tailored marketing and price strategies, such as discounts during slow months.
2. **Melbourne's secret:** What factors contribute to Melbourne's high performance? Can other regions adopt its successful strategies?
3. **Compare with revenue:** A comparison of this occupancy heatmap with the revenue

heatmap (**Figure 5.16**) is necessary. Do high occupancy months also produce high revenue? Or could some regions with high occupancy increase their revenue with different price points?

The **Figure 5.16** heatmap identifies the *when* and *where* of revenue generation. Further analysis should focus on the *why*.

1. **NSW spikes:** What causes the revenue peaks in January and October in NSW? Are they a result of special events or school holidays? Can this success be replicated in other months or regions?
2. **Melbourne's consistency:** What factors allow Melbourne to maintain a consistent revenue stream throughout the year? Is its market less dependent on seasonal tourism?
3. **WA's unique pattern:** Why does Western Australia's revenue pattern differ from the other states.

5.3.5.2 Reveal geographical distribution with geospatial maps

Does location play a key role in your insights? Following Andrew Abela's data visualisation taxonomy (Abela 2008), geospatial maps display how values spread across geographic locations. Geospatial maps offer a geographical visualisation of data, similar to heatmaps but placed on geographic space. This visualisation type illustrates location-based information and identifies spatial trends and patterns. This section demonstrates geospatial maps with Australian Airbnb data to illustrate geographical trends and insights.

When to use geospatial maps? Use geospatial maps whenever location has a significant effect on your data. For example, a map can show Airbnb reservation density across cities or the average reservation price in various neighbourhoods.

How to use geospatial maps? Modern BI tools provide mapping tools. Plot data points on the map and use colour or size to represent different numerical values.

Best practices for Geospatial Maps

- Select an appropriate map for your data scope. For example, avoid a global map for city-specific data.
- Choose colours that are distinct and suitable for people with colour blindness.
- Include a clear legend that explains the meaning of colours or sizes.

Geospatial Map Examples from Airbnb data

Figure 5.17 displays a basic map of Australia with bubbles for each region. The size of each bubble corresponds to the total revenue for that region. A user can hover over any bubble to see the exact region name and revenue. The map provides a quick overview of the geographic sources of Airbnb revenue and shows that the largest revenue sources cluster around coastal regions and major cities.

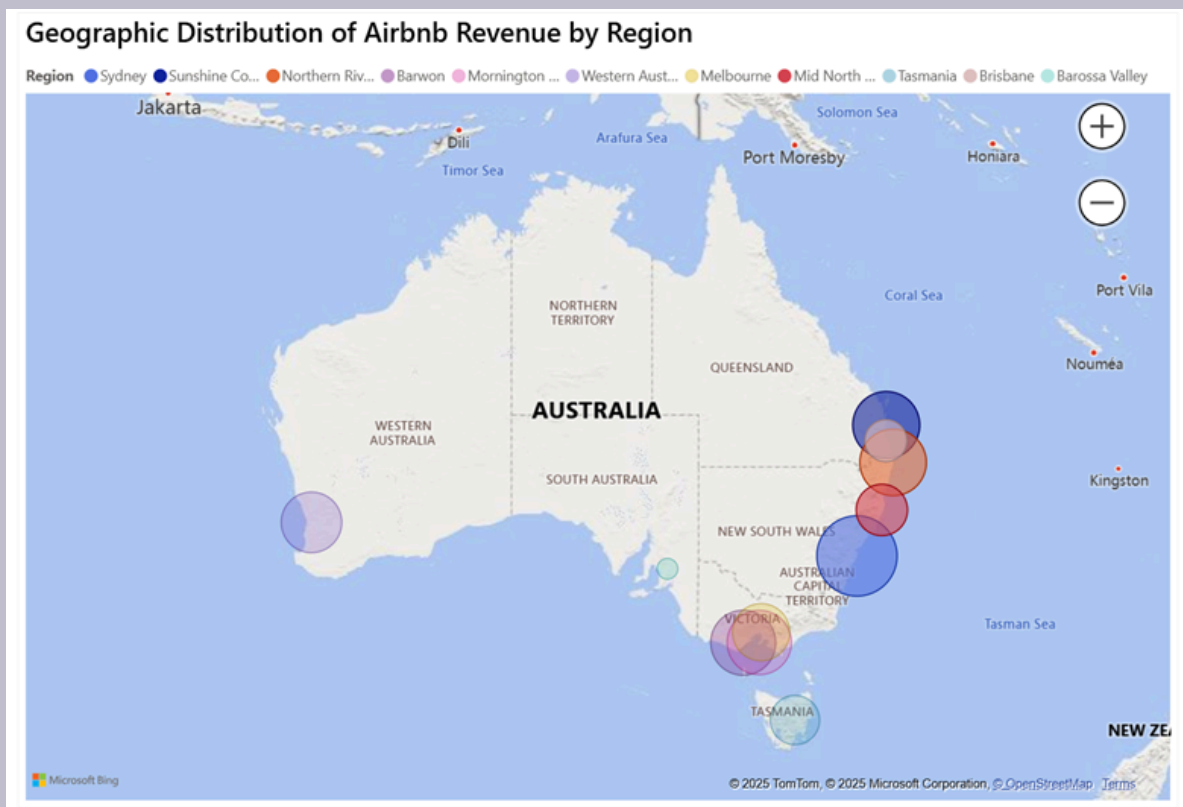


Figure 5.17: Geospatial map showing geographic distribution of Airbnb revenue. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.17

Map titled 'Geographic Distribution of Airbnb Revenue by Region' showing Australia with colored circles representing regional Airbnb revenue. Larger circles appear along the east coast especially around Sydney, Brisbane, and Melbourne indicating higher revenue, while smaller circles in Western Australia, Tasmania, and inland regions indicate lower revenue.

Figure 5.18 is a similar map where bubble size also represents total revenue. This version, however, includes a custom tooltip with a line chart that shows revenue over time and the occupancy rate. When a user hovers over a bubble, a small line chart appears for that region with its occupancy rate. This enhancement allows an analyst to see the revenue source, its trend over time and its occupancy rate in a single view. For example, the large bubble for the Sunshine Coast indicates high revenue and the tooltip reveals its monthly revenue trend and occupancy rate.

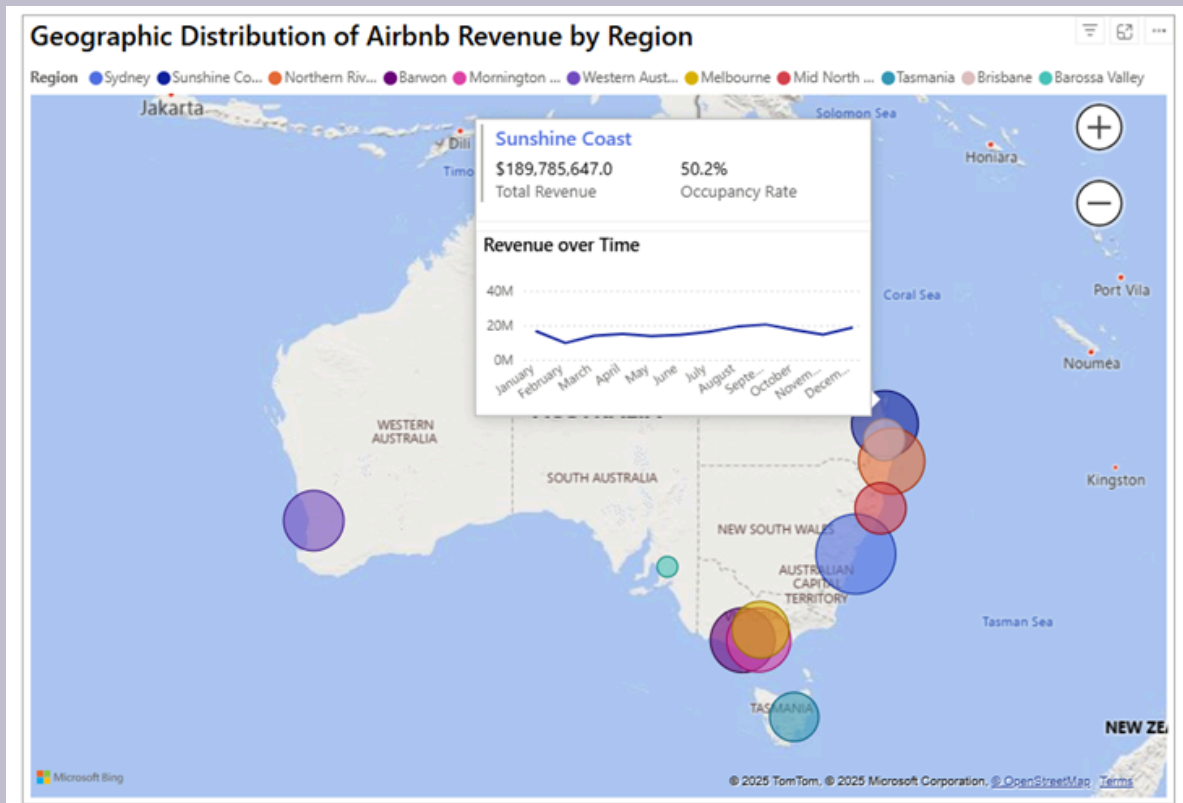


Figure 5.18: Geospatial map showing geographic distribution of Airbnb revenue (with a visual tooltip). Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.18

Interactive map titled 'Geographic Distribution of Airbnb Revenue by Region' showing Australia with colored circles sized by regional Airbnb revenue. A tooltip for the Sunshine Coast displays total revenue of about \$190M, an occupancy rate of 50%, and a small line chart of monthly revenue trends, while larger circles cluster along the east coast around Sydney, Brisbane, and Melbourne.

Key insights

- High revenue regions are concentrated along the coast, which aligns with popular tourist destinations.
- Sydney and Melbourne are major revenue drivers, as indicated by their large bubble sizes.
- The Sunshine Coast also stands out as a high revenue region.

Additional questions to explore

1. **Drill down for more granularity:** Zoom in on specific areas to understand the drivers of high revenue. For example, is the high revenue on the Sunshine Coast a result of a few expensive listings or many moderately priced ones?
2. **Compare to other data:** Compare the revenue map to other data sources, such as maps of population density or tourist attractions, to identify potential opportunities in new areas.
3. **Time series:** Use the custom tooltip on the map to analyse trends in revenue and occupancy over time.

5.3.5.3 Display composition of components with treemaps

Do you need to visualise multiple levels of category breakdown? Following Andrew Abela's data visualisation taxonomy (Abela 2008), treemaps illustrate hierarchical data, categories and their respective subcategories. Tree maps offer a visual representation of a total divided into parts, with each part further segmented into smaller components.

When to use treemaps? Use tree maps to show contributions of different elements within multiple category levels. An example is a visual of the contribution of product categories and their subcategories to overall sales. For our Airbnb data, a treemap could illustrate how each region contributes to total revenue and further, how room types within those regions contribute as individuals. Treemaps provide both high-level overviews and detailed insights.

How to use treemaps? Show each category as a rectangle. The rectangle size indicates its value, with larger rectangles representing higher values. Smaller rectangles for subcategories appear nested within larger rectangles for main categories. This structure shows hierarchical relationships in a visual format.

Best practices for Treemaps

- Use clear labels and colours to aid the differentiation of categories.
- Do not overcrowd the treemap with too many levels or it gets hard to read.
- Ensure the rectangle sizes reflect the values they represent with accuracy.

Treemap Examples from Airbnb data

A treemap is a good alternative to a pie chart when a dataset contains many categories. The pie chart in **Figure 5.10**, for example, becomes cluttered with numerous slices, which makes a comparison between regions difficult.

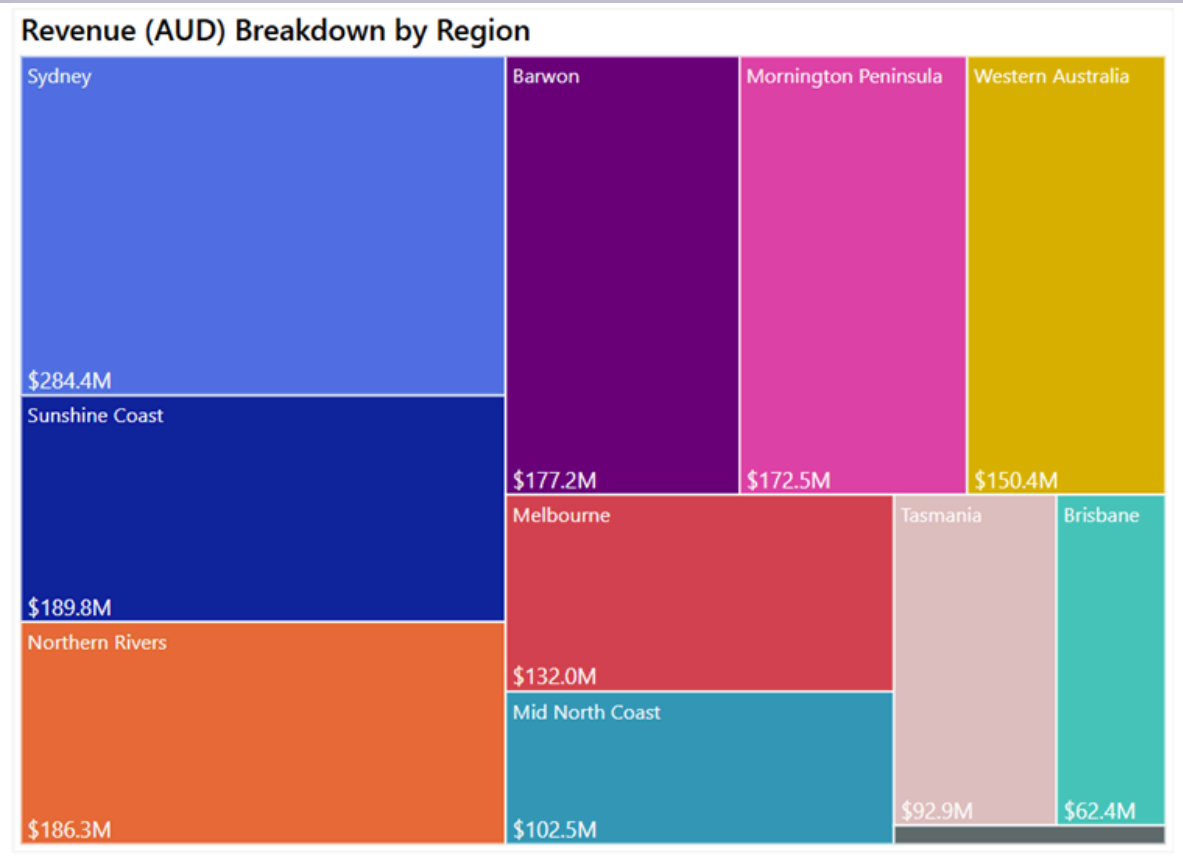


Figure 5.19: Overall revenue distribution by region (using treemaps instead of pie charts).
Source: <http://insideairbnb.com>

Alternative Text for Figure 5.19

Treemap titled ‘Revenue (AUD) Breakdown by Region’ showing Sydney as the largest contributor at about \$284M, followed by Sunshine Coast (\$190M), Northern Rivers (\$186M), Barwon (\$177M), Mornington Peninsula (\$173M), Western Australia (\$150M), and smaller contributions from Melbourne, Mid North Coast, Tasmania, and Brisbane.

Figure 5.19 shows total revenue for each region in a treemap. The rectangle size corresponds to its value; Mornington Peninsula, Victoria, is the largest rectangle and thus the top revenue performer. This format provides a simple overview of top performers and is a good alternative to a pie chart with many categories.

Figure 5.20 gets more detailed. The treemap provides a more detailed perspective with a nested treemap. This visual reveals how different room types perform within each state and shows a two-level hierarchy. Level one, the outer rectangles, represents the room type. Level two, the inner rectangles, represents the states within each room type. The chart first breaks

down revenue by room type and highlights “Entire home/apt” as the primary revenue source. Within that category, states like Victoria and NSW stand out.

Hierarchical treemaps like **Figure 5.20** combine multiple treemaps into one visual. They function similarly to four separate pie charts, one for each room type, which allows for performance comparisons across both room types and states. The visual also highlights that shared rooms generate the least revenue. This breakdown could extend to cities or individual listings to provide more granular insights. This nested treemap illustrates revenue composition and identifies areas that require further attention.

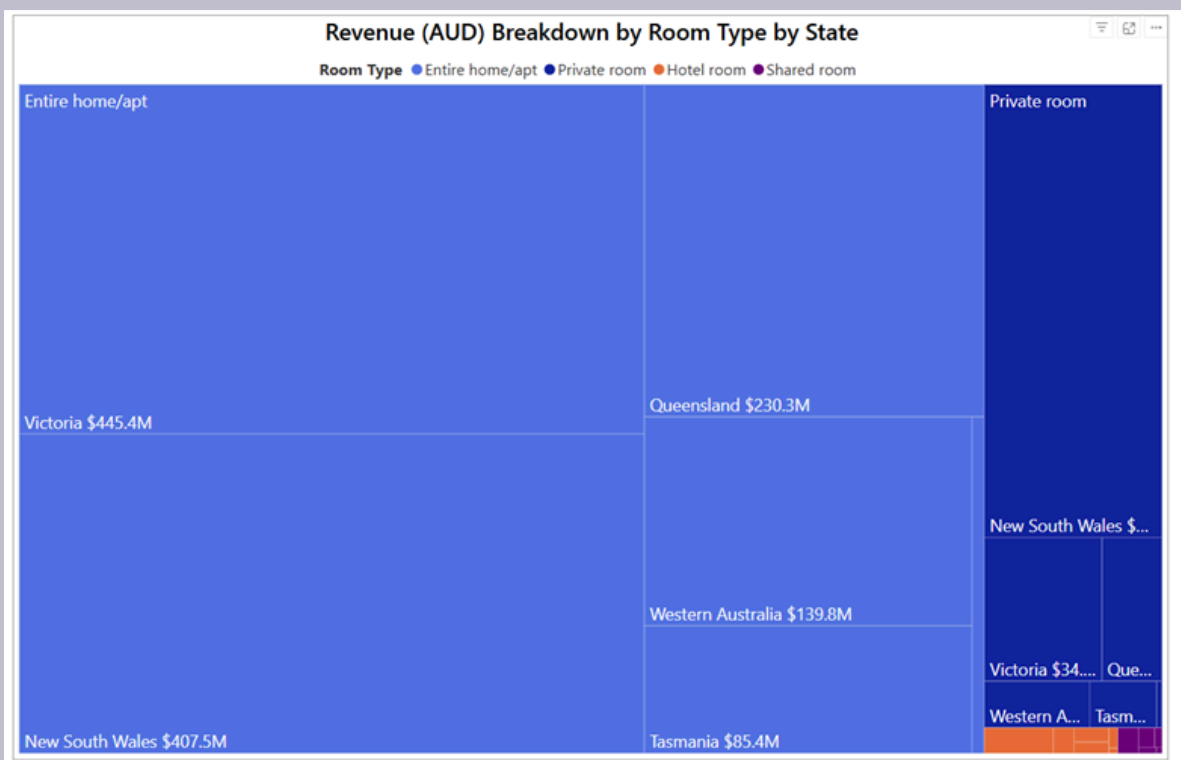


Figure 5.20: Overall revenue distribution by room type by state (using treemaps instead of pie charts). Source: <http://insideairbnb.com>

Alternative Text for Figure 5.20

Treemap titled ‘Revenue (AUD) Breakdown by Room Type by State’ showing that entire homes/apartments generate the majority of Airbnb revenue across all states, with Victoria and New South Wales contributing the largest amounts. Smaller sections represent private rooms, hotel rooms, and shared rooms, which account for a much smaller share of total revenue.

Key insights

- Mornington Peninsula, Northern Rivers and Barwon are among the regions with the highest revenue (**Figure 5.19**).
- “Entire home/apt” is the room type with the largest revenue. Victoria and NSW stand out in this category (**Figure 5.20**).
- An analysis of **Figure 5.19** shows which regions are strong overall and **Figure 5.20** explains which room types drive that revenue.

Additional questions to explore

These treemaps give us a great overview of revenue distribution. Here’s what we could dig into next:

1. **Entire home/apt focus:** Since the “Entire home/apt” category is so important, a closer analysis is required. Which specific regions within Victoria and NSW drive that revenue? We can follow with a drill down to provide more granular detail.
2. **Low-performing regions:** We can investigate further factors that are behind the low revenue in regions like Barossa Valley.
3. **Consider other data:** How does this revenue breakdown compare to occupancy rates, average prices and the number of listings in each region and for each room type?

5.3.5.4 Reveal potential relationships with key influencers

Do you need to determine which factors most influence your target metric? The Key Influencers visual in Power BI uses machine learning (ML) to analyse data and identify factors associated with a particular outcome. Following the Abela data visualisation taxonomy (Abela 2008), visuals that show relationships or correlations help users understand associations between variables. The Key Influencers visual highlights variables, such as room type, region or month, that have the strongest association with your target metric. It reveals important correlations and potential factors that affect your outcomes.

Think of the Key Influencers visual as a built-in data analyst. It identifies influential factors that affect your chosen metric, ranks them in order of importance and helps you decide where to focus your analysis.

When to use Key Influencers? Use Key Influencers to examine a specific key metric such as occupancy rate and potential variables that influence it. For example, you might want to explore which factors affect Airbnb occupancy rates including region, room type and price.

How to use Key Influencers? In Power BI, add the Key Influencers visual and select the metric you wish to analyse, for example Overall Occupancy Rate. Next, drag potential influential variables like region

and room type into the Explain by section. Power BI will perform an automated analysis and identify the variables with the greatest impact on your selected metric and will also show whether their influence increases or decreases the metric.

Best practices for Key Influencers

- Do not place too many variables into the Explain by section at once. Start with a manageable number of potential influencers.
- Ensure you understand the calculation of the metric you analyse.
- Remember, the visual identifies correlations not causal relationships. To identify a key influencer does not prove direct causation.

Key Influencer Examples from Airbnb data

The Key Influencers visual in **Figure 5.21** and **Figure 5.22** explain the Airbnb occupancy rates. The visual has two main parts; the left side shows factors that increase the occupancy rate and the right side shows factors that decrease it. An analyst can select each factor to see more detail. This provides a way to explore the data and identify potential factors that influence occupancy.

The chart identifies “Room Type is Shared room” as the top factor. This indicates that, on average, shared rooms have an occupancy rate 0.24 points higher than other room types. The visual also lists other factors, such as the months of May and April, which suggests occupancy rates are higher during that period.

The right side of the visual shows that “Month is November” is the top factor to decrease occupancy, with an average negative impact of 0.13. Other factors that lower the rate include “Month is October” and “Room Type is Entire home/apt”.

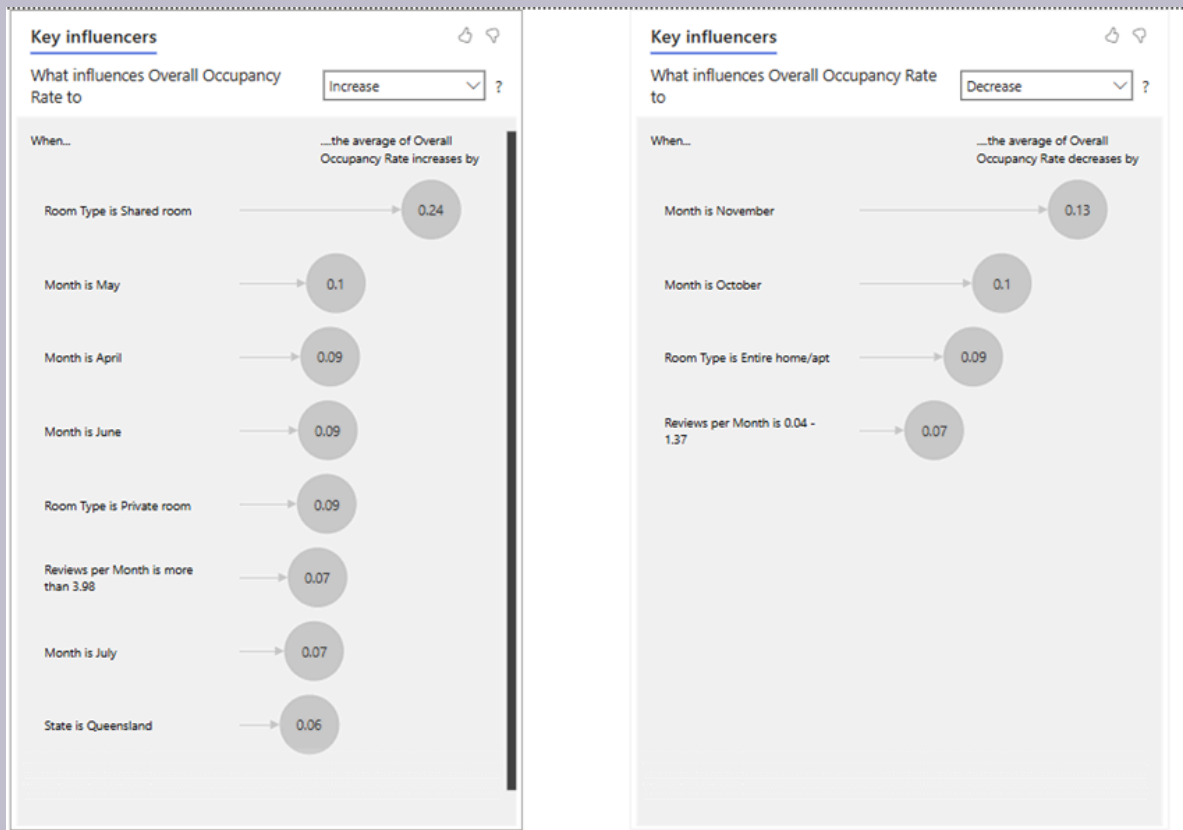


Figure 5.21: Using key influencers to investigate factors that influence occupancy rate. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.21

Two 'Key influencers' panels showing factors that increase or decrease overall Airbnb occupancy rate.

On the left, occupancy increases most when the room type is a shared room, followed by months such as May, April, and June, and higher review frequency.

On the right, occupancy decreases most in November and October, for entire home/apartment listings, and when reviews per month are low.

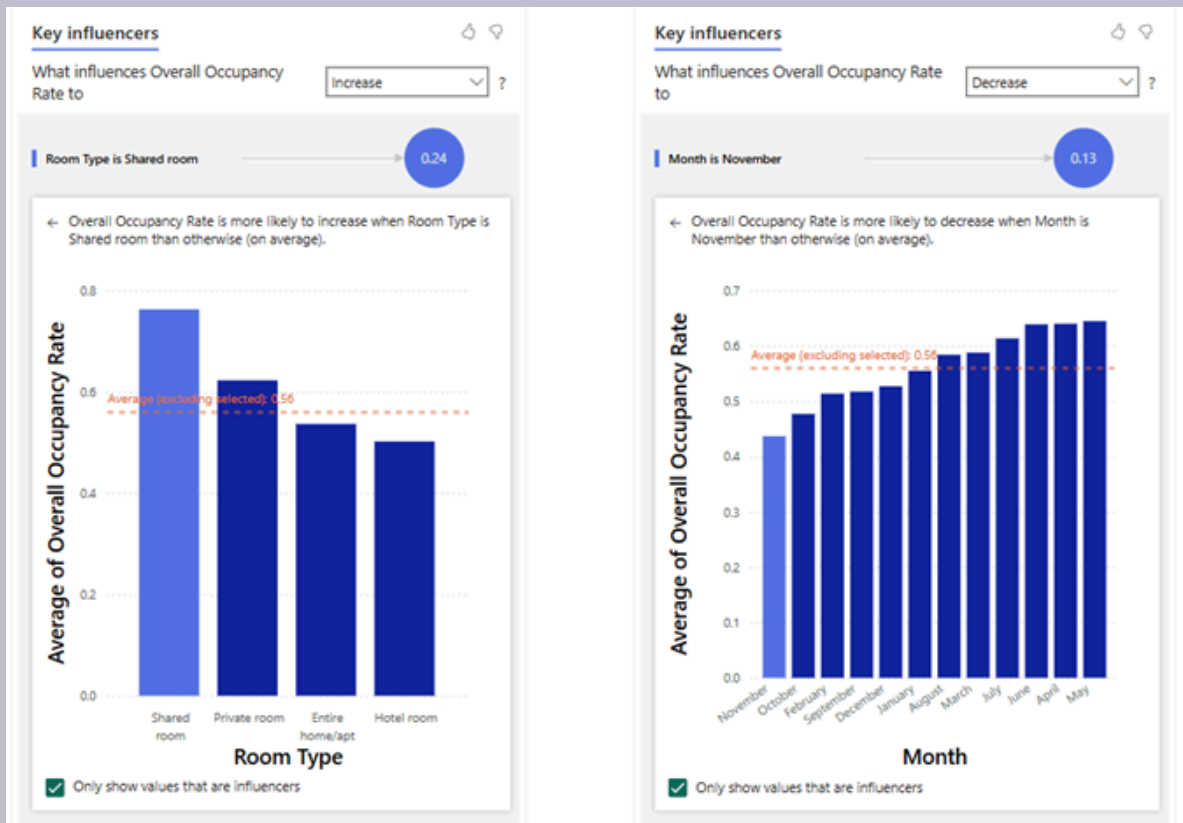


Figure 5.22: Using key influencers to investigate factors that influence occupancy rate. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.22

Two 'Key influencers' visuals explaining factors affecting overall Airbnb occupancy rate. The left chart shows occupancy is higher for shared rooms compared with private rooms, entire homes/apartments, and hotel rooms. The right chart shows occupancy is lower in November than in other months, with monthly averages rising from late-year lows to higher levels earlier in the year.

Key insights

- Room type is a key influencer; shared rooms tend to increase occupancy.

- Certain months appear to correlate with occupancy changes. May, April and June correlate with higher occupancy, while November and October associate with lower occupancy.
- The consistent appearance of reviews per month in both figures suggests a possible influence on reservations.

Additional questions to explore

This Key Influencers visual gives us some good starting points. Now we can investigate potential influences in more detail.

1. **Room types:** Why do shared rooms have higher occupancy? Is the cause related to price or do they attract a different type of traveller?
2. **Confirm seasonality:** The visual confirms the seasonal patterns suggested in other charts. Further analysis should focus on why specific months have high or low occupancy.
3. **Reviews per month:** What is it about this metric? May need to explore further to understand its specific impact on occupancy.

5.4 Crafting narratives with visualisations

In BI, we use data storytelling to turn raw data into narratives. When combined with suitable visuals, data storytelling helps people understand complex data and allows non experts to translate insights into decisions. This section explains how to create data-driven stories.

5.4.1 The power of a data story

When supported by strong narratives, data visualisations are not just for simple analysis, they aim to inspire action. For example, Florence Nightingale famously created a rose diagram to investigate causes of deaths in military hospitals. Her diagrams highlighted the negative effect of sanitary conditions. This resulted in public health reforms and became an early example of an impactful visual story (Brasseur, 2005). In this section, you will learn techniques to create clear, understandable data stories.

5.4.2 Techniques for effective storytelling

Data storytelling extends beyond the simple presentation of charts. A good data story communicates insights audiences can understand and act on easily. This section demonstrates how to craft data-driven stories that combine insights into messages that connect with your audience and inspire action.

The following framework, inspired by Cole Nussbaumer Knaflic's *Storytelling with Data* (Knaflic 2015,

2019), uses Airbnb data to demonstrate these techniques. Think of this approach as a story arc that guides your narrative.

5.4.2.1 Building our story with the narrative arc framework, using Barossa Valley data

The Barossa Valley story

Analysis of Airbnb performance across Australian regions can identify growth opportunities. The bar chart of total revenue by region (**Figure 5.3**) provides the initial context and shows that Sydney generates the most revenue while Barossa Valley generates the least. The low number of listings in Barossa Valley results in this underperformance. A simple forecast shows that if current trends continue, its revenue will remain low relative to other regions. The goal is to understand the cause of this underperformance.

A look at occupancy rates in **Figure 5.12**, however, reveals a key insight. Barossa Valley's occupancy rate stays stable throughout the year, and the revenue line chart (**Figure 5.7**) shows a distinct revenue jump in October. This evidence suggests that while visitor demand exists, particularly in October, the region fails to capture potential revenue due to an insufficient number of listings.

This analysis reveals a potential opportunity. The stable occupancy rate indicates consistent demand in Barossa Valley. The region needs more listings, with a focus on availability around the October peak. Potential solutions include host incentives, partnerships with local property managers or targeted marketing campaigns.

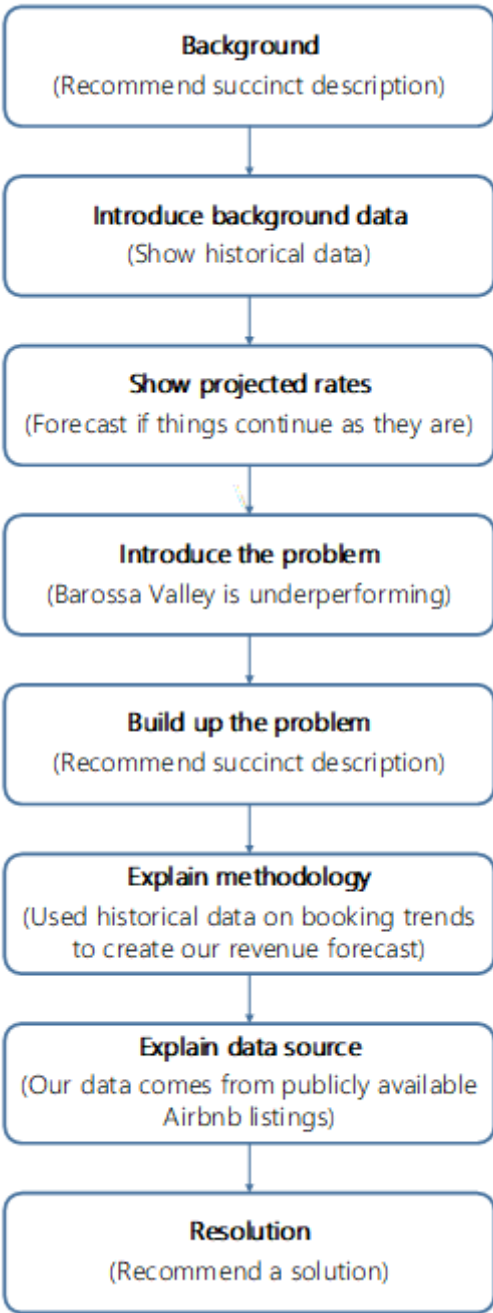


Figure 5.23: Framework for building our story

Alternative Text for Figure 5.23

Vertical flowchart outlining a data storytelling structure: background, introduction of historical data, projected rates, identification and buildup of a problem (Barossa Valley

underperformance), explanation of methodology and data sources, and a final resolution proposing a solution.

How the story maps to our narrative arc framework

1. **Background and introduce background data.** The first paragraph establishes the business objective and uses **Figure 5.3** to present the historical context.
2. **Show projected rates and introduce the problem.** The first paragraph includes the forecast and identifies Barossa Valley's underperformance as the story's central problem.
3. **Build up the problem (climax).** The second paragraph develops the story's climax. It introduces a conflicting piece of data (stable occupancy from **Figure 5.12**) to reveal the true, nuanced issue – a lack of supply, not a lack of demand.
4. **Explain methodology and data sources used.** In a formal report, you would place these details in a footnote or appendix to avoid a disruption of the narrative flow. (*Note:* Our data comes from publicly available [Airbnb listings](#), provided quarterly for the past 12 months.)
5. **Resolution (recommend a solution).** The final paragraph provides the resolution. It states the business opportunity clearly and recommends specific, actionable solutions.

5.4.2.2. Data story challenge: a re-creation of Nightingale's rose diagram

In this section, we examine Florence Nightingale's original rose diagram (**Figure 5.24**) and show how to create a similar data story with modern visuals. We will use stacked column and line charts and apply the narrative arc framework from [Section 5.4.2.1](#).

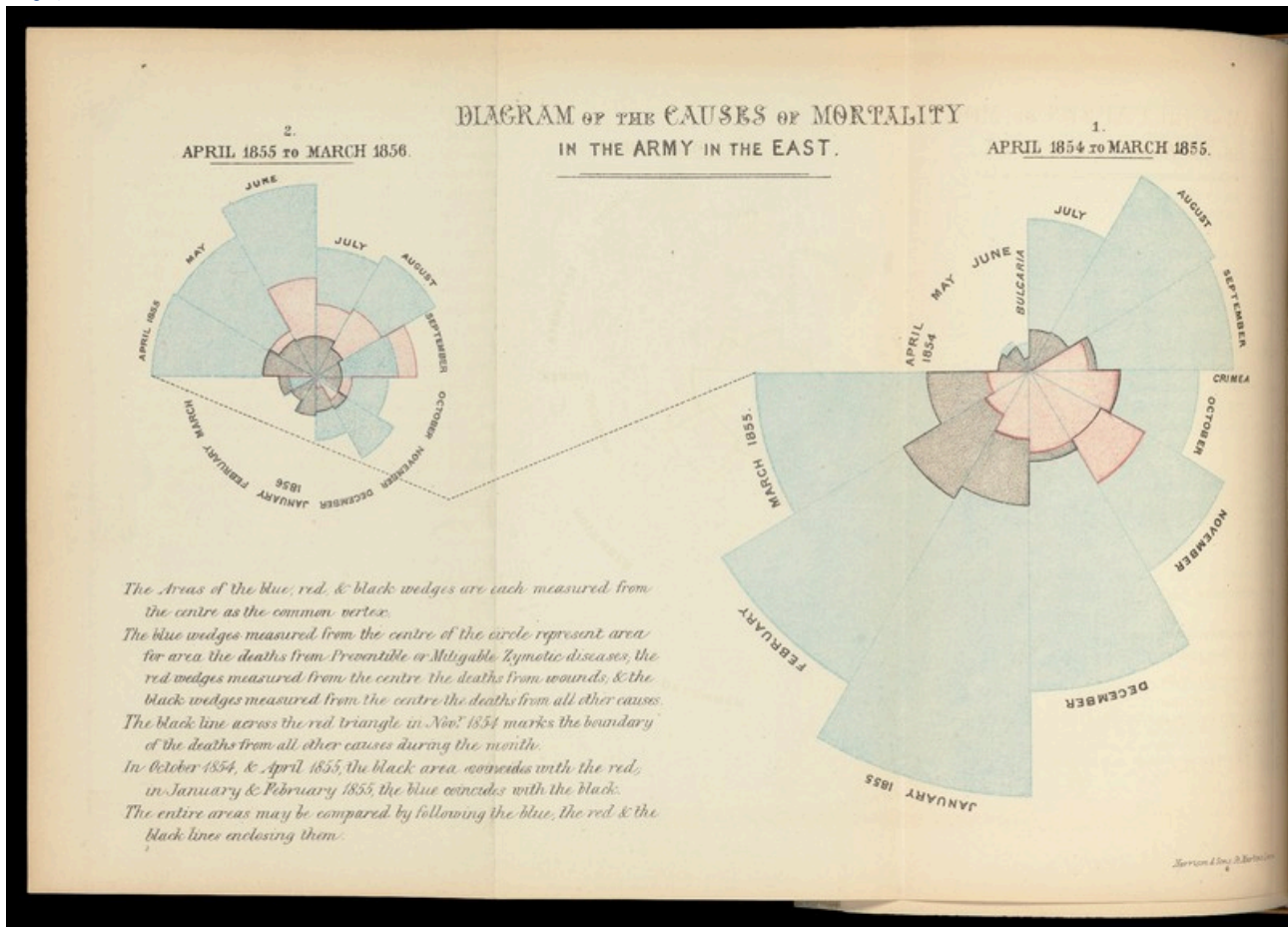


Figure 5.24: Florence Nightingale's original rose diagram (Nightingale, 1858). This work is out of copyright and in the public domain. Image via the [Wellcome Collection, Library](#)

Alternative Text for Figure 5.24

Florence Nightingale's polar area diagram titled 'Diagram of the Causes of Mortality in the Army in the East,' showing monthly deaths during the Crimean War. Radial segments represent causes of death, with larger blue areas indicating deaths from preventable disease compared to smaller red and black areas for wounds and other causes, highlighting seasonal patterns and excess mortality.

Building Florence Nightingale's story: Using modern visuals and the narrative arc framework

During the Crimean War (1854–1856), Florence Nightingale observed that disease, not combat, was the primary cause of death in military hospitals. The modern charts in **Figure 5.25** and **Figure 5.26** use Nightingale's data. The stacked column chart breaks down the monthly percentage of deaths by three causes: preventable disease (zymotic), wounds and other causes.

The chart for the period from April 1854 highlights the scale of the problem. The tall blue sections, which

represent deaths from disease, show that disease was widespread and unchecked. In the first year, disease caused the majority of deaths and took far more lives than combat. Over the course of the war, disease accounted for 81.57% of all deaths, while wounds caused only 10.95% (**Figure 5.25**). These numbers underscored the urgent need for the sanitary reforms Nightingale advocated.

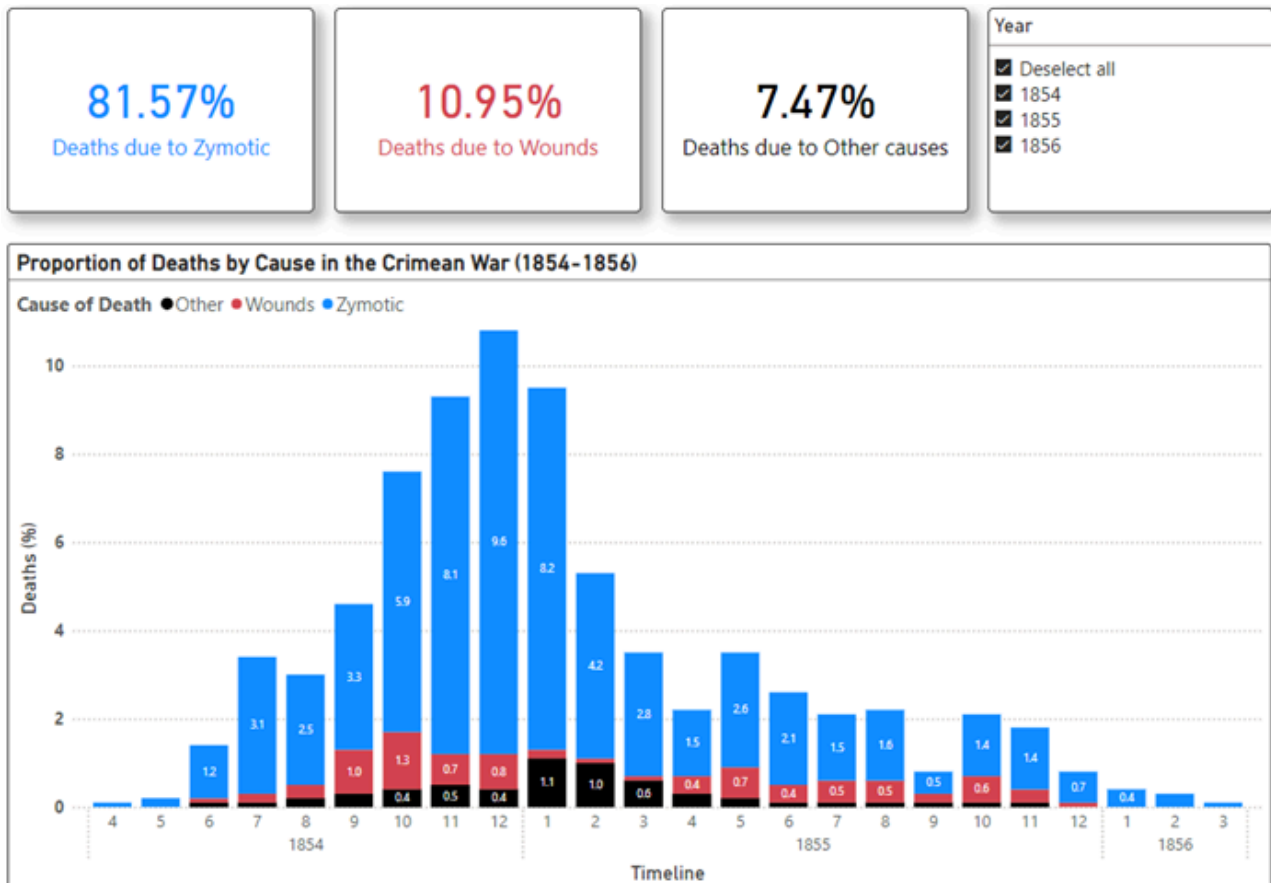


Figure 5.25: A re-creation of Florence Nightingale's data with contemporary visuals and a corresponding story. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.25

Dashboard showing causes of death in the Crimean War (1854–1856). Summary cards indicate that 81.6% of deaths were due to zymotic disease, 11.0% to wounds, and 7.5% to other causes. A stacked bar chart below displays monthly proportions of deaths over time, with blue bars (zymotic disease) dominating most months, especially in late 1854 and early 1855.

The line chart in **Figure 5.26** illustrates the resolution. It shows a sharp drop in the blue line, which represents disease-related deaths, after March 1855. This decline followed the arrival of the Sanitary Commission, sent by the British government in response to Nightingale's work (Fee & Garafalo, 2010;

National Army Museum, n.d.). The commission improved sanitation, ventilation and overcrowding. The chart demonstrates how these changes saved soldiers' lives and how we can use data stories to inspire change.

Trends in Causes of Death During the Crimean War (1854-1856)

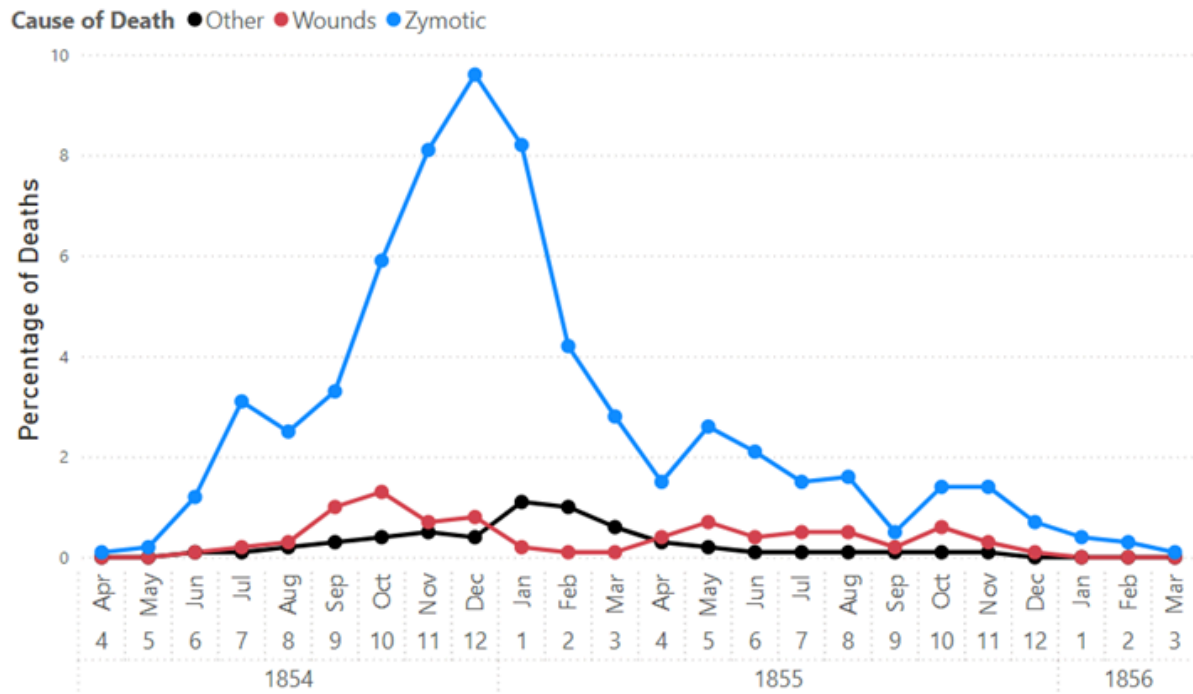


Figure 5.26: From proportions to trends, a line chart that shows trends in the causes of death between 1854 and 1856. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.26

Line chart titled 'Trends in Causes of Death During the Crimean War (1854-1856)' showing monthly percentages of deaths from zymotic disease, wounds, and other causes. Zymotic deaths peak sharply in late 1854 and early 1855 before declining, while deaths from wounds and other causes remain comparatively low and steady throughout the period.

How the Florence Nightingale data story aligns with the narrative arc framework?

1. **Background and introduce background data.** The first paragraph establishes the historical context of the Crimean War and introduces the modern data visuals that tell the story.
2. **Introduce the problem and build up the problem (climax).** The second paragraph defines the story's central problem and builds to its climax. It uses the data to show that disease was the primary killer, culminating in the statistic that disease accounted for over 81% of all deaths.

3. **Explain methodology and data source.** The details of the methodology and data source provide credibility. In a formal report, you would place them in a footnote or appendix to maintain the narrative flow. We use percentages to show the share of deaths from each cause, similar to Nightingale's approach. We use Death Rate (Scaled) on the y-axis, allowing us to compare segment areas instead of just height.
4. **Resolution (recommend a solution).** The third paragraph provides the resolution. It uses the line chart to show the direct, positive impact of the Sanitary Commission's intervention and delivers a clear conclusion to the narrative.

Why did we use these charts? We used two charts to present Nightingale's data, a stacked column chart and a line chart. Each chart has a different purpose and aligns with Andrew Abela's data visualisation taxonomy (Abela 2008).

- **Stacked column chart for composition:** This chart shows proportion. Abela's taxonomy categorises stacked column charts under composition, since they show parts of a whole at specific points in time. Each column represents a month, and the coloured sections show the percentage of deaths from disease, wounds or other causes. The chart is great for a quick comparison of categories and for quick look at the large blue section for disease, especially at the start of the war. Nightingale used this chart type to show that disease was the real issue.
- **Line chart for comparison over time:** This chart shows trends. Abela's taxonomy positions line charts in the comparison category, for a comparison of data points over time. Each line represents a cause of death and shows changes on a monthly basis. The blue line for disease drops after March 1855, an indication of the positive effect of the sanitary reforms. The line chart shows how the sanitation issues improved speedily after the reforms.

The stacked column chart explains *what happened*: disease was the largest killer. The line chart shows *when and how fast* the situation changed. Using both charts provides a complete and powerful story.

5.5 Effective visualisation design and best practices

Data visualisation isn't just about choosing a chart type; it is also about good design. This section covers key principles to make your visuals clear and easy to understand.

5.5.1 Best practice and rules for creating meaningful visualisations

Before we discuss the cognitive and perceptual principles of good design, let's review the basic rules that every effective chart must follow. At a minimum, the visuals must be clear, accurate and have a clear purpose. You must start with a clear business question and then choose the simplest chart that answers it. Every element on the page, from axes and labels to colours and annotations, should have a clear reason to exist.

Applying basic guidelines, such as starting bar charts at zero, limiting colours and avoiding unnecessary 3D effects, reduces cognitive load and helps your audience see patterns. From there, we look at how clutter can get into your visuals, how the human memory shapes what we see and how we can apply Gestalt principles to guide the eye. But none of that matters if your visualisation is not clear or accurate. Keep these best practices in mind as the foundation for the following sections.

5.5.2 Cognitive load and keeping it simple

Cognitive load is the mental effort required to understand new information. If your visuals have a high cognitive load, people will struggle to see your key message. Good visual design reduces cognitive load. Your audience should focus on the story, not deciphering the chart.

5.5.2.1 Clutter

Clutter is any element that complicates your visualisation without adding value, overwhelms readers and hides your key message. Clutter increases cognitive load. Examples include:

- too many gridlines
- redundant labels
- unnecessary 3D effects
- too many colours
- overly complex charts.

Practical tip how to identify clutter

A practical tip to identify clutter is to question every element. Does each item add to the story? If the answer is no, remove it.

For example, look at the bar chart in **Figure 5.27**. It shows total Airbnb revenue for different regions, with different colours for 2024 and 2025. The problem is not the data, but the extra visual elements. The gridlines and unnecessary decimal places increase cognitive load by forcing the brain to work harder to understand the chart. Reduce cognitive load to help the audience focus on the message in the data. What could we simplify? The next few sections will demonstrate, step by step, how to simplify this chart, eliminate clutter and use colour to highlight insights.

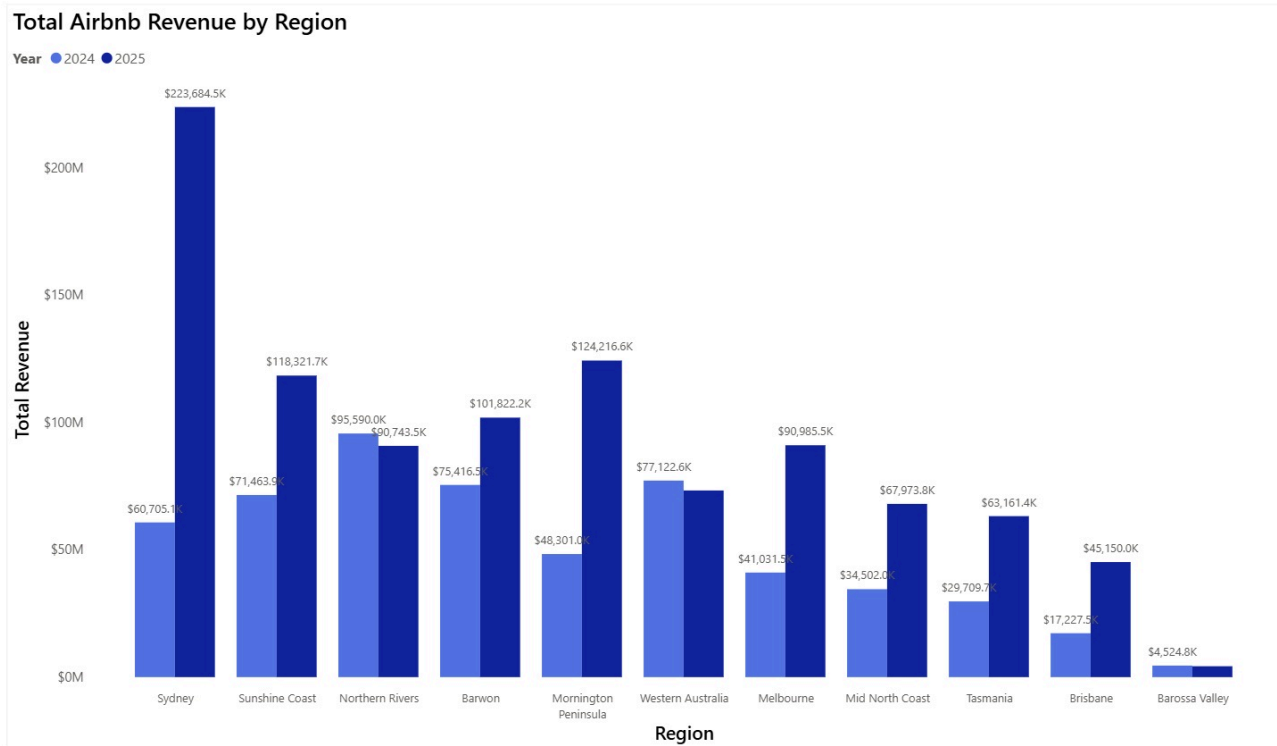


Figure 5.27: Let's eliminate clutter (Example 1): What clutter would you eliminate in the bar chart? Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.27 Grouped bar chart titled 'Total Airbnb Revenue by Region' comparing 2024 and 2025 revenues. Most regions show higher revenue in 2025, with Sydney and Sunshine Coast recording the largest increases, while smaller regions such as Barossa Valley and Brisbane generate the lowest totals.

Here, we show how removing one element makes a difference. We make each change step by step, to show its impact on the visual. **Figure 5.28** shows the first two small changes. The units now appear in millions ("M") and unnecessary trailing zeros after the decimal have been removed. These small changes make the chart cleaner.

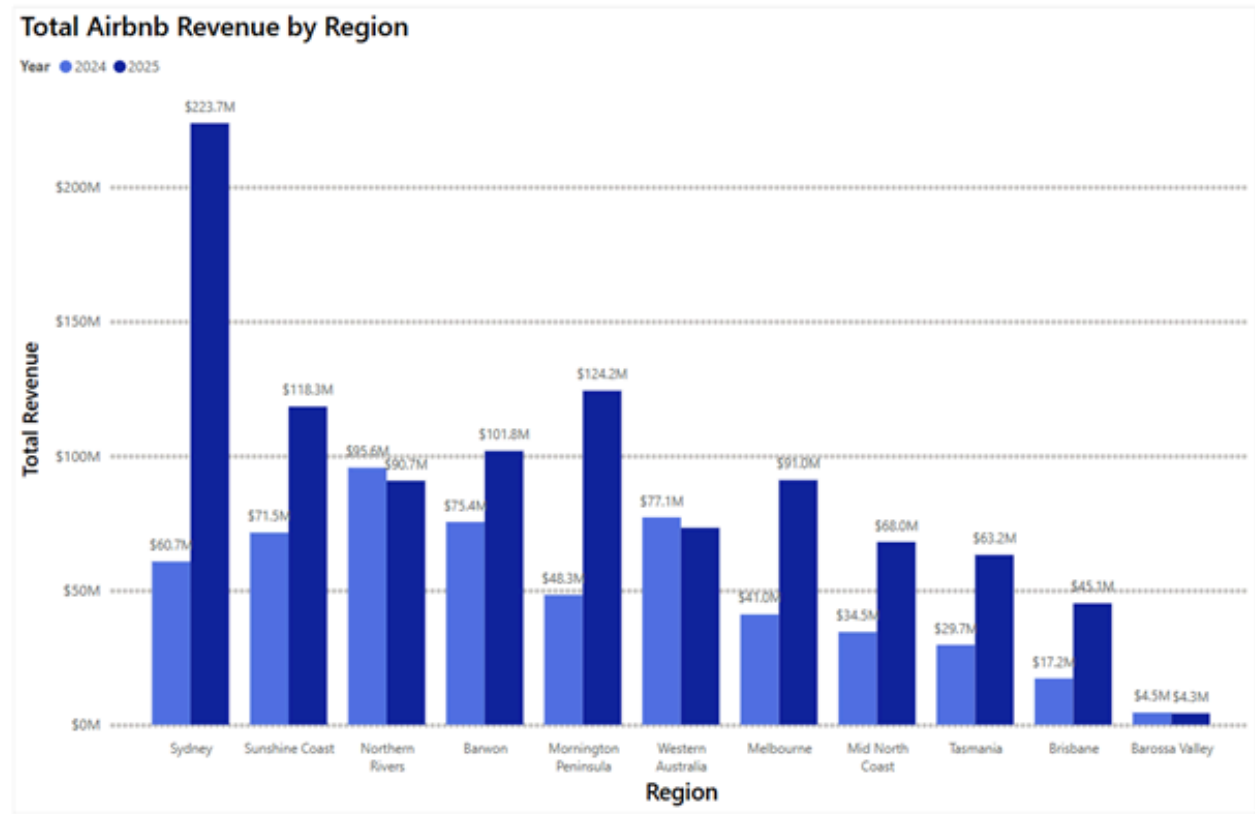


Figure 5.28: Let’s eliminate clutter (Example 1): First steps – change units to millions and remove trailing zeros. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.28

Grouped bar chart titled ‘Total Airbnb Revenue by Region’ comparing 2024 and 2025. Revenue increases across most regions in 2025, led by Sydney (about \$224M) and followed by Mornington Peninsula and Sunshine Coast, while Barossa Valley remains the lowest in both years.

The next step is to address gridlines and data labels. Do they help, or do they add clutter? We will remove them to see the improvement. **Figure 5.29** continues the analysis of total Airbnb revenue by region. In this version, the gridlines are removed and the data labels now appear inside the bars to connect the numbers to their corresponding bars with more clarity. Notice how much cleaner it looks. We also moved data labels inside the bars. This way, your eyes easily connect the numbers to the bars. What do you think? Is it easier to read now?

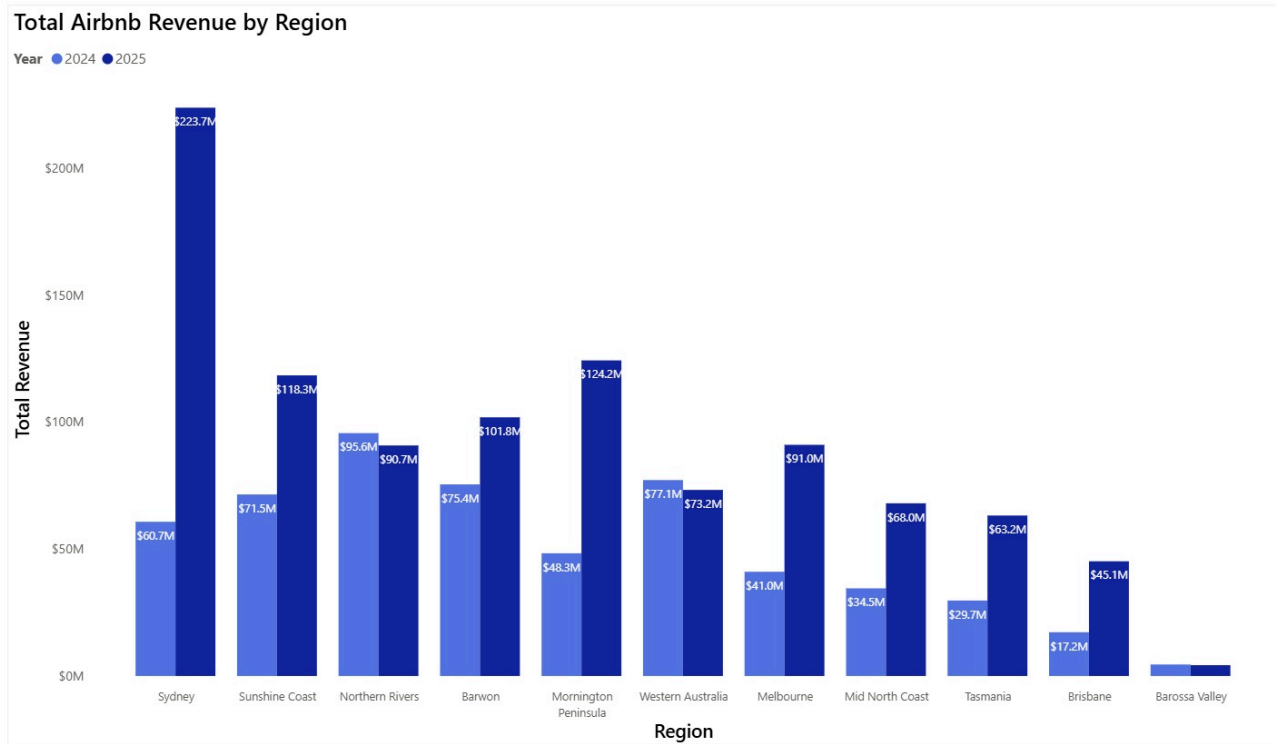


Figure 5.29: Let's eliminate more clutter (Example 1): Remove gridlines and move data labels into the bars. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.29

Grouped bar chart titled 'Total Airbnb Revenue by Region' comparing 2024 and 2025. Revenue increases across most regions in 2025, led by Sydney (about \$224M) and followed by Mornington Peninsula and Sunshine Coast, while Barossa Valley remains the lowest in both years.

Now, **Figure 5.30** takes this simplification a step further with the complete removal of data labels. The result is a pure visual comparison, where the audience can see differences in bar lengths without the distraction of numbers. The choice between these two versions depends on the story's objective.

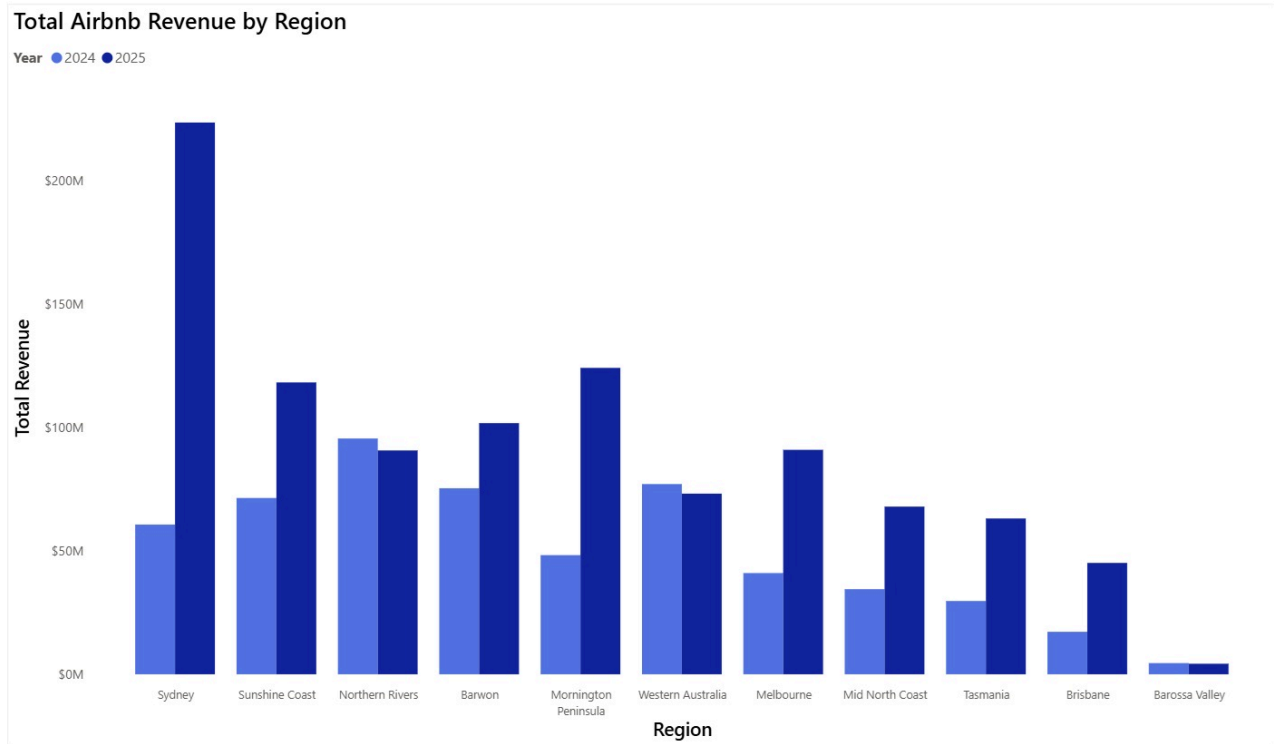


Figure 5.30: Let's eliminate more clutter (Example 1): Remove data labels. Source: <http://insideairbnb.com>

Alternative Text for Figure 5.30

Grouped bar chart titled 'Total Airbnb Revenue by Region' comparing 2024 and 2025. Revenue rises across nearly all regions in 2025, led by Sydney with the highest total, followed by Mornington Peninsula and Sunshine Coast, while Barossa Valley remains the smallest contributor.

Of course, there isn't a single correct way to do this. It depends on what you want to emphasise. **Figure 5.29** is better if exact revenue figures are important. **Figure 5.30** is more effective if the intention is to compare regions and years.

Here's another example of how we simplify visuals. **Figure 5.31** shows another common design problem. It shows revenue by room type over time, but the chart is cluttered. The diagonal month labels at the bottom are hard to read. Furthermore, the values for "Hotel room" and "Shared room" are so small they are difficult to see. This lack of a clear visual hierarchy makes it hard for a viewer to know where to focus their attention. Where do your eyes go first? We need to think about drawing attention thoughtfully in our charts. Is there a better way?

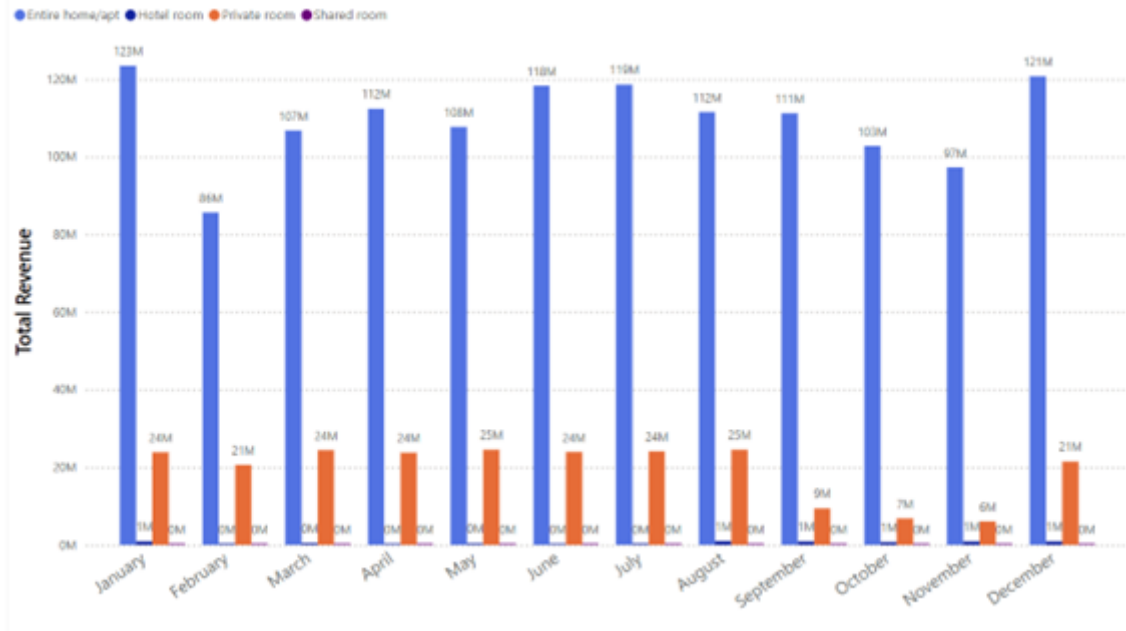


Figure 5.31: Cluttered bar chart (Example 2): What clutter would you eliminate in the bar chart? (revenue by room type)

Alternative Text for Figure 5.31

Monthly grouped bar chart showing Airbnb revenue by room type. Entire homes/apartments dominate revenue every month (roughly \$86M–\$123M), private rooms contribute a much smaller share (\$6M–\$25M), while hotel rooms and shared rooms generate minimal revenue throughout the year.

Let's take the next step towards improvement in **Figure 5.32**.

In **Figure 5.32**, we removed gridlines and the month labels are horizontal, which improves readability. However, the chart still includes all room types, even those with very low values. This raises the question of whether a bar chart is the best option for this data. For a better comparison of room type revenue over time, another chart type may be more effective. Should we try a line chart?

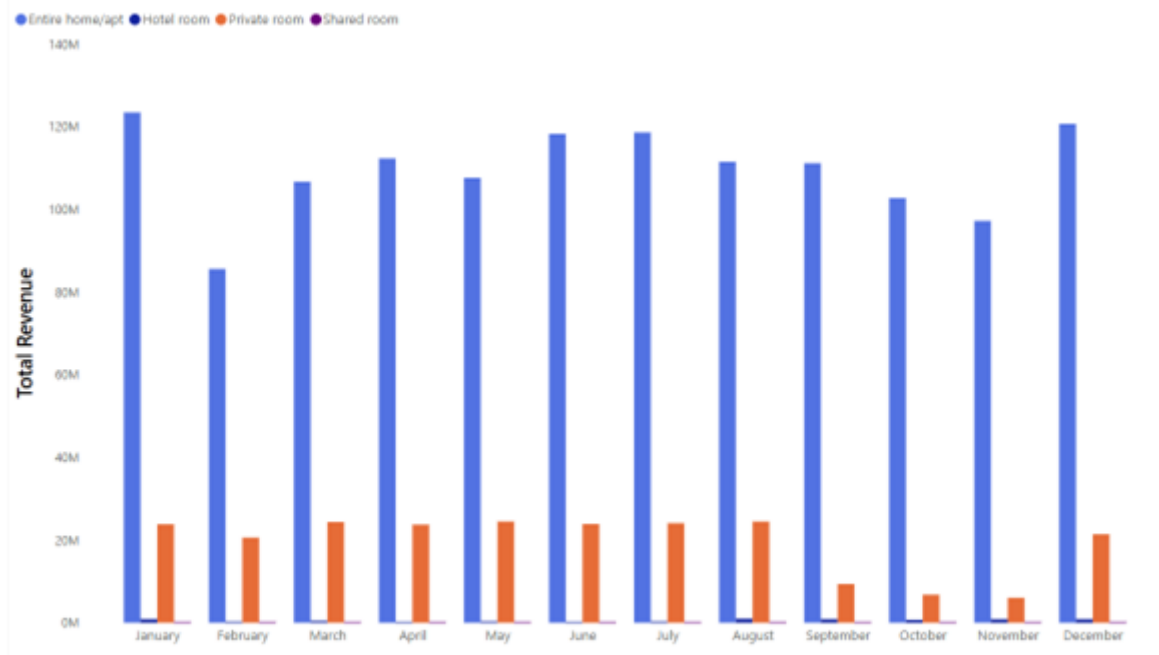


Figure 5.32: Less cluttered bar chart (Example 2): Eliminate data labels and gridlines and make month labels horizontal (revenue by room type). Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.32

Monthly bar chart showing Airbnb revenue by room type. Entire homes/apartments generate the majority of revenue each month, peaking in January, mid-year, and December, while private rooms contribute a smaller but steady share and hotel and shared rooms remain negligible.

In **Figure 5.33** we switch to a line chart. Why a line chart? As Andrew Abela’s taxonomy (Abela 2008) recommends, line charts are the best choice if we want to compare trends over time. The line chart allows us to see how revenue changes on a monthly basis for each room type. However, the chart still has a problem. The revenue for “Hotel room” and “Shared room” is so low that their trends appear as flat lines at the bottom of the chart.

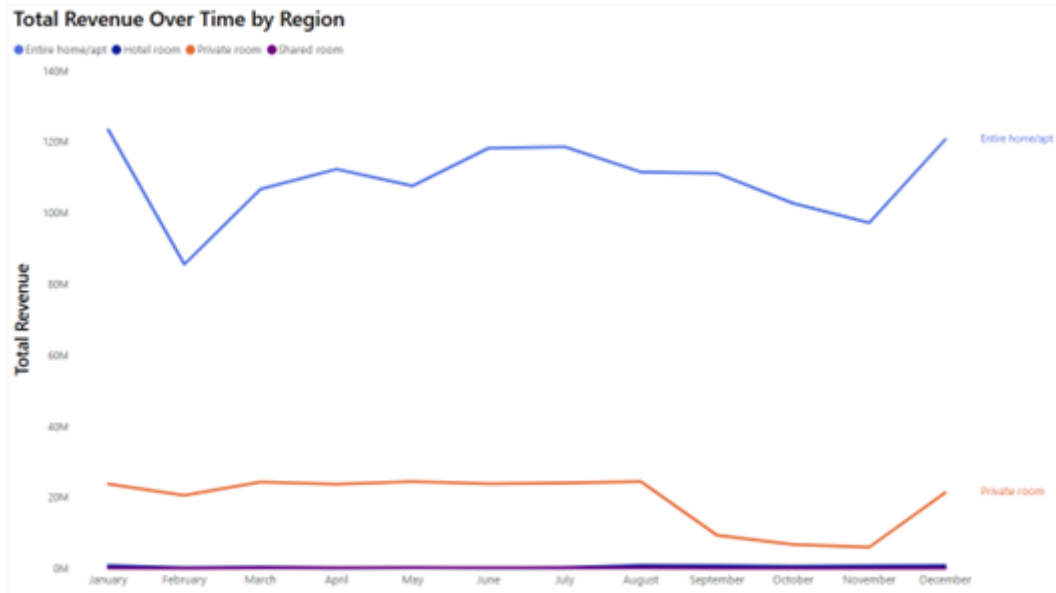


Figure 5.33: Switching to a line chart. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.33

Line chart titled 'Total Revenue Over Time by Region' showing monthly Airbnb revenue by room type. Entire homes/apartments consistently generate the highest revenue, dipping in February and peaking mid-year and again in December, while private rooms contribute a smaller share with a notable decline in early spring before rising at year end.

What else can we try? Let's see a better approach in **Figure 5.34**.

Figure 5.34 focuses on only the two main room types, Entire home/apt and Private room, which makes the comparison between them easier. The chart also uses colour to guide the eye: revenue numbers above \$100 million appear in blue while lower ones are orange. Colour is a preattentive attribute, which means the brain notices the difference in colour before thinking starts. The combination of colour and the Gestalt principle of similarity separates the data into two different categories.

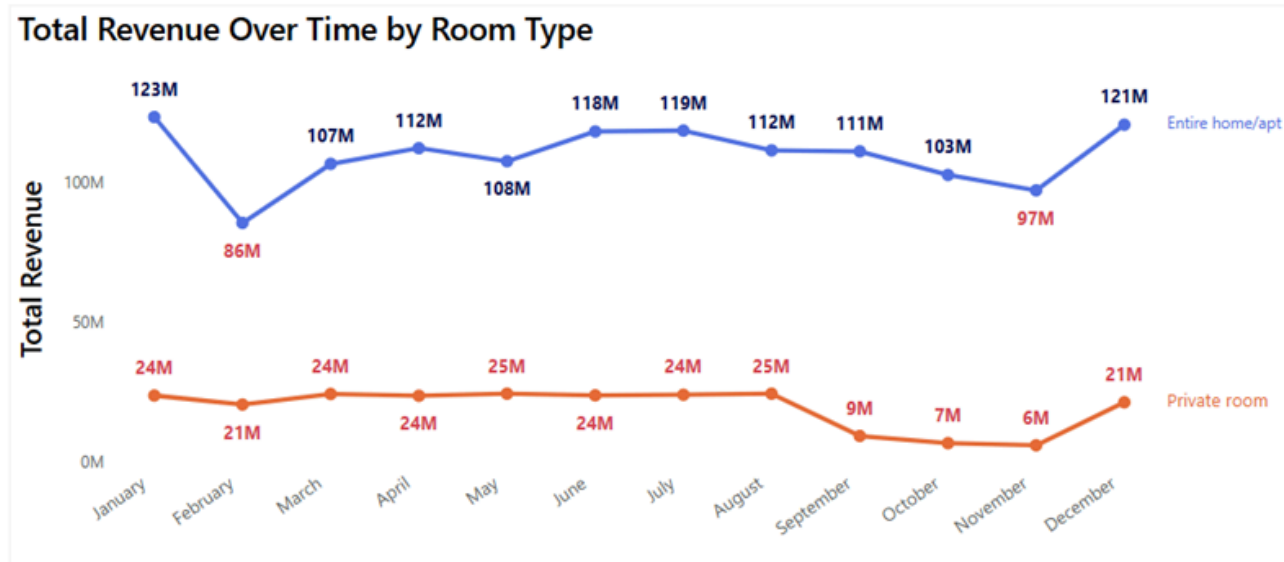


Figure 5.34: Focusing on the key trends. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.34

Line chart titled 'Total Revenue Over Time by Room Type' showing monthly Airbnb revenue. Entire homes/apartments dominate revenue throughout the year, dipping in February and peaking in mid-year and December, while private room revenue remains much lower and declines sharply from September to November before recovering in December.

Figure 5.34 allows us to compare revenue trends for the two main room types. The “Entire home/apt” category generates more revenue than the “Private room” category, but both show noticeable changes during the year. A focus on fewer categories and thoughtful use of colour will simplify the visual and help highlight the most important insights.

5.5.2.2 Reduce clutter by separating important information from noise

Think of your visualisation as a message. Your data is the important information you want to communicate. Clutter is the noise that interferes with your message. Our goal is to maximise information and minimise noise.

The Gestalt principles, which we discuss in [Section 5.5.3.1](#) help here. These principles describe how our brains will first group then translate visual information. Principles like proximity, the group of related elements and similarity, the use of consistent colours and shapes, help reduce clutter.

Practical tip

Gridlines add clutter and interfere with your message. For more on gridlines, see [Gridlines are Gratuitous](#).

The gridlines in **Figure 5.35** make it difficult to separate the important information (revenue) from the noise (gridlines). Compare **Figure 5.35** to **Figure 5.34**. The same line graph without gridlines is much easier to read.

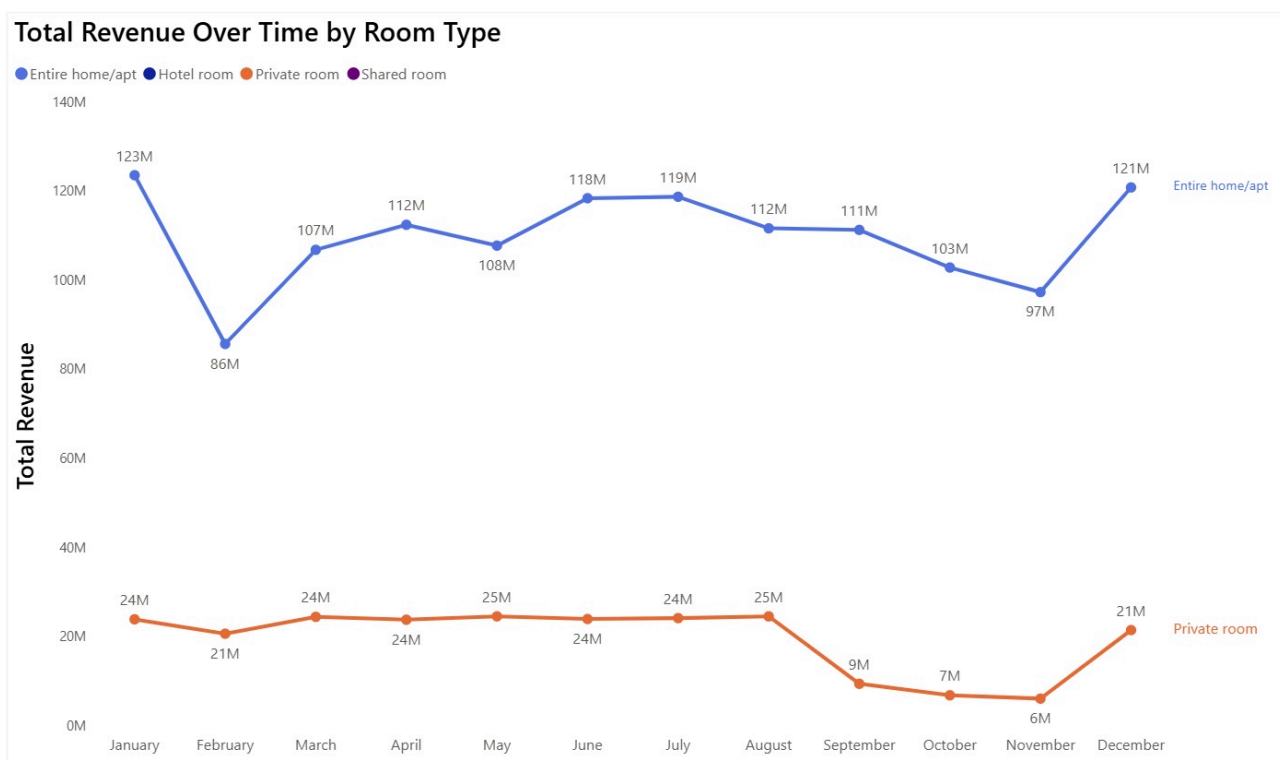


Figure 5.35: Separating information from noise: Gridlines are often noise. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.35

Line chart titled 'Total Revenue Over Time by Region' showing monthly Airbnb revenue by room type. Entire homes/apartments generate the highest revenue each month, dropping sharply in February and peaking again in mid-year and December, while private room

revenue remains much lower and declines noticeably from September to November before recovering in December.

5.5.2.3 Use memory to your advantage

We can apply some basic memory principles to make charts more memorable and impactful.

- **Iconic memory:** This is your first impression, what people notice immediately. Keep it simple. Bold charts, clear titles and strong colours will grab attention.
- **Short-term memory:** People can only hold a few items in short-term memory at once. Do not overload charts with too many data points. Focus on your main message.
- **Long-term memory:** Want your insights to stick? Use consistent visual elements like the same colours, chart types and layouts. Consistency helps people to build familiarity and makes your data story easier to recall.

For more on visual perception and memory, see [Visual Perception](#).

Practical tip

Think about those Airbnb occupancy trends again. If you use the same colour consistently for each city across all your charts, people begin to associate each colour with a specific city. This consistency makes it easier to follow that city's performance across different charts. See how we maintain colour consistency across multiple visualisations (e.g., recall **Figure 5.6** and **Figure 5.7**).

5.5.3 Visual perception in visualisation design

This section introduces two important design principles, the Gestalt principles of visual perception and preattentive attributes. Gestalt principles help to explain how people group and organise visual information by human instinct. Preattentive attributes are visual features our brains process, even before we start to think. Both ideas come from psychology and help us to design better BI visuals.

5.5.3.1 Gestalt principles

Here, we discuss six key Gestalt principles and how they apply to BI visualisations. Each principle includes examples to show how they help create easier visuals, reduce cognitive load and bring out your message.

- **Proximity:** Elements close together appear related.
- **Similarity:** Elements that look alike, with the same colour, shape, size or orientation, appear related.
- **Enclosure:** Elements physically enclosed together, such as inside a box or border, appear grouped.
- **Closure:** Our minds tend to see open shapes as complete when it makes sense.
- **Continuity:** Elements that align or look continuous appear grouped.
- **Connection:** Elements connected, such as by a line, appear grouped.

5.5.3.1.1 Gestalt principles in practice (Example 1): Similarity

Figure 5.36 shows the monthly occupancy rate and is a good example of Gestalt principles. The use of red and blue bars is an application of **similarity**. Our brains group these bars by colour. The red bars and red text bring attention to specific data points. We use similarity to create two distinct groups, which makes it simple to see which months are above or below the 50% threshold. Gestalt principles do not just make visuals attractive, they help our brains process information with less effort.

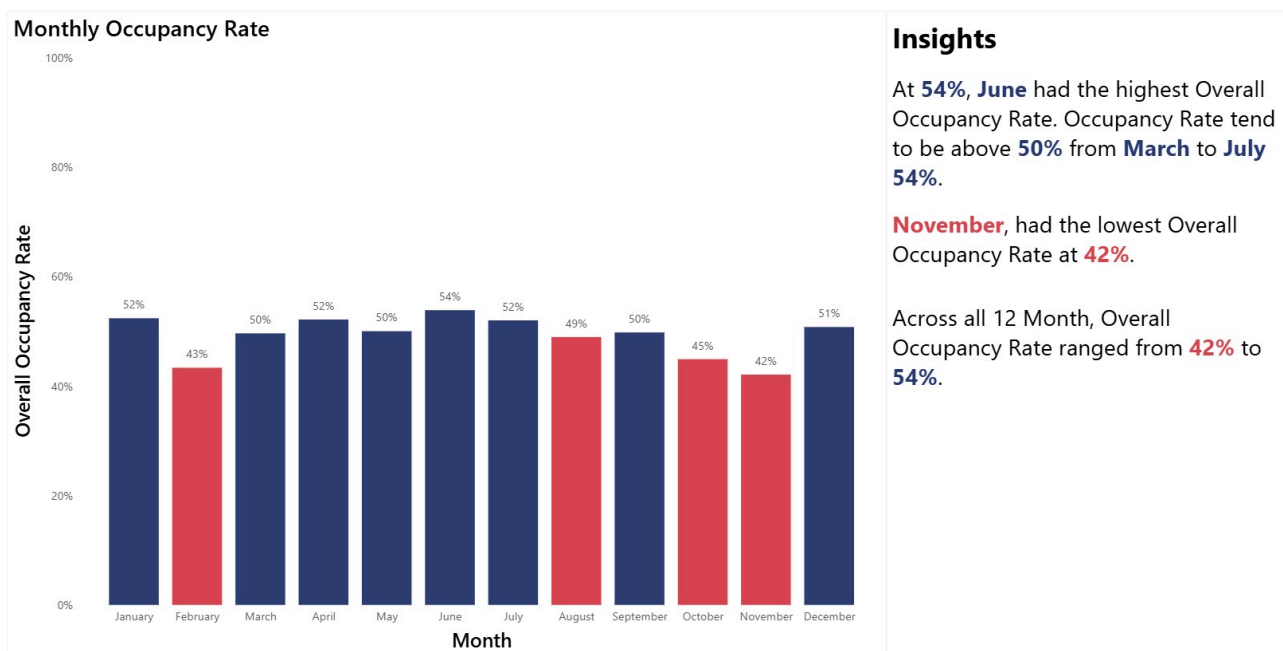


Figure 5.36: Gestalt principal (Example 1): Similarity. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.36

Bar chart titled 'Monthly Occupancy Rate' showing overall Airbnb occupancy by month, ranging from 42% to 54%. Occupancy peaks in June at 54%, stays above 50% from March

to July, and is lowest in November at 42%, with an accompanying text panel summarizing these insights.

5.5.3.1.2 Applying Gestalt principles (Example 2): Similarity and connection

Figure 5.37 also shows the monthly occupancy rate, but we use a different design. We use red and blue bars and match them with the red and blue text, another application of **similarity**. The visual links the bars and text with colour. The lines that connect the text to the bars demonstrate the principle of **connection**, leaving no doubt about which text describes which bar.

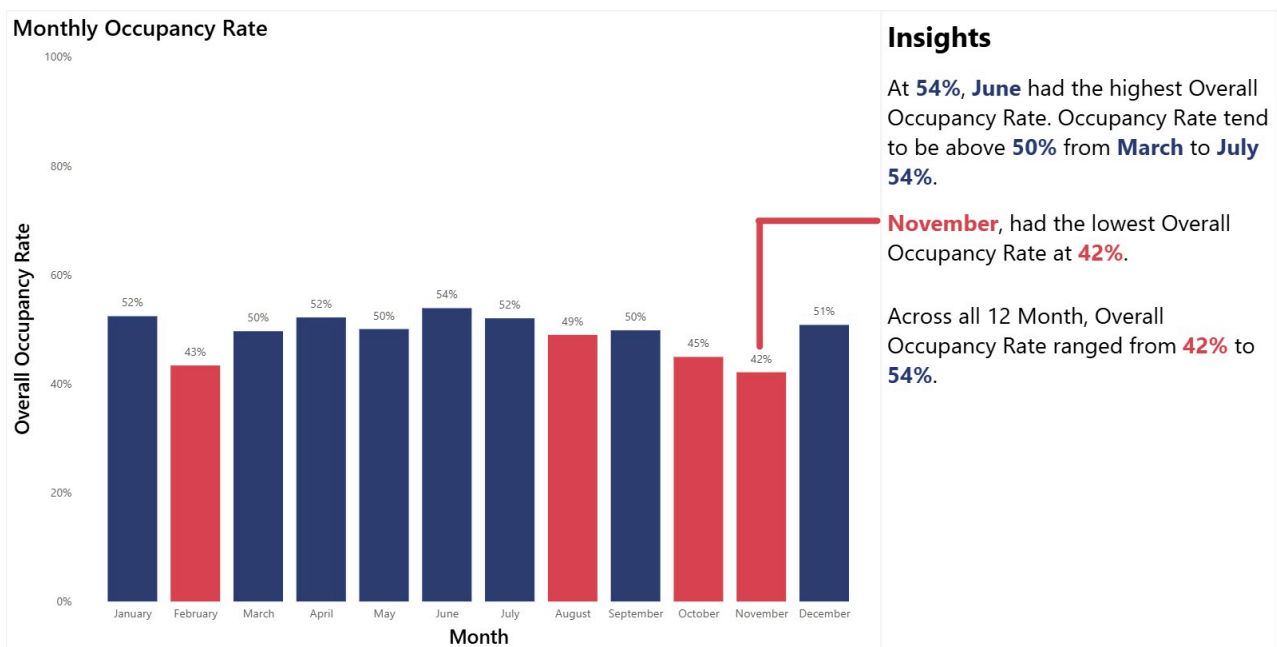


Figure 5.37: Gestalt principal (Example 2): Similarity and connection. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.37

Bar chart titled 'Monthly Occupancy Rate' showing overall Airbnb occupancy by month. Rates range from 42% to 54%, peaking in June at 54% and staying above 50% from March to July, while November has the lowest occupancy at 42%, highlighted with an annotation and supporting insights text.

Use with caution, as too many lines can make a visual messy. When you combine text and data, you must help your audience to know where to look. Gestalt principles like similarity and connection are effective tools to achieve this goal.

5.5.3.2 Preattentive attributes

Preattentive attributes are visual elements like colour, shape or position that our brains process before we think about them. These attributes make parts of a visualisation stand out and help to draw attention to key information.

This section introduces preattentive attributes and shows how they help to highlight important insights in BI dashboards and reports. We cover three main categories, form, colour and spatial position, and discuss limitations and best practices.

Categories of preattentive attributes

There are three main categories of preattentive attributes relevant to data visualisation (Few, 2012).

- **Form:** Includes length, width, orientation, shape, size and enclosure.
- **Colour:** Includes hue and intensity.
- **Spatial position:** Includes 2D position and grouping.

Best practices for Preattentive Attributes

- Use preattentive attributes carefully.
- Limit the number of differences, about 8 colours, 4 orientations or 4 sizes at once.
- Less is usually better.
- Choose attributes strategically to highlight important insights.

5.5.3.2.1 Preattentive attributes in practice (Example 1): Colour

Remember, preattentive attributes are features our brains notice immediately. Colour is a powerful one. **Figure 5.38** shows the use of colour to highlight Sydney's revenue. In the bar chart of total revenue by region, Sydney's bar is bright red while the rest are blue. The eye goes straight to Sydney's red bar because our brains quickly spot differences in colour.

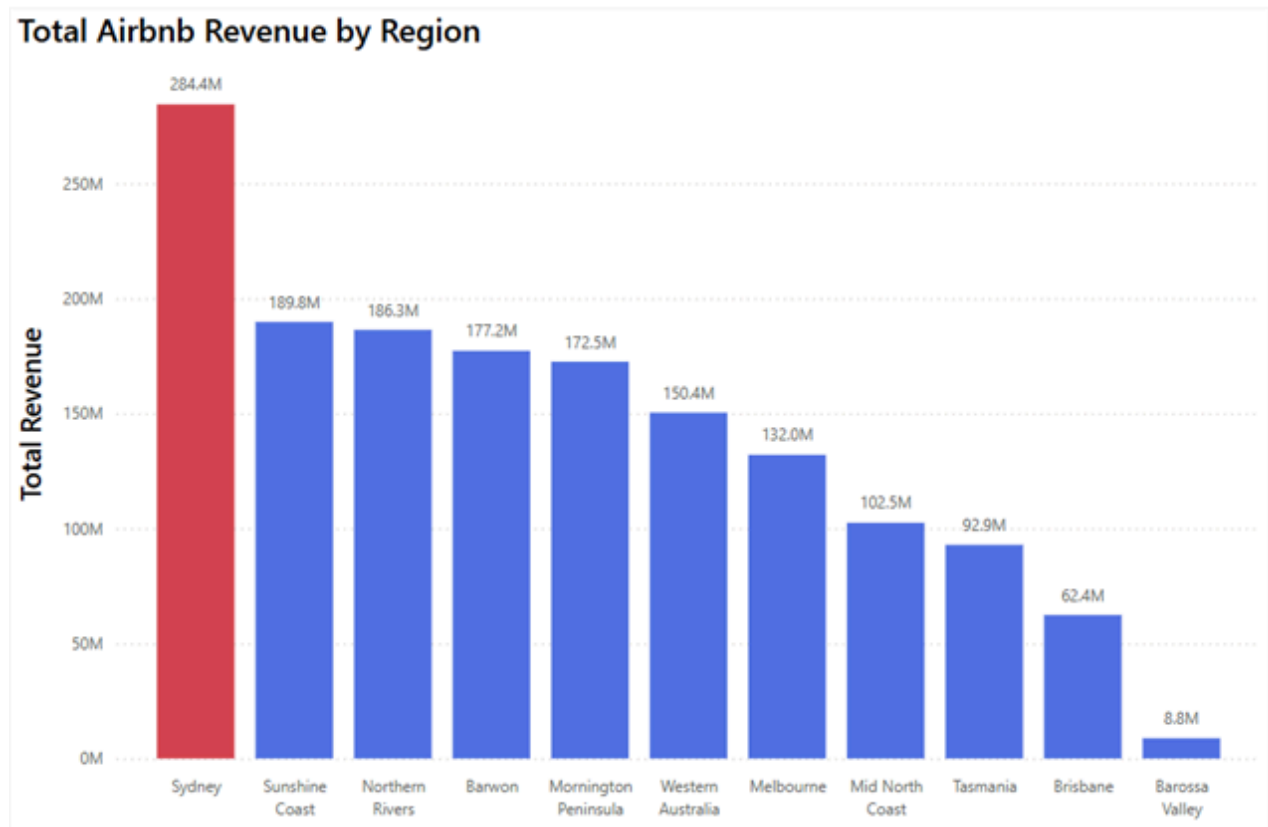


Figure 5.38: Preattentive attribute (Example 2): Using colour to draw attention. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.38

Bar chart titled 'Total Airbnb Revenue by Region' showing Sydney as the highest-revenue region at about \$284M, followed by Sunshine Coast, Northern Rivers, Barwon, and Mornington Peninsula, with progressively lower revenues across Western Australia, Melbourne, Mid North Coast, Tasmania, Brisbane, and Barossa Valley.

We use colour here to highlight Sydney and to focus attention on its revenue. The other blue bars are important as they provide context, but Sydney is the focus. This simple method helps our audience to see the key message: Sydney generates the most revenue. You can use this same method to highlight any important category. The key is the strategic use of colour for your message to pop out.

5.5.3.2.2 Applying preattentive attributes in practice (Example 2): Colour and form

Figure 5.39 shows how to combine colour and form, in this case enclosure, to emphasise parts of a chart. Here, we use colour to highlight Sydney. The chart also draws rectangles around each quarter, which is

an application of the Gestalt principle of enclosure. Combining colour and enclosure draws the eye to the highlighted sections. Enclosure helps to group the months within each quarter while preattentive use of colour makes Sydney pop out.

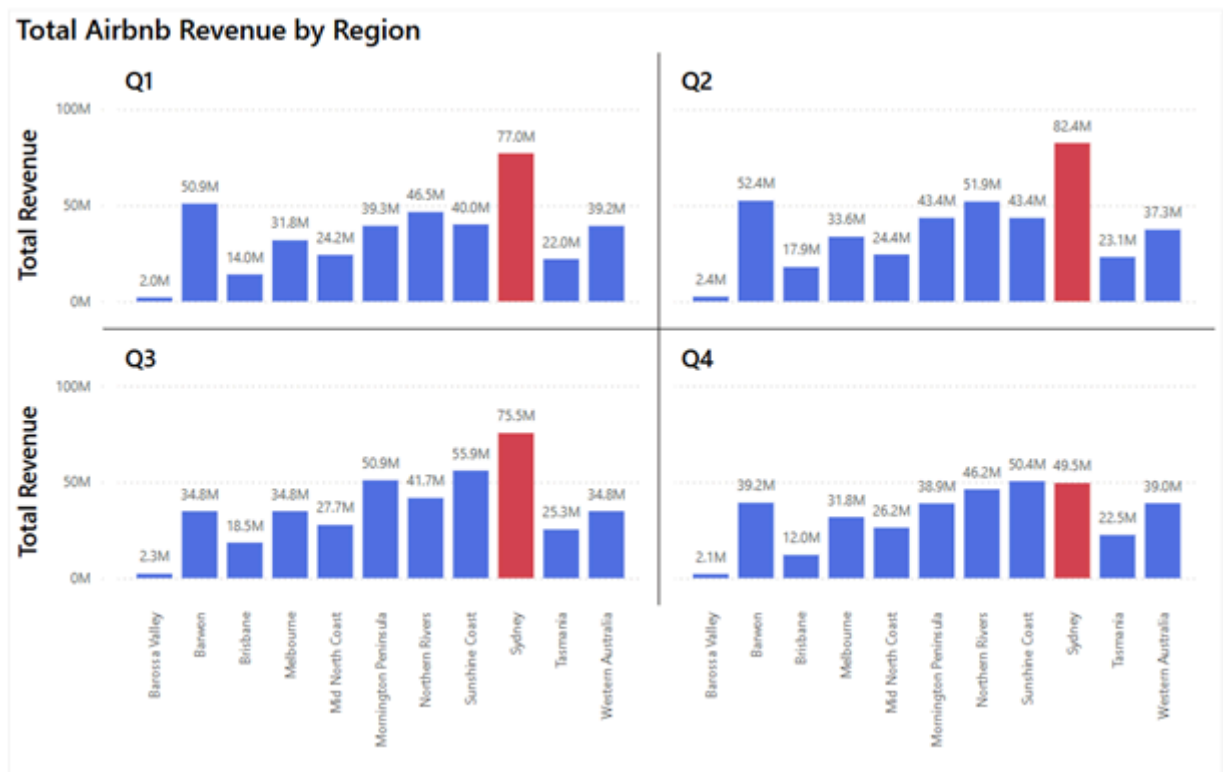


Figure 5.39: Preattentive attribute (Example 2): Using form (enclosure) and colour to draw attention. Source: <http://insideairbnb.com>.

Alternative Text for Figure 5.39

Four-panel bar chart titled 'Total Airbnb Revenue by Region' showing quarterly revenue (Q1–Q4) across Australian regions. Sydney is the top-revenue region in every quarter, peaking in Q2, while Sunshine Coast and Northern Rivers follow, and Barossa Valley remains the lowest contributor throughout the year.

5.6 Chapter summary and end of chapter questions

5.6.1 Why data visualisation matters

This chapter explored how data visualisation helps people understand complex data to make better

decisions. We reviewed key historical moments, from the first graphs of the 10th century to modern interactive dashboards. We saw how pioneers of data visualisations like William Playfair invented bar and line charts, Charles Minard created a narrative graphic of Napoleon's Russian Campaign and how Florence Nightingale used her rose diagram to bring about public health reform. Today, BI tools make data exploration accessible to non-technical users, leading to their widespread use in making decisions in healthcare, finance and logistics.

5.6.2 Types of charts and graphs: From basic to specialised

This chapter used Andrew Abela's taxonomy as a framework to select the correct charts. The framework organises visuals based on four goals: comparison, relationship, distribution and composition. We covered basic charts like bar charts to compare categories and line charts to show trends over time. We also explored specialised visuals like heatmaps to show intensity, treemaps for hierarchical data and geospatial maps for geographic insights. A key example was the Key Influencers visual, which uses machine learning to identify factors that have a strong association with a specific outcome.

5.6.3 Crafting narratives with visualisations

We then looked at how to build a data story. We introduced the narrative arc framework, which we used with Airbnb data to show how to structure a data story. This story structure begins with the background and introduction of a problem, builds to a climax with a key insight and finishes with a clear resolution or recommendation. We showed how a data story can give context to the visuals and motivate an audience to act.

5.6.4 Effective visualisation design and best practices

Finally, we reviewed key design principles. The primary goal is to reduce cognitive load and eliminate clutter so the audience can focus only on the message. The chapter explained how to use Gestalt principles, like proximity and similarity, and preattentive attributes like colour and form to guide an audience's attention to specific elements and messages.

5.7 Test your knowledge

Drag the words into the correct boxes.

1) A chart is best for comparing categories such as revenue by region.

2) A

chart is most appropriate for showing trends over time, such as monthly revenue.

3) A

plot helps reveal relationships between two numerical variables, like price and occupancy.

4) A

highlights intensity patterns across two dimensions, such as region and month.

5) A

is useful for showing how parts contribute to a whole when there are many categories.

scatter

line

treemap

bar

heatmap



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=715#h5p-6>

An alternative source for this activity: <https://uts.h5p.com/content/1292858509124492209>

Additional readings and resources

- Abela, A. (2013). *Advanced presentations by design: Creating communication that drives action* (2nd ed). John Wiley & Sons.
- Al-Ansi, A. M., Jaboob, M., Garad, A., & Al-Ansi, A. (2023). Analysing augmented reality (AR) and virtual reality (VR) recent development in education. *Social Sciences & Humanities Open*, 8(1), 100532.
- Baigabulov, S., Ipalakova, M., & Turzhanov, U. (2025). Virtual reality enabled immersive data visualisation for data analysis. *Procedia Computer Science*, 265, 602–609.
- Brasseur, L. (2005). Florence Nightingale's visual rhetoric in the rose diagrams. *Technical Communication Quarterly*, 14(2), 161–182.
- Cairo, A. (2016). *The truthful art: Data, charts, and maps*.
- Codd, E. F. (1970). A relational model of data for large shared data banks. *Communications of the ACM*, 13(6), 377–387.
- Cossich, V. R., Carlgren, D., Holash, R. J., & Katz, L. (2023). Technological breakthroughs in sport: Current practice and future potential of artificial intelligence, virtual reality, augmented reality and modern data visualisation in performance analysis. *Applied Sciences*, 13(23), 12965.
- Fee, E., & Garafalo, M. E. (2010). Florence Nightingale and the Crimean War: The power of data visualization. *American Journal of Public Health*, 100(12), 2267. <https://doi.org/10.2105/AJPH.2009.188607>

- Few, S. (2007). Save the pies for dessert. https://www.perceptualedge.com/articles/visual_business_intelligence/save_the_pies_for_dessert.pdf
- Few, S. (2012). *Show me the numbers: Designing tables and graphs to enlighten* (2nd ed.). Analytics Press.
- Friendly, M. (2008). A brief history of data visualisation. In C. Chen, W. Härdle, & A. Unwin (Eds.), *Handbook of data visualisation* (pp. 15–56). Springer.
- Inmon, W. H. (1992). *Building the data warehouse*. Wiley.
- Knafllic, C. N. (2015). *Storytelling with data: A data visualisation guide for business professionals*. John Wiley & Sons.
- Knafllic, C. N. (2019). *Storytelling with data: Let's practice!* John Wiley & Sons.
- Kolokolov, A., & Zelensky, M. (2024). *Data visualization with Microsoft Power BI: How to design savvy dashboards*. O'Reilly Media.
- Mărcuță, C., & MoldStud Research Team. (2025, May 20). Top 5 challenges businesses face when implementing AR and VR solutions. MoldStud. <https://moldstud.com/articles/p-top-5-challenges-businesses-face-when-implementing-ar-and-vr-solutions>
- National Army Museum. (n.d.). *Florence Nightingale: The Lady with the Lamp*, from <https://www.nam.ac.uk/explore/florence-nightingale-lady-lamp>
- Nightingale, F. (1858). Mortality of the British Army. Wellcome Collection. <https://wellcomecollection.org/works/jxwtskzc/images?id=s9eq8yvx>
- Olshannikova, E., Ometov, A., Koucheryavy, Y., & Olsson, T. (2015). Visualizing Big Data with augmented and virtual reality: Challenges and research agenda. *Journal of Big Data*, 2(1), 22.
- Song, Y., Kalacska, M., Gašparović, M., Yao, J., & Najibi, N. (2023). Advances in geocomputation and geospatial artificial intelligence (GeoAI) for mapping. *International Journal of Applied Earth Observation and Geoinformation*, 120, 103300. <https://doi.org/10.1016/j.jag.2023.103300>
- Theoris. (2024). *Tracing the history of business intelligence (BI)*. <https://www.theoris.com/tracing-the-history-of-business-intelligence-bi/>
- Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Graphics Press.
- Yau, N. (2024). *Visualise this: The flowing data guide to design, visualisation and statistics* (2nd ed.). Wiley.

About the author



Dr Florence Mwagwabi

MURDOCH UNIVERSITY

https://researchportal.murdoch.edu.au/esploro/profile/florence_mwagwabi/overview

Dr Florence Mwagwabi is a Senior Lecturer in Information Technology and the Academic Chair of Transnational Education (IT) at Murdoch University, Singapore. She has taught Business Intelligence application development for more than nine years and has led curriculum design that integrates real-world datasets, BI tools and current industry trends. Her teaching spans databases and systems analysis, data communications, business intelligence, machine learning foundations and information security across undergraduate and postgraduate programs. Her research in this area includes the use of SDG data in analytics education, the impact of real-world projects on BI skill development and the design of data literacy programs.

Her broader research spans cybersecurity, AI applications in learning, user security behaviour and ICT use across different population groups. Her work has been featured in national media and supported by competitive research funding. She is a Fellow of the Higher Education Academy (FHEA), a member of IEEE and the Australian Computer Society (ACS) and a recipient of Murdoch University's Citation for Excellence in Enhancing Student Learning.

6. INFORMATION DASHBOARDS

Amara Atif

Learning Objectives (LO)

By the end of this chapter, you will be able to:

LO1 Explain the role of information visualisation and visual analytics in analysing, interpreting and communicating data, and understand their differences and interconnections in business contexts.

LO2 Define business reporting, its components, and its importance in decision-making and organisational performance.

LO3 Differentiate between various types of managerial reports (Metric Management Reports, Dashboard-Type Reports, and Balanced Scorecard Reports) and their use in tracking and analysing performance.

LO4 Define information dashboards, understand their role in visualising and communicating key metrics, and distinguish between static and dynamic dashboards.

LO5 Explain how to design effective dashboards that capture attention and deliver actionable insights through best practices such as layout, interactivity and appropriate use of visualisations, negative space, colour and typography.

LO6 Explain the concept of data storytelling, its importance in communicating insights effectively, and how to create data stories that combine visuals, context and narrative to engage audiences and support decision-making.

6 Overview: the importance of understanding Information Visualisation and Visual Analytics

As we dive into the world of information dashboards, it is important to understand two foundational concepts: *information visualisation* and *visual analytics*. These terms are not just technical jargon; they represent essential skills that will empower you to create impactful dashboards, interpret data effectively and make informed decisions.

Information visualisation helps you present data clearly and intuitively, allowing you to explore patterns, trends and structures. Visual analytics goes a step further, combining visualisation with advanced analysis to uncover deeper insights and predict future outcomes. Together, these concepts form the backbone of modern data reporting and dashboard design.

By mastering these ideas, you will learn how to build business reports that effectively communicate key insights and design dashboards that are visually appealing, interactive and informative. You will also gain a deeper understanding of how dashboards play an important role in data interpretation, problem-solving and decision-making. These skills are not only valuable in the classroom but are also highly sought after in the professional world. In the remaining chapter, you will learn how dashboards, powered by information visualisation and visual analytics, have the ability to transform **raw data** into **actionable insights**.

Information Visualisation

What is it?

Focuses on the static or dynamic representation of data using visual techniques such as charts, graphs and dashboards. It is about understanding and communicating “what happened” or “what is happening” (descriptive analytics).

Purpose:

- Helps users find patterns, trends and structures in data with simple and clear visuals.
- Makes complex data easy to understand quickly.
- Mainly used for **data exploration, reporting** or **storytelling**.

How it works?

- Often used with Business Intelligence (BI) tools such as Tableau, Power BI, Excel, Matplotlib/Seaborn (Python libraries), D3.js and R ggplot2.
- Works best with smaller or medium-sized, well-structured datasets.
- User interaction is basic, e.g., zooming, filtering or sorting.
- The user looks at visual patterns and finds insights from them.

Examples:

- A bar chart showing sales trends over time.
- A pie chart showing the market share of companies.
- A scatter plot showing the relationship between advertising spend and sales revenue.
- A heatmap displaying website user activity.
- Real-time dashboards for tracking sales performance.

Visual Analytics

What is it?

Focuses on predictive and prescriptive analytics, answering “why it is happening” and “what might happen next” by *combining visualisation techniques with analytical algorithms* to support complex data analysis tasks.

Purpose:

- Helps users uncover hidden patterns, trends or relationships in large and complex datasets.
- Provides deeper insights and supports decision-making by integrating computational models with user interaction.
- Focuses on interactive and iterative exploration of data to solve complex problems.

How it works?

- Often used with Business Analytics tools such as SAS Visual Analytics, IBM Cognos, advanced features of Tableau/Power BI, Qlik and Python libraries like Plotly integrated with machine learning frameworks.
- Designed to handle large-scale, high-dimensional and unstructured datasets.
- High level of user interaction and iterative analysis allows users to explore data deeply and flexibly.
- Involves model-driven techniques (e.g., clustering, forecasting) combined with human intuition for data exploration.

Examples:

- A forecasting dashboard predicting future sales based on historical trends and market conditions.
- Segmentation analysis to understand customer groups and target them effectively.
- A system analysing social media data to predict customer sentiment and visualise trends.
- A tool monitoring network traffic, identifying anomalies and visualising attack patterns.
- An interactive dashboard that uses machine learning to predict sales trends and allows users to filter data.

It's important to understand that **visual analytics doesn't replace information visualisation**, but rather **builds upon it**.

6.1 Business reporting

A **report** is a *communication artifact* created with the specific purpose of presenting information, typically derived from *data* and *personal experiences*, in a clear and organised format. This information can be conveyed through *narrative*, *graphics* and/or *tables*, tailored to meet the needs of the intended audience¹. Reports can be delivered on a *recurring basis* (periodic) or *as needed* (ad hoc), ensuring the information is available *whenever* and *wherever* it is required.

If it concerns *business matters*, then it is called a **business report** (also called enterprise reporting). In business contexts, reports come in various forms, such as memos, minutes, lab reports, sales reports, progress reports, justification reports, compliance reports, annual reports, and policies and procedures. Business reporting is a key component of the business intelligence initiative, which focuses on enhancing managerial decision-making and organisational knowledge management. These reports are built upon data collected from both *internal* and *external* sources of the organisation.

Modern business reports have become increasingly visual, leveraging colours, graphical icons and dashboards to improve the presentation and comprehension of information. The *reporting process* often involves querying structured data sources, which are developed using logical data models and data dictionaries, to produce reports that are human-readable and easily digestible. These reports facilitate managers and teams to remain informed, evaluate options and make informed decisions by converting data from various sources into actionable information. Through effective business reporting, organisations enhance their ability to analyse data, improve processes and drive better outcomes.

The key to any successful report lies in the 4Cs: *clarity*, *conciseness*, *completeness* and *correctness*. Business reports can generally be categorised into three types based on content and format: informal, formal and short.

- **Informal reports:** Typically, up to 10 pages long, these reports are routine, internal, and follow a letter or memo format. They often use personal pronouns and casual language, making them less

1. The audience can include novice users, generalists, management, experts, and executives. *Novices* are new to a subject and require clear explanations without oversimplification. *Generalists* are familiar with the topic and seek a broad overview and key themes. *Management* looks for actionable insights, detailed understanding, and connections between various elements. *Experts* prefer deeper exploration and discovery with minimal focus on simplifications. *Executives*, on the other hand, focus on the significance of the data and conclusions, emphasising weighted probabilities and outcomes.

formal and more conversational.

- **Formal reports:** These reports, typically 10 to 100 pages long, are based on research or analysis and are meant for internal or external audiences who need the information. They exclude personal pronouns and casual language and include formal sections like a title page, table of contents and executive summary.
- **Short reports:** These are designed to quickly inform stakeholders about events or system status changes. They are often periodic, investigative, compliance-focused or situational.

The purpose and audience of the report significantly influence its nature. Internal reports typically aim to inform stakeholders and decision-makers within an organisation, while external reports serve as a means of communication between businesses and/or external entities such as the government. For instance, businesses may submit reports for tax compliance or regular filings with regulatory bodies e.g., the Australian Securities and Investment Commission (ASIC).

6.1.1 Managerial reports

Managerial reports are essential tools in business decision-making, providing structured information to help managers monitor performance, analyse data and make informed decisions. **They serve as a bridge between raw data and the decision-making process**, ensuring that managers have a clear understanding of key performance areas and trends.

Managerial reports are designed to address different levels of decision-making – *operational, tactical* and *strategic*. At the **operational level**, reports may focus on day-to-day activities, such as tracking production or sales figures. **Tactical reports** provide information for mid-level managers to plan and allocate resources effectively. **Strategic reports**, on the other hand, support top-level executives in shaping long-term goals and policies.

Below are the three primary types of managerial reports, along with Australian examples and scenarios from various industries.

6.1.1.1 Metric management reports

Metric management reports focus on monitoring business performance through measurable, outcome-oriented metrics. These metrics help organisations ensure they meet both external and internal performance expectations. For external groups, these are service-level agreements (SLAs). For internal management, they are key performance indicators (KPIs). Typically, there are enterprise-wide agreed targets to be tracked over a period of time. They may be used as part of other management strategies, such as Six Sigma or Total Quality Management (TQM), to improve efficiency, enhance quality and achieve operational excellence.

Australian example: Healthcare sector

Metric management reports play a crucial role in the Australian healthcare industry, particularly in monitoring hospital performance and patient care quality. For example, an Australian hospital may have an enterprise-wide target to achieve 95% patient satisfaction and maintain an emergency department wait time under 4 hours. These targets are tracked over time, and metric management reports are used to monitor trends and highlight areas for improvement. Table 6.1 highlights the critical role of **metric management reports** in the Australian healthcare industry, particularly in monitoring hospital performance and patient care quality. It demonstrates how these reports are used both for **external accountability** (via Service-Level Agreements with state health departments) and for **internal performance management** (via Key Performance Indicators). The table illustrates examples of measurable targets, such as emergency department wait times, patient satisfaction, and staff-to-patient ratios, emphasising their importance in tracking trends, ensuring compliance, and driving improvement initiatives.

Table 6.1: Applications of metric management reports in Australian healthcare for external accountability and internal performance monitoring

Usage	
For external groups	For internal management
Service-Level Agreements (SLAs) Australian state health departments (e.g., NSW Health) often have SLAs with hospitals. These agreements specify performance targets such as: • Maximum emergency department wait times. • Patient admission times for elective surgeries (e.g., within 30 days for urgent cases).	Key Performance Indicators (KPIs) Internal management at hospitals uses KPIs to monitor operational performance, such as: • Average length of patient stays. • Staff-to-patient ratios. • Patient satisfaction scores.

Scenario in practice

A hospital in Sydney, **SouthernCross Hospital**, might generate weekly metric management reports that detail:

- Monitoring patient flow to ensure compliance with set targets like the 4-hour rule for emergency care.
- Tracking the percentage of patients who return for treatment, which can indicate the quality of care and discharge planning.
- Ensuring that there are adequate staff members available to meet patient needs and maintain operational efficiency.
- As part of a Six Sigma initiative, identifying and reducing errors or complications in

surgeries to improve outcomes and achieve near-perfect performance.

- **Under TQM**, monitoring hospital-acquired infections to maintain high standards of hygiene and patient care aligns with continuous improvement goals.

6.1.1.2 Dashboard-type reports

Dashboard-type reports have become a popular approach in business reporting because they provide a visual overview of performance indicators on a single screen, much like a car's dashboard displaying key metrics. They allow businesses to track progress in real-time, identify problem areas and make informed decisions quickly. Dashboards typically combine predefined elements with customisable widgets, allowing organisations to tailor their dashboards to specific needs. Colour-coded indicators (e.g., red, orange, green) are often used to signal performance levels and draw attention to critical areas.

Australian example: Retail sector

A grocery retail chain might use a dashboard-type report to monitor store performance across various locations in real time. The dashboard may have:

Customisable widgets: To display key metrics such as daily sales, inventory levels, customer footfall and employee shift coverage, with the ability for store managers to customise the view to focus on specific regions or categories (e.g., fresh produce vs. packaged goods).

Colour-coded indicators: *Green* indicates that store sales are meeting or exceeding targets, *orange* signifies sales are slightly below target but within acceptable variance, and *red* highlights sales that are significantly below target, requiring immediate action.

Real-time updates: The dashboard provides live data feeds, so managers can see changes in sales or inventory levels and adjust staffing or promotional efforts accordingly.

Scenario in practice

At the end of each day, the regional manager for a retail chain such as Coles or Woolworths reviews the dashboard. One store in Sydney is flagged red for low sales performance in the fresh produce section. Upon drilling down into the data, the manager notices a stockout issue due to delayed deliveries. The manager uses the dashboard to:

- Reallocate inventory from nearby stores (tracked via the same dashboard).
- Assign additional staff to prioritise restocking.
- Monitor changes in real time to ensure the issue is resolved promptly.

6.1.1.3 Balanced scorecard-type reports

Balanced scorecard-type reports, developed by Kaplan and Norton in the early 1990s², provide an integrated view of organisational success by measuring performance across four key perspectives: *financial performance*, *customer satisfaction*, *business processes*, and *learning and growth*. This approach ensures that organisations balance their focus between short-term financial outcomes and long-term strategic goals.

Australian example: Higher education sector

An Australian university might use a balanced scorecard to monitor and improve its performance across these four perspectives:

Financial performance:

- **Metrics:** Revenue from tuition fees, research grants and alumni donations.
- **Example:** The university tracks income generated from international students to ensure financial sustainability while managing cost efficiencies in operations.

Customer (student) satisfaction:

- **Metrics:** Student satisfaction survey results, graduate employability rates and retention rates.
- **Example:** The university measures satisfaction scores from student surveys to assess the quality of teaching, facilities and campus experience.

Internal business processes:

- **Metrics:** Time taken to process admissions, efficiency of course scheduling and research output quality.
- **Example:** The university evaluates how quickly it processes student applications or publishes research papers in high-impact journals to streamline internal operations.

Learning and growth:

- **Metrics:** Staff development programs, faculty qualifications and innovation in teaching methods.
- **Example:** The university tracks professional development initiatives for academic staff

2. First introduced in their 1992 Harvard Business Review article, “The Balanced Scorecard – Measures That Drive Performance”. The article can be accessed at: <https://hbr.org/1992/01/the-balanced-scorecard-measures-that-drive-performance-2>.

and invests in digital learning tools to enhance teaching quality.

Scenario in practice

At the end of each month, the senior academic leaders, such as the Provost and Senior Vice-President, the faculty Associate Deans (Teaching and Learning) and the Heads of Schools at the University of Technology Sydney (UTS), review the balanced scorecard dashboard. For example, the learning and growth perspective might highlight a low faculty satisfaction score, flagged in red for the School of Computer Science.

Drilling into the data, the senior academic leaders discover that faculty members are dissatisfied due to limited access to professional development programs and outdated teaching resources. Using the balanced scorecard, the leadership team takes the following actions:

- Allocate additional funding to the School of Computer Science for professional development workshops and training opportunities.
- Engage with faculty members to identify specific resources or tools needed to enhance teaching effectiveness.
- Set measurable improvement goals for faculty satisfaction and teaching innovation, monitoring progress on a monthly basis via the dashboard.

6.2 Information dashboards

An **information dashboard** is a **visual interface** designed to provide a concise and organised display of KPIs, metrics and data points, often in real time, to enable users to monitor and analyse performance at a glance.

Dashboards can be either **static**, presenting fixed data snapshots, or **dynamic**, offering real-time updates and interactive features. Over time, dashboard tools have evolved to become more visually appealing and user-friendly, allowing a wider range of users to create high-quality outputs. Popular tools include Excel, Tableau, Power BI and SAS Visual Analytics.

6.2.1 Designing effective dashboards: capturing attention and delivering insight

With a human attention span shorter than that of a goldfish – roughly eight seconds – designing a dashboard that captures and retains the viewer’s attention is important. In today’s data-driven world, all

insights compete for user focus, and dashboards must provide an immediate and compelling reason for users to stay engaged.

The best dashboards are **layered**, allowing users to navigate through multiple tabs or pages linked together. These layers offer context and guide the audience seamlessly through their questions. Actionable insight is important, meaning that every click or tap should give more details or a new view while keeping the bigger picture clear. The transition from highlights to detailed views should be effortless and intuitive.

Designers must also consider how interactive or static a dashboard needs to be, based on its intended use. Static dashboards are often used for things like posters or social media posts. They need to be simple and easy to understand. Interactive dashboards, on the other hand, let users explore the data in more detail, allowing them to click and see different views or go deeper into the information.

6.2.2 Best practices in dashboard design

6.2.2.1 Planning and preparation

- **Know your audience. What do they need?**

Understand who will use the dashboard (e.g., executives, analysts or general staff) and tailor the design to their level of expertise and information needs. For example, executives may require high-level summaries, while analysts need detailed insights.

- **What is the main idea you want to convey to your audience?**

Ensure your dashboard communicates a single, unified story. Each visualisation or section should support the main purpose and avoid conveying conflicting or unrelated messages, which can confuse users.

- **Sketch on paper first.**

Create an initial draft of your dashboard on paper. Select appropriate visualisations (e.g., bar charts, line graphs or pie charts) and outline the structure. This low-cost iteration process allows for easy revisions before moving to a digital platform.

- **Iterate frequently.**

Revisit and refine your dashboard design based on user feedback to ensure it meets the audience's evolving needs.

6.2.2.2 Data

The real estate saying “location, location, location” emphasises the importance of property location.

Similarly, for dashboards, the mantra is “data, data, data” – reliable data is the foundation of any effective dashboard.

- **Validate your data.**

Even if your dashboard’s appearance looks professional and follows accepted design standards, the data’s quality determines its usefulness. Always ask: is the data reliable? Is it up to date (timely)? Is any critical data missing? Is the data consistent across dashboards and other reporting tools?

- **Apply filters and interactivity.**

Allow users to interact with the data through filters, drill-downs or hover-over tooltips to reduce the cognitive load and provide tailored insights. This feature makes the dashboard dynamic and user-centric.

- **Choose the right visualisations.**

Match the type of data to the visualisation method that best conveys the message. For example: Use line charts for trends over time, use bar charts for comparisons, and avoid pie charts when comparing many categories as they can be harder to read.

- **Limit the amount of data per visualisation.**

Too much data in a single chart can overwhelm users. Focus on showing only the most relevant data points to maintain clarity and effectiveness.

6.2.2.3 Layout

- **Design a clear layout.**

Use an approach such as the Z-pattern, golden triangle or emphasis quadrant to guide the viewer’s eye through the dashboard logically and efficiently. These design principles help users process information faster.

- **Embrace white space (negative space).**

Negative space between elements such as charts, graphs and text enhances balance, organisation and readability. It ensures the dashboard isn’t visually cluttered, making it easier for users to focus on key insights.

- **Group related elements together.**

Place related data visualisations and text close to each other to help users understand the connections. For example, group KPIs at the top and detailed charts below.

6.2.2.4 Style

- **Use colours with care.**

Colours have different meanings in various cultures, so be mindful of your audience's cultural context. For example, red may signal danger or loss in some cultures, but it can represent luck in others.

- **Keep colours consistent.**

Use a consistent colour scheme across the dashboard to create a cohesive look. For example, always use green for positive trends and red for negative trends to avoid confusion.

- **Avoid overusing colours.**

Too many colours can overwhelm users and distract them from the data. Stick to a limited palette of 3–5 complementary colours to maintain simplicity and focus.

- **Pay attention to font choice and alignment.**

Use clear, legible fonts that align with your organisation's branding guidelines. Maintain consistent font sizes, weights and alignment across the dashboard for a polished and professional appearance. Avoid decorative fonts that may reduce readability.

6.2.2.5 Additional suggestions for clarity

- **Test the dashboard.**

Before deploying, test the dashboard with a sample audience to ensure it is intuitive, accurate and meets user needs.

- **Provide a clear title and labels.**

Every chart, graph or KPI should have a clear title and well-labelled axes or data points to avoid ambiguity.

- **Ensure accessibility.**

Consider accessibility features such as colour-blind-friendly palettes, scalable fonts and keyboard navigation to make the dashboard usable for everyone.

6.3 Data storytelling

Effective data storytelling transforms raw data into meaningful narratives that inform and engage audiences. Key recommendations include:

6.3.1 Define your story's purpose

Determine whether your story aims to prompt action, present a narrative or build a case, and structure it appropriately by deciding whether to lead with conclusions followed by supporting data, particularly for time-constrained audiences.

6.3.2 Choose the type of story you want to tell

Table 6.2 provides an overview of various data story types, detailing their primary functions, examples of application, and the key discussions they are designed to initiate, offering valuable insights into how data can be effectively used to drive meaningful conversations and decisions.

Table 6.2: Types of data stories

Data Story Type	What it does?	Examples	Discussion it starts
Change over time	Uses a chronology to illustrate a trend, emphasising how metrics, behaviors, or events evolve over a given period.	- Showing monthly sales trends over the past five years. - Analysing temperature changes across decades to highlight climate change.	- Why did this change happen during this specific period? - Are the changes seasonal, cyclical, or irregular? - What strategies can we implement to sustain or reverse these trends? - How might future trends look if current patterns continue?
Drill down	Sets context so that your audience better understands what's happening within a specific category or subset.	- Breaking down a company's revenue by product line or region. - Comparing customer satisfaction scores across different age groups.	- Why does this specific category perform differently? - Are there common traits within this subset that explain its behaviour? - How does performance within this category compare to the overall average? - Should resources be allocated differently to improve or leverage this category?
Zoom out	Describes how a specific subject relates to the bigger picture, helping your audience contextualise local or individual phenomena.	- Showing how a region's GDP growth compares to global economic trends. - Relating a company's market share to the overall size of the industry.	- How does this subject fit into or impact the broader context? - Are there external forces (e.g., industry trends, macroeconomic factors) driving these results? - What role does this area play in shaping the overall picture?

Contrast	Highlights the differences between two or more subjects, offering insights into why one outperforms or underperforms relative to another.	• Comparing customer churn rates for two subscription models. • Analysing demographic differences in political voting patterns.	• What factors explain the differences between these subjects? • Can the strengths of one be applied to improve the other? • Should efforts focus on maintaining high performance in one area or closing the gap in another?
Intersections	Highlights key moments or shifts when one category overtakes another, often signaling an important inflection point.	• Charting the point when mobile device usage surpassed desktop usage. • Showing when renewable energy capacity overtook coal-powered capacity.	• What caused these shifts or intersections to occur? • Are these shifts temporary or indicative of long-term trends? • How might this shift impact related systems or strategies? • Are there opportunities or risks associated with these transitions?
Factors	Breaks down a subject into types or categories to explain its components and their contributions to the whole.	• Disaggregating total customer acquisition cost into paid ads, events, and organic leads. • Categorising employee turnover rates by department or job level.	• Which categories have the greatest impact on the overall metric? • Is there a category that's overperforming or underperforming? • Should specific types or categories receive more focus or resources?

Outliers	Focuses on anomalies or areas where something is exceptionally different, either positively or negatively.	- Identifying a region with unusually high sales compared to others. - Highlighting a specific product that vastly outperformed or underperformed expectations.	- What makes this item or data point different from the norm? - Is this outlier indicative of a unique opportunity, a red flag, or an error? - Should this anomaly influence broader strategies or future plans? - Are there lessons to be learned from this exception that can be applied elsewhere?

6.4 Case study from the South Australia (SA) Health Emergency Department dashboard

Figure 6.1 shows the [SA Health Emergency Department \(ED\) dashboard](#), which provides near real-time updates on the current status of all public emergency hospital departments in South Australia. The dashboard displays key information such as current waiting times, triage categories of patients, expected arrivals and departures, and a rolling 24-hour summary of patient presentations and discharges across all emergency departments. The data is refreshed every 30 minutes. While the dashboard may lack visual appeal, it serves as a strong example of how public services can use near real-time data to enhance accountability, transparency and consumer choice. Features such as colour coding (yellow and red) allow users to quickly identify areas of concern.

Below is a detailed explanation of the data shown in this dashboard, how it provides insights into patient flow and ED operations, and why these insights are important for resource management and decision-making.

Status table (top left): This table provides a real-time key statistic for multiple hospitals, including:

- **Hosp:** Abbreviation for the hospital name.
- **EA (Expected Arrivals):** Number of anticipated arrivals, including ambulances and other notifications.
- **WTBS (Waiting to be Seen):** Patients currently waiting for medical attention in the ED.
- **Com Treat (Commenced Treatment):** Patients currently undergoing treatment in the ED.
- **CAP (Capacity):** Total available treatment areas within the ED.
- *** Alert:** Current ED status, calculated by comparing patient numbers with total capacity. The normal state is blank. Where there is one or more patients with a wait time over the defined threshold, a red alert icon is displayed.

- **Avg wait (Average wait time):** Average waiting time for patients in the ED (in minutes).

Additional note: The table uses colour-coded circles (green, amber, red and white) to indicate patient load.

- Green: Less than 80% capacity.
- Amber: 80%–95% capacity.
- Red: 95%–125% capacity.
- White: Above 125% capacity.

Streams table (top right): This table breaks down patient data by streams (care categories) for the selected hospital, including:

- **Streams:** Assigned based on patient needs (e.g., Acute, Ambulatory, Mental Health, Pediatrics, Unclassified).
- **Wait to See:** Time patients wait after triage to see a doctor.
- **Tot (Total):** The total number of patients “waiting to be seen” for the specific stream at the date and time the data was last updated.
- **WOT (Waiting Over Threshold):** Patients exceeding the clinically recommended wait time for their urgency category (e.g., Category 1 is for resuscitation cases that require immediate attention).
- *** Alert:** Indicates that one or more patients have exceeded the wait time threshold for their assigned triage category. In the normal state, this field is blank. When a patient surpasses the defined threshold, a red alert icon is displayed to highlight the delay. For example, for general patients – amber for >8 hours, red for >12 hours and for mental health patients – amber for >18 hours, red for >24 hours.
- **Treat (Being treated):** Patients currently receiving treatment, without a decision on admission.
- **Wait for Bed:** Time between admission decision and transfer to an inpatient ward.
- **Tot (Total):** The total number of patients with a current status of “waiting for a bed” at the time the data was last updated.
- **WOT (Waiting Over Threshold):** Refers to the number of “waiting for a bed” patients who have exceeded the clinically defined wait time threshold for admission.
- *** Alert:** Same description as in the “Wait to See” column.
- **Total Time in ED:** Total time spent in ED, categorised into intervals (e.g., 0–2 hours, 2–3 hours, 3–4 hours, 4–8 hours and 8+ hours).
- **EECU (Extended Emergency Care Unit):** The total number of patients currently admitted to the EECU within the Emergency Department. This unit is specifically designed for patients who require monitoring for up to 24 hours before being discharged or transferred. Patients moved to the EECU typically need extended care that exceeds standard emergency department stays but does not yet require inpatient ward admission.
- **Tot (Total):** The total number of patients currently assigned to this specific stream within the ED at the time the data was last updated.

Expected Arrivals table: This table tracks anticipated arrivals to help ED staff plan resources, including:

- **Ambulance Categories (AMB 1 to AMB 6):** These reflect urgency levels, from life-threatening cases (AMB 1) to non-urgent monitoring cases (AMB 6).

Departures (Last 1 Hour) table: This table summarises departures from the ED in the last hour, categorised by reason (e.g., admitted, discharged, transferred or other). Data updates every 30 minutes and synchronises with the 24-hour activity graph.

Triage Categories table: This table displays patient distribution by triage categories, with urgency levels ranging from:

- **Cat (Category):** Represents the triage category, numbered from 1 to 5, based on urgency: 1 is Resuscitation (requires immediate attention), 2 is Emergency, 3 is Urgent, 4 is Semi-urgent, and 5 is Non-urgent.
- **WTS (Wait to See):** Same description as in the Streams table.
- **WOT (Waiting Over Threshold):** The number of patients in the WTS status who have exceeded the nationally defined wait time threshold for their triage category.
- *** Alert:** Indicates that one or more patients in the triage category are **waiting beyond the threshold time**. The normal state is blank, but a red alert icon appears when the threshold is exceeded.
- **Oth (Other):** The total number of patients in the triage category who do not have a status of WTS.
- **Tot (Total):** Total number of patients currently in ED for the given triage category, including both WTS and Other statuses.

Activity in the Last 24 Hours (Graph): This bar chart visualises patient flow over the past 24 hours, showing:

- Arrivals (ARR) and Departures (DEP) per hour.
- Patients Waiting to Be Seen (WTBS) and Commenced Treatment (Com Treat).

Colour coding (red, amber, green) highlights fluctuations in patient activity and helps identify peak hours.

Waiting Times table and graph: This component shows a breakdown of patient wait times across key ED stages (e.g., 0–2 hours, 2–4 hours, 4–8 hours, 8–12 hours, 12–24 hours, and 24+ hours) as a table and colour-coded graph. It highlights bottlenecks in the patient journey. The Tot (Total) column shows the total number of patients per waiting category.

Percentage Admissions (% Admissions): This bar graph shows the percentage of patients admitted to the hospital out of the total number of ED departures.

LOS (Length of Stay) Under 4 Hours: These bar graphs track the percentage of ED patients discharged

or treated within 4 hours, with a target of 60%. One graph covers all patients, and the other focuses on discharged patients.

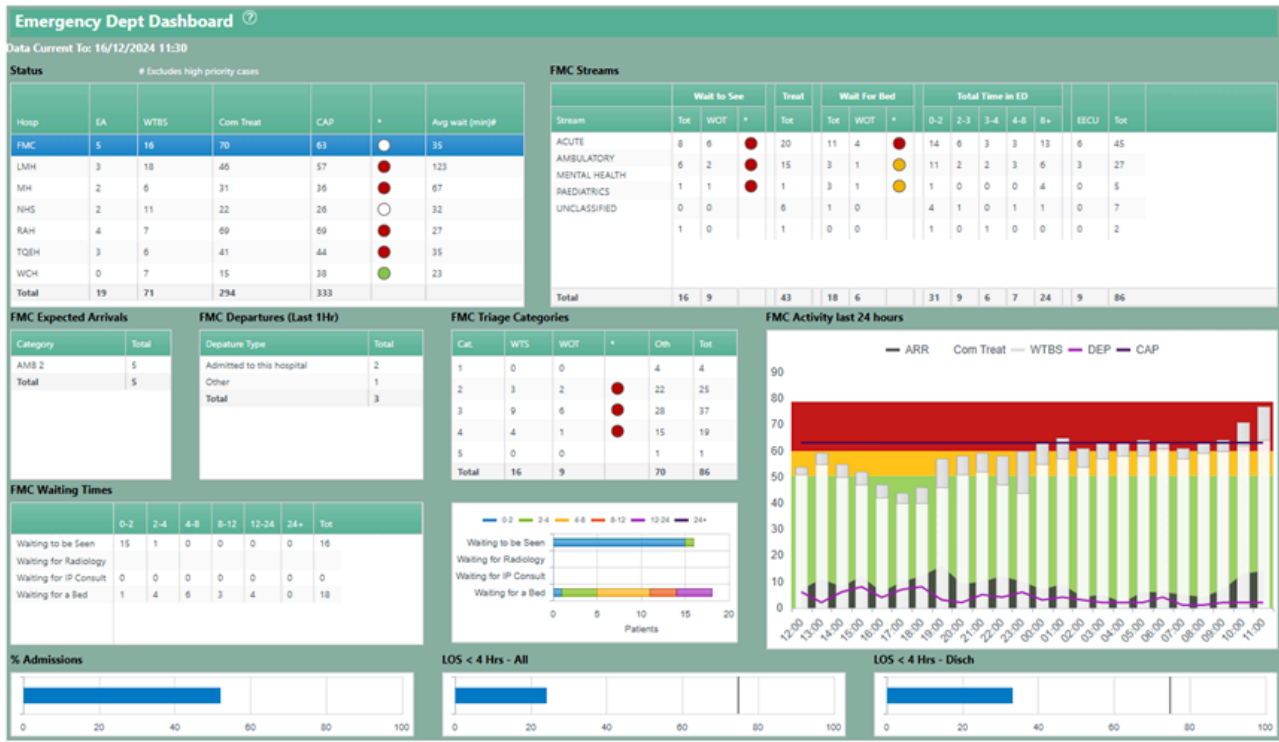


Figure 6.1: South Australia (SA) Health Emergency Department dashboard from the Government of SA Health website

Alternative Text for Figure 6.1

Emergency Department Dashboard – Real-Time Status Overview

This dashboard provides a **real-time snapshot** of emergency department (ED) performance across multiple hospitals. It includes key indicators such as **waiting times**, **patient flow**, **triage categories**, and **resource availability**.

Key Sections:

- **Status Table (Top Left):** Displays patient counts for each hospital, including expected arrivals, patients waiting to be seen, commenced treatments, and overall capacity, with **colour-coded alerts** indicating congestion levels.
- **FMC Streams (Top Right):** Categorises patients by care streams (Acute, Ambulatory, Mental Health, Pediatrics, Unclassified) and tracks waiting times, treatment progress, and bed allocation.
- **Expected Arrivals & Departures (Middle Left):** Shows upcoming ambulance arrivals and recent ED departures, helping monitor patient flow.
- **Triage Categories (Middle Right):** Displays patient distribution across **triage**

levels 1-5, highlighting critical patients exceeding wait time thresholds with **red alerts**.

- **Waiting Times Breakdown (Bottom Left)**: Summarises wait durations for different patient statuses, including waiting for treatment, radiology, or a bed.
- **24-Hour Activity Graph (Bottom Right)**: A stacked bar chart visualising **trends in patient arrivals, departures, and ongoing treatments** over the past 24 hours, with **colour-coded congestion zones** (red, yellow, and green).
- **Performance Metrics (Bottom Row)**: Includes **% Admissions** and **Length of Stay (LOS) Under 4 Hours**, measuring ED efficiency in admitting and discharging patients within the recommended timeframe.

This dashboard helps **hospital administrators, clinicians, and decision-makers** monitor ED performance, identify bottlenecks, and optimise resource allocation in real time.

6.5 Test your knowledge

Drag the words into the correct boxes.

1) An is management tool that allows information, whether it be numerical or text, to be represented in a visual dashboard.

2) Information dashboards provide particular information that businesses need in order to run

and .

3) Many of the information visualisation vendors are adding the capabilities to call themselves

solution providers.

4) Dashboards provide visual displays of important information that is consolidated and arranged on a single

.

5) A data story of type

explains how a specific subject fits into a larger context or relates to the bigger picture.

screen

efficiently

Zoom Out

information dashboard

visual analytics

effectively



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=333#h5p-4>

An alternative source for this activity: <https://uts.h5p.com/content/1292858509721476829>

Additional readings

- McCandless, D. (n.d.). *Colours in culture*. Information Is Beautiful. Retrieved January 28, 2025, from <https://informationisbeautiful.net/visualizations/colours-in-cultures/>
- SAS. (n.d.). *Business reports and dashboards*. Retrieved January 28, 2025, from <https://communities.sas.com/html/assets/breports/index.html>
- SAS. (2018). *Are your colour choices ruining your reports?* SAS Global Forum 2018. Retrieved from <https://support.sas.com/resources/papers/proceedings18/2608-2018.pdf>
- Tableau. (n.d.). *Dashboards in Tableau Desktop*. Retrieved January 28, 2025, from <https://help.tableau.com/current/pro/desktop/en-us/dashboards.htm>
- Zencos. (n.d.). *Data visualization: Color and cultural influence*. Retrieved January 28, 2025, from <https://www.zencos.com/blog/data-visualization-color-culture-influence>

About the author



Dr Amara Atif
UNIVERSITY OF TECHNOLOGY SYDNEY
<https://profiles.uts.edu.au/Amara.Atif>

Dr Amara Atif is a Senior Lecturer in the School of Computer Science, Faculty of Engineering and IT (FEIT), at the University of Technology Sydney (UTS), Australia. Amara is an expert in educational technology, with a comprehensive skill set that includes research and development in the design of technology-enhanced learning environments, human-computer interaction, artificial intelligence in education, and learning analytics/educational data mining.

At UTS, she developed the undergraduate subject “Business Intelligence”, integrating culturally informed Indigenous knowledge systems through the FEIT Indigenous Graduate Attribute. She has received multiple UTS Vice Chancellor’s Learning and Teaching Awards, including recognition for data-informed teaching practices and inclusive assessment. Passionate about the accessibility and sustainability of educational resources, Amara is an advocate for Open Educational Resources (OER). *Decoding Data* continues this mission, bringing rigorous Business Intelligence concepts to learners through an open, practitioner-oriented resource grounded in Australian business contexts.

7. PREDICTIVE ANALYTICS

Marcella Papini

Learning Objectives (LO)

By the end of this chapter, you will be able to:

- LO1** Understand the fundamental concepts and purpose of predictive analytics in modern business contexts.
- LO2** Distinguish between key predictive analytics techniques, including data mining, text mining, web analytics, and social media analytics.
- LO3** Explore practical applications of predictive analytics across various industries, such as finance, healthcare, and retail.
- LO4** Identify the challenges and best practices in implementing predictive analytics solutions, with a focus on data quality, scalability, and ethical considerations.
- LO5** Apply your understanding of predictive analytics to design data-driven strategies for anticipating and addressing organisational challenges.

7 Introduction to predictive analytics

Few technologies have the potential to transform the way businesses operate as profoundly as predictive analytics. This forward-looking, data-driven technology leverages historical data, statistical algorithms, and machine learning to forecast future outcomes and trends. At its core, predictive analytics seeks to answer the crucial business question: “*What will happen?*” By uncovering patterns within vast datasets, predictive analytics empowers organisations to make smarter, proactive decisions. This technology has become indispensable across industries such as finance, healthcare, retail, and beyond.

Predictive analytics isn’t just for large corporations, its applications scale to businesses of all sizes. Small and medium-sized enterprises (SMEs) can use predictive analytics to streamline operations, personalise customer experiences, and make data-backed decisions without requiring the extensive infrastructure of larger firms. Meanwhile, for larger organisations, predictive analytics serves as a strategic tool to navigate complex markets, optimise resources, and manage risks effectively.

In this chapter, we delve into the types, techniques, and applications of predictive analytics. You will explore key methods like data mining (section 7.2), text mining (section 7.3), web analytics (section 7.4), and social media analytics (section 7.5), each tailored to address specific types of data and business questions. To illustrate the real-world potential of predictive analytics, this chapter also includes examples that demonstrate how organisations have successfully used these techniques to solve practical problems.

7.1 Types of predictive analytics

Algorithms for predictive modelling are often divided into two groups: **supervised learning methods** and **unsupervised learning methods**.

Supervised learning methods work by training a model on historical data with known outcomes, often called “**labelled data**”.

Labelled data means each example in the dataset has been tagged with the correct answer or category, helping the model learn what it needs to predict. This labelled information helps the model learn patterns that can apply to new data.

Common supervised methods include **classification** and **regression**, which are explored in Section 7.2.

Unsupervised learning methods do not learn from historical data with known labels, and so, they perform without any supervision. These models look for natural patterns and connections in the data on their own.

Standard unsupervised techniques include **clustering** and **association rule learning**, discussed in Section 7.2.2.3 and Section 7.2.2.4.

Key Takeaway

While supervised learning is about making specific predictions, unsupervised learning helps uncover patterns, giving businesses insights.

7.2 Data mining

Data mining is the process of uncovering hidden patterns, trends, or relationships within large datasets. In today's world, data is everywhere, websites track user clicks, sensors monitor weather conditions, businesses record every transaction, and social media platforms log interactions. However, raw data by itself doesn't hold much value. Data mining transforms this avalanche of information into actionable knowledge, enabling smarter decisions, improved efficiency, and better predictions about the future.

A clear and accessible overview of these concepts is presented in the YouTube video “[What is Data Mining?](#)” by IBM Technology. This video provides a concise explanation of how data mining works and highlights its importance in modern data-driven decision-making.

In this chapter we present some of the most widely used data mining techniques. These techniques have unique applications and play an essential role in tackling different data challenges. The following sections will explore each of these techniques in more detail. But before diving into them, it's important to understand how the process of data mining works.

7.2.1 How data mining works?

A common framework for guiding data mining process is **Cross-Industry Standard Process for Data Mining (CRISP-DM)**. CRISP-DM provides a structured and systematic approach to turning raw data into useful insights and is applicable across many industries and problem domains.

The data mining process typically involves working with data through a series of **iterative steps**, beginning with understanding the business problem and collecting relevant data, and progressing through data preparation, modelling, and evaluation. The final stage involves deploying the resulting model or insights so they can be used in real-world decision-making.

Under the **CRISP-DM framework**, the data mining process consists of **six** main phases:

1. **Business understanding** – defining the problem and objectives from a business perspective
2. **Data understanding** – collecting and exploring the data to assess its quality and relevance
3. **Data preparation** – cleaning, transforming, and selecting data for modelling
4. **Modelling** – applying data mining techniques to build predictive or descriptive models
5. **Evaluation** – assessing model performance and ensuring it meets business goals
6. **Deployment** – implementing the model or insights into operational use

Key Takeaways

It is important to note that CRISP-DM is **not a linear process**. Insights gained at later stages may require revisiting earlier steps, such as refining the business objective or preparing additional data. This iterative nature helps ensure that data mining outcomes remain aligned with organisational needs and practical constraints.

7.2.2 Data mining techniques

Data mining involves a range of techniques. Each technique addresses specific types of data analysis problems. In the following sections, we'll explore four major data mining techniques: ***Classification***, ***Regression***, ***Clustering***, and ***Association Rule Learning***. Each technique is explained in detail, including its applications, strengths, and limitations, to give a complete picture of how data mining works in practice.

7.2.2.1 Classification

Classification involves building a model that distinguishes and describes data classes or categories based on a set of labelled data. This process analyses training data to construct a model that can then be used to predict the class label of new, unlabelled data.

Classification algorithms typically follow three main phases: ***training***, ***testing***, and ***deployment***. The training and testing phases both rely on labelled data.

- During the **training phase**, a subset of the labelled data is used to develop a model that aims to classify instances accurately.
- In the **testing phase**, the trained model is applied to the remaining labelled data, and its predicted class labels are compared with the actual labels to evaluate performance. This evaluation helps determine whether the model is sufficiently accurate or requires refinement.
- Once a model demonstrates satisfactory performance during testing, it is **deployed** and used to classify new, unlabelled data in real-world applications.

Although all classification methods share the same objective, i.e., predicting the category of new data based on patterns learned from historical data, there are multiple techniques for achieving this.

The following sections describe some of the most commonly used classification methods.

7.2.2.1.1 Decision trees

A decision tree is a graphical representation of a decision-making process, with **nodes** representing data features and **branches** showing decisions based on those features. Decision trees are a popular technique for classification in data mining because they are straightforward and easy to understand. The way they work is by breaking down a dataset into smaller groups based on certain features. Decision trees are especially useful for datasets with mixed types of data, like numbers and categories, and are used in areas that benefit from easy-to-interpret results, such as finance and healthcare.

Decision trees use several key terms to describe their structure. Figure 7.1 illustrates a **decision tree model** used to predict the binary outcome **PLAY** versus **DON'T PLAY** based on weather-related attributes. The root node shows the overall class distribution (9 “Play” and 5 “Don’t Play”). The first split is on **Outlook**, which branches into *sunny*, *overcast*, and *rain*.

- When the outlook is **overcast**, all instances result in *Play*, making this a pure leaf node.
- For **sunny** conditions, the decision depends on **Humidity**: values ≤ 70 lead to *Play*, while values > 70 lead to *Don't Play*.
- For **rainy** conditions, the decision depends on **Windy**: if it is *TRUE*, the outcome is *Don't Play*; if *FALSE*, the outcome is *Play*.

Each node displays the number of instances classified as *Play* and *Don't Play*, demonstrating how the dataset is progressively partitioned until a final decision is reached.

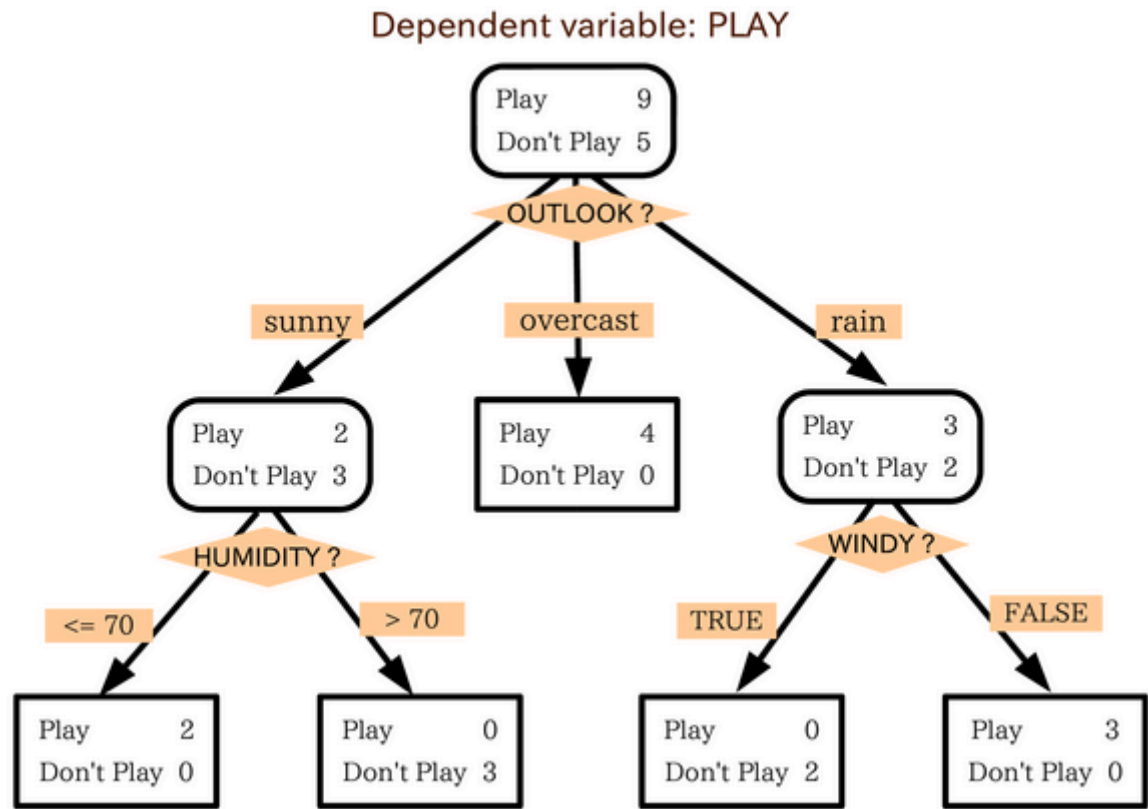


Figure 7.1: Decision tree illustrating the classification of the target variable **PLAY based on weather attributes including outlook, humidity, and wind conditions.**
Attribution: T-kita at English Wikipedia, Public domain, via Wikimedia Commons

Alternative Text for Figure 7.1

Decision tree diagram predicting whether to play based on weather conditions. The root node splits by outlook (sunny, overcast, rain), with further splits on humidity for sunny conditions and wind for rainy conditions, showing counts of play versus don't play at each node.

Table 7.1 summarises the key advantages and disadvantages of decision trees used in classification.

Table 7.1: Advantages and disadvantages of decision tree classification

Advantages	Disadvantages
Easy to understand and interpret.	Prone to overfitting.
Handles both numerical and categorical data.	Sensitive to small data changes.
Requires minimal data preparation.	Biased towards dominant classes.
Highlights important features.	Less accurate alone.
Scales well with large datasets.	Struggles with complex relationships.

Decision trees are widely used in business to support **classification tasks**, such as predicting customer churn. By analysing factors including **contract type, usage patterns, billing history, service quality, and customer support interactions**, a decision tree model can classify customers into groups such as **“likely to churn”** or **“unlikely to churn.”** This enables organisations to take proactive retention actions before customers decide to leave.

A hypothetical telecommunications business scenario: AustraiTel

A hypothetical Australian telecommunications provider, **AustraiTel**, identifies an increase in customer cancellations in several regional markets. To address this issue, the company applies a decision tree model to its customer data. The model indicates that churn risk is highest among customers on **short-term contracts** who experience **frequent service disruptions** and have made **multiple support complaints**.

Based on these insights, AustraiTel implements targeted retention actions for high-risk customers, including **temporary bill credits, additional data allowances, and priority technical support**. In this scenario, the decision tree supports prescriptive analytics by helping the organisation decide **which customers should receive intervention and what actions should be taken** to reduce churn and protect revenue.

7.2.2.1.2 Support Vector Machines (SVM)

SVMs are also a popular tool for classification tasks. They work by finding the best line (or *hyperplane*) that separates two classes of data in a way that leaves the widest possible gap between the closest points from each class. If there are only two features in the data, the hyperplane is a line, but if there are more, it can be a plane or a more complex boundary in higher dimensions. Since there are usually multiple ways to draw this boundary, SVMs choose the one with the largest margin between classes, helping the model generalise better to new data.

Figure 7.2 illustrates a maximum-margin classifier in a two-dimensional feature space (X_1 , X_2). The solid line represents the optimal separating hyperplane between two classes, while the dashed parallel lines indicate the margin defined by the nearest data points from each class. The classifier maximises this margin to achieve better generalisation.

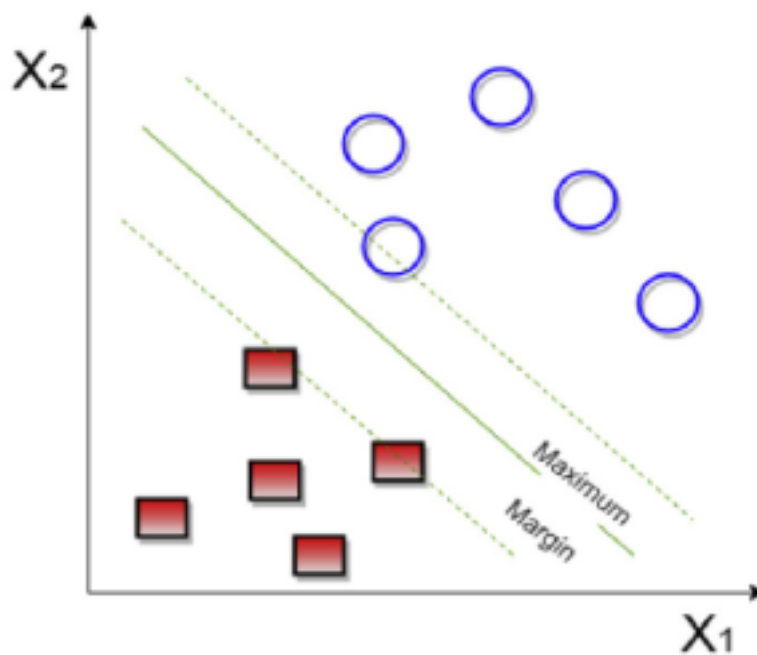


Figure 7.2: SVM showing the optimal separating hyperplane and the maximum margin between two linearly separable classes in a two-dimensional feature space. Attribution: Adrian Stetco a, *, Fateme Dinmohammadi b, Xingyu Zhao b, Valentin Robu b, David Flynn b, Mike Barnes c, John Keane a, Goran Nenadic a, CC0, via Wikimedia Commons

Alternative Text for Figure 7.2

Two-dimensional scatter plot with red squares and blue circles representing two classes. A diagonal separating line with parallel dashed lines shows the maximum margin between the classes along the X_1 and X_2 axes.

A concise visual explanation of this concept is provided in the YouTube video “[Support Vector Machine \(SVM\) in 2 minutes](#)” by Visually Explained, which clearly demonstrates how SVMs determine the optimal separating boundary.

The **kernel function** is essential in SVMs for tackling complex classification tasks where a simple linear separation isn't possible. The kernel in SVMs is a function that helps transform input data into a higher-dimensional space, making it possible to separate classes even when they aren't linearly separable in their original form.

This transformation allows SVMs to capture non-linear patterns and gives more flexibility in creating decision boundaries between classes.

Some common kernel types include the **linear kernel**, which is best for straightforward, linearly separable data, and the **polynomial kernel**, which can model more complex, non-linear relationships by creating curved decision boundaries.

Table 7.2 summarises the key advantages and disadvantages of SVM used in classification.

Table 7.2: Advantages and disadvantages of SVM classification

Advantages	Disadvantages
Performs well with high-dimensional data.	Computationally intensive.
Reduces overfitting risks.	Sensitive to kernel choice.
Adaptable using different kernels.	Hard to interpret results.

SVMs are widely used in the financial sector for **fraud detection**, where the objective is to classify transactions as **legitimate** or **fraudulent**. SVM models analyse multiple transaction attributes such as **transaction value**, **transaction frequency**, **geographic location**, **time of purchase**, and **device**

information to identify patterns that distinguish normal customer behaviour from suspicious activity. Because SVMs perform well with **high-dimensional data** and can separate classes even when relationships are complex and non-linear, they are particularly effective for detecting fraudulent transactions and reducing financial losses.

A hypothetical banking business scenario: MyBank

A hypothetical Australian bank, **MyBank**, experiences an increase in card-not-present fraud associated with online purchases. To address this issue, the bank applies a **SVM** model to analyse transaction data. The model evaluates features such as **unusual transaction amounts, rapid repeat purchases, overseas transaction locations, and unfamiliar devices** to identify suspicious behaviour.

Transactions classified as high risk are automatically **flagged for additional verification or temporarily blocked**. In this scenario, the SVM supports prescriptive analytics by helping MyBank decide **which transactions require intervention and what action should be taken**. As a result, the bank reduces fraud-related losses, improves customer trust, and maintains a smooth experience for legitimate customers.

7.2.2.1.3 K-Nearest Neighbours (k -NN)

k -NN is a simple, yet effective algorithm often used for classification and regression tasks. The basic idea behind k -NN is that it makes predictions based on the similarity of new data points to points in a known labelled dataset. k -NN is known as a “lazy learner” because it doesn’t build a model or learn parameters in advance. Instead, k -NN simply stores the entire training dataset and makes predictions by comparing new data points directly to this stored data.

A clear and intuitive overview of how this process works is presented in the YouTube video “[K Nearest Neighbors | Intuitive explained | Machine Learning Basics](#)” by Intuitive Machine Learning, which visually demonstrates how k -NN identifies neighbours and uses them to make predictions.

When using k -NN, there are two important choices to make: **the number of neighbours (k)** and **the distance metric**. Together, the choices of k and distance metric significantly influence k -NN’s performance and accuracy, so careful tuning is essential.

- **The number of neighbours (k)** determines how many nearby points the algorithm considers when making a prediction. With a smaller k (e.g., $k=1$), the model only looks at the single closest neighbour to make its prediction. This can make the model very responsive to small changes or “noise” in the data, as it relies on just one example. Noise includes any data points that don’t fit the

general trend or that are caused by factors outside the main focus of the analysis. On the other hand, a larger k means the model considers more neighbours (like the 5 or 10 closest points) and makes its prediction based on the majority, which smooths out the effect of any unusual points. However, using a larger k can sometimes make the predictions less precise because it averages across more points rather than focusing on just the closest one.

- **The distance metric** defines how similarity is measured between data points. Common metrics include Euclidean distance, which calculates the straight-line distance between points and works well for data on similar scales, and Manhattan distance, which measures distance along the feature axes and can suit grid-like data.

Table 7.3 summarises the key advantages and disadvantages of k -NN used in classification.

Table 7.3: Advantages and disadvantages of k -NN used in classification

Advantages	Disadvantages
Simple and easy to understand.	Slow with large datasets.
No training needed.	Impacted by irrelevant features.
Works with diverse data types.	Can be biased with imbalanced classes.

For anyone looking to get started with classifying data using k -NN in Python, [W3Schools provides a clear and hands-on tutorial](#). This guide covers everything you need, from setting up Python and loading your data to implementing k -NN with Scikit-Learn. It also dives into key aspects like how to select the best number of neighbours and measure distances between data points.

A hypothetical retail business scenario: ShopNow AU

A hypothetical Australian online retailer, **ShopNow AU**, uses prescriptive analytics to decide **what products should be recommended** to each customer in real time. When a customer logs in and browses items such as running shoes and sportswear, ShopNow AU applies a k -NN model to compare this customer's behaviour with that of similar customers.

The model identifies a group of customers with similar browsing and purchase histories who

previously bought **running socks and fitness trackers**. Based on this similarity, the prescriptive system **recommends these specific products** to the customer.

In this context, k -NN supports prescriptive analytics by moving beyond predicting customer interests to answering the question **“What should we recommend right now to maximise relevance and sales?”** The output of the model directly informs an action, displaying targeted product recommendations, thereby improving the customer experience and increasing the likelihood of additional purchases.

7.2.2.1.4 Artificial Neural Networks (ANNs)

ANNs are a powerful classification technique inspired by the way the human brain processes information. ANNs are made up of interconnected units called “neurons” or “nodes”, which are organised in layers. Typically, an ANN has an input layer for receiving data, one or more hidden layers for processing, and an output layer for producing predictions. When a new data point is input, it travels through each layer, with the neurons in each layer performing calculations to detect patterns in the data.

A concise introduction to these ideas is presented in the YouTube video “[Neural Networks Explained in 5 minutes](#)” by IBM Technology, which clearly illustrates how neural networks learn and make predictions.

Figure 7.3 presents a simplified schematic representation of a ANN. The circles represent individual neurons, arranged into distinct vertical layers. The leftmost column corresponds to the **input layer**, which receives the input features. The middle column represents a **hidden layer**, where intermediate computations and feature transformations take place. The rightmost column represents the **output layer**, which produces the final model output. Although the connections between neurons are not explicitly shown, the diagram illustrates the layered structure typical of neural networks, where information flows from the input layer through one or more hidden layers to the output layer.

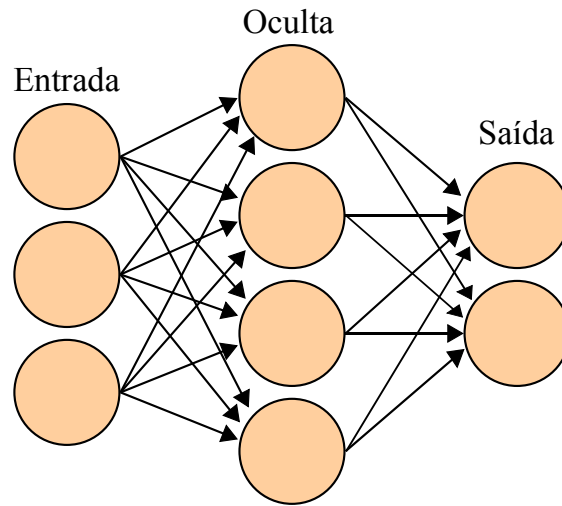


Figure 7.3: ANN architecture, illustrating input, hidden, and output layers.

Attribution: en>User:Cburnett, CC BY-SA 3.0
<http://creativecommons.org/licenses/by-sa/3.0/>, via Wikimedia Commons

Alternative Text for Figure 7.3

Simplified diagram of a neural network showing three vertical layers of circular nodes representing input, hidden, and output neurons.

Activation functions play a key role in ANNs by determining whether a neuron should be activated based on the input it receives, allowing the network to capture complex, non-linear relationships. There are several common types of activation functions, each with unique characteristics suited to different tasks.

- The sigmoid function squashes inputs into a range between 0 and 1, which is useful for binary classification tasks.
- The Rectified Linear Unit (ReLU) function has become one of the most popular activation functions, outputting zero for any negative input and the original input value for positive values.

Choosing the right activation function is essential, as it directly affects the network’s ability to learn and make accurate predictions.

Learning in ANNs refers to the process by which the network improves its accuracy over time by adjusting its internal parameters based on the data it processes. This learning is achieved through a method called **backpropagation**, which involves two key steps: **forward propagation** and **backward propagation**.

- In **forward propagation**, the input data moves through the network layers, with each layer applying a transformation and generating predictions. These predictions are then compared to the actual outputs from the labelled data, allowing the network to calculate the error.
- In **backward propagation**, this error is used to adjust the weights of the connections between neurons. The network updates these weights to reduce the error, aiming to make the model more accurate with each iteration. This iterative adjustment through backpropagation is the foundation of learning in ANNs.

Table 7.4 summarises the key advantages and disadvantages of ANNs used in classification.

Table 7.4: Advantages and disadvantages of ANNs used in classification

Advantages	Disadvantages
Captures complex patterns.	Often need a significant amount of labelled data to achieve good performance.
Scalable and flexible.	Training ANNs, especially deep networks, requires substantial computational resources, including powerful GPUs
Generalises well with enough data.	Difficult to interpret, making it hard to understand how they make predictions.

ANNs are widely used in **medical decision support**, particularly for disease risk assessment and diagnosis. ANNs analyse large volumes of patient data including **symptoms, medical history, laboratory results, and medical imaging** to learn complex, non-linear relationships that are difficult to capture using simpler models. These capabilities make ANNs well suited to classification tasks such as identifying patients at different levels of disease risk.

A hypothetical healthcare business scenario: HealthCare AU

An Australian healthcare provider, **HealthCare AU**, uses an ANN-based system to support early disease detection in hospital settings. The model is trained on historical patient records and learns patterns associated with specific medical conditions. When data from a new patient is entered, the ANN classifies the patient as **“high risk”** or **“low risk.”**

Based on this classification, the prescriptive analytics system recommends **specific clinical actions**, such as prioritising further diagnostic tests, scheduling specialist consultations, or initiating early treatment. In this context, the ANN supports prescriptive analytics by helping clinicians decide **what action should be taken next** to improve patient outcomes and allocate medical resources effectively.

7.2.2.2 Regression analysis

While classification predicts categories such as “yes/no” or “high/low,” regression predicts continuous numerical outcomes such as sales, prices, or demand.

Regression in data mining is a statistical technique used to model the relationship between a **dependent (response) variable** and one or more **independent (predictor) variables**. In practice, the terms *dependent* and *response*, as well as *independent* and *predictor*, are often used interchangeably.

A clear introductory explanation of this concept is provided in the YouTube video “[Regression Analysis | CLOSER Learning Hub](#)” by Closer, which outlines how regression models are built and how they help uncover relationships between variables.

Regression is widely used in applications such as financial forecasting, inventory management, and risk assessment. In data mining, several regression techniques exist, with the most common distinction being between **linear regression** and **multiple regression**. The key difference lies in the number of independent (predictor) variables used to make a prediction. Linear regression uses a single predictor, whereas multiple regression uses two or more predictors to model and explain the dependent (response) variable.

7.2.2.2.1 Linear regression

Linear regression is one of the simplest and most widely used regression techniques in data mining. It models the relationship between a single independent (predictor) variable and a dependent (response) variable by fitting a straight line to the data. This line, known as the **regression line**, represents the predicted value of the dependent variable for a given value of the independent variable.

The mathematical equation for simple linear regression is: $y = \beta_0 + \beta_1 x + \varepsilon$

where y is the dependent variable, x is the independent variable, β_0 is the intercept (the point where the line crosses the y -axis), β_1 is the slope (indicating how much y changes for a one-unit increase in x), and ε represents the error term, also known as the residual.

Figure 7.4 illustrates the concept of **prediction error in a regression model**. The blue points represent observed data points, while the green line indicates the model's predicted relationship between the input features and the target variable. For a given input value x_i , the model produces a predicted output $\hat{y}_i$, which may differ from the observed value y_i .

The vertical dashed line highlights the **prediction error**, defined as the difference between the observed value and the predicted value for that data point. This error is used during model training to update parameters and improve the model's fit to the data, typically by minimising an overall loss function.

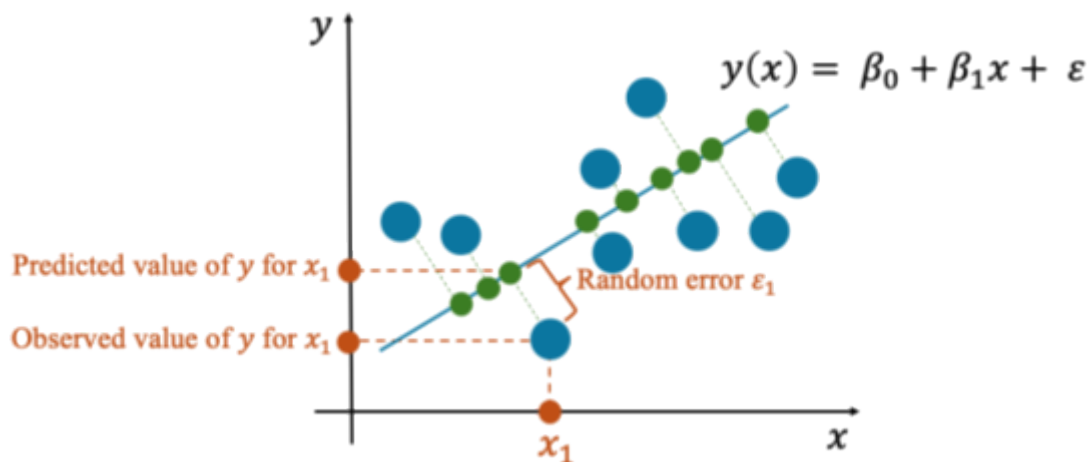


Figure 7.4: Regression model showing the difference between the observed value and the model's predicted value for a given input.

Alternative Text for Figure 7.4

Scatter plot with blue data points and a green regression line, showing the predicted value and observed value for a data point, with a dashed line indicating the prediction error.

Model evaluation metrics

To evaluate the performance of a regression model, several metrics are commonly used to measure how well the model fits the data and predicts outcomes. Two widely used metrics are **Mean Squared Error (MSE)** and **Root Mean Squared Error (RMSE)**. MSE is calculated as the average of the squared residuals, ensuring that positive and negative errors do not cancel each other out and placing greater emphasis on larger errors. However, because MSE is expressed in squared units, it can be difficult to interpret.

RMSE addresses this issue by taking the square root of MSE, returning the error measure to the same units as the dependent variable. A lower RMSE value indicates better model performance and provides an intuitive measure of average prediction error. RMSE is particularly useful when comparing different regression models.

Another important metric is **R-squared (R^2)**, also known as the coefficient of determination. R^2 measures the proportion of variance in the dependent variable that is explained by the independent variable. Its value ranges from 0 to 1, with higher values indicating a stronger explanatory power.

Table 7.5 summarises the key advantages and disadvantages of linear regression.

Table 7.5: Advantages and disadvantages of linear regression

Advantages	Disadvantages
Simple and easy to interpret.	Assumes a linear relationship between variables.
Works well when relationships are linear.	Sensitive to outliers.
Provides clear coefficient estimates.	Assumes constant variance of residuals.
Low computational cost.	

A hypothetical energy utilities business scenario: EnergyCo AU

A hypothetical Australian electricity provider, **EnergyCo AU**, aims to forecast daily energy demand to improve grid planning and resource allocation. To achieve this, the company uses a **linear regression model** to predict electricity consumption based on variables such as **daily temperature, historical electricity usage, and time of year**.

By analysing past data, the model learns how changes in temperature influence electricity demand, particularly during heatwaves and colder months. Using these demand forecasts, EnergyCo AU can plan electricity generation and distribution in advance, reducing the risk of power shortages while avoiding unnecessary overproduction.

In this scenario, linear regression supports prescriptive analytics by informing decisions about **how much electricity should be generated and when**, helping the provider optimise energy distribution, improve operational efficiency, and deliver a more stable and cost-effective energy supply.

While **linear regression** uses a single predictor variable, many real-world problems involve multiple factors influencing the outcome. In such cases, **multiple regression** is used to model the combined effect of several independent variables.

7.2.2.2 Multiple regression

Multiple regression is an extension of linear regression that models the relationship between a dependent (response) variable and **two or more independent (predictor) variables**. Instead of examining the effect of a single predictor, multiple regression allows analysts to understand how several factors collectively influence an outcome. This makes it more suitable for real-world problems, where outcomes are rarely driven by only one variable.

The mathematical equation for multiple linear regression is: $y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \dots + \beta_nx_n + \epsilon$

where y is the dependent variable, $x_1, x_2, \dots, x_n$ are the independent variables, β_0 is the intercept, $\beta_1, \beta_2, \dots, \beta_n$ are the coefficients representing the contribution of each predictor, and ϵ is the error term (residual). Each coefficient indicates how much the dependent variable is expected to change when the corresponding independent variable increases by one unit, assuming all other variables remain constant.

In multiple regression, the model estimates a multidimensional regression surface rather than a simple straight line. The goal remains the same as in linear regression: to minimise the residuals and achieve the best possible fit between the observed data and the predicted values.

Model evaluation metrics

The same evaluation metrics used in linear regression **Mean Squared Error (MSE)**, **Root Mean Squared Error (RMSE)**, and **R-squared (R₂)** are also used to assess multiple regression models. RMSE remains particularly useful because it expresses prediction error in the same units as the dependent variable.

In the context of multiple regression, **R-squared** represents the proportion of variance in the dependent variable explained collectively by all independent variables. As additional predictors are included, R₂ typically increases; however, this does not always indicate a better model. To address this, **Adjusted R-squared** is often used, as it penalises unnecessary predictors and provides a more reliable measure of model performance.

Table 7.6 summarises the key advantages and disadvantages of multiple regression.

Table 7.6: Advantages and disadvantages of multiple regression

Advantages	Disadvantages
Models complex relationships involving multiple factors.	Assumes linear relationships between predictors and the response.
Provides insight into the relative importance of different predictors.	Sensitive to multicollinearity among predictors.
More realistic representation of real-world scenarios.	More complex to interpret than simple linear regression.
Extends the interpretability of linear regression.	Requires larger datasets to produce reliable estimates.

A hypothetical property analytics business scenario: PropertyInsights AU

A hypothetical Australian property analytics firm, **PropertyInsights AU**, uses **multiple regression** to estimate residential property prices in metropolitan areas. The firm builds a model using historical sales data and considers multiple factors simultaneously, including

property size, suburb location, number of bedrooms, property age, and proximity to public transport and amenities.

When a new property is listed, PropertyInsights AU inputs these characteristics into the model to generate an estimated market price. This estimate supports real estate agents in setting competitive listing prices, assists investors in evaluating potential returns, and helps urban planners analyse housing market trends across different regions.

In this scenario, multiple regression supports prescriptive analytics by informing decisions such as **what price a property should be listed at** and **which areas present strong investment opportunities**, providing more realistic and accurate guidance than single-factor pricing models.

7.2.2.3 Clustering

Clustering is a data mining technique used to group data points into clusters or segments, where points within the same cluster share similar traits, while points in different clusters are distinct.

Unlike classification, clustering is an unsupervised learning method. This makes it particularly useful for exploratory analysis, especially when there's no prior knowledge about categories.

The clustering process typically begins with data preparation, where the dataset is cleaned and pre-processed to ensure that features are on similar scales, as many clustering algorithms are sensitive to differences in feature magnitude. Once the data is prepared, a clustering algorithm is applied to group the data points. The algorithm evaluates the similarity or distance between points, using metrics like Euclidean or Manhattan distance, to form clusters. These clusters are refined iteratively to ensure that points within a cluster are as similar as possible, while the clusters themselves remain distinct. After clustering, results are often visualised using tools like scatter plots to interpret the groupings.

Key clustering algorithms include **K-Means** and **Hierarchical Clustering**, which are described as follows.

7.2.2.3.1 K-Means clustering

K-Means is one of the most popular clustering algorithms. The algorithm works by dividing the dataset into a predefined number of clusters, K, where each cluster is represented by its centre, or centroid. The

goal of K-Means is to minimise the total distance between the data points in a cluster and their centre/centroid, creating compact and well-separated clusters.

The algorithm works iteratively through two main steps: **assignment** and **update**. In the **assignment step**, each data point is assigned to the cluster whose centroid is closest to it, based on a chosen distance metric. In the **update step**, the centroids are recalculated as the mean position of all the points within each cluster. These steps are repeated until the centroids no longer change significantly, indicating that the algorithm has converged.

K-Means effectiveness depends on choosing the right number of clusters (K). One common method for selecting the optimal number of clusters is the Silhouette Score, which measures how well-defined the clusters are by assessing the distance between points within the same cluster and points in other clusters.

Table 7.7 summarises the key advantages and disadvantages of K-Means clustering.

Table 7.7: Advantages and disadvantages of K-Means clustering

Advantages	Disadvantages
Easy to implement.	Sensitive to initial centroids.
Efficient for large datasets.	Requires predefined clusters.
Performs well with clear clusters.	Struggles with irregular clusters.
Flexible cluster number.	Affected by outliers.

A hypothetical waste management business scenario

A hypothetical Australian local council uses **K-Means clustering** to analyse waste generation data across different suburbs. The dataset includes features such as **total waste volume, dominant waste type (organic, plastic, electronic), population density, and collection**

frequency. By applying K-Means, the council groups suburbs with similar waste disposal behaviours into distinct clusters.

The analysis reveals that densely populated suburbs generate large volumes of organic waste, while other areas produce higher levels of electronic waste. Based on these clusters, the council adjusts its operations by scheduling **more frequent organic waste collection** in high-density areas and establishing **specialised e-waste drop-off points** in regions where electronic waste is concentrated. These data-driven decisions enable more efficient collection routes, improved placement of recycling facilities, and better resource allocation.

In this context, **K-Means clustering supports prescriptive analytics by helping decision-makers determine what waste management actions should be taken in each area**, allowing collection strategies to be tailored to the specific needs of different suburbs.

7.2.2.3.2 Hierarchical clustering

Hierarchical clustering is a method used to group data points into a hierarchy of clusters. Unlike “flat” clustering methods like K-Means, hierarchical clustering organises data into a tree-like structure called a dendrogram, which visually represents how data points are grouped together at different levels of similarity. This hierarchical structure makes it different from K-Means clustering, which assigns data points to a fixed number of clusters without providing information about relationships between clusters.

The Hierarchical clustering process begins by **calculating the similarity or distance between all data points** using a distance metric such as Euclidean distance. Based on these distances, the algorithm groups points into clusters iteratively.

There are two main types of hierarchical clustering: **agglomerative** and **divisive**. In agglomerative clustering (a “bottom-up” approach), each data point starts as its own cluster, and clusters are progressively merged based on their similarity until all points belong to a single cluster. In contrast, divisive clustering (a “top-down” approach) starts with all data points in one large cluster, which is then repeatedly split into smaller clusters. While agglomerative clustering is more commonly used due to its simplicity, both methods provide a detailed hierarchy that allows users to analyse data at various levels of granularity.

Figure 7.5 illustrates this process. The **left panel** shows the original data points (A-K) plotted in feature space, with no grouping applied. The **middle panel** depicts the formation of clusters as points are progressively grouped based on similarity, shown by nested contours representing different cluster levels. The **right panel** presents the corresponding **dendrogram**, which visualises the hierarchical relationships between data points and clusters. This figure demonstrates how hierarchical clustering builds a tree-like

structure of nested clusters, allowing the data to be partitioned at different levels of granularity depending on the chosen cut of the dendrogram.



Figure 7.5: Hierarchical clustering.

Alternative Text for Figure 7.5

Three-panel diagram showing hierarchical clustering: raw data points, grouped clusters with nested contours, and a dendrogram illustrating cluster hierarchy.

Table 7.8 summarises the key advantages and disadvantages of hierarchical clustering.

Table 7.8: Advantages and disadvantages of hierarchical clustering

Advantages	Disadvantages
No need to define cluster number.	Computationally demanding.
Handles complex cluster shapes.	Sensitive to outliers.
Reveals data hierarchies.	Hard to scale.
	Merges can't be undone.

A hypothetical hospitality business scenario: MyHotel AU

A hypothetical Australian hotel group, **MyHotel AU**, uses **hierarchical clustering** to segment its guests based on data such as **booking frequency, length of stay, room preferences, spending patterns, and demographic information**. Unlike clustering methods that require a predefined number of clusters, hierarchical clustering allows MyHotel AU to explore guest groupings at different levels of detail.

By analysing the resulting hierarchy, MyHotel AU identifies distinct guest segments, such as **frequent business travellers, price-sensitive leisure guests, and high-spending premium guests**. These insights enable the hotel group to design tailored marketing campaigns, personalise room and service offerings, and optimise loyalty programs for each segment.

In this context, hierarchical clustering supports prescriptive analytics by helping MyHotel AU decide **how services and promotions should be tailored for different guest segments**, leading to improved guest satisfaction and increased revenue.

For anyone interested in learning how to perform hierarchical clustering in Python, W3Schools provides a [clear and practical tutorial](#). This guide walks you through the entire process, from preparing your dataset to implementing hierarchical clustering using Python libraries like SciPy.

7.2.2.4 Association rule learning

The last data mining technique that we are going to cover in this Chapter is **Association rule learning**. Association Rule Learning focuses on discovering relationships between variables in a dataset, particularly patterns of co-occurrence. Unlike classification or regression tasks, which predict a specific outcome, association rule learning identifies connections, like which items are often purchased together in a retail setting. This task involves analysing large datasets to uncover rules that indicate relationships.

The process of **association rule learning** typically begins with **data preparation**, where the dataset is cleaned, formatted, and transformed into a structure suitable for analysis. After preparation, the next step is **rule generation**, which involves identifying frequent itemsets. These are groups of items that appear together in the data more often than a specified minimum threshold, known as support.

The most popular associate rule learning method is the Apriori algorithm which is detailed below.

7.2.2.4.1 Apriori algorithm

The Apriori works by iteratively identifying frequent itemsets based on a minimum support threshold, which measures how often a combination of items occurs in the dataset. Initially, it scans the dataset to find individual items that meet the minimum support. Then, it combines these frequent items to form larger itemsets, checking each for support. The process continues until no more frequent itemsets can be generated. From these itemsets, the algorithm generates association rules in the form of “if-then” statements, such as “*If a customer buys bread, they are likely to buy butter.*”

These rules are evaluated using metrics like **confidence** and **lift**. Together, these metrics ensure that the rules generated by the Apriori algorithm are not only frequent but also meaningful and actionable.

Confidence measures the likelihood of the “then” part of the rule occurring, given that the “if” part has already occurred. For example, in the previous rule “*If a customer buys bread, they are likely to buy butter,*” confidence would calculate the percentage of transactions containing bread that also contain butter. A higher confidence value indicates a stronger relationship between the items.

Lift, on the other hand, assesses the strength of the rule compared to what would be expected by random chance. It calculates the ratio of the observed confidence to the confidence expected if the two items were independent of each other. For instance, if bread and butter are purchased together more often than their individual probabilities suggest, the lift value will be greater than 1, indicating a meaningful association.

Table 7.9 summarises the key advantages and disadvantages of Apriori algorithm.

Table 7.9: Advantages and disadvantages of Apriori algorithm

Advantages	Disadvantages
Easy to use and understand.	Can be slow with big data.
Identifies frequent itemsets.	Requires lots of memory.
Adjustable thresholds.	Requires lots of memory.
Useful across industries.	Struggles with long itemsets.

A hypothetical e-commerce business scenario: ShopNow AU

A hypothetical Australian online retailer, **ShopNow AU**, uses the Apriori algorithm to analyse transaction data and identify common purchasing patterns. The system finds that customers who purchase a **smartphone** often also buy a **protective case and screen protector**. Based on these associations, ShopNow AU automatically recommends these related items to customers during checkout.

In this context, Apriori supports prescriptive analytics by answering the question “**What additional products should we recommend to this customer?**” By presenting relevant suggestions at the point of purchase, the retailer improves cross-selling effectiveness, enhances the customer shopping experience, and increases overall sales.

7.3 Text mining

Text mining is the process of extracting meaningful information and patterns from unstructured text data. While data mining focuses on structured data, text mining tackles the challenge of analysing the vast amounts of unstructured text data generated every day.

Text mining combines techniques from natural language processing (NLP), machine learning, and statistics to transform raw text into actionable insights. The goal is to understand what the text is communicating, uncover trends, and sometimes even predict future outcomes based on those insights.

A helpful introduction to these concepts is presented in the YouTube video “[What is Text Mining](#)” by IBM Technology, which provides a clear overview of how text mining works and why it is essential for making sense of language-based data. It’s worth noting that text mining encompasses both supervised and unsupervised techniques, depending on the specific task and the availability of labelled data.

In this chapter, we explore some of the most widely used text mining techniques, including topic modelling, named entity recognition (NER), and sentiment analysis. The following sections will delve into each of these techniques, providing an overview of how they work and their real-world applications. But before diving into the specifics, it’s important to understand how the process of text mining works.

7.3.1 How text mining works?

Like data mining, text mining follows a systematic process to ensure meaningful insights are extracted from raw text data. This process typically involves several stages, starting with data collection and progressing through data preprocessing, text analysis, and evaluation and application.

1. **Data collection and preprocessing.** Text mining begins with data collection, where text is gathered from various sources such as emails, social media, survey responses, or web pages. Text data is often messy and unstructured, so it needs to go through preprocessing to prepare it for analysis. This involves cleaning the text by removing irrelevant elements like stop words (“the,” “and,” “is”), punctuation, and HTML tags. Techniques like stemming or lemmatisation reduce words to their base forms, such as “running” to “run”, ensuring consistency in the analysis.
2. **Text analysis and feature extraction.** Once pre-processed, the text is analysed to extract features or patterns. A common approach is creating a Term-by-Document Matrix (TDM), where rows represent unique terms (words or phrases), and columns represent individual documents. Each cell in the matrix indicates the frequency of a term in a document. This matrix provides a foundation for many text mining techniques, as it captures the relationship between terms and documents numerically.
3. **Evaluation and application.** The final step involves evaluating the results to ensure they are accurate and useful. Once validated, the insights can be applied to solve real-world problems.

7.3.2 Text mining techniques

Text mining involves a variety of techniques that transform unstructured text into meaningful insights. Each technique has its unique methods, applications, and strengths, making them valuable tools for solving distinct challenges in text analysis.

The following subsections explore three widely used text mining techniques: *topic modelling*, *named entity recognition (NER)*, and *sentiment analysis*. Each technique is explained in detail, with real-world applications, advantages, and limitations discussed to provide a clear understanding of their roles in text mining.

7.3.2.1 Topic modelling

Topic modelling is a popular text mining technique used to discover hidden topics or themes within large collections of unstructured text data (e.g., customer reviews, survey responses, and social media comments). Given that topic modelling is an unsupervised learning technique, it does not require labelled data, making it particularly useful for exploratory analysis.

The technique relies on the assumption that documents are composed of a mixture of topics, and each topic is defined by a specific distribution of words. For instance, a topic on “sports” might heavily feature words like “team”, “game”, and “score”, while a topic on “technology” might include “AI”, “software”,

and “innovation”. Popular algorithms like Latent Dirichlet Allocation (LDA) use mathematical models to identify these word distributions.

The **topic modelling** process begins with preprocessing the text. Each document is then represented as a probability distribution over a set of topics, and each topic is represented as a probability distribution over a set of words.

For example, in LDA, the algorithm iteratively assigns words in the documents to topics based on their likelihood of belonging to that topic. Over multiple iterations, the algorithm refines these assignments to generate coherent topics that maximise the likelihood of the observed data. This allows each document to be represented as a mixture of topics, and the dominant topic in each document can provide a high-level summary of its content. The resulting topics and their distributions can then be visualised using tools like word clouds, heatmaps, or topic-document matrices, helping to provide intuitive insights into the dataset.

Table 7.10 summarises the key advantages and disadvantages of topic modelling.

Table 7.10: Advantages and disadvantages of topic modelling

Advantages	Disadvantages
Reveals hidden topics.	Needs parameter tuning.
Scales to large text data.	Struggles with short texts.
Helps organise unstructured text.	Topics can overlap.
Topics are easy to interpret.	Requires heavy preprocessing.

A hypothetical customer feedback analysis scenario: ServicePlus AU

A hypothetical Australian small-to-medium business, **ServicePlus AU**, collects large volumes of customer feedback through online reviews and post-purchase surveys. To better understand

customer opinions, the business applies **topic modelling using Latent Dirichlet Allocation (LDA)** to analyse the text data. The model automatically identifies recurring themes related to **pricing concerns, delivery speed, and positive customer service experiences**.

Based on these insights, ServicePlus AU revises its pricing structure to address perceived affordability issues while reinforcing staff training and recognition programs to maintain strong customer service performance. In this context, topic modelling supports prescriptive analytics by helping the business decide **what actions should be taken to improve customer satisfaction**.

7.3.2.2 Named entity recognition (NER)

NER is a supervised text mining technique that focuses on identifying and categorising specific entities within text. Entities can include names of people, organisations, locations, dates, and quantities.

NER systems typically work in two stages: **identification** and **classification**. During **identification**, the algorithm scans the text to detect potential entities. Once identified, the **classification** stage assigns each entity to a predefined category.

For example, in the sentence, “*Google announced its new product in California on January 1st*” an NER system would identify and classify “Google” as an organisation, “California” as a location, and “January 1st” as a date.

Modern NER systems often use machine learning techniques, trained on labelled datasets where entities are already annotated. These models rely on features such as word context, capitalisation, and part-of-speech tags to make predictions. Word context plays a key role, as the words surrounding a potential entity provide important clues about its category.

For instance, the phrase “*Apple announced its new product*” suggests that “Apple” is likely an organisation, while in “I ate an apple,” it’s clear that “apple” refers to some fruit. Capitalisation is another important feature, especially in languages like English. Proper nouns, such as names of people, places, or organisations, are often capitalised, providing a strong signal for entity identification. Part-of-speech tags identify whether a word is a noun, verb, adjective, etc.

Table 7.11 summarises the key advantages and disadvantages of NER.

Table 7.11: Advantages and disadvantages of NER

Advantages	Disadvantages
Extracts structured info quickly.	Depends on language and domain.
Applies across many sectors.	Misclassifies ambiguous terms.
Boosts other text mining tasks.	Needs large training datasets.

A hypothetical financial analysis scenario: MyFinance AU

A hypothetical Australian financial services firm, **MyFinance AU**, uses **Named Entity Recognition (NER)** to analyse annual financial reports submitted by listed companies. The NER system scans lengthy documents and automatically extracts key entities such as **company names, revenue figures, profit values, reporting periods, and risk-related terms**.

By structuring this information, MyFinance AU can quickly assess financial performance and identify potential compliance or risk issues. In a prescriptive analytics context, the extracted data supports decisions such as **which companies require further review, where compliance checks should be prioritised, and which investment opportunities warrant deeper analysis**. This approach significantly reduces manual data extraction effort while improving speed and accuracy in financial decision-making.

7.3.2.3 Sentiment analysis

Sentiment analysis, also known as **opinion mining**, is a text mining technique used to identify the emotional tone expressed in textual data. It categorises text as **positive, negative, or neutral**, enabling organisations to understand opinions, attitudes, and perceptions expressed in sources such as customer reviews, survey responses, and social media posts. Sentiment analysis combines techniques from **Natural Language Processing (NLP)** (see Section 7.3.3) and machine learning to transform unstructured text into actionable insights.

The sentiment analysis process typically begins with **text preprocessing**, where raw text data is cleaned by removing stop words, punctuation, and irrelevant symbols, and then tokenised into smaller units such as words or phrases. Next, **feature extraction** is applied to convert the text into a structured numerical format, such as a **term-document matrix (TDM)**. Once the text is represented in this form, sentiment classification can be performed using either **rule-based** or **machine learning** approaches.

Rule-based approaches rely on predefined dictionaries of positive and negative words. Each word contributes a sentiment score based on its polarity. For example, words such as “*great*” or “*excellent*” contribute positively, while words such as “*poor*” or “*hate*” contribute negatively. The overall sentiment of a text is determined by aggregating these scores.

Machine learning approaches, in contrast, use algorithms such as **Support Vector Machines (SVMs)** or **neural networks**, which are trained on labelled datasets where the sentiment of each text instance is already known. These models learn patterns and relationships within the data and can then predict the sentiment of new, unseen text. Machine learning approaches are generally more flexible and accurate than rule-based systems, particularly when dealing with varied language use.

A hypothetical product launch scenario: BrandCo AU

A hypothetical Australian consumer goods company, **BrandCo AU**, launches a new product and uses sentiment analysis to evaluate customer reactions across online reviews, customer feedback forms, and social media discussions. The sentiment analysis system categorises customer opinions as **positive, negative, or neutral** and identifies key themes driving sentiment, such as **product quality, pricing, packaging, and customer support**.

The analysis reveals strong positive sentiment around product quality but increasing negative sentiment related to pricing. Based on these insights, BrandCo AU adjusts its marketing messages to better communicate value and introduces targeted promotional offers to address pricing concerns. In this context, sentiment analysis supports prescriptive analytics by helping the company decide **what actions should be taken to improve customer perception and satisfaction**.

Table 7.12 summarises the key advantages and disadvantages of sentiment analysis.

Table 7.12: Advantages and disadvantages of sentiment analysis

Advantages	Disadvantages
Quickly processes large texts.	Can miss complex language.
Offers actionable insights.	May carry training data bias.
Works in many industries.	Performance varies by domain.
Understands context better with modern models.	Oversimplifies nuanced opinions.

Modern sentiment analysis: Transformers and BERT

Modern NLP systems increasingly rely on **transformer-based models**, which are designed to understand language by analysing the **context of all words in a sentence simultaneously**. Unlike earlier approaches that processed text sequentially, transformers consider how each word relates to every other word, allowing them to capture meaning more accurately.

One widely used transformer-based model is **BERT (Bidirectional Encoder Representations from Transformers)**. BERT is pre-trained on large volumes of text and learns general language patterns, including grammar, word meaning, and contextual relationships. Because it analyses text in both directions (left-to-right and right-to-left), BERT is particularly effective at understanding **nuance, ambiguity, and intent**. As a result, BERT performs well in tasks such as sentiment analysis, topic classification, and named entity recognition.

Modern Natural Language Processing (NLP) systems increasingly rely on **transformer-based models**, which are designed to understand text by analysing the **context of all words in a sentence at the same time**. Unlike earlier approaches that processed text word by word in sequence, transformers consider how each word relates to every other word, allowing them to capture meaning more accurately. One widely used transformer-based model is **BERT (Bidirectional Encoder Representations from Transformers)**. BERT is pre-trained on large volumes of text and learns general language patterns, such as grammar, word meaning, and context. Because it analyses text in both directions (left-to-right and right-to-left), BERT is particularly effective at understanding nuance, ambiguity, and intent in language. As a result, BERT performs well in tasks such as **sentiment analysis, topic classification, and named entity recognition**.

A hypothetical service company business scenario: InsightCo AU

A hypothetical Australian service company, **InsightCo AU**, receives large volumes of unstructured customer feedback through emails, online reviews, and support tickets. To better understand customer concerns, the company uses a **BERT-based NLP model** to analyse the text data.

The model processes customer messages and identifies sentiment, key topics, and important entities such as product names and service issues. For example, BERT correctly interprets feedback like *“The service wasn’t bad, but delivery took too long”* as mixed sentiment rather than purely negative, due to its ability to understand context.

Based on these insights, InsightCo AU prioritises delivery-related issues for operational improvement while maintaining current service practices. In this context, transformer-based models such as BERT support **prescriptive analytics** by helping the business decide **what actions should be taken based on nuanced customer feedback**, leading to more accurate decisions and improved customer satisfaction.

7.3.3 Natural language processing (NLP)

When studying text mining, it’s essential to talk about NLP because it provides the foundational tools and techniques that make text mining possible. While text mining focuses on extracting patterns and insights from unstructured text, NLP enables machines to process and understand the language itself. Tasks like tokenising sentences, identifying key entities, or analysing sentiment rely heavily on NLP.

Essentially, **NLP** acts as the bridge between raw text data and the advanced analytical techniques used in text mining, ensuring the language is structured and ready for deeper analysis.

NLP is a field of AI that allows computers to understand, interpret, and generate human language in text or speech form. It’s used in a wide range of applications, from virtual assistants like Alexa and Siri to machine translation tools, spam detection systems, and customer service chatbots. Businesses also rely on NLP to analyse customer feedback, automate document processing, and monitor brand sentiment on social media.

NLP process begins with text preprocessing, which involves cleaning the text data to make it suitable for analysis. This includes tasks like tokenisation (splitting sentences into words), removing stop words like “the” or “and”, and stemming or lemmatisation (reducing words to their root forms). Preprocessing ensures that the text is consistent and ready for further processing. After preprocessing, NLP systems analyse the text to extract meaning. This often involves methods like part-of-speech tagging (identifying

whether a word is a noun, verb, etc.), named entity recognition, and dependency parsing (understanding the grammatical structure of a sentence).

A hypothetical customer service scenario: SupportCo AU

A hypothetical Australian service company, **SupportCo AU**, receives thousands of customer enquiries each week through emails, online chat, and support tickets. To manage this volume efficiently, SupportCo AU uses NLP techniques to automatically analyse incoming messages.

The NLP system processes the text to identify **key topics, intent, and urgency**, categorising enquiries into areas such as **billing issues, technical problems, delivery delays, and general enquiries**. Based on this analysis, the system automatically routes messages to the appropriate support team or generates suggested responses for common issues.

In a prescriptive analytics context, NLP helps SupportCo AU decide **what action should be taken next for each enquiry**, such as escalating urgent complaints, prioritising technical faults, or providing automated responses. This improves response times, reduces operational costs, and enhances overall customer satisfaction.

7.4 Web analytics

Web analytics involves the collection and analysis of data from online platforms to understand user behaviour, improve website performance, and support data-driven business decision-making.

An introduction to these ideas is presented in the YouTube video “[What is Web Analytics](#)” by Google Garage, which explains the fundamentals of tracking user behaviour and turning online activity into actionable insights.

The first record of web analytics dates back to the early 1990s with the introduction of server log file analysis. As the internet gained traction, website administrators began using server logs to track visitor activity, such as page views, visits, and IP addresses. The field continued evolving pioneered by tools like Urchin Software, which Google acquired in 2005 and developed into Google Analytics. This shift allowed for more accurate data collection by capturing user interactions directly in the browser rather than relying on server logs, enabling more detailed insights into user journeys and behaviours.

7.4.1. How web analytics works?

The web analytics process is all about gathering, preparing, and making sense of data from websites to understand how users interact with them.

The first step, data collection, usually happens in two main ways: **log files** and **page tagging**. *Log files* method records each user action, like a page load, in server logs, providing a solid history of all interactions. As the web evolved, *page tagging* became popular. It works by embedding JavaScript code into each webpage, capturing real-time data directly in the browser. However, log files remain highly relevant. As data is captured directly on the server, log files record every interaction with the website, including visits from bots, users with JavaScript disabled, and server-side errors, which page tagging often misses. This makes log files particularly valuable for industries requiring detailed compliance records or where a complete view of all interactions including non-human traffic is essential.

Once data is collected, it's often raw and messy, so data preprocessing and cleaning is essential. With clean data, the process moves to analysis and interpretation. Here, analysts dive into metrics, looking at things like bounce rate, time on site, and conversion paths. They interpret patterns and trends to make informed recommendations for improving the website and user experience.

7.4.2 Web analytics techniques

Web analytics techniques are essential for understanding and optimising online user behaviour. They are diverse and tailored to address different aspects of online performance. The following sections explore two critical techniques, Clickstream Analysis and Search Engine Optimisation (SEO) Analytics, in more detail, highlighting their processes, applications, and benefits. Together, they form the backbone of web analytics strategies.

7.4.2.1 Clickstream analysis

Clickstream analysis is a web analytics technique used to track and analyse the sequence of clicks visitors make as they navigate through a website.

This “click path” data provides insights into user behaviour, such as the pages they visit, the order in which they navigate, how long they spend on each page, and at what point they exit the site. By understanding these patterns, businesses can identify popular content, optimise user journeys, and improve overall website performance.

Clickstream analysis starts with collecting data about user interactions on a website, and this process relies on tools like web server logs and tracking scripts.

- **Web server logs** capture details like the URL of each page visited, the IP address of the visitor, the browser and device used, and the time and date of the request.
- **Scripts** embedded in web pages refer to small pieces of code, often written in JavaScript, added to the website's code to track user activity in real-time, such as clicking buttons, scrolling, hovering over

elements, or filling out forms. Tools like Google Analytics use such scripts to collect and send data about user interactions to analytics platforms for further processing.

Timestamps are critical components of both server logs and tracking scripts. A *timestamp* records the exact time a specific action occurred, such as when a user loaded a page or clicked a link. These timestamps allow analysts to sequence events, calculate time spent on pages, and understand user navigation patterns. Visitor identifiers are unique markers used to track individual users during a browsing session or across multiple visits. These identifiers might be cookies, small files stored on a user's device, or session IDs, which temporarily track users until they leave the website.

Once data is collected, preprocessing is essential to clean and organise it for analysis. After preprocessing, the cleaned data is visualised using tools like flowcharts, which provide intuitive representations of user journeys.

A hypothetical digital business scenario: ClickCo AU

A hypothetical Australian e-commerce company, **ClickCo AU**, uses **clickstream analysis** to understand how visitors navigate its website. By analysing data such as page views, time spent on pages, click paths, and drop-off points, the company identifies where users abandon the site during the checkout process.

The analysis reveals that a large number of customers exit the website after reaching the payment page, indicating a potential bottleneck. Based on this insight, ClickCo AU simplifies the checkout flow, reduces the number of required steps, and adds clearer progress indicators. As a result, the company improves conversion rates and reduces cart abandonment.

In this context, clickstream analysis supports prescriptive analytics by helping ClickCo AU decide **what changes should be made to the website to guide users toward desired actions**, such as completing a purchase or signing up for a newsletter.

7.4.2.2 Search Engine Optimisation (SEO) analytics

SEO analytics focuses on monitoring and improving how a website ranks on Search Engine Results Pages (SERPs), which are the lists of results displayed by search engines like Google, Bing, or Yahoo in response to a user's query. SERPs consist of both organic results, which are earned through SEO efforts, and paid advertisements. The higher a website ranks on the SERP, especially on the first page, the more likely it is to receive traffic.

SEO analytics involves tracking metrics like keyword rankings, organic traffic, which measures visitors arriving via search engines, and click-through rates (CTR), which show how often users click on a result

after seeing it in the SERP. SEO analytics' goal is not only to rank high but to rank for the right terms that align with user intent. For instance, a business selling fitness equipment might aim to rank for terms like *“best home gym equipment”* rather than unrelated keywords like *“free fitness classes.”*

The process of SEO analytics begins with identifying opportunities for improvement in both on-page SEO, which focuses on optimising individual pages (e.g., content, headings, meta tags), and off-page SEO, which looks at external factors like backlinks and social signals. Backlinks, also known as inbound or incoming links, are links from one website to another, whereas social signals refer to the interactions content receives on social media platforms. The SEO process also involves evaluating technical SEO, which ensures the website's infrastructure supports search engine crawling and indexing.

Search engine **crawling** and **indexing** are fundamental processes that ensure websites appear in search engine results.

Crawling is the initial step, where search engines send out automated programs called “crawlers” or “spiders” to discover and scan web pages across the internet. These crawlers navigate through links on web pages to find new or updated content, gathering information about each page, such as keywords, meta tags, and other elements. Once crawlers retrieve this information, the **indexing** process begins.

Indexing involves organising and storing the collected data in a vast database, where it's analysed and categorised based on relevance and quality. When a user performs a search, the search engine retrieves and ranks indexed pages that match the query.

SEO techniques are generally categorised into white, grey, and black hat strategies based on their compliance with search engine rules. These are described as follows.

- **White hat SEO** focuses on ethical, search engine-approved strategies that enhance user experience. These techniques comply with search engine guidelines, ensuring long-term results without the risk of penalties. Examples include creating high-quality, original content that provides value to users and building legitimate backlinks by guest posting on reputable websites.
- **Grey hat SEO** falls in a middle ground, using techniques that are not explicitly prohibited but may push the boundaries of search engine guidelines. Examples of this class include creating spun or lightly modified content to target multiple keywords and buying backlinks or using private blog networks for link building.
- **Black hat SEO** involves unethical practices designed to manipulate search engine rankings. These techniques violate guidelines and can lead to severe penalties, including being banned from search engine listings. Examples are overloading content with keywords to manipulate rankings and using bots to generate fake clicks or traffic to inflate engagement metrics.

A hypothetical web analytics business scenario: EcoMart AU

A hypothetical Australian online retailer, **EcoMart AU**, uses web analytics to improve the performance of its digital marketing campaigns and increase conversion rates. The company analyses user behaviour data such as ad clicks, page visits, and checkout drop-off points to understand how visitors move through its website.

The analysis shows that many users leave the site before completing a purchase, particularly on the landing and checkout pages. Based on these insights, EcoMart AU tests different ad messages, simplifies page layouts, and improves calls-to-action using A/B testing.

In this context, web analytics supports prescriptive analytics by helping EcoMart AU decide **which campaign elements should be changed to guide users toward conversion goals**, such as making a purchase or signing up for a newsletter.

7.5 Social media analytics

Social media analytics is the process of gathering, measuring, and interpreting data from social media platforms to uncover patterns, trends, and insights. With billions of users active on platforms like Instagram and TikTok, social media has become a goldmine of data reflecting customer sentiments, preferences, and behaviours.

Advanced social media analytics techniques often use NLP to analyse sentiment (see Sections 7.3.2.3) and detect key themes in user-generated content. Social media analytics has applications across industries, from monitoring public opinion during political campaigns to assessing the effectiveness of a product launch or identifying emerging customer needs.

7.5.1 How social media analytics works?

The process of social media analytics involves several interconnected steps, each designed to transform raw social media data into actionable insights.

1. **Data collection.** It begins with data collection, where information is gathered from social media platforms. To collect this data, organisations often rely on APIs provided by social media platforms,

which allow for automated data retrieval. Additionally, third-party tools like Hootsuite, Sprout Social, or Brandwatch offer pre-built solutions for collecting and organising social media data efficiently.

2. **Data pre-processing.** Once the data is collected, it undergoes pre-processing, when text data, which forms a significant portion of social media content, may also be tokenised or stemmed to prepare it for analysis. Depending on the goals of the analysis, data is then organised into structured formats, such as databases or dataframes, for easier handling.
3. **Data analysis.** The analysis phase applies advanced techniques to extract meaningful patterns and trends. For example, sentiment analysis determines whether the tone of posts or comments is positive, negative, or neutral, providing insights into customer opinions. Trend detection helps identify emerging topics or hashtags that are gaining traction, which can inform timely marketing strategies or risk management actions.
4. **Data visualisation and reporting.** Finally, the results are communicated through visualisation and reporting. Dashboards and reports, often created using tools like Tableau or Power BI, display key metrics like engagement rates, audience demographics, and sentiment trends. These visualisations make complex data more understandable and actionable for decision-makers. By presenting insights clearly, organisations can quickly identify opportunities to enhance their campaigns, address potential reputational risks, or refine their customer engagement strategies.

7.5.2. Social media analytics techniques

Social media analytics relies on various techniques to transform raw social media data into actionable insights. In this section, we focus on two widely used techniques: **network analysis** and **trend analysis**.

7.5.2.1 Network analysis

Network analysis is a technique used to explore and understand relationships and interactions within social media networks.

These networks consist of **nodes**, representing entities such as users, accounts, or pages; and **edges**, representing interactions such as likes, shares, comments, or follows.

Network analysis begins with constructing a network graph using data collected from social media platforms. Nodes and edges are extracted from this data to form a graph. Visualisation tools are often used to create and analyse these graphs, making it easier to understand complex relationships.

Once the network is mapped, various analytical techniques are applied to uncover meaningful insights. Centrality metrics provide insights into how individuals or entities are positioned within a social structure and how they contribute to the flow of information. Some of these metrics are described as follows. Degree

centrality reflects how many direct connections a node has, often indicating active users such as influencers who frequently engage through likes, comments, or shares. Betweenness centrality identifies users who act as bridges between different groups, enabling information to flow across otherwise disconnected parts of the network, like a journalist linking niche communities to broader audiences. Closeness centrality measures how quickly a node can reach others in the network, making users with high closeness ideal for rapidly spreading information, particularly useful in situations like crisis communication. By analysing these centrality metrics, businesses and organisations can strategically engage with high-impact users, optimise influencer marketing campaigns, or identify key accounts for partnership opportunities.

Network analysis in social media is particularly useful for tracking the spread of information, such as viral posts or trending hashtags, and detecting anomalies, like bot activity or coordinated misinformation campaigns. By understanding how users interact and form connections, businesses can optimise their engagement strategies, amplify their reach through influencers, and enhance customer relationships.

A hypothetical social media business scenario: ConnectRetail AU

A hypothetical Australian retail brand, **ConnectRetail AU**, uses **network analysis** to understand how information about its products spreads on social media. By analysing interactions such as likes, shares, mentions, and follows, the company maps a social network where users are represented as nodes and interactions as edges.

The analysis identifies a small group of users with high **degree and betweenness centrality**, indicating that they are influential and act as bridges between different online communities. Based on this insight, ConnectRetail AU prioritises these users for influencer partnerships and targeted engagement. In this context, network analysis supports prescriptive analytics by helping the company decide **which users should be engaged to maximise reach and influence**.

7.5.2.2 Trend analysis

Trend analysis is a key technique in social media analytics used to identify patterns, shifts, or emerging topics over time within vast datasets of social media content. It helps businesses uncover what topics are gaining traction, monitor user sentiment changes, and identify potential opportunities or risks. By understanding trends, organisations can make data-driven decisions to optimise their strategies, align their messaging with audience interests, and react swiftly to evolving market dynamics. For example, trend analysis can highlight viral hashtags, rising customer concerns, or seasonal engagement spikes, offering valuable insights for marketing campaigns and brand management.

The process of trend analysis begins with collecting data from social media platforms using APIs, third-

party tools, or web scraping techniques. This data includes posts, hashtags, comments, likes, and shares, often organised into time series for chronological tracking. Pre-processing is then applied to clean and structure the data. Techniques like keyword analysis and clustering (see Section 7.2.2.3) are used to group related terms and identify recurring topics or themes. The analysis phase often employs time-series analysis, NLP, and statistical models to uncover patterns and detect changes in user behaviour. Visualisation tools like trend graphs, heatmaps, or word clouds make the findings accessible, enabling decision-makers to interpret the data quickly and act accordingly.

Trend analysis is particularly valuable for proactive planning, such as launching marketing campaigns aligned with emerging topics or identifying and addressing negative trends before they escalate.

A hypothetical social media business scenario: TrendWatch AU

A hypothetical Australian consumer brand, **TrendWatch AU**, uses **trend analysis** to monitor social media conversations related to its industry. By tracking hashtags, keywords, and post volumes over time, the company identifies a rapid increase in discussions around sustainable packaging.

Recognising this emerging trend, TrendWatch AU adjusts its marketing strategy to highlight its environmentally friendly practices and launches a campaign aligned with the topic while interest is growing. In this context, trend analysis supports prescriptive analytics by helping the company decide **what topics to focus on and when to act** to align with audience interests and market dynamics.

7.6 Test your knowledge

Drag the words into the correct boxes.

1) The main business question predictive analytics aims to answer is .

2) A decision tree is a

representation of a decision-making process.

3) Multiple regression estimates the

influence of multiple independent variables on an outcome.

4) Clustering groups data points based on

among their features.

5) Text mining focuses on analysing

text data to extract meaningful information and patterns.

6) Clickstream analysis provides insights into user

by tracking the sequence of clicks on a website.

unstructured

similarity

behaviour

graphical

combined

What will happen



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=717#h5p-8>

An alternative source for this activity: <https://uts.h5p.com/content/1292858510354700519>

References

- Assef, F. M., Steiner, M. T. A., & Lima, E. P. (2022). A review of clustering techniques for waste management. *Heliyon*, 8(2), e08911. <https://doi.org/10.1016/j.heliyon.2022.e08911>
- Change Management Insight. (2017). *Pepsi crisis management case study*. Retrieved November 7, 2024, from <https://changemanagementinsight.com/pepsi-crisis-management-case-study-2017/>
- Collabstr. (2024). *2024 influencer marketing report: Influencer marketing statistics, trends, and predictions*. <https://collabstr.com/2024-influencer-marketing-report>
- Core dna. (n.d.). *Modern all-in-one hybrid CMS and eCommerce platform*. Retrieved November 5, 2024, from <https://www.coredna.com>
- DataFloq. (n.d.). *The future of predictive analytics: Trends and innovations to watch*. <https://datafloq.com/read/the-future-of-predictive-analytics-trends-and-innovations-to-watch/>
- Factspan. (n.d.). *Sentiment analysis and its importance in brand reputation*. Factspan Analytics. <https://www.factspan.com/case-studies/sentiment-analysis-and-its-importance-in-brand-reputation/>
- Forbes Technology Council. (2023). *Five key trends shaping the future of predictive*

- analytics*. *Forbes*. <https://www.forbes.com/councils/forbestechcouncil/2023/10/05/five-key-trends-shaping-the-future-of-predictive-analytics/>
- Gwynnie Bee. (n.d.). *Gwynnie Bee: Unlimited clothing rental subscription for women*. Retrieved November 5, 2024, from <https://closet.gwynniebee.com>
 - InfometryKS. (2023). *The future of predictive analytics in 2024*. *Medium*. <https://medium.com/@infometryks/the-future-of-predictive-analytics-in-2024-244f1cd24535>
 - KDnuggets. (2020). *Decision tree algorithm, explained*. <https://www.kdnuggets.com/2020/01/decision-tree-algorithm-explained.html>
 - Mashayekhi, A. (2017). *Support vector machine — Introduction to machine learning algorithms*. *Towards Data Science*. <https://towardsdatascience.com/support-vector-machine-introduction-to-machine-learning-algorithms-934a444fca47>
 - McCloskey, B. (2020). *Topic modeling analysis for small businesses*. *Towards Data Science*. <https://towardsdatascience.com/topic-modeling-analysis-for-small-businesses-73ba23474261/>
 - Paxcom. (n.d.). *The future of predictive analytics: What's next in 2024?* <https://paxcom.ai/blog/the-future-of-predictive-analytics-whats-next-in-2024/>
 - Sharma, M. (2019). *How to accurately predict energy savings using linear regression and high frequency energy data*. *kW Engineering*. <https://kw-engineering.com/predict-energy-audit-savings-linear-regression-high-frequency-building-data-analytics/>
 - Snorkel AI. (2022). *10-Ks & information extraction case studies: Automating document processing with AI*. <https://snorkel.ai/blog/10-ks-information-extraction-case-studies/>
 - The Honest Company. (n.d.). *Honest – Health & wellness brand*. Retrieved November 1, 2024, from <https://www.honest.com/>
 - Think with Google. (n.d.). *How The Honest Company uses campaign experiments to drive more qualified traffic*. Retrieved November 1, 2024, from <https://www.thinkwithgoogle.com/marketing-strategies/app-and-mobile/honest-company-trusts-campaign-experiments-drive-more-qualified-traffic/>
 - van Leeuwen, R., & Koole, G. (2021). *Data-driven market segmentation in hospitality using unsupervised machine learning* [Preprint]. arXiv. <https://arxiv.org/abs/2111.02848>

Additional Readings/Podcasts

- The [What It Means](#) podcast episode “The Business Case For Green Consumer Segmentation” talks about customer segmentation.
- The [Youtube video from the University of Iowa Health Care – Informatics – “Pathology Data Mining”](#) features a discussion by Prof Nitin Karandikar and A/Prof John Blau on the intersection of pathology and data mining.
- In the [Stats + Stories](#) podcast episode “Analysing Art Through Text Mining”, Dr Julia Silge, the Engineering Manager at Posit PBC, discusses how text mining and data science can be applied to analyse various forms of media, including literature and screenplays.

- [The Capacity Interactive podcast episode “Web Analytics 101”](#) covers the fundamentals of web analytics and its impact on digital marketing.

Acknowledgements

Thank you to Dr Amara Atif for contributing to the hypothetical Australian business scenarios used in this chapter.

About the author



Dr Marcella Papini
UNIVERSITY OF NEWCASTLE AUSTRALIA
<https://www.newcastle.edu.au/profile/marcella-papini>

Dr Marcella Papini is an expert in mathematical optimisation with a strong commitment to advancing the United Nations Sustainable Development Goals (SDGs). She actively aligns her research, teaching, and engagement efforts with the SDGs, using mathematics as a tool to support evidence-based decision-making for a more sustainable and equitable world.

Marcella has led several publicly and privately funded research projects focused on improving the management of operations across diverse domains—optimising performance not only in terms of economic costs but also environmental and social impacts. Her work has been supported by prestigious institutions such as Australian universities, Transport for NSW, and the Brazilian National Research Council (CNPq). Her research has resulted in high-quality publications in leading international journals, and she regularly collaborates with academic partners around the globe. She brings a truly international perspective to her work, having studied, taught, and conducted research in Brazil, Australia, and Germany.

8. PRESCRIPTIVE ANALYTICS

Meena Jha

Learning Objectives (LO)

By the end of this chapter, you will be able to:

- LO1** Understand the role and significance of prescriptive analytics in the broader context of data-driven decision-making, distinguishing it from descriptive and predictive analytics.
- LO2** Identify and apply core techniques and methodologies used in prescriptive analytics, including mathematical optimisation, and heuristics.
- LO3** Evaluate data and technological requirements for implementing prescriptive analytics solutions across various domains.
- LO4** Analyse real-world applications and case studies to understand how prescriptive analytics drives effective decision-making.
- LO5** Critically assess challenges, limitations, and future trends in prescriptive analytics, including ethical concerns, model transparency, and integration with AI.

8 Introduction to prescriptive analytics

This chapter is designed to provide a comprehensive overview of prescriptive analytics, positioning it within the broader landscape of data analytics and decision science. By the end of the chapter, you will be equipped with both the theoretical insight and practical understanding to leverage prescriptive analytics in a variety of decision-making contexts, ultimately supporting more informed, effective, and data-driven strategies. In the era of data-driven decision-making, analytics has become an indispensable tool across industries, enabling organisations to extract actionable insights from vast and complex datasets. Analytics can be broadly categorised into four types: **descriptive**, **diagnostic**, **predictive**, and **prescriptive**. Each type plays a unique role in the decision-making process, contributing to a comprehensive understanding of business problems and opportunities.

Descriptive analytics is concerned with summarising and interpreting historical data to answer the question, “**What has happened?**”. It uses techniques such as data aggregation, reporting, and visualisation to provide a clear picture of past performance. Common tools include dashboards, charts, and key performance indicators (KPIs).

For example, a sales dashboard showing last quarter’s revenue by region is a form of descriptive analytics.

Diagnostic analytics seeks to understand “**Why did it happen?**” by examining relationships and patterns in historical data. It often employs methods such as correlation analysis, root cause analysis, and drill-down techniques to uncover underlying factors.

For instance, if sales dropped in a region, diagnostic analytics would help determine whether the decline was due to reduced marketing spend, customer churn, or competitor actions.

While descriptive and diagnostic analytics are retrospective, **predictive analytics** looks forward, aiming to answer, “**What is likely to happen?**” It uses statistical models and machine learning algorithms to forecast future outcomes based on historical data. Applications include demand forecasting, credit scoring, and churn prediction.

For example, a predictive model might forecast next month’s customer demand for a retail product using seasonality and purchasing trends.

Taking this a step further, **prescriptive analytics** addresses the question, “**What should we do?**” It goes beyond analysing past trends and predicting future outcomes by recommending actions to achieve desired objectives. Prescriptive analytics integrates data, business rules, mathematical models, and computational algorithms such as mathematical optimisation, simulation, and heuristic methods to generate and evaluate decision options within defined constraints.

While mathematical optimisation seeks to identify an optimal solution, this is not always

computationally feasible in complex, large-scale, or time-sensitive problems. In such cases, heuristic approaches are used to produce near-optimal, good-enough solutions efficiently, enabling timely and practical decision-making. Unlike predictive analytics, which focuses on probabilities and forecasts, prescriptive analytics provides decision-makers with concrete guidance on the most effective course of action.

For example, an airline may use prescriptive analytics to determine pricing and scheduling strategies by analysing demand forecasts, operating costs, and capacity constraints. An optimisation model may be used to compute an optimal schedule under stable conditions, while heuristic rules or metaheuristic algorithms can be applied to rapidly adjust prices or reallocate aircraft in response to disruptions such as weather events or sudden demand changes, ensuring operational feasibility and responsiveness..

8.1 Foundations of prescriptive analytics

8.1.1 Key characteristics

- **Decision orientation.** It focuses on suggesting or automating the best actions rather than just providing insights or forecasts.
- **Optimisation-centric.** At its core, prescriptive analytics relies on mathematical optimisation models to evaluate multiple decision scenarios.
- **Constraint-aware.** It takes into account operational, financial, and regulatory constraints, making recommendations that are feasible in practice.
- **Forward-looking.** While it uses historical and real-time data like predictive analytics, it explicitly evaluates the consequences of decisions on future outcomes.
- **Iterative and adaptive.** It allows scenario simulation and sensitivity analysis, enabling decision-makers to adjust strategies as data and objectives evolve.

The value of prescriptive analytics lies in its **actionability** enabling organisations to systematically explore the impact of various choices and select the one that aligns best with business goals.

8.1.2 Comparison with predictive analytics

While both predictive and prescriptive analytics support decision-making, they differ in purpose and application. Table 8.1 summarises the key differences between these two approaches.

Table 8.1 Key differences between predictive analytics and prescriptive analytics in terms of objectives, methods, inputs, and outputs

Feature	Predictive analytics	Prescriptive analytics
Objective	Forecast likely future outcomes	Recommend optimal actions
Methodologies	Statistical modelling, machine learning	Optimisation, simulation, heuristics
Data input	Historical and current data	Includes predictive outputs + business constraints
Output	Probabilities, trends, forecasts	Decisions, action plans, policies
Tools	Regression, decision trees, neural networks	Linear programming, genetic algorithms, simulation models

Predictive analytics is akin to a weather forecast that tells you the likelihood of rain tomorrow. Prescriptive analytics is like a travel planner that recommends whether you should carry an umbrella, take a car, or change your schedule, considering the forecast and your constraints.

The use of predictive and prescriptive analytics in Australia’s Energy Market

An Australian example of this interplay can be seen in energy demand forecasting and optimisation. The **Australian Energy Market Operator (AEMO)** uses predictive models to forecast electricity demand. Prescriptive models then help determine optimal generation and distribution plans to meet this demand while minimising cost and emissions (AEMO, n.d.).

Australian Energy Market Operator (AEMO). (n.d.). *Forecasting and planning approach*. <https://aemo.com.au/energy-systems/electricity/national-electricity-market-nem/nem-forecasting-and-planning/forecasting-approach>

The Forecasting Approach diagram below represents how forecasting components build up to the high-level forecasting deliverables that AEMO publishes. The components are colour-coded to indicate their relevant documentation.

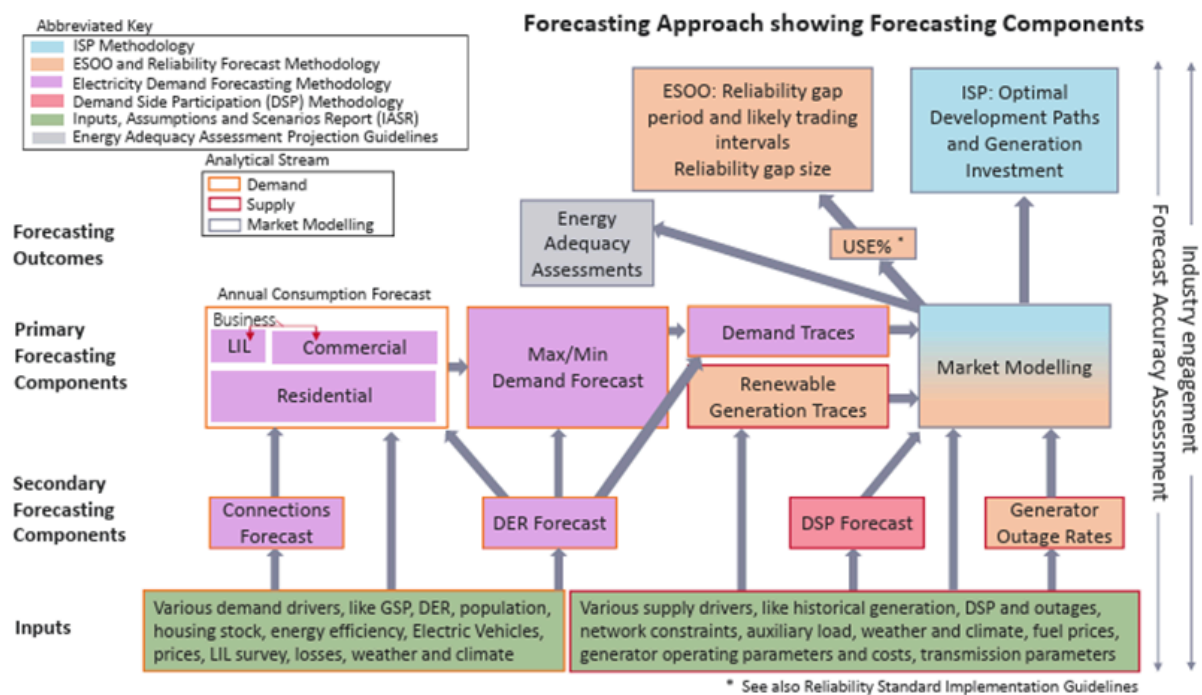


Figure 8.1. Forecasting approach showing forecasting components

Alternative Text for Figure 8.1

Forecasting approach showing forecasting components: Outlining the process of energy forecasting.

It shows **inputs flowing** into forecasting components and leading to forecasting outcomes. Inputs include demand drivers (e.g., population, electric vehicles, energy efficiency) and supply drivers (e.g., historical generation, outages, fuel prices).

Primary forecasting components include forecasts for annual consumption (split into Business, Large industrial loads (LIL), Commercial, Residential), Max/Min Demand Forecast, Demand Traces, and Renewable Generation Traces. These components feed into Energy Adequacy Assessments and Market Modelling.

Secondary Forecasting Components include Connections Forecast, Distributed Energy Resource (DER), Demand Side Participation (DSP), and Generator Outage Rates.

Outcomes of the process inform ESOO (Electricity Statement of Opportunities) and ISP (Integrated System Plan) methodologies, which cover reliability gaps, optimal generation investment, and generator outage rates. The diagram uses coloured boxes to differentiate methodologies and analytical streams for demand, supply, and market modelling. A vertical arrow on the right side emphasizes the importance of industry engagement and forecast accuracy assessment.

Guidelines and methodologies evolve according to improvement opportunities noted by AEMO and

stakeholders, and to reflect rule changes. Each document describes the approach used for determining modelling inputs, incorporating/publishing data, reporting forecast uncertainties, incorporating exogenous factors, and representing resource/network constraints affecting energy delivery.

8.1.3 Types of decisions supported

Prescriptive analytics plays a vital role in enhancing strategic decision-making by enabling organisations to evaluate complex, long-term scenarios and select optimal courses of action aligned with their overarching goals. In the context of strategic decisions, which often involve high-stakes commitments such as market entry, mergers and acquisitions, or infrastructure investments, prescriptive analytics integrates advanced modelling techniques, including optimisation algorithms and simulation, to recommend actions that maximise outcomes while minimising risks. Prescriptive analytics supports a wide spectrum of decision-making across three tiers.

8.1.3.1 Strategic decisions

These are high-level, long-term decisions that shape the direction of an organisation.

Prescriptive tools enable scenario analysis and optimal resource allocation under uncertainty, supporting robust strategic planning. Prescriptive analytics informs strategic decisions at governmental and societal levels, especially in sectors like urban planning, healthcare, and transport. Australian case studies, such as the use of digital twins for city-scale infrastructure management, demonstrate how prescriptive models simulate various future scenarios and recommend optimal investments in housing, mobility, and environmental resilience. These tools enable policymakers to weigh trade-offs among economic growth, environmental sustainability, and social equity, ensuring long-term strategies are both evidence-based and adaptive. By translating predictive insights into concrete recommendations, prescriptive analytics transforms data into actionable intelligence, supporting decisions that shape industries, cities, and communities for decades to come.

Scenario: Beyond data, towards sustainability – A Sydney case study on urban Digital Twins

As urban areas grapple with unprecedented challenges stemming from population growth and climate change, the emergence of urban digital twins offers a promising solution. A case study focusing on Sydney's urban digital twin, is a virtual replica integrating diverse real-time and historical data, including weather, crime, emissions, and traffic. Through advanced visualisation and data analysis techniques, the study has explored some applications of this digital twin in urban sustainability, such as spatial ranking of suburbs and automatic identification of

correlations between variables. Additionally, this also looks into modelling, employing machine learning to forecast traffic crash risks using environmental data, showcasing the potential for proactive interventions. The contributions of this case lie in the comprehensive exploration of a city-scale digital twin for sustainable urban planning, offering a multifaceted approach to data-driven decision-making.

New South Wales (NSW), Australia, digital twin¹ digitally represents the state's physical and geo-spatial features through satellite imagery, LiDAR, and other data sources. Like other live digital twins, the NSW digital twin aids in urban planning and asset monitoring, including roads, bridges, and utilities. It also contributes to emergency management by simulating natural disasters such as bushfires or flood (<https://arxiv.org/abs/2406.04902>).

¹<https://nsw.digitaltwin.terria.io/>

8.1.3.2 Tactical decisions

Prescriptive analytics supports tactical decisions by providing data-driven recommendations for medium-term planning and efficient resource allocation. These decisions typically focus on optimising processes over months or quarters, such as supply chain adjustments, workforce scheduling, inventory management, and marketing strategies. These relate to medium-term planning and resource management.

Scenario: Australian retailer Woolworths use prescriptive analytics to address competition in grocery retailing

Summary of Woolworths' position

- **Competition in Australian grocery retailing is vigorous and effective in practice.** The drivers of this intense competition include:
 - Fundamental changes in the way customers shop for groceries.
 - The significant expansion and success of “hard discounter”, ALDI.
 - The expansion into sale of groceries by a broad variety of non-supermarket retailers.
 - The entry and scaleup of Amazon.
- **Woolworths has clear financial incentives to compete to retain all of the items in every customer's basket** by setting competitive prices each week.

- **Woolworths competes with many effective rivals in Australian grocery retail**, and is not part of a “supermarket oligopoly” in a “highly concentrated supermarket sector”.
 - To attract a broad range of customers, Woolworths must compete hard on price and non-price dimensions including convenience and access, value, quality, range/choice and customer experience against a broad set of competitors.
 - This includes Coles, ALDI, independents, online retailers such as Amazon, category specialists and food specialty retailers who are competing for the same customers in relation to many of the same grocery products.
- **The quality and dynamism of competition in grocery retailing** cannot be gleaned from a narrow focus on supermarkets or a simplistic focus on concentration.
 - A grocery sector with four major supermarket-format players, a growing number of non-supermarket “category killers”, and multiple new and established online players indicates effective, dynamic and increasing competition.
- **Entry and expansion in the Australian grocery market is viable and attractive to committed players.** ALDI, Amazon, Costco, Bunnings, Chemist Warehouse, The Reject Shop and Catch provide compelling evidence of this.
- **New and growing grocery retailers can establish strong supplier networks.**
- **Access to retail sites is not a barrier to retail business establishment and growth.**
 - ALDI, Chemist Warehouse and other competitors (outlined above) are prime examples.
 - Woolworths acquires sites with the intention to develop new stores (or expand or refurbish existing stores) or eCommerce fulfilment capacity to drive growth and support its existing customer base or population growth.

Reference: Section 2 from <https://www.accc.gov.au/system/files/woolworths-group-supplementary-submission-acc-interim-report-9-december-2024.pdf>

8.1.3.3 Operational decisions

Prescriptive analytics enhances operational decisions by offering precise, real-time recommendations that support day-to-day activities with immediate impact. These decisions often involve scheduling,

dispatching, pricing adjustments, or inventory replenishment and require rapid responses to dynamic conditions. These are short-term, day-to-day decisions with immediate impact.

- In industries like retail, prescriptive analytics is used to optimise daily stock levels, pricing strategies, and promotional offers based on real-time sales data and customer behaviour.
- In logistics, it helps companies dynamically adjust delivery routes and fleet utilisation to account for traffic patterns and urgent orders.

By providing actionable insights grounded in current data and operational constraints, prescriptive analytics ensures organisations make informed, agile decisions that enhance efficiency, reduce costs, and improve customer satisfaction in the short term.

8.2.3 Theoretical underpinnings

Prescriptive analytics is grounded in several mathematical and computational disciplines. The four principal theoretical pillars are **optimisation**, **simulation**, **decision theory**, and **heuristics and metaheuristics**.

In this version of the book, the focus is on **optimisation**, as it provides a foundational and widely used framework for prescriptive analytics in business and information systems contexts. The remaining pillars are acknowledged as important but are beyond the scope of this edition and are not covered in detail. Readers interested in advanced or large-scale prescriptive analytics applications are encouraged to explore simulation, decision theory, and heuristic approaches in specialised texts.

8.2.3.1 Optimisation

Optimisation is at the core of prescriptive analytics. It involves finding the best solution from a set of feasible alternatives that maximise or minimise an objective function under constraints. Common types of optimisations include:

- **Linear Programming (LP)**. Solves problems where both the objective function and constraints are linear. It is widely used for resource allocation, production planning, and transportation problems.
- **Integer Programming (IP)**. Similar to LP but requires some or all decision variables to take integer values. This is useful in scheduling, workforce planning, and logistics where discrete units (like people or trucks) are involved.
- **Nonlinear Programming (NLP)**. Addresses problems with at least one nonlinear objective function or constraint. NLP is used in fields such as energy systems, engineering design, and finance where relationships between variables are not linear.

Linear Programming (LP)

Linear programming deals with problems where both the objective function and constraints are linear. It is widely used in operations such as transportation, scheduling and resource allocation.

Key components:

1. **Objective function.** What you are trying to optimise (e.g., profit, cost).
2. **Decision variables.** Quantities to determine (e.g., number of products to make).
3. **Constraints.** Restrictions on resources (e.g., labour hours, materials).
4. **Feasible region.** The set of all possible solutions that satisfy the constraints.
5. **Optimal solution.** The point in the feasible region where the objective is maximised or minimised.

The general form of an LP problem is:

Maximise or Minimise $Z = c_1x_1 + c_2x_2 + \dots + c_nx_n$
 Subject to:
 $a_{11}x_1 + a_{12}x_2 + \dots + a_{1n}x_n \leq b_1$
 $x_1, x_2, \dots, x_n \geq 0$

Excel comes with the ***add-in Solver***, which can readily obtain an optimal solution. The Excel Solver add-in is a tool used for what-if analysis, allowing users to find optimal values for a formula in a cell (the objective cell) subject to constraints on other cells. It is not enabled by default but can be easily added in Excel.

How to load the Solver add-in?

1. Open Excel: Launch the Microsoft Excel application.
2. Go to File: Click on the “File” tab.
3. Open Options: Select “Options”.
4. Navigate to Add-ins: Choose “Add-ins” from the left-hand menu.
5. Manage Excel Add-ins: In the “Manage” box at the bottom, select “Excel Add-ins” and click “Go”.

6. Enable Solver: In the “Add-Ins available” box, check the box next to “Solver Add-in” and click “OK”.
7. Verify installation: The Solver command should now be visible in the “Analyse” group on the “Data” tab.

Case Study – Western NSW Local Health District

This [case study](#) explains how **Western NSW LHD (WNSWLHD)** partnered with the Ministry of Health (MoH) to deliver the Rostering Proficiency Program to achieve some significant improvements through more effective rostering.

Right people, right skills, right place to deliver excellent patient care every time

Rostering is a critical enabler to patient safety, quality of care, staff wellbeing and organisational efficiency. Yet it is complex, requiring managers to balance often competing demands. Recognising this, WNSWLHD in partnership with MoH embarked upon a rostering improvement initiative to improve rostering practice and upskill roster managers.

The initiative comprised a series of structured educational webinars focused on different aspects of the rostering cycle underpinned by online training available via the NSW Health [Rostering Capability Framework \(RCF\)](#). Alongside this, roster managers received one-on-one coaching and support from a team of local rostering experts. Roster Managers embarked upon a 90-day challenge to demonstrate improved performance against key rostering metrics.

Sponsorship and local leadership

The initiative had strong sponsorship from the LHD Executive and facility managers which was critical in building momentum. Local expertise with an understanding of the rostering environment and specific challenges faced by roster managers was also invaluable in driving rostering improvements focused on individual roster characteristics.

A commitment to continuous improvement

WNSWLHD have recognised that to continue to achieve excellence in rostering, a commitment to continuous improvement is required. The WNSWLHD Rostering Proficiency Program is part of an ongoing effort to continue to develop and mature rostering capability across the LHD making use of the many rostering resources now available via the NSW Health Rostering Portal.

Key outcomes

The WNSWLHD Rostering Proficiency Program has shown demonstrable improvements in the timeliness, quality and effectiveness of rostering. Proven that with executive commitment, local engagement and dedicated support, significant improvements in rostering outcomes can be made. Key outcomes include:

- **8% (\$106k) decrease** in nursing agency fees resulting from more proactive planning of staffing needs and improved recruitment strategies
- **8.86 FTE reduction** in casual staffing resulting from improved utilisation of the permanent workforce
- **2.9% (\$45k) decrease** in nursing overtime due to a more proactive approach to managing shift vacancies in targeted facilities
- **2% decrease** in annual leave equating to a reduction of \$696,574 in leave liability
- **\$1.2m reduction** in leave liability resulting from a 3704 reduction in excessive leave days due to more proactive leave planning
- **66% (\$247k) decrease** in 'no minimum break' payments resulting from improved education and awareness
- An **increase in the timeliness** of rosters being published from an average of 2 to 6 days in advance of the roster period commencing.

To find out more

Go to [NSW Health Rostering Portal](#) or search for 'RCF' on [MyHealth Learning](#)

Prescriptive Analytics Objective

Determine the optimal number of nurses to assign to each shift per day so that:

- Every shift has enough nurses.
- Overtime is minimised.
- Total staffing cost is minimised.
- Constraints (legal and operational) are respected.

How Prescriptive Analytics Works Here

- Descriptive analytics provides past staffing patterns and current nurse availability.
- Predictive analytics forecasts patient inflow and workload for each day.
- Prescriptive analytics uses this data and LP to generate the best possible schedule.

The work is licensed under CCY BY 4.0: © State of New South Wales NSW Ministry of Health.

Integer Programming (IP)

Similar to LP but requires some or all decision variables to be integers. In many real-world situations, decision variables must take integer values (e.g., number of trucks, machines, or employees). IP problems are harder to solve but more realistic.

The three aspects of any model can be categorised based on the following questions:

- ***What do we control?*** refers to the decision variables that must take integer (whole number) values typically representing discrete choices or countable resources. These are the variables you are optimising through your model, subject to constraints.
- ***What do we want to achieve?*** refers to the objective of the optimisation model. This is the goal or target you are trying to reach by selecting the best possible combination of integer decisions. We want to optimise a measurable outcome usually by minimising a cost or maximising a benefit while satisfying all constraints.
- ***What constraints us?*** refers to the real-world limitations or rules that restrict what decisions we can make. Constraints are the conditions, limits, or rules that must be satisfied when making decisions. These could be resource limitations, policy rules, legal requirements, or operational capacities.

Case Study – Australian public transport

An Australian public transport operator may use **integer programming (IP)** to determine the number of buses to deploy on different routes during peak hours. Because buses are indivisible resources, decision variables must take **integer values**, it is not possible to assign a fraction of a bus to a route. The objective may be to minimise operating costs while ensuring passenger demand and service-level requirements are met. Although integer programming problems are generally harder to solve than linear programming problems, they provide more **realistic models** of real-world decision-making involving discrete resources such as vehicles, machines, or staff.

Nonlinear Programming (NLP)

Deals with nonlinear relationships in objective or constraints. When the relationships among variables are not linear, NLP is employed. These models are common in engineering, economics, and energy sectors.

Examples

Example 1: Renewable energy (solar power)

An Australian solar energy provider may use **nonlinear programming (NLP)** to optimise energy output while accounting for **nonlinear efficiency losses** caused by factors such as temperature, shading, and panel degradation. The relationship between solar irradiance, temperature, and power output is not linear, making NLP suitable for modelling and optimising system performance under varying environmental conditions.

Example 2: Emergency aviation and healthcare logistics

The Royal Flying Doctor Service (RFDS) may use **nonlinear programming (NLP)** to schedule aircraft for emergency medical retrievals across rural and remote Australia. Response time, fuel consumption, and aircraft performance vary nonlinearly with distance, weather conditions, and payload weight. NLP models help minimise response times while respecting aircraft availability, patient priority, and operational constraints.

8.2.4 Problem solving search methods

Problem-solving search methods encompass analytical techniques, algorithmic approaches, blind search, and heuristic search. These approaches differ in how they explore the solution space and determine stopping conditions, ranging from optimal solutions to satisficing outcomes. The key algorithmic search approaches and solution paths discussed in this section are illustrated in Figure 8.2.

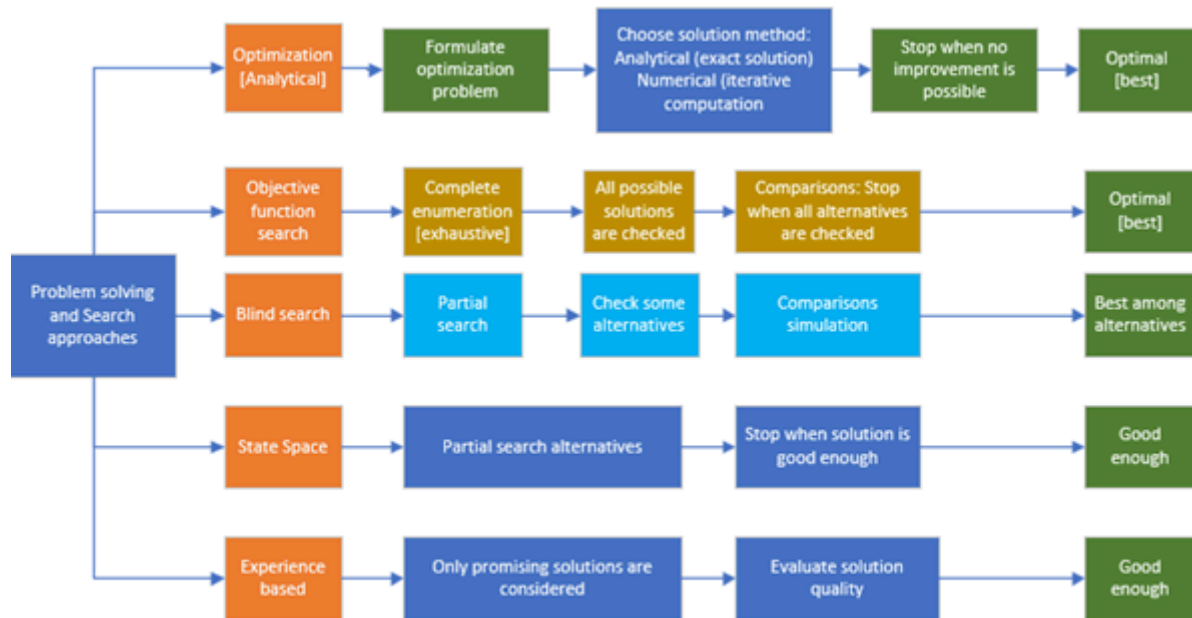


Figure 8.2. Algorithmic Search Approaches

Alternative Text for Figure 8.2

Flowchart titled 'Problem solving and Search approaches' outlining different optimisation and search strategies. Paths include analytical optimisation and exhaustive search leading to an optimal solution, as well as blind, state-space, and experience-based searches that evaluate partial or promising alternatives and stop when a solution is good enough.

8.2.4.1 Analytical techniques

Analytical techniques are mathematical and logical methods used to solve structured decision-making problems. As discussed in Section 8.3.2.1, Linear Programming (LP), Integer Programming (IP), and Non-Linear Programming (NLP) are widely used to derive optimal solutions under well-defined constraints.

In addition to these methods, several other analytical and search-based techniques are commonly applied, particularly when solution spaces are small, discrete, or computationally manageable.

Remaining analytical approaches include:

- **Exact analytical methods.** Derive an optimal solution using closed-form mathematics where feasible.
- **Numerical (iterative) optimisation.** Uses computational iteration to converge toward an optimal

solution.

- **Dynamic programming.** Solves complex problems by breaking them into smaller overlapping subproblems.
- **Branch and bound.** Systematically prunes solution regions using upper and lower bounds on solution quality.
- **Sensitivity analysis.** Tests how changes in inputs or constraints affect the optimal solution.

8.2.4.2 Objective function search (Complete enumeration)

Objective function search refers to methods that evaluate solutions by directly measuring them against an objective function. In its exhaustive form, it involves systematically checking every feasible alternative and selecting the best-performing option. This approach guarantees an optimal solution when all possibilities can be evaluated, but it becomes impractical when the number of alternatives is very large.

Common objective function search approaches include:

- **Complete enumeration (exhaustive search).** Evaluates every feasible alternative to identify the best solution.
- **Grid search.** Searches across a defined range of parameter combinations at fixed intervals.
- **Brute-force optimisation.** Tests all candidate solutions without shortcuts, typically used for small search spaces.
- **Comparative scoring search.** Assigns a score to each alternative based on the objective function and ranks results.
- **Full-factorial evaluation.** Examines all combinations of decision variables to determine the optimal configuration.

8.2.4.3 Blind search (Uninformed search)

Blind search methods explore the search space without using domain knowledge, heuristics, or cost estimates to guide decision-making. They treat all possible actions equally and expand states according to predefined rules. Blind search methods are primarily used as baseline techniques or when no heuristic information is available.

Common blind search approaches include:

- **Breadth-First Search (BFS).** Explores all nodes at the current depth before moving to the next level.
- **Depth-First Search (DFS).** Explores as deeply as possible along a branch before backtracking.
- **Uniform-Cost Search.** Expands the least-cost node first without heuristic guidance.
- **Iterative Deepening Search (IDS).** Combines BFS completeness with DFS memory efficiency.
- **Bidirectional Search.** Searches simultaneously from the start and goal states.

While blind search guarantees completeness in many cases, it suffers from high computational and memory costs and is rarely suitable for large-scale or real-time problems.

8.2.4.4 State space search (Heuristic search)

State space search represents a problem as a set of states and transitions between them. Unlike blind search, it often incorporates heuristic guidance to prioritise more promising states, reducing the number of alternatives explored. This makes it suitable for large or complex problems where exhaustive exploration is computationally infeasible.

Common heuristic and state-space search approaches include:

- **Greedy Best-First Search.** Expands the node that appears closest to the goal based on heuristic value.
- **A* Search.** Combines path cost and heuristic estimates to guarantee optimal solutions when the heuristic is admissible.
- **Hill Climbing.** Iteratively improves the current solution by selecting better neighbouring states.
- **Beam Search.** Limits the number of nodes expanded at each level to reduce computational cost.
- **Branch and Bound.** Prunes large portions of the search space using bounds on solution quality.

These approaches are well suited to routing, scheduling, and optimisation problems where speed is critical and exact solutions may be infeasible.

8.2.4.5 Experience-based search (Metaheuristic search)

Experience-based search methods use prior knowledge, probabilistic rules, or learning-based strategies to focus on promising solutions. Rather than exploring the full search space, these approaches evaluate solution quality iteratively and stop when further improvement is unlikely or unnecessary. They are useful for complex, non-linear, or high-dimensional problems where exact optimisation is impractical.

Common experience-based search approaches include:

- **Genetic algorithms.** Evolve solutions using selection, crossover, and mutation.
- **Simulated annealing.** Allows occasional worse moves to escape local optima and improve long-run results.
- **Particle swarm optimisation.** Uses a population of candidate solutions that collectively move toward better regions.
- **Tabu search.** Uses memory structures to avoid revisiting recently explored poor solutions.
- **Ant colony optimisation.** Mimics pheromone-based path selection to identify strong solution paths over time.

8.3 The steps involved in prescriptive analytics

Prescriptive analytics follows a structured, multi-stage process that transforms raw data into actionable decision recommendations. Each step builds on the previous one, ensuring that recommendations are data-driven, analytically sound, and practically feasible.

8.3.1 Data collection

The first and foundational step in the prescriptive analytics process involves collecting and aggregating data from a diverse range of sources. These include both internal organisational data such as sales records, supply chain logs, financial statements, customer transactions, and operational metrics – and external data obtained from third-party providers, government agencies, industry benchmarks, social media platforms, and sensor networks.

The richness, quality, and relevance of these data sources are critical, as they provide the context and input parameters required for downstream predictive modelling and optimisation. At this stage, it is also important to consider data accessibility, licensing constraints, update frequency, and interoperability across systems to ensure that the analytics process can scale effectively and produce reliable recommendations.

8.3.2 Data analysis

Once data has been collected and prepared, the next critical step is data analysis, during which raw information is transformed into meaningful insights. In this phase, analysts apply exploratory data analysis (EDA), statistical techniques, and advanced computational methods to uncover patterns, relationships, anomalies, and trends within the data.

Depending on the nature of the problem, techniques such as regression analysis, clustering, time-series forecasting, and machine learning algorithms (e.g., decision trees, support vector machines, and neural networks) may be employed. This stage supports a deeper understanding of the factors driving observed outcomes and provides the analytical foundation required for predictive and prescriptive modelling. Increasingly, automated and augmented analytics tools are used to accelerate this process while maintaining statistical rigour.

8.3.3 Modelling

Following data analysis, a mathematical or computational model is developed to convert analytical insights into actionable decisions. This model forms the core of the prescriptive analytics process and is designed to simulate, optimise, or recommend specific courses of action based on defined objectives and constraints.

The modelling stage typically incorporates financial, operational, regulatory, and strategic considerations to ensure that recommendations are both analytically sound and organisationally relevant. Model accuracy depends on how well real-world conditions such as uncertainty, interdependencies, and business rules—are

represented. As a result, effective modelling requires collaboration between data scientists, domain experts, and decision-makers. Models are commonly validated using historical data and stress-tested under multiple scenarios to assess robustness and reliability.

8.3.4 Recommendation generation

Once the model has been built and validated, it is used to generate data-driven recommendations that guide decision-makers toward optimal or near-optimal actions. Rather than producing a single outcome, prescriptive analytics often generates a set of alternative actions, each accompanied by trade-offs, projected outcomes, and risk assessments.

For example, in transportation planning, a prescriptive model may recommend multiple delivery routes, one minimising cost, another reducing delivery time, and a third balancing both objectives. Each option may include quantitative projections such as fuel consumption, delivery time, resource utilisation, and carbon emissions. In strategic planning contexts, models may evaluate competing investment strategies under uncertain market conditions, highlighting expected returns, risk exposure, and compliance implications.

This stage enables organisations to move beyond predicting what will happen to determining what should be done, supporting decisions that are transparent, defensible, and adaptable. Advanced systems may also provide interactive dashboards or what-if analysis tools, allowing users to explore alternative scenarios and refine decisions in real time.

8.3.5 Recommendation evaluation

The final stage of the prescriptive analytics process involves evaluating the recommended actions to assess their feasibility, impact, and alignment with organisational objectives. While models may identify mathematically optimal solutions, it is essential to consider practical constraints such as budget limitations, operational capacity, regulatory requirements, and organisational culture.

This evaluation phase often involves stakeholder review, cost–benefit analysis, risk assessment, and pilot testing before full implementation. Feedback from this stage may lead to adjustments in the model or recommendations, creating an iterative improvement loop. By incorporating human judgement alongside analytical outputs, organisations ensure that prescriptive analytics delivers solutions that are not only optimal in theory but also effective in practice.

8.4 Related Australian case studies

Several Australian organisations have embraced prescriptive analytics to improve efficiency and competitiveness.

Australian Rail Industry: Timetable optimisation and delay management

Sydney Trains collaborated with the University of Technology Sydney (UTS) to develop a software toolkit aimed at improving timetable updates, train maintenance scheduling, and staff rostering, particularly at congested locations. The project focused on understanding the ripple effects of primary delays and their escalation into secondary delays across the network.

Application of prescriptive analytics:

By integrating diverse data sources including Opal card usage, signal records, CCTV footage, staff rosters, and weather data the team standardised the data to feed into advanced AI models. These models predicted system performance and identified potential bottlenecks, enabling the development of optimised timetables and maintenance schedules. The prescriptive analytics approach allowed for proactive decision-making to minimise delays and improve overall network efficiency

Australian Banking Sector

Scenario 1: Personalised financial product recommendations

Commonwealth Bank Australia (CBA) utilises its Customer Engagement Engine (CEE), to deliver personalised financial product recommendations to customers. This system processes vast amounts of data daily to understand individual customer needs and preferences.

Application of prescriptive analytics:

The CEE analyses customer data to predict needs and prescribes tailored financial solutions, such as loan options or savings plans, aligning with each customer's financial situation and goals. This approach ensures that recommendations are both personalised and compliant with regulatory standards.

Scenario 2: Risk and compliance management

To maintain regulatory compliance and manage risks, CBA has implemented AI-driven systems that monitor transactions and customer interactions for potential issues.

Reference: <https://www.commbank.com.au/articles/newsroom/2022/06/CBA-artificial-intelligence-usages.html>

Application of prescriptive analytics:

By continuously analysing transaction data, the system can detect anomalies or patterns indicative of fraud or non-compliance. When such patterns are identified, the system prescribes specific actions, such as flagging transactions for review or alerting compliance officers, ensuring timely and appropriate responses.

Australia Post optimising last-mile delivery with prescriptive analytics

Australia Post faced unprecedented challenges during the COVID-19 pandemic, with a significant surge in parcel deliveries due to increased online shopping. The organisation needed to enhance its last-mile delivery efficiency to meet customer expectations across both urban and regional areas.

To address these challenges, Australia Post partnered with e2open to implement advanced last-mile delivery solutions. These solutions aimed to improve route planning, optimise delivery sequencing, and provide accurate estimated times of arrival (ETAs) for parcels. The integration of these technologies was crucial in managing the complexities of increased parcel volumes and ensuring timely deliveries.

Application of prescriptive analytics:

Prescriptive analytics played a pivotal role in transforming Australia Post's last-mile delivery operations:

- **Data Integration:** By consolidating data from various sources, including parcel tracking systems, traffic conditions, and delivery schedules, Australia Post could gain a comprehensive view of its delivery network.
- **Route Optimisation:** Advanced algorithms analysed the integrated data to determine the most efficient delivery routes, considering factors like traffic patterns and delivery time windows.
- **Dynamic Sequencing:** The system provided drivers with optimised stop lists, allowing for manual adjustments when necessary, enhancing flexibility and responsiveness.

These prescriptive analytics applications allowed Australia Post to navigate the complexities of last-mile delivery effectively, ensuring efficiency and reliability in service.

Australian Pharmaceutical sector

The Australian pharmaceutical industry is a critical sector contributing to national health outcomes and economic growth. To maintain competitiveness and drive innovation, the industry must make strategic decisions about where to invest in R&D, which technologies to prioritise, and how to respond to global trends in pharmaceutical innovation.

This case is based on a real-world study published by IP Australia, which analysed pharmaceutical patents filed in Australia over a 20-year period. The study aimed to help policymakers and businesses make data-driven, forward-looking decisions using prescriptive analytics based on patent data.

Application of prescriptive analytics:

Prescriptive analytics in this case is applied to recommend specific actions based on insights drawn from historical patent data. After identifying patterns and projecting trends using descriptive and predictive analytics, the report goes a step further to prescribe strategic choices for the future.

Example recommendations of prescriptive insights from the case include:

1. Invest more in biologics R&D, as biologic-related patents have grown steadily, indicating global and local importance.
2. Encourage domestic patenting by Australian companies and universities, since a majority of pharmaceutical patents in Australia come from foreign entities.
3. Support collaboration between universities and industry to commercialise promising pharmaceutical technologies based on strong academic patent portfolios.
4. Improve innovation policy and infrastructure to address Australia's relatively low ranking in high-value pharmaceutical patents.

These recommendations are not just reactive but forward-looking, offering actionable strategies aligned with Australia's national interests in health, research, and economic development.

Reference: *A Patent Analytics Study on the Australian Pharmaceutical Industry*. IP Australia. Creative Commons Attribution 4.0 International (CC BY 4.0)

8.5 Test your knowledge

Drag the words into the correct boxes.

- 1) Prescriptive analytics integrates predictive outputs, business rules, and models to recommend feasible actions under real-world .
- 2) Prescriptive analytics builds on insights from analytics and forecasts from analytics to recommend actions.
- 3) Prescriptive analytics is most valuable when decisions involve multiple options, clear constraints, and competing .
- 4) When exact optimisation is too slow for real-time decisions, prescriptive analytics may use methods to produce near-optimal solutions.
- 5) A transport operator allocating whole buses to routes while minimising operating cost under capacity constraints should use programming rather than programming.

predictive

heuristic

linear

constraints

objectives

descriptive

optimisation

integer



The interactive version of this H5P content is available at:

<https://oercollective.caul.edu.au/decoding-data/?p=736#h5p-7>

An alternative source for this activity: <https://uts.h5p.com/content/1292858510784816029>

Additional readings

- CGI. (n.d.). A prescriptive analytics project for maximising healthcare value generation: Sir Mortimer B. Davis Jewish General Hospital. Retrieved 8th July 2025, from <https://www.cgi.com/au/en-au/media/viewpoint/prescriptive-analytics-project-maximizing-healthcare-value-generation-sir-mortimer-b>
- Dataversity. (2025). Prescriptive analytics use cases. Retrieved 1st July 2025, from <https://www.dataversity.net/prescriptive-analytics-use-cases/>
- Indeed Editorial Team. (2025). What is prescriptive analytics? Benefits, uses and examples. *Indeed Australia*. Retrieved 6th July 2025, from <https://au.indeed.com/career-advice/career-development/prescriptive-analytics>
- IP Australia. (2018). *A patent analytics study on the Australian pharmaceutical industry*. Retrieved 7th July 2025 from <https://www.ipaustralia.gov.au/tools-and-research/professional-resources/data->

[research-and-reports/publications-and-reports](#)

- TBM Consulting Group. (2018). Prescriptive analytics case study: Make profitable decisions. Retrieved 3rd July 2025 from <https://tbmcg.com/resources/blog/prescriptive-analytics-and-profit-based-decision-making/>

About the author



Associate Professor Meena Jha
CENTRAL QUEENSLAND UNIVERSITY

<https://staff-profiles.cqu.edu.au/home/view/1248>

With a degree in Electronics and Communication Engineering and a Master and Ph.D. in the area of Computer Science, I am a researcher and educator in the field of computer science and IT in the School of Engineering and Technology (SET), at Central Queensland University (CQUniversity) in Sydney, Australia. My research interests include in particular Big Data and Artificial Intelligence, Expert Systems, Business Intelligence, Real-Time Analytics, Legacy System Modernization, the Internet of Things (IoT), and Academic Integrity in education. I have developed courses on Information Systems Analysis and Business Intelligence using Big Data and Data Science. I am also working in the area of learning analytics, which is an emerging field and uses sophisticated analytic tools to improve learning and education. It draws from and is closely tied to, a series of other fields of study including business intelligence, web analytics, academic analytics, educational data mining, and action analytics. I am passionate about women in STEM and working on how to improve and enhance women's power in the STEM field.

9. BRINGING IT ALL TOGETHER

A Summary of Key Concepts in Business Intelligence

Amara Atif

9 Introduction

You have reached the end of *Decoding Data: A Complete Guide to Business Intelligence*. Across eight chapters, this book has guided you through the core ideas, tools, frameworks, and practical skills that define modern business intelligence. Each chapter has built upon the last, beginning with the fundamental nature of data and progressing to the sophisticated analytical techniques organisations use to inform and optimise decision-making.

This final chapter offers an opportunity to step back and view the discipline as an integrated whole. It revisits the central concepts from each chapter, highlights the connections between them, and reinforces why each idea matters in real-world contexts. Whether you are a student preparing for a career in data, analytics, or management, or a professional seeking to strengthen your BI capabilities, use this chapter to consolidate your understanding and reflect on the broader narrative and strategic importance of business intelligence.

9.1 What each chapter covered

Chapter 1 – Data to Knowledge Continuum (Amara Atif)

Everything in business intelligence begins with data. Chapter 1 grounded you in what data actually represents: raw, unprocessed facts that, on their own, provide little meaning. The central framework introduced was the DIKW hierarchy, the progression from Data to Information to Knowledge to Wisdom illustrating how each stage adds context, interpretation and value. Without this foundation, the concepts explored in later chapters would lack coherence and direction. The chapter also examined the different types and sources of data: structured and unstructured, quantitative and qualitative, primary and secondary, internal and external. You explored what makes data analytics-ready, including qualities such as accuracy, completeness, consistency, timeliness and security. In addition, you worked through the four key stages of data pre-processing: consolidation, cleaning, transformation and reduction. Australian case examples from Woolworths, Coles, Toll Group and Telstra demonstrated that these principles are not merely theoretical constructs, but essential components of everyday business operations.

Chapter 2 – Rethinking Business Decisions: The Synergy of Decision Support System, Analytics, Business Intelligence,

Artificial Intelligence, and Data Science (Amara Atif & M.A. Qureshi)

If Chapter 1 explained what data is, Chapter 2 examined how organisations use it to make better decisions. The chapter introduced Decision Support Systems (DSS) as the foundation of data-driven decision-making, tracing their evolution from simple database reporting tools in the 1960s to today's AI-enabled intelligent systems. You explored the major forces reshaping business decision-making, including rapid technological advancement, globalisation, the exponential growth of data, shifting consumer expectations, and the volatility triggered by events such as the COVID-19 pandemic. The chapter also positioned the broader ecosystem of data-driven technologies — Analytics, Business Intelligence, Artificial Intelligence, and Data Science demonstrating how these capabilities operate synergistically rather than independently. Australian case studies featuring Barossa Valley wine exporters, The Iconic, and Woolworths brought these concepts to life, illustrating how data-driven tools translate into strategic insight, competitive advantage, and real-world organisational impact.

Chapter 3 – Architecture of Business Intelligence (Amara Atif)

Understanding why organisations need BI naturally leads to the next question: how does it actually work? Chapter 3 took you inside the infrastructure of business intelligence systems, explaining the components that transform raw data into meaningful reports, dashboards, and actionable insights. The chapter outlined the core building blocks of BI architecture: data sources, the ETL (Extract, Transform, Load) process, data warehouses and data lakes, analytics tools, and reporting systems. A crucial distinction was drawn between OLTP systems, which manage real-time operational transactions, and OLAP systems, which enable complex analytical queries across historical data. The fictional ShopSmart retail scenario made this technical architecture tangible and accessible. You also examined different BI architectural frameworks, from traditional on-premise data warehouses to modern cloud-based solutions and considered how organisations evaluate scalability, flexibility, and cost when selecting the structure that best fits their strategic needs.

Chapter 4 – Statistical Modelling for Business Analytics (Yang Yang)

Business intelligence is not merely about collecting and organising data, it is about analysing it rigorously and drawing conclusions that withstand scrutiny. Chapter 4 provided the statistical foundations necessary to do precisely that. Using Sydney's rental market as a running example, the chapter introduced the distinction between cases and variables, examined the four core data types (nominal, ordinal, discrete, and continuous), and explained how to select appropriate visualisations for each. You were also introduced to probability distributions particularly the normal distribution and its importance in modelling real-world phenomena, as well as correlation analysis for measuring linear relationships between variables. These statistical concepts form the foundation of advanced analytics techniques explored in later chapters,

including predictive modelling and machine learning, and they underpin sound, evidence-based decision-making in practice.

Chapter 5 – Data/Information Visualisation (Florence Mwagwabi)

Chapter 5 explored the rich history and evolving practice of data visualisation, from the earliest statistical graphics of the seventeenth century including William Playfair's foundational charts and Florence Nightingale's influential rose diagrams to today's AI-assisted and augmented reality visualisation tools. More importantly, the chapter emphasised that effective visualisation is not simply about creating charts, but about communicating insight with clarity and integrity. You learned to apply Gestalt principles and preattentive attributes to direct attention strategically, to select appropriate chart types from the broader visualisation taxonomy, and to construct persuasive, evidence-based data narratives. The chapter also examined common visualisation errors and cognitive biases, alongside design principles that ensure data is represented accurately, ethically, and without misleading the audience. Consistent with the applied focus of this book, the chapter used Australian Airbnb data to illustrate business intelligence concepts and visualisation techniques. The dataset included information on reservation patterns, pricing, and regional performance, allowing you to see how BI tools extract actionable insights from real-world data. This practical application reinforced the critical role of visualisation in supporting informed decision-making.

Chapter 6 – Information Dashboards (Amara Atif)

Chapter 6 brought visualisation into its most common organisational setting: the business dashboard. It distinguished between two foundational concepts: information visualisation (the descriptive view that shows what is happening) and visual analytics (which combines visualisation with analytical algorithms to explore why something is happening and what may occur next). You examined different types of business reports (metric management reports, dashboard-style reports, and balanced scorecard reports), and learned when each is most appropriate. The chapter then focused on effective dashboard design, exploring the distinction between static and dynamic dashboards, and demonstrating how layout, colour, typography, negative space, and interactivity influence clarity and usability. The principles of data storytelling were reinforced, showing how dashboards can move beyond displaying metrics to communicating strategic insight. The chapter incorporated a range of hypothetical business scenarios across sectors including healthcare and higher education to demonstrate how dashboards support both operational control and strategic oversight. A detailed case study of a South Australia health emergency department further illustrated how well-designed dashboards can improve visibility of patient flow, resource allocation, and performance monitoring in time-sensitive environments. These examples emphasised the practical impact of dashboard design on organisational effectiveness and decision quality.

Chapter 7 – Predictive Analytics (Marcella Papini)

Chapter 7 moved from describing the past and present to forecasting the future. Predictive analytics

uses historical data, statistical algorithms, and machine learning to answer the central business question: *What will happen?* The chapter incorporated a range of hypothetical business scenarios across sectors including retail, finance, healthcare, and logistics to demonstrate how predictive models are applied in real organisational contexts. The chapter introduced the two fundamental families of predictive modelling: supervised learning methods (including classification and regression, which learn from labelled historical data) and unsupervised learning methods (including clustering and association rule learning, which identify patterns without predefined labels). You explored four major predictive techniques in depth: data mining (uncovering hidden patterns in large datasets using the CRISP-DM framework), text mining (extracting insights from unstructured text), web analytics (understanding user behaviour online), and social media analytics (leveraging social data for business intelligence). Applications across finance, healthcare, and retail illustrated how these techniques scale across organisations of different sizes and levels of analytical maturity.

Chapter 8 – Prescriptive Analytics (Meena Jha)

The final chapter completed the analytics journey by addressing the most action-oriented question of all: *What should we do?* Prescriptive analytics moves beyond forecasting outcomes to recommending optimal or near-optimal courses of action within defined objectives and constraints. The chapter positioned prescriptive analytics within the four-stage analytics framework: descriptive (what happened?), diagnostic (why did it happen?), predictive (what will happen?), and prescriptive (what should we do?), demonstrating how each stage builds logically upon the insights of the previous one. Core techniques included mathematical optimisation, simulation modelling, and heuristic methods. You examined their applications across industries, including airline scheduling, supply chain coordination, pricing strategy, and resource allocation. A range of Australian scenarios illustrated these concepts in practice: Woolworths using prescriptive analytics to respond to competitive pressures in grocery retailing; Australia Post optimising last-mile delivery routes; timetable optimisation and delay management in the Australian rail industry; and decision optimisation challenges in the Australian banking and pharmaceutical sectors. A detailed case study of the Western NSW Local Health District further demonstrated how prescriptive analytics can support resource planning and service delivery in complex public health environments. Throughout the chapter, you explored the critical distinction between identifying mathematically optimal solutions (when computationally feasible) and generating near-optimal solutions using heuristics when time, complexity, and scale demand practical trade-offs. The chapter also addressed important considerations surrounding model transparency, ethical implications, and the integration of prescriptive analytics with AI-driven systems, reinforcing that the ultimate goal of business intelligence is not simply insight, but informed, responsible, and effective action.

9.2 How the chapters connect

One of the most important insights you can take from this book is that these eight chapters are not separate topics — they form a single, integrated system. The table below shows how the BI journey flows from one chapter to the next:

Table 9.1: Chapter focus and core questions answered

Chapter	Title	Core question answered	Key concepts
Chapter 1	Data to Knowledge Continuum	What is data and how do we prepare it?	DIKW hierarchy, data taxonomy, data quality metrics, pre-processing
Chapter 2	Rethinking Business Decisions: The Synergy of Decision Support System, Analytics, Business Intelligence, Artificial Intelligence, and Data Science	How do organisations use data to make better decisions?	DSS evolution, drivers of change, analytics/BI/AI/data science synergy
Chapter 3	Architecture of Business Intelligence	How does the infrastructure move data to decision-makers?	BI components, OLTP vs OLAP, ETL, data warehouses, architecture frameworks
Chapter 4	Statistical Modelling for Business Analytics	How do we analyse data rigorously?	Data types, distributions, correlation, statistical reasoning
Chapter 5	Data/Information Visualisation	How do we represent data meaningfully?	Visualisation history, design principles, chart taxonomy, Gestalt principles, data storytelling
Chapter 6	Information Dashboards	How do we communicate insights effectively?	Info visualisation vs visual analytics, report types, dashboard design, data storytelling
Chapter 7	Predictive Analytics	What will happen?	Supervised/unsupervised learning, data mining, text mining, web & social analytics
Chapter 8	Prescriptive Analytics	What should we do about it?	Optimisation, simulation, decision theory, heuristics, ethical considerations

The journey moves from raw data (Chapter 1) through the decision-making systems that use it (Chapter 2) and the infrastructure that powers those systems (Chapter 3), then builds the analytical capabilities needed to generate insight (Chapters 4-8). Visualisation threads through the middle of this journey (Chapters 5 and 6) as the bridge between analysis and action. Predictive and prescriptive analytics (Chapters 7 and 8) represent the frontier, where BI moves from understanding the past to shaping the future.

9.3 The analytics framework at a glance

One of the most useful mental models from this book is the four-stage analytics framework woven throughout the book:

Table 9.2: The analytical framework at a glance

Type	Question	Chapters
Descriptive	What happened?	1, 3, 4, 5, 6
Diagnostic	Why did it happen?	2, 4
Predictive	What will happen?	7
Prescriptive	What should we do?	8

A note on the Australian context

Throughout this book, Australian organisations including Woolworths, Coles, Toll Group, Telstra, The Iconic, Barossa Valley wine exporters, Airbnb, Sydney real estate agencies, and many others have been used to ground abstract concepts in familiar and practical realities. Australia's business landscape provides rich and diverse examples of how business intelligence is applied across retail, logistics, healthcare, finance, agriculture, property, and public services.

As you move forward in your studies or career, continue seeking out these local applications. The global frameworks and analytical methods you have learned here translate directly into the Australian context. In an increasingly data-driven economy, organisations across Australia are investing heavily in analytics capabilities, and demand for strong BI skills remains high and growing. The ability to apply these concepts thoughtfully within your own industry and environment will distinguish you as a capable and confident data professional.

9.4 Key takeaways from the book

Here are the most important ideas to carry with you from *Decoding Data*:

- **Data quality underpins everything.** Decisions are only as good as the data behind them. Every chapter in this book rests on the foundation established in Chapter 1: accurate, complete, timely, secure, and well-prepared data.
- **BI is a system, not a tool.** Business intelligence is the integrated result of high-quality data, robust architecture, rigorous analysis, effective visualisation, and sound managerial judgement. No single component delivers value in isolation.
- **Analytics is a progression.** Descriptive, diagnostic, predictive, and prescriptive analytics each address a distinct question and build upon one another. The full power of BI emerges when all four stages are aligned and integrated.
- **Visualisation is communication.** A chart or dashboard is not merely a technical output – it is a message. Thoughtful design, clarity, and audience awareness transform insight into influence and

action.

- **Statistics is your analytical language.** The ability to describe distributions, measure relationships, assess uncertainty, and interpret results forms the foundation for advanced methods such as machine learning, simulation, and optimisation.
- **Context shapes application.** The Australian case studies throughout the book demonstrate that BI principles must be adapted to industry, organisational scale, regulatory environments, and strategic priorities. Understanding context is itself a critical analytical skill.
- **The field is constantly evolving.** Artificial intelligence, real-time streaming data, augmented reality, and cloud-native architectures are expanding what organisations can achieve. The conceptual foundations developed in this book equip you to adapt confidently to that ongoing transformation.

9.5 A final thought

Business intelligence is ultimately about one thing: *enabling better decisions*. Every concept in this book, from the DIKW hierarchy and data pre-processing to BI architecture and prescriptive optimisation serves that single purpose. Data is abundant, but value is created only when it is transformed into insight and translated into thoughtful action. Organisations and individuals who can navigate the full journey from raw data to informed, wise decisions will be the ones who create lasting impact in an increasingly data-driven world.

You now have the foundations to be one of them.

About the author



Dr Amara Atif
UNIVERSITY OF TECHNOLOGY SYDNEY
<https://profiles.uts.edu.au/Amara.Atif>

Dr Amara Atif is a Senior Lecturer in the School of Computer Science, Faculty of Engineering and IT (FEIT), at the University of Technology Sydney (UTS), Australia. Amara is an expert in educational technology, with a comprehensive skill set that includes research and development in the design of technology-enhanced learning environments, human-computer interaction, artificial intelligence in education, and learning analytics/educational data mining.

At UTS, she developed the undergraduate subject “Business Intelligence”, integrating culturally informed Indigenous knowledge systems through the FEIT Indigenous Graduate Attribute. She has received

multiple UTS Vice Chancellor's Learning and Teaching Awards, including recognition for data-informed teaching practices and inclusive assessment. Passionate about the accessibility and sustainability of educational resources, Amara is an advocate for Open Educational Resources (OER). *Decoding Data* continues this mission, bringing rigorous Business Intelligence concepts to learners through an open, practitioner-oriented resource grounded in Australian business contexts.

REVIEW STATEMENT

UTS ePress is committed to publishing high-quality open textbooks that meet the needs of students and educators. Quality assurance is central to our mission to provide reliable, inclusive, and pedagogically sound open educational resources.

Each textbook undergoes a rigorous, multi-stage review process designed to ensure accuracy, usability and the meeting of publication standards for OERs (Open Educational Resources). This process includes:

- Peer review by subject matter experts
- Editorial review for clarity, consistency, and structure
- Copyright and licensing checks
- Accessibility checks
- Professional proofreading and copyediting by an accredited editor
- Final quality assurance and usability testing by the publisher and editor

Peer Review

Each chapter was independently reviewed by at least one expert with specialist knowledge in business intelligence. Reviews were guided by a structured [template](#) adapted from the Rebus Community framework, covering criteria such as:

- Comprehensiveness and accuracy
- Clarity and consistency
- Modularity, structure, and flow
- Diversity and inclusion

We recognise that peer reviewing is a generous act of service. This book would not be the robust, valuable resource it is without the feedback and input of our reviewers. We gratefully acknowledge the contributions of:

- [Dr Li Guan](#), University of Technology Sydney, Australia
- [Dr Víctor G. Alfaro-García](#), Universidad Michoacana de San Nicolás de Hidalgo, México
- [Associate Professor Dr Walayat Hussain](#), Australian Catholic University, Australia
- [Dr Betzabe Ruiz Morales](#), Universidad Michoacana de San Nicolás de Hidalgo, México
- [Dr Meena Jha](#), Central Queensland University, Australia
- [Mohammad Sadegh Khorshidi](#), University of Technology Sydney, Australia

Ongoing Commitment

In addition to pre-publication review, UTS ePress is committed to continuous improvement through post-publication evaluation, including student feedback and holistic reviews of the text. We welcome comments and feedback to utsepress@uts.edu.au.

ACCESSIBILITY STATEMENT

At UTS ePress, we believe that education must be available to everyone which means supporting the creation of free, open, and accessible educational resources. We are actively committed to increasing the accessibility and usability of the textbooks we produce.

Accessibility features of the web version of this resource

The web version of this resource has been designed with accessibility in mind by incorporating the following features:

- It has been optimized for people who use screen-reader technology.
 - all content can be navigated using a keyboard
 - links, headings, and tables are formatted to work with screen readers
 - images have alt tags
- Information is not conveyed by colour alone.

Other file formats available

In addition to the web version, this book is available in a number of file formats including PDF, EPUB (for eReaders), and various editable files. Choose from the selection of available file types from the 'Download this book' drop-down menu. This option appears below the book cover image on the [eBook's landing page](#).

Third-Party Content

In some cases, our open text includes third-party content. In these cases, it is often not possible to ensure accessibility of this content.

Accessibility Improvements

While we strive to ensure that this resource is as accessible and usable as possible, we know there's more that can be done. We are always looking for ways to make our resources more accessible. If you have problems accessing this resource, please contact utsepress@uts.edu.au to let us know so we can fix the issue.

Copyright Note: This accessibility disclaimer is adapted from [BCampus's Accessibility Toolkit](#), licensed under a [Creative Commons Attribution 4.0 International license](#) and University of Southern Queensland's [Enhancing Inclusion, Diversity, Equity and Accessibility \(IDEA\) in Open Educational Resources \(OER\)](#) licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License](#).

VERSIONING HISTORY

This page provides a record of changes made to this textbook. Each set of edits is acknowledged with a 0.01 increase in the version number. The exported files for this toolkit reflect the most recent version.

If you find an error, please contact utsepress@uts.edu.au

Version	Date	Change	Details
1.01	14 Feb 2025	Published five chapters	Published the Chapters: Data to Knowledge Continuum; Rethinking Business Decisions; Architecture of Business Intelligence; Statistical Modelling for Business Analytics; Information Dashboards
1.02	3 Dec 2025	Removed glossary, updated book cover, addition of authors, addition of Review Statement, updated Acknowledgements	In preparation for publication of additional chapters: updated book cover, added authors, removed glossary, added Review Statement, updated Acknowledgements
2.01	2 Mar 2026	Published additional 4 chapters and forewords.	Published the additional Chapters: Data/Information Visualisation, Predictive Analytics, Prescriptive Analytics, Bringing it all together. Updated About this book, About the authors, Acknowledgement of Country.
2.02	13 Mar 2026	Minor edits and disclaimer added	Added paragraph about currency of content to 'About this book' in front matter Minor fixes: table widths for PDFs; internal links from front matter; styling options in the book theme
2.03	26 Mar 2026	h5p links updated	For ongoing access, provided alternative links for h5p activities on the UTS h5p hub

FEEDBACK AND CORRECTIONS

If you'd like to report a problem, share feedback or request corrections, please contact the Editor amara.atif@uts.edu.au or Publisher utsepress@uts.edu.au.

Thank you for your help!