Data for Decision Makers: Data Concepts and Applications

Course Handbook

Proochista Ariana   Ernest Guevarra

07 June 2025

Table of contents

Preface

In today's data-driven world, the responsibility of public service demands more than experience and intuition; it requires evidence-based decision-making grounded in a deep understanding of data. For government officials at all levels, from local administrators to national policymakers, data is not just a tool - it is an indispensable asset in crafting policies that are effective, equitable, and accountable. Data for Decision Makers is developed with you in mind: to support those entrusted with public leadership in leveraging data to serve communities more effectively.

Across the domains of public health, education, transportation, environmental policy, and beyond, the availability of data has never been greater. But with this abundance comes complexity. Making sense of it - identifying relevant patterns, understanding root causes, evaluating outcomes, and anticipating future trends - requires more than access. It demands a strong foundation in the principles and practices of modern data use.

This course highlights how data literacy empowers government officials to navigate uncertainty, combat misinformation, and design policies that truly respond to the needs of the public. From statistical reasoning and geographic information systems to predictive modelling and real-time dashboards, the tools of data are transforming governance. Understanding these tools is essential to strengthening transparency, accountability, and public trust.

This course bridges the gap between technical expertise and policy leadership. It offers clear, accessible explanations of core data concepts alongside practical examples from the public sector. Whether your role involves strategic planning, budget allocation, programme evaluation, or legislative development, this course will help you make more informed, timely, and impactful decisions.

Public service is a profound responsibility. By embracing the potential of data, government leaders can enhance their ability to meet that responsibility with clarity, foresight, and integrity.

Part I

Data Concepts

1 Introduction

In an era defined by information, the ability to make sound decisions increasingly hinges on the intelligent use of data. Across sectors and industries, from healthcare and education to finance and public policy, decision-makers are confronted with unprecedented volumes of information. Yet, it is not the sheer quantity of data that holds value, but our capacity to interpret, understand, and apply it effectively.

Data is more than numbers on a spreadsheet; it is the language of modern insight. When approached with the right tools and understanding, it becomes a powerful asset for identifying patterns, predicting outcomes, evaluating strategies, and ultimately, improving results. For decision-makers, this means developing fluency not just in reading reports, but in questioning assumptions, validating sources, and interpreting results within context.

Understanding modern data concepts - from statistical reasoning and data visualisation to machine learning and real-time analytics - is no longer optional. It is foundational. These concepts empower leaders to move beyond intuition and anecdote, and toward evidence-based action. As data continues to shape the world around us, the ability to engage with it critically and creatively is becoming an essential skill.

This course aims to equip its participants with both the conceptual grounding and practical knowledge to navigate this landscape. Whether you are a seasoned executive, a policy analyst, or an emerging leader, this course is designed to bridge the gap between data science and decision-making. It demystifies the tools and techniques of modern data analysis and offers real-world applications that demonstrate how data can drive progress and innovation.

Good decisions are not just supported by data; they are shaped by those who know how to use it wisely.

1.1 Data-driven decision-making

Data-driven decision-making or DDDM refers to the process of making decisions based on data and information rather than intuition or experience alone. It involves collecting, analysing, interpreting, and presenting data to support decision-making processes[1–3].

In this approach, decisions are made by relying on facts, figures, trends patterns, and insights derived from data. The goal is to make objective, evidence-based decisions that are more accurate, consistent, and transparent.

> **Note 1: Features of data-driven decision-making**
>
> Data-driven decision-making is widely used in various fields such as business, healthcare, finance, education, and government. It allows organisations and individuals to:
>
> 1. **Informed Decisions** - make decisions based on data rather than assumptions or guesswork;
> 2. **Improved Accuracy** - educe errors and biases by relying on objective information;
> 3. **Efficiency** - Optimise resources and processes by identifying trends, patterns, and inefficiencies;
> 4. **Transparency** - ensure that decisions are made in an open and transparent manner; and,
> 5. **Scalability** - Apply to large-scale operations or complex problems where traditional methods may be insufficient.

Data-driven decision-making often involves the use of tools, techniques, and technologies such as data analytics, machine learning, artificial intelligence, and visualisation software. By leveraging these tools, organisations can transform raw data into actionable insights that drive better outcomes.

In today's organisations, this approach has become increasingly important as it allows for more objective and accurate decision-making. The process typically includes identifying relevant data sources, applying analytical techniques, and leveraging technologies like machine learning, artificial intelligence, and visualisation tools to transform raw data to actionable insights that drive better outcomes.

An organisation that is data-driven also benefits in being able to spot opportunities and threats early. By analysing data regularly, organisations can anticipate changes and act before problems arise.

Saving costs is another advantage. In a survey of executives of Fortune 1000 companies regarding their data investments since 2012 commissioned by the Harvard Business Review, nearly half (48.4%) of respondents report that they are documenting measurable results from their investments in big data and 80.7% of the executives describing their investments in big data as being successful[1,4].

1.2 About this course

In this course, we will explore everything from the basics such as what data is and why it matters to more advanced topics like data collection, storage, analysis, and visualisation. Through practical examples and real-world applications, you'll learn how to harness the power of data to drive insights, solve problems, and make informed decisions in fields ranging from business and technology to healthcare and beyond. By the end of this course, you'll not only understand the importance of data but also be prepared to apply these concepts in your own work.

1.2.1 Objectives

All these towards the overall objective of making a case for shifting to more data-driven decision-making processes.

Specifically, by the end of the course, participants are expected to be able to:

1. Articulate the value of data driven decision making and programming;
2. Critically assess a data by it source, format, structure, types, and classes;

3. Critically evaluate the state of their own dataset based on stated best practices;
4. Outline the strengths and weaknesses of various types of data tools;
5. Demonstrate capacity to use spreadsheet software to clean, process, and structure data; and,
6. Demonstrate capacity to use spreadsheet software to perform data analysis.

1.2.2 Case studies

To achieve these objectives, the course employs the **case-study method**, an approach that involves in-depth examination of a specific individual, group, organisation, or event to understand a complex issue in its real-life context.

For this course, the **five case studies** (one for each of the next five chapters) provide a more nuanced narrative of opportunities and challenges of adopting a data-driven approach to decision-making specifically in the context of governance within governments (rather than just in businesses).

1.2.3 The who, what, when, where, how, and why framework

When going through these five case studies, it is recommended to first go through them using the *who*, *what*, *when*, *where*, *how*, and *why* framework as a way to get a firm grounding on the case study details.

The **"who, what, when, where, how, and why" framework** is a systematic approach to understanding and analysing data. Another term that can be used for this framework is **descriptive metadata** which is data that provides information about other data, but not the content itself. So, if I have an image, the metadata wouldn't be the actual picture, but the details about who took it, when, or where.

Here's a structured explanation of each component within this framework:

Who

Refers to the individuals or entities involved with the data. This includes stakeholders, users, customers, employees, or business partners who interact with or are affected by the data. More specifically, this may include, among others, information on:

- who owns the data;
- who manages the data;
- who collects the data;
- who stores the data; and,
- who protects/safeguards the data.

What

Describes what the data is about and its type, nature, and provenance. It specifies what information is available, such as numerical data, text, images, etc., which helps in understanding the scope and relevance of the data, and how to work with the data.

When

Pertains to the timing, period, and/or frequency in which the data was/is being collected, recorded, or analysed.

Where

Indicates the location where the data is stored or accessed. This could be within a database, on a server, or even from external sources like devices or sensors, providing context about data accessibility and storage.

How

Focuses on the methods used to collect, process, or extract the data. This includes techniques such as surveys, sensor readings, or existing records, which helps in understanding how reliable and comprehensive the data is.

Why

Asks for the purpose behind collecting and analysing the data. It clarifies why this information is being gathered i.e., whether it's for reporting, decision-making, monitoring performance, or other objectives. This in turn guides appropriate actions based on the data insights.

Summary

Using this structured approach helps clarify each aspect of data, ensuring clarity and focus. It is particularly useful for complex datasets and can help address varying questions based on the user's role, such as an analyst versus a stakeholder.

In summary, using the "who, what, when, where, how, and why" framework provides a systematic method to identify key elements of data, ensuring clarity and focus in data management and analysis.

2 Case study: Data use and analytics in water quality management

This is a case study about the Division of Water (DOW), a local government agency in the State of New York, which has attempted to improve its analytic capabilities by developing efficient data management practices, suggest governance models, and identify analytic techniques potentially beneficial to addressing harmful algal blooms (HABs; see Figure 2.1) and high chloride concentrations[3].

Figure 2.1: Harmful algal blooms (HABs) may look like green dots, clumps or globs on the water surface.

The DOW faces challenges in using its legacy systems and traditional analytical methods effectively in addressing the problems of HABs and high chloride levels. DOW aims to enhance its decision-making processes through DDDM by improving its ability to gather and analyse data more effectively, beyond their current capabilities, to better inform policy decisions.

From this process, nine key factors across four overarching determinants have been observed and articulated as being crucial to consider by an organisation in implementing a comprehensive strategy for DDDM (see Note 2). These factors interrelate and influence each other, requiring a holistic approach to ensure successful adoption.

ℹ Note 2: Nine key factors for an effective DDDM strategy

Data determinants
DOW bases its decisions on internal water data from sampling and assessments, supported by a quality assurance process ensuring reliability and compliance with federal standards like those of the Environmental Protection Agency (EPA). Despite these strengths, challenges include manual sampling processes, incomplete data coverage, missing values, compatibility issues, and interoperability problems that hinder seamless data exchange and system integration.

1. *Data quality and coverage* Ensuring robust data infrastructure is foundational, as it supports the collection, storage, and accessibility of high quality data necessary for effective analysis.

2. *Compatibility and operability*

   DOW manages water-related data through interconnected teams responsible for producing and analysing information from various sources like lakes and streams. While collaboration is facilitated by multiple analysts and teams, this setup poses challenges in maintaining consistent and compatible datasets due to differing file versions and a lack of field locking in their proprietary Filemaker system, risking data integrity. Additionally, varying levels of observation across systems complicate integration efforts.

Data compatibility and interoperability ensure that information flows freely, efficiently, and accurately across different systems, which is vital for organisations to function well, innovate, comply with regulations, and adapt as needed.

3. *External data*

   DOW utilises external datasets to address complex environmental and social issues beyond its internal data. While this approach enhances knowledge creation by incorporating charts and maps that combine water chemistry with geographical data, it faces challenges. These include potential quality issues due to lack of control over external sources and incompatibility with specific analytical needs, as seen with United States Geological Survey (USGS) land-cover data not providing sufficient detail on farm types affecting water bodies.

Utilisation of external data potentiates and enriches an organisation's existing information which can lead to better and richer insights that can be derived from them.

Technological determinants

4. *Information systems and software*

5. *Analytical techniques* Investment in both skilled personnel and advanced tools is essential to transform raw data into actionable insights.

Organisational determinants

6. *Cooperation*

7. *Culture*

Institutional determinants Engaging with external institutions and navigating legal frameworks can provide resources and support, or pose restrictions, respectively.

8. *Privacy and confidentiality* Addressing legal requirements regarding data protection is crucial to ensure comprehensive analyses.

9. *Public procurement* Navigating bureaucratic processes efficiently can accelerate tool adoption without unnecessary delays.

These key determinants are interrelated and interdependent. For example, if an organisation has strong data infrastructure (determinant 1) but lacks the right analytical tools or skilled personnel (determinant 2), their DDDM efforts will be hampered. Similarly, even with good internal structures (determinant 3), if external regulations make it hard to access necessary tools or collaborate externally (determinants 7 and 9), progress is still limited. Without proper stakeholder engagement (determinant 6) and user involvement (determinant 5), the organisation might develop solutions in isolation, leading to less effective decisions. Moreover, privacy constraints (determinant 8) can affect data availability, which in turn impacts analytical capabilities since data is a key input.

While DDDM is often seen as a technical issue involving tools and data, it's also deeply influenced by organisational and institutional factors. This makes sense because any significant change requires not just new technology but also cultural shifts within the organisation to embrace these changes.

These determinants also influence the ability of an organisation to adapt over time. For example, if the organisation faces challenges in public procurement, which is a structural issue, this could create delays that affect the organisation's overall strategy. Conversely, strong stakeholder engagement might mitigate some of these delays by providing alternative solutions or resources.

2.1 Leadership role

Leadership plays a critical part in driving organisational change. Without supportive leadership, many of these determinants could be obstacles rather than opportunities. For instance, if leaders aren't committed to DDDM, they might not push for necessary cultural shifts or investment in new tools.

2.2 Balancing existing practices

The balance between existing practices and new methods is important. While the state agency was implementing DDDM, traditional approaches were still relied upon. This blend can be beneficial initially but may need careful management to avoid conflicts or inefficiencies as newer methods prove their worth.

2.3 Measuring success

How would this state agency assess its progress in implementing DDDM? They might look at metrics like the quality and timeliness of decisions, reduction in issues (like HABs), efficiency improvements, and user satisfaction. These outcomes can help gauge whether their efforts are paying off despite facing various challenges.

2.4 Conclusion

A tailored strategy that evaluates specific organisational strengths and weaknesses across these determinants is essential for effective DDDM implementation. This approach ensures that each organisation maximises opportunities while minimising challenges, leading to more informed and efficient decision-making processes.

3 Case Study: Enhancing Local Governance Through Data-Driven Decision-Making in Indonesia

In an era where technology and data are transforming governance, adopting a data-driven approach is crucial for improving decision-making and fostering transparency. This case study explores Indonesia's journey toward integrating data into local governance, highlighting both challenges and opportunities, and offers recommendations for mid-level government officials to enhance their governance strategies[5].

3.1 Context

Indonesia, the largest archipelagic nation in the world, operates under a federalist system with provinces and regencies. With a diverse population of over 270 million people, it faces significant challenges such as inequality, environmental degradation, and sustainable development. These issues necessitate effective local governance to ensure equitable growth and environmental preservation.

3.2 Current Situation

Currently, Indonesia's policy-making is often influenced by top-down directives rather than data-driven insights. Decisions are frequently based on the instructions of superior officials due to a history of autocratic administration. Additionally, there is a lack of standardised data quality frameworks, leading to fragmented and siloed data systems. Limited analytics capacity and reliance on outdated technologies further hinder effective decision-making.

3.3 Challenges

1. **Autocratic Administration** A cultural tendency towards hierarchical decision-making limits the use of data in governance.
2. **Fragmented Data Systems** Siloed systems across different levels of government result in data inconsistencies and inefficiencies.
3. **Lack of Skilled Personnel** Insufficient training and expertise in data analysis impede effective data utilisation.
4. **Public Distrust** Concerns about data accuracy and misuse erode public confidence in data-driven decisions.

3.4 Opportunities

1. **Recent Regulations** The 2022 Data Governance Regulation provides a framework to standardize data collection and use.
2. **International Collaboration** Partnerships with international organizations offer resources for capacity-building and technological support.
3. **Available Data Sources** Rich datasets on demographics, environment, and economy can enhance policy-making, such as managing forest fires or coral reef preservation.
4. **Capacity-Building** Training programs can equip officials with data analysis skills, fostering a culture of evidence-based decision-making.

3.5 Recommendations

1. **Develop Data Quality Frameworks** Establish standardized protocols to ensure data accuracy and consistency across all levels of government.
2. **Enhance Analytical Skills** Implement training programs to build expertise in data analysis and visualisation tools.
3. **Foster Public Trust** Promote initiatives that demonstrate the benefits of data-driven decisions, such as improving public services or environmental outcomes.
4. **Encourage Collaboration** Facilitate intergovernmental cooperation to share best practices and resources for effective data use.
5. **Adopt Technology** Invest in integrated digital platforms to streamline data collection and sharing processes.
6. **Establish Feedback Mechanisms** Create channels for public input to ensure that data-driven policies reflect community needs and concerns.

3.6 Conclusion

Indonesia's shift towards data-driven governance presents a transformative opportunity to address pressing challenges and enhance decision-making effectiveness. By overcoming existing barriers and leveraging available resources, Indonesia can set a precedent for other developing nations. Mid-level officials worldwide are encouraged to consider these insights in their own governance strategies, fostering a global culture of transparency, collaboration, and innovation in public service.

4 Case Study: The Use of Data in Local Governance - A Michigan Perspective

This case study explores the current state of data-driven decision-making in Michigan's local governments, highlighting challenges and opportunities for integrating data into policy and governance based on the results of the Michigan Public Policy Survey (MPPS)[2].

4.1 Michigan Public Policy Survey

The MPPS, established post the 2009 Great Recession, is the first ongoing survey of local leaders across an entire state in the United States, involving over 1,856 jurisdictions in Michigan. It addresses a critical gap by providing insights into local officials' perspectives, crucial for informed policymaking. Conducted biannually, it tracks long-term trends on fiscal and operational policies while addressing current issues like the COVID-19 pandemic and infrastructure. Collaborations with key associations enhance its credibility and scope.

4.2 Current Situation of Policy and Decision-Making in Michigan Local Governments

Michigan's local governments have seen significant growth in data-driven decision-making (see Figure 4.1 and Figure 4.2).

This approach is now widespread across jurisdictions of all population sizes (see Figure 4.3) and across regions, with

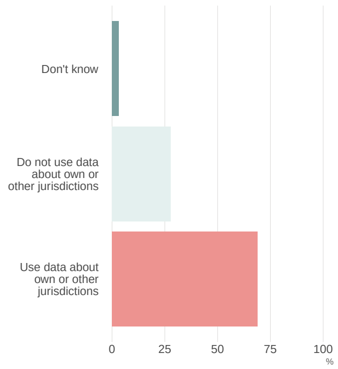
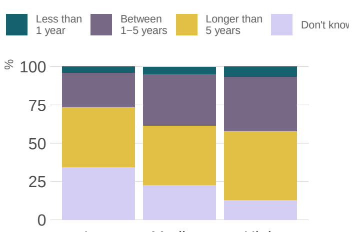


Figure 4.1: Percentage of Michigan jurisdictions reporting use of performance data



many jurisdictions using data to inform budgeting and resource allocation.

Despite this progress, most data use remains informal or ad hoc (see Figure 4.4), particularly among smaller communities (see Figure 4.5).

The MPPS reveals that while larger jurisdictions are more likely to engage in formal performance measurement, over half of the state's smallest jurisdictions also incorporate some form of data into their decision-making processes (see Figure 4.2). This indicates a trend towards broader adoption, albeit at varying levels of formality.

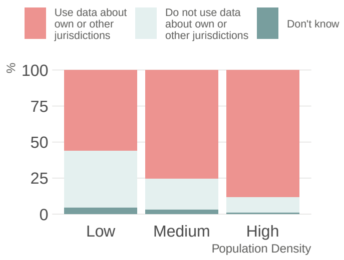


Figure 4.3: Percentage of Michigan jurisdictions reporting data use, by population density

4.3 Challenges and Concerns

1. **Cost Concerns**

Many local governments, especially smaller ones with limited resources, perceive data use as costly. The MPPS found that 62% of non-data users cited cost concerns, though only 28% of current users reported significant issues, suggesting costs may be manageable.

2. **Informal Practices**

The reliance on informal methods can lead to inconsistent outcomes and less accountability. Only about 16% of jurisdictions have formal performance measurement practices, indicating a gap in structured data use.

3. **Resource Constraints**

Smaller jurisdictions often face limitations in staff and financial resources, hindering their ability to adopt more formal data practices.

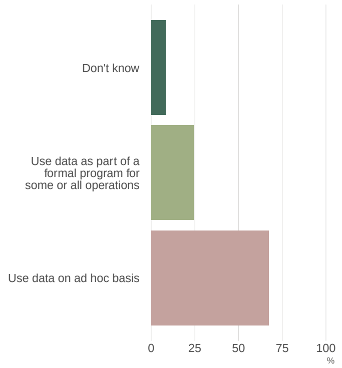


Figure 4.4: Percentage of Michigan jurisdictions reporting ad hoc vs. systematic data use (among data users)

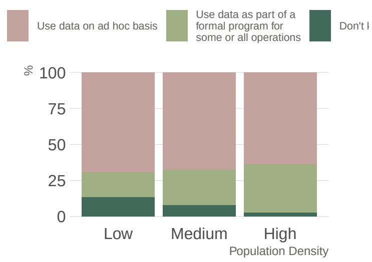


Figure 4.5: Percentage of Michigan jurisdictions reporting ad hoc vs. systematic data use (among data users), by population size

4.4 Opportunities and Benefits

1. **Fiscal Efficiency**

Data-driven approaches help identify cost savings and program efficiencies, crucial for jurisdictions grappling with fiscal challenges.

2. **Improved Service Delivery**

By aligning services with community needs, data can enhance service quality and responsiveness.

3. **Enhanced Transparency and Trust**

Effective use of data fosters transparency, improving public trust in government decisions.

4. **Policy Communication**

Data provides a clear evidence base for policy-making, aiding communication between governments and stakeholders.

4.5 Conclusion

The integration of data into Michigan's local governance has proven valuable despite challenges like cost concerns and resource limitations. The broader adoption of data-driven practices, even informally, highlights its potential to improve decision-making and service delivery.

4.6 Recommendations

1. **Capacity Building** - Invest in training to enhance technical and analytical skills among local officials.
2. **Encourage Collaboration** - Foster partnerships with academic institutions or tech firms to support data initiatives.
3. **Leverage Resources** - Utilise available tools and frameworks, such as those provided by Michigan's MPPS, to guide data practices.
4. **Promote Leadership and Cultural Change** - Champion leadership roles that prioritize data use and cultivate a culture of evidence-based decision-making.

By adopting these strategies, countries can effectively integrate data into local governance, enhancing policy outcomes and public trust.

5 Case Study: Enhancing Data-Driven Decision-Making in Local Governance - A Focus on Turkana County

This case study describes the significant steps that the Turkana County local government are taking to modernising early childhood development and education services management through the use of digital technology[6].

5.1 Introduction

Turkana County, located in northwest Kenya, is a region marked by significant natural resource wealth and cultural diversity. However, it faces challenges such as poverty, infrastructure gaps, and governance inefficiencies. The county's recent efforts to embrace data-driven decision-making offer valuable insights for enhancing local governance through improved policy formulation and implementation.

5.2 Background

Turkana County was established under Kenya's devolution framework in 2013, with its administrative structure comprising several wards and sub-counties. The county has made strides in adopting digital tools like the Turkana Early Childhood Development and Education (ECDE) Management Information System or TECDEMIS and the Continuous Database Updating System or CODUSYS for education management, reflecting a commitment to modernise governance.

5.3 Current Situation of Policy and Decision-Making

Policy-making in Turkana County is characterised by structured processes involving the County Assembly and Executive. Data utilisation is integral to planning and budgeting, with systems like TECDEMIS facilitating real-time data collection and analysis. These tools support decision-makers in tracking program outcomes and resource allocation efficiency.

5.4 Challenges in Data Utilisation for Governance

Despite progress, several challenges impede effective data use:

- **Technological Barriers:** Limited internet access hampers system functionality.
- **Institutional Weaknesses:** Insufficient skilled personnel affect system implementation.
- **Financial Constraints:** Inadequate funding limits infrastructure development and capacity building.
- **Socio-Political Factors:** Resistance to change and lack of awareness about data's value.

5.5 Opportunities for Enhancing Data Use

The county presents several opportunities:

- **Investments in Digital Infrastructure:** Initiatives like TECDEMIS and CODUSYS provide a solid foundation.
- **Partnerships with Development Agencies:** Collaborations with organisations like the Japan International Cooperation Agency or JICA offer resources and expertise.

- **Capacity Building:** Training programs enhance staff skills in data management and analysis.
- **Community Engagement:** Involving citizens fosters trust and ownership of data initiatives.

5.6 Conclusion

Embracing data-driven governance is crucial for Turkana County to overcome challenges and achieve sustainable development. Effective data use aligns with broader goals of accountability, service efficiency, and inclusive growth.

5.7 Recommendations

1. **Invest in IT Infrastructure:** Expand internet access and upgrade digital tools.
2. **Enhance Training Programs:** Prioritise skills development in data management and analysis.
3. **Foster Multi-Sectoral Partnerships:** Strengthen collaborations with development agencies and the private sector.
4. **Improve Stakeholder Engagement:** Involve communities to build trust and ownership of data initiatives.
5. **Establish Monitoring Frameworks:** Develop systems to evaluate the impact of data-driven policies.

6 Case Study: Indigenous Data Governance in the United States

This case study describes the challenges and opportunities with regard to Indigenous data governance in the United States[7].

6.1 Introduction

Indigenous nations in the United States exercise sovereignty over their data, recognising their right to control and manage their own information. This sovereignty is supported by federal laws such as Native American Graves Protection and Repatriation Act or NAGPRA which provide frameworks for protecting Indigenous rights, including those related to data governance.

6.2 Current Strategies

1. **Tribal Data Sovereignty**

Indigenous nations establish policies and institutions, like tribal councils, to own and control their data, ensuring it aligns with cultural values.

2. **Collaboration**

Partnerships with federal and state governments are facilitated through initiatives like the National Historic Preservation Act or NHPA promoting shared goals in data governance.

3. **Capacity Building**

Training programs and technological infrastructure development enhance technical skills, though resources vary among tribes.

4. **Legal Frameworks**

Treaties and international agreements, such as the United Nations Declaration on the Rights of Indigenous Peoples or UNDRIP[8], provide legal backing for data governance, ensuring respect for Indigenous rights.

6.3 Challenges

- **Legal Complexities:** Overlapping jurisdictions complicate data governance, requiring clear resolution mechanisms.
- **Resource Limitations:** Financial and technical constraints affect smaller tribes' ability to implement strategies.
- **Cultural Preservation:** Balancing modern data practices with cultural preservation is complex but crucial.

6.4 Opportunities

- **Global Networks:** Engagement with international bodies like the UNDRIP offers support and recognition, enhancing governance effectiveness.
- **Capacity Building:** Support through grants and partnerships can bridge resource gaps.
- **Collaboration:** Inter-tribal agreements strengthen collective data management efforts.

6.5 Conclusion

Indigenous data governance in the U.S. is advancing through sovereignty assertion collaboration, capacity building, and international frameworks. While challenges persist, opportunities for improvement are significant. Government officials must support Indigenous nations by respecting their sovereignty, providing resources, and fostering international engagement to enhance data governance effectively.

7 All about data

In this chapter, we go further into data concepts with a discussion on the **sources**, **formats**, **structures**, **types**, **classes**, and **systems** of data.

7.1 Data Sources

Data can be classified as either being of **primary** or **secondary** source.

- **Primary data** includes original data collected directly from primary sources such as experiments surveys, or interviews.
- **Secondary data** exists in various forms like reports, government statistics, or academic publications which are data that have been already collected primarily by some other person and/or organisation/entity who make such data available for others to use for either the same purpose or a totally different use-case altogether from the original purpose.

Data sources also refer to where data was obtained or sourced from. These encompass a wide range of information repositories, from traditional databases and files to emerging online platforms and application programming interfaces (APIs).

7.2 Data Formats

Data formats define how information is organised, stored, and accessed within a file or database. They determine the structure of data, such as text, numbers, or multimedia, using common formats like CSV, JSON, and XML, each with unique methods for representing data.

Data formats may specifically refer to the following:

- **Recording format** - a format for encoding data for storage on a storage medium
- **File format** - a format for encoding data for storage in a computer file
- **Container format (digital)** - a format for encoding data for storage by means of a standardised audio/video codecs file format
- **Content format** - a format for representing media content as data
- **Audio format** - format for encoded sound data

7.3 Data Structures

A data structure is an organised format for storing data, designed to allow efficient access and modification. It encompasses not just the storage of data but also the relationships between data elements and the operations that can be performed on them. These operations are structured with defined behaviors where operations have specific properties.

Examples of data structures include:

- **Relational Databases** - Organised into tables with defined relationships (e.g., SQL).
- **NoSQL Systems** - Flexible storage solutions like document stores or key-value systems.
- **Hierarchical Structures** - Data organised in a tree-like structure, such as XML or JSON.
- **Flat Structures** - All data resides at the same logical level without hierarchy (e.g., JSON arrays).
- **Semi-Structured Formats** - Use tags and nested structures for complex data (e.g., JSON).

7.4 Data Types

- **Categorical** - Data divided into categories (e.g., gender, color).
- **Numerical** - Involves numbers, which can be discrete or continuous.
- **Temporal** - Data with time-based attributes (e.g., dates, times).

- **Textual** -Includes natural language text and speech data.
- **Binary** - Represents presence/absence of a feature.
- **Spatial** - Geospatial data indicating locations (e.g., coordinates).
- **Multimedia** - Combines multiple types like images, audio, and video.

7.5 Data Systems

- **Databases** - Platforms for managing and querying structured data, including relational (SQL) and NoSQL systems.
- **Data Lakes** - repositories storing raw, unstructured, or semi-structured data in a lake-like structure.
- **Big Data Systems** - Designed to handle large-scale datasets with distributed processing.
- **Business Intelligence Tools** - Provide analytics capabilities for transforming data into actionable insights.

7.6 Integration and Considerations

7.6.1 Data flow

Data is collected from sources, processed or formatted as needed, organised into appropriate types and structures, and managed by suitable systems.

7.6.2 Interconnected Components

Each component (sources, formats, structures) plays a role in ensuring data compatibility with various systems, which are then used for classification based on specific needs.

Part II

Data Management

8 Data privacy, security, and protection

The increasing volume of digital data collected in today's world necessitates robust protection mechanisms. Breaches can lead to devastating consequences, such as identity theft, financial loss, and potential public health risks, particularly in sectors like healthcare where patient privacy is paramount under regulations such as the General Data Protection Regulation (GDPR).

For institutions, safeguarding sensitive data is crucial for maintaining customer trust, preventing identity theft, and avoiding the loss of valuable customers due to data breaches.

8.1 Definitions

8.1.1 Data privacy

Data privacy is about controlling how your personal information is collected, used, and shared. It's about protecting your right to know who has your data, how it's being used, and who else it's being shared with. Essentially, it's the right to privacy in the digital world.

8.1.2 Data protection

Data protection encompasses the security strategies and processes designed to safeguard sensitive data against unauthorised access, misuse, corruption, and loss. It aims to maintain the integrity, availability, and confidentiality of data, while also ensuring compliance with relevant regulations and ethical standards.

8.1.3 Data security

Data privacy and data security are distinct but related disciplines. Both are core components of an institution's broader data governance strategy.

Data privacy focuses on the individual rights of data subjects or the users who own the data. For organisations, the practice of data privacy is a matter of implementing policies and processes that allow users to control their data in accordance with relevant data privacy regulations.

8.2 Legal frameworks

The General Data Protection Regulation (GDPR) is a European Union law that controls how organizations handle the personal data of EU residents. It was adopted in 2016 and became effective on May 25, 2018. It aims to give individuals more control over their personal data and to ensure that organizations are more transparent and accountable for how they process that data.

Since then, other countries have followed suit in creating their own legislation similar to the GDPR. Generally, countries created such laws as a response to the EU's rollout of GDPR given that the GDPR applies to any organisation that processes the personal data of EU residents, regardless of whether the organisation is located within the EU.

The Seychelles has passed into law the **Data Protection Act, 2023** otherwise entitled as

> An act for the protection of individuals with regard to the processing of personal data, to recognise the right to privacy envisaged in article 20 of the constitution, to promote and facilitate responsible and transparent flow of information by private and public entities and to provide for other related matters.

8.3 Principles of data protection

Article 5 of the GDPR sets out key principles which lie at the heart of the general data protection regime. These key principles are set out right at the beginning of the GDPR and they both directly and indirectly influence the other rules and obligations found throughout the legislation. Therefore, compliance with these fundamental principles of data protection is the first step for controllers in ensuring that they fulfil their obligations under the GDPR. The following is a brief overview of the Principles of Data Protection found in article 5 GDPR:

8.3.1 Lawfulness, fairness, and transparency

Any processing of personal data should be lawful and fair. It should be transparent to individuals that personal data concerning them are collected, used, consulted, or otherwise processed and to what extent the personal data are or will be processed. The principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used.

8.3.2 Purpose Limitation

Personal data should only be collected for specified, explicit, and legitimate purposes and not further processed in a manner that is incompatible with those purposes. In particular, the specific purposes for which personal data are processed should be explicit and legitimate and determined at the time of the collection of the personal data. However, further processing for archiving purposes in the public interest, scientific, or historical research purposes or statistical purposes (in accordance with Article 89(1) GDPR) is not considered to be incompatible with the initial purposes.

8.3.3 Data Minimisation

Processing of personal data must be adequate, relevant, and limited to what is necessary in relation to the purposes for which they are processed. Personal data should be processed only if the purpose of the processing could not reasonably be fulfilled by other means. This requires, in particular, ensuring that the period for which the personal data are stored is limited to a strict minimum (see also the principle of 'Storage Limitation' below).

8.3.4 Accuracy

Controllers must ensure that personal data are accurate and, where necessary, kept up to date; taking every reasonable step to ensure that personal data that are inaccurate, having regard to the purposes for which they are processed, are erased or rectified without delay. In particular, controllers should accurately record information they collect or receive and the source of that information.

8.3.5 Storage Limitation

Personal data should only be kept in a form which permits identification of data subjects for as long as is necessary for the purposes for which the personal data are processed. In order to ensure that the personal data are not kept longer than necessary, time limits should be established by the controller for erasure or for a periodic review.

8.3.6 Integrity and Confidentiality

Personal data should be processed in a manner that ensures appropriate security and confidentiality of the personal data, including protection against unauthorised or unlawful access to or use of personal data and the equipment used for the processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures.

8.3.7 Accountability

Finally, the controller is responsible for, and must be able to demonstrate, their compliance with all of the above-named Principles of Data Protection. Controllers must take responsibility for their processing of personal data and how they comply with the GDPR, and be able to demonstrate (through appropriate records and measures) their compliance.

8.4 Types of data security

To enable the confidentiality, integrity and availability of sensitive information, organizations can implement the following data security measures:

8.4.1 Encryption

By using an algorithm to transform normal text characters into an unreadable format, encryption keys scramble data so that only authorized users can read it. File and database encryption software serve as a final line of defense for sensitive volumes by obscuring their contents through encryption or tokenization. Most encryption tools also include security key management capabilities.

8.4.2 Data erasure

Data erasure uses software to completely overwrite data on any storage device, making it more secure than standard data wiping. It verifies that the data is unrecoverable.

8.4.3 Data masking

By masking data, organizations can allow teams to develop applications or train people that use real data. It masks personally identifiable information (PII) where necessary so that development can occur in environments that are compliant.

8.4.4 Data resiliency

Resiliency depends on how well an organization endures or recovers from any type of failure—from hardware problems to power shortages and other events that affect data availability. Speed of recovery is critical to minimize impact.

8.5 Summary

Data privacy, security, and protection are fundamental concerns in today's digital landscape. They involve protecting personal information from unauthorised access, implementing robust security measures, and upholding ethical standards in data handling. Addressing these challenges effectively requires a holistic approach that integrates technical safeguards with ongoing education and ethical practices to maintain trust and prevent significant risks.

9 Data tools

Working with data is a multi-faceted endeavour that involves collecting, storing, analysing, visualising, securing, and managing data across various domains. The various steps in the data pathway (see Figure 9.1) often require specific tools that are best-suited for the task at hand.

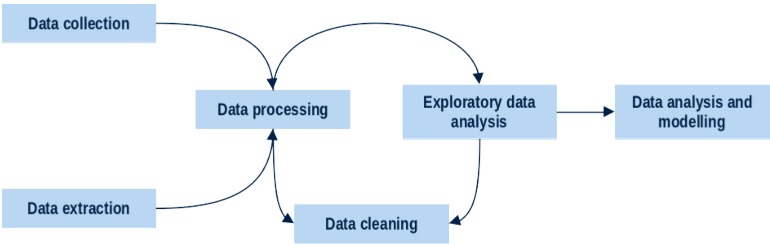


Figure 9.1: Data pathway

In this section, we present the most common data tools, describe their key functionalities, and discuss what each tool is best suited for in relation to the steps in the data pathway.

9.1 Microsoft Excel and other Excel-like spreadsheet software

Microsoft Excel is versatile spreadsheet software with robust formula capabilities, pivot tables for quick data summarisation, and Power Query for advanced data cleaning. Other than for data collection, it is also suited for detailed analysis, budgeting, and financial tracking. Suitable for complex data management. On the other hand, some may find that using it presents a steeper learning curve and costs of the subscription-based software-as-a-service (SaaS) model as part of Microsoft 365 can be prohibitive.

An estimated 750 million up to 1.5 billion people[1] use Microsoft Excel. It has numerous applications, including data entry, analysis, accounting, financial modelling, and reporting. It's used in various fields like business, education, and personal finance to organise, manage, and visualise data.

Other than Microsoft Excel, there are Excel-like applications available as part of a suite of office applications that use the Open Document Form (ODF), an open file format for word processing documents, spreadsheets, presentations and graphics and using ZIP-compressed XML files. It was developed with the aim of providing an open, XML-based file format specification for office applications. ODF was based on the Sun Microsystems specification for OpenOffice.org XML, the default format for OpenOffice.org and LibreOffice. This standard was originally developed to provide an open standard for office documents. Versions of Microsoft Excel since 2003 use the ODF XML standard to afford compatibility to other spreadsheets that use the standard. A number of free and proprietary software use the ODF XML standard hence there are various Excel-like spreadsheet alternatives available that use the standard[2] and are mostly compatible with Microsoft Office/Microsoft 365 applications including Excel. Although generally compatible in almost all of the basic features, Excel-like spreadsheet applications may not fully implement highly customised Excel spreadsheets that use Visual Basic for Applications (VBA) macros as there are significant differences in syntax and implementation to LibreOffice Calc's Basic macro system and environment.

9.2 Google Sheets

Google Sheets, a free and web-based spreadsheet application, is a versatile data tool used for organising, managing, and analysing data, as well as creating visualisations. It's part

[1] It is challenging to make more precise estimates for this. The lower end of this estimate is most likely very conservative and based on historical information. The upper end of this estimate is based on Microsoft's own estimation based on subscription to Microsoft 365. These estimates likely don't include unlicensed or unauthorised usage of the software.

[2] To see a list of free and proprietary software that use the ODF XML standard, see https://en.wikipedia.org/wiki/OpenDocument.

of the Google Workspace suite, along with Google Docs and Google Slides. Sheets offers features like pivot tables, formulas, conditional formatting, and data validation for a variety of data-related tasks.

Google Sheets is technically not an Excel-like spreadsheet (although general use and behaviour is similar to Excel) as it doesn't use the ODF XML standard but rather has its own proprietary format called the Google Sheets format which can only be accessed or utilised through a web browser rather than through a standalone installer for your computer. In order to access/open a Google Sheets format outside of a browser, one h as to download it as either an Excel file or as a comma-separated value (CSV) file which can then be opened in Excel. Google Sheets has similar features and functionalities as Excel but because of its indirect compatibility with Excel and Excel-like ODF XML-compliant software, advanced features of both applications are not interoperable.

9.3 Google Forms

Google Forms is a tool for creating online forms, surveys, and quizzes that can be shared with others to collect data. It allows users to create and edit these forms online, collaborate in real-time, and have the collected data automatically entered into a spreadsheet. Google Forms is part of the free, web-based Google Suite and the software-as-a-service (SaaS) Google Workspace which includes Google Docs, Google Sheets, Google Slides, Google Drawings, Google Sites, and Google Keep. Google Forms is only available as a web application.

9.4 Airtable

Airable is a **spreadsheet-database hybrid**, with the features of a database but applied to a spreadsheet. The fields in an Airtable table are similar to cells in a spreadsheet, but have types such as *'checkbox'*, *'phone number'*, and *'drop-down list'*, and can reference file attachments like images.

Users can create a database, set up column types, add records, link tables to one another, collaborate, sort records and publish views to external websites. Users cannot download their database in full, but can download some of the data by manually downloading CSVs for each table.

Airtable is user-friendly and is designed for ease of use, making it accessible to a wide range of users, including those without technical backgrounds. It also enables users to build and customise applications for various purposes, such as managing product roadmaps, launching marketing campaigns, and tracking job applications. It facilitates collaboration by allowing multiple users to access and work on the same database. Airtable integrates with various other platforms, enabling data to be shared and workflows to be automated.

9.5 QuickBooks and other accounting-specific software

QuickBooks is a popular accounting software designed to help businesses manage their finances, including tasks like bookkeeping, invoicing, expense tracking, and payroll.

QuickBooks is a widely used accounting software known for its ease of use and automation capabilities. It's a solution for small to medium-sized businesses (SMEs), offering features like invoicing, expense tracking, inventory management, and payroll processing.

9.5.1 Other Accounting Software

Beyond QuickBooks, several other software options exist, each with its strengths and weaknesses:

- Xero: Offers a user-friendly interface and strong integration capabilities, making it popular among small businesses.
- Sage 50: A desktop accounting software with robust reporting and features for larger businesses.

- Wave Accounting: A free option that provides basic accounting features, suitable for startups and small businesses.
- Zoho Books: A comprehensive online accounting software with various features, including project management.
- FreshBooks: A popular choice for freelancers and sole proprietors, known for its simplicity.

9.5.2 Key Features of Accounting Software

Common features across different accounting software include bookkeeping and recording of financial transactions, invoicing, expense tracking and managing and categorising business expenses, payroll processing financial reporting to generate reports like income statements and balance sheets, and inventory management to track and manage inventory levels.

9.6 Business intelligence and analytics platforms

Power BI, Tableau, and Qlik are classified as business intelligence (BI) tools or data analytics platforms. They share the common goal of enabling users to interact with data, visualise it, analyse it, and ultimately, make data-driven decisions. However, they each have unique strengths and features that cater to different needs and preferences.

9.6.1 PowerBI

Microsoft's BI platform offers a wide range of functionalities, including data connectivity, data modelling, interactive visualisations, and dashboard creation. It's known for its ease of use and seamless integration with other Microsoft products.

9.6.2 Qlik

This platform focuses on its associative data model, allowing users to explore relationships within data freely. It also offers strong data integration capabilities and is well-suited for large, complex datasets.

9.6.3 Tableau

Tableau is highly regarded for its visual analytics capabilities, enabling users to create stunning and interactive dashboards. It's known for its user-friendly interface and strong visualisation options.

9.6.4 Comparison

9.6.4.1 Ease of Use

Power BI is generally considered to have a more intuitive interface, while Qlik Sense is more powerful but can have a steeper learning curve. Tableau's drag-and-drop interface is known for its ease of use.

9.6.4.2 Data Integration:

Qlik is particularly strong in data integration and can handle diverse data sources, while Tableau offers a dedicated tool (Tableau Prep) for data preparation. Power BI's data integration capabilities are also robust, particularly when used in conjunction with other Microsoft products.

9.6.4.3 Visualisation

Tableau is renowned for its visual analytics, offering a wide array of visual options and a focus on storytelling through data. Power BI also offers extensive visualisation capabilities, and Qlik provides a unique approach with its associative model.

9.6.4.4 Scalability and performance

All three tools are scalable, but Qlik is particularly well-suited for large, real-time datasets. Power BI is strong for smaller to medium datasets and can leverage Microsoft Azure for scalability. Tableau's performance depends on the complexity of the dashboards, but it's generally robust for complex visualisations.

9.6.4.5 Pricing

Power BI is known for its affordable pricing, while Tableau and Qlik Sense can be more expensive, particularly for enterprise users.

9.7 Cloud-based data storage

In today's digital age, efficient data storage and quick access are crucial, particularly as remote work becomes more prevalent. Cloud storage solutions like Google Drive, Dropbox, and OneDrive have become vital tools for both businesses and individuals due to their ease of use and collaborative features.

9.7.1 Google Drive

Google Drive is a cloud storage service included in the Google Suite or Google Workspace of tools that allows users to store, sync, and access files across multiple devices and platforms via an internet connection. It also offers features like collaboration tools, document creation, and sharing capabilities.

9.7.2 OneDrive

OneDrive is a cloud storage service by Microsoft included in the Microsoft 365 set of applications that provides collaboration, document creation, and sharing tools. It allows users to store and sync files across multiple devices and offers 5GB of free storage. Paid plans are available for additional storage ranging from 50GB to 1TB.

9.7.3 Dropbox

Dropbox is a cloud storage service that allows users to store, share, and sync files across multiple devices. Available on Windows, Mac, iOS, and Android, it offers document creation, collaboration, and sharing tools. With 2GB of free storage, paid plans range from 200GB to 3TB for additional needs.

9.8 Databases

A database is an organised collection of structured and/or unstructured data, typically stored electronically in a computer system. It's a system for storing and managing data, and it's managed by a Database Management System (DBMS). Databases are used to store, retrieve, and manipulate data efficiently.

Hence, the concept of a database is both software, which deals with the handling and management of the data, and hardware, which deals with the physical storage of the data.

9.8.1 SQL and other relational databases

SQL databases, also known as *relational databases*, are systems that store collections of tables and organise structured sets of data in a tabular columns-and-rows format, similar to that of a spreadsheet. The databases are built using **structured query language (SQL)**, the query language that not only makes up all relational databases and relational database management systems (RDBMS), but also enables them to "talk to each other".

The history of database technology/relational databases SQL was invented as a language in the early 1970s, which means SQL databases have been around for as long as the Internet itself. Dubbed the structured English query language (SEQUEL), SQL was originally created to streamline access to relational database systems and to assist with the processing of information. Today, SQL remains one of the most popular and widely used query languages in open-source database technology due to its flexibility, ease of use, and seamless integration with a variety of different programming languages. You'll find SQL being used throughout all types of high-performing, data-centric applications.

9.8.2 NoSQL

NoSQL stands for *"Not Only SQL."* It refers to a type of database that doesn't rely on the traditional relational database models, which are organised into tables with fixed schemas and use SQL for querying. NoSQL databases offer a more flexible approach to data storage and querying, often using document, graph, key-value, or other data models. NoSQL databases are equipped to handle large volumes of structured, semi-structured, and unstructured data from non-traditional sources.

Popular database management systems include Microsoft SQL Server, PostgreSQL, MongoDB, Redis, Elasticsearch, SQLite, MariaDB, IBM Db2, Oracle Database, and MySQL. In essence, databases are fundamental to modern IT infrastructure, enabling organisations to store, manage, and analyse data efficiently for various applications, including websites, apps, and business processes.

9.9 Management information systems

A Management Information System (MIS) is an integrated system that collects, processes, stores, and disseminates information to support managerial decision-making and improve operational efficiency. It essentially acts as a tool for gathering and analysing data, converting it into actionable insights, and making those insights available to the right people within an organisation.

9.9.1 Key Features

- **Data Collection and Storage** - MIS systems gather data from various sources, both internal (e.g., sales records, inventory) and external (e.g., market trends, competitor information).

- **Data Processing and Analysis** - The collected data is processed and analysed to identify trends, patterns, and opportunities, often using sophisticated tools and techniques.
- **Information Dissemination** - The analysed information is then formatted and delivered to managers and other stakeholders in a way that is easy to understand and use.
- **Decision Support** - MIS provides the information that managers need to make informed decisions about various aspects of their business, such as sales, marketing, finance, and operations.
- **Improved Efficiency** - By providing timely and accurate information, MIS helps organisations to operate more efficiently, reduce costs, and improve decision-making.

9.9.2 Examples of MIS applications

- **Sales and Marketing** - Tracking sales figures, analysing marketing campaign effectiveness, and identifying customer trends.
- **Accounting and Finance** - Managing financial records, generating financial statements, and tracking investments.
- **Human Resources** - Managing employee information, tracking performance, and supporting recruitment and training activities.
- **Inventory Management** - Tracking inventory levels, managing warehouses, and forecasting demand.
- **Health records** - tracking of patients and clients of various health services. This is often called a Health Management Information System (HMIS).
- **Customer-relationship manager** - tracking of clients/customers data and interactions with company (see Section 9.10).

9.10 Customer-relationship Manager

Customer Relationship Management (CRM) systems are software applications that help businesses manage and analyse customer data and interactions. They are used to collect, organise, and process information about customers, including their interactions, preferences, and purchase history. The goal is to improve customer service, increase customer retention, and drive sales growth.

9.10.1 Key Features and Functionality

9.10.2 Data Management

CRMs store and organise customer data from various sources, like sales interactions, customer service inquiries, marketing campaigns, and social media.

9.10.2.1 Sales Management

CRMs help track sales opportunities, pipeline management, and sales activities, enabling sales teams to improve efficiency and close deals faster.

9.10.2.2 Customer Service

CRMs facilitate communication with customers, track service requests, and help resolve issues, leading to improved customer satisfaction.

9.10.2.3 Marketing Automation

CRMs can be integrated with marketing automation tools, allowing businesses to personalise and automate marketing campaigns.

9.10.2.4 Reporting and Analytics

CRMs provide insights into customer behaviour, sales performance, and overall business trends, enabling data-driven decision-making.

9.10.3 Types of CRM Systems

- **Operational CRM** - Focuses on day-to-day customer interactions, such as sales and customer service.
- **Analytical CRM** - Analyses customer data to identify trends, patterns, and opportunities.
- **Collaborative CRM** - Facilitates communication and collaboration between different departments, such as sales, marketing, and customer service.
- **Strategic CRM** - Uses customer insights to make strategic decisions about product development, pricing, and market positioning.

9.10.4 Benefits of using a CRM

- **Improved Customer Service** - By having a centralised database of customer information, companies can provide better and more personalised service.
- **Increased Sales** - CRMs help sales teams manage leads, track opportunities, and close deals more effectively.
- **Enhanced Customer Retention** - By understanding customer preferences and needs, businesses can build stronger relationships and retain customers.
- **Data-Driven Decision Making** - CRMs provide valuable insights into customer behaviour and business performance, enabling data-driven decision-making.
- **Increased Efficiency** - Automating tasks and streamlining processes can free up employees to focus on more strategic initiatives.

9.11 Statistical packages

SAS, SPSS, and Stata are popular statistical software packages used for data analysis, but have distinct strengths and target industries. SPSS is known for its user-friendly interface, making it popular in social sciences and market research. Stata is a general-purpose statistical software, often favoured for econometrics, and known for its command-line interface and strong data management features. SAS is a powerful system for advanced analytics, business intelligence, and data management, and is widely used in various industries due to its scalability and robustness.

9.11.1 SPSS

Statistical Package for the Social Sciences or SPSS has a ser-friendly interface and intuitive data management making it suitabile for social sciences and market research. It focuses on descriptive and inferential statistics, data exploration, and model building. Its common uses are for surveys, market research, data mining, and other social science applications. The interface is Menu-driven with a graphical user interface.

9.11.2 Stata

Stata is a general-purpose software with strong data management capabilities, and a command-line interface. It is used commonly in econometrics, time series analysis, and statistical modelling. Its most commong uses are in economics, biomedicine, and political science research. It has some graphical user interface but full capability is accessed via the command-line. It has a graphical output.

9.11.3 SAS

SAS or Statistical Analysis System is robust, scalable, and suitable for advanced analytics, business intelligence, and data management. It can be used for Multivariate analysis, predictive analytics, and large-scale data processing. It's common sses are for business analytics, data warehousing, and industry-specific applications. The interface to SAS is primarily as a procedural language.

9.12 Programming languages

R, Python, and Julia are powerful programming languages frequently used in data science, scientific computing, and related fields. They offer distinct advantages, making them suitable for various tasks.

9.12.1 R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different implementation of S. There are some important differences, but much code written for S runs unaltered under R.

R provides a wide variety of statistical (linear and non-linear modelling, classical statistical tests, time-series analysis, classification, clustering, etc.) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

R is available as Free Software under the terms of the Free Software Foundation's GNU General Public License in source code form. It compiles and runs on a wide variety of UNIX platforms and similar systems (including FreeBSD and Linux), Windows and MacOS.

R is unique in that it is not general-purpose. It does not compromise by trying to do a lot of things. It does a few things very well, mainly statistical analysis and data visualisation. While you can find data analysis and machine learning libraries for languages like Python, R has many statistical functionalities built into its core. No third-party libraries are needed for much of the core data analysis you can do with the language.

But even with this specific use case, it is used in every industry you can think of because a modern business runs on data. Using past data, data scientists and data analysts can determine the health of a business and give business leaders actionable insights into the future of their company.

Just because R is specifically used for statistical analysis and data visualisation doesn't mean its use is limited. It's actually quite popular, ranking 12th in the TIOBE index of the most popular programming languages.

Academics, scientists, and researchers use R to analyse the results of experiments. In addition, businesses of all sizes and in every industry use it to extract insights from the increasing amount of daily data they generate.

9.12.2 Python

Python is an interpreted, interactive, object-oriented programming language. It incorporates modules, exceptions, dynamic typing, very high level dynamic data types, and classes. It supports multiple programming paradigms beyond object-oriented programming, such as procedural and functional programming. Python combines remarkable power with very clear syntax. It has interfaces to many system calls and libraries, as well as to various window systems, and is extensible in C or C++. It is also usable as an extension language for applications that need a programmable interface. Finally, Python is portable: it runs on many Unix variants including Linux and macOS, and on Windows.

Python is a high-level general-purpose programming language that can be applied to many different classes of problems.

The language comes with a large standard library that covers areas such as string processing (regular expressions, Unicode, calculating differences between files), internet protocols

(HTTP, FTP, SMTP, XML-RPC, POP, IMAP), software engineering (unit testing, logging, profiling, parsing Python code), and operating system interfaces (system calls, filesystems, TCP/IP sockets). Look at the table of contents for The Python Standard Library to get an idea of what's available. A wide variety of third-party extensions are also available. Consult the Python Package Index to find packages of interest to you.

9.12.3 Julia

Julia is a high-level, open-source, general-purpose programming language designed for technical and scientific computing. It's known for its fast performance, approaching that of languages like C and Fortran, while remaining relatively easy to use. Julia is particularly well-suited for tasks like numerical analysis, data science, and machine learning.

10 Project-based workflow

Part III

Exploratory Data Analysis

11 Exploratory data analysis

12 Univariate statistics

13

References

1. Stobierski, T. The advantages of data-driven decision-making. Business insights. https://online.hbs.edu/blog/post/data-driven-decision-making (2019).
2. Ivacko, T. M., Horner, D. & Crawford, M. Q. Data-driven decision-making in michigan local government. *SSRN Journal* (2013) doi:10.2139/ssrn.2351916.
3. Choi, Y. *et al.* Towards data-driven decision-making in government: Identifying opportunities and challenges for data use and analytics. in (2021). doi:10.24251/HICSS.2021.268.
4. Bean, R. How companies say they're using big data. *Harvard Business Review* (2017).
5. Sayogo, D. S., Yuli, S. B. C. & Amalia, F. A. Data-driven decision-making challenges of local government in indonesia. *TG* **18**, 145–156 (2024).
6. Onunga, J. & Odongo, P. Digital transformation in public administration and data-driven decision-making: A review of turkana county government. *IJRIAS* **IX**, 234–240 (2025).
7. Carroll, S. R., Rodriguez-Lonebear, D. & Martinez, A. Indigenous data governance: Strategies from united states native nations. *CODATA* **18**, 31 (2019).
8. United Nations General Assembly. United nations declaration on the rights of indigenous peoples : Resolution / adopted by the general assembly. (2007).

Index