Data for Decision Makers: Data Concepts and Applications

Course Handbook

Proochista Ariana Ernest Guevarra

20 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(Stobierski, 2019; Ivacko, Horner & Crawford, 2013; Choi et al., 2021).

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 (Bean, 2017; Stobierski, 2019).

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 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.

2.1 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).

2.2 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

2.3 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).

2.4 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.

2.5 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.

2.6 Integration and Considerations

2.6.1 Data flow

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

2.6.2 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 Case Studies

3 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 3.1) and high chloride concentrations(Choi et al., 2021).

Figure 3.1: 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 infrastruc-

ture (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.

3.1 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.

3.2 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.

3.3 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.

3.4 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.

4 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(Sayogo, Yuli & Amalia, 2024).

4.1 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.

4.2 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.

4.3 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.

4.4 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.

4.5 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.

4.6 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.

5 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)(Ivacko, Horner & Crawford, 2013).

5.1 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.

5.2 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 5.1 and Figure 5.2).

This approach is now widespread across jurisdictions of all population sizes (see Figure 5.3) and across regions, with

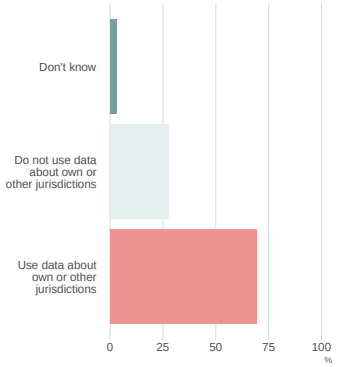
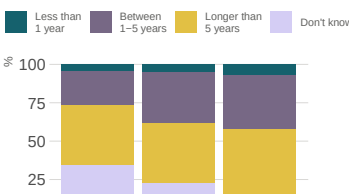


Figure 5.1: Percentage of Michigan jurisdictions reporting use of performance data



many jurisdictions using data to inform budgeting and resource allocation.

Despite this progress, most data use remains informal or ad hoc (see Figure 5.4), particularly among smaller communities (see Figure 5.5).

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 5.2). This indicates a trend towards broader adoption, albeit at varying levels of formality.

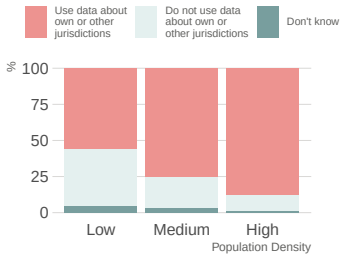


Figure 5.3: Percentage of Michigan jurisdictions reporting data use, by population density

5.3 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.

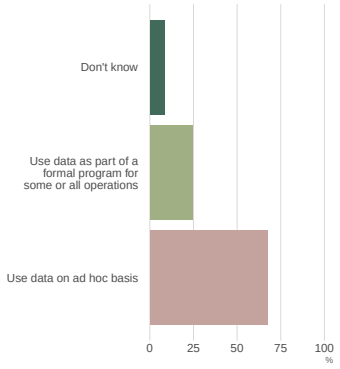


Figure 5.4: Percentage of Michigan jurisdictions reporting ad hoc vs. systematic data use (among data users)

5.4 Opportunities and Benefits

1. **Fiscal Efficiency**

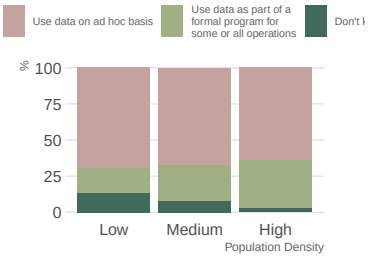


Figure 5.5: Percentage of Michigan jurisdictions reporting ad hoc vs. systematic data use (among data users), by population size

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.

5.5 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.

5.6 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.

6 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(Onunga & Odongo, 2025).

6.1 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.

6.2 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.

6.3 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.

6.4 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.

6.5 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.

6.6 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.

6.7 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.

7 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(Carroll, Rodriguez-Lonebear & Martinez, 2019).

7.1 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.

7.2 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(United Nations General Assembly, 2007), provide legal backing for data governance, ensuring respect for Indigenous rights.

7.3 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.

7.4 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.

7.5 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.

Part III

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.

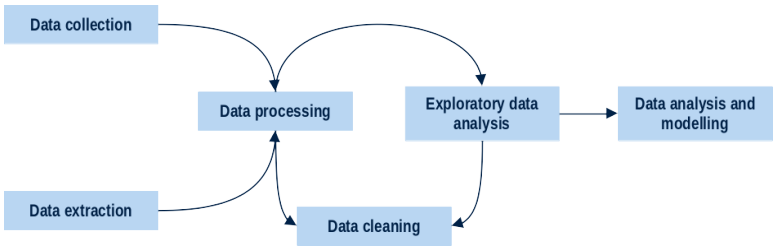


Figure 9.1: Data pathway

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 suitable 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 common 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 uses 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 All about spreadsheets

A spreadsheet is a digital tool for organising, analysing, and storing data in tables, originally developed as an electronic version of paper accounting worksheets. It allows users to enter numerical or textual data and formulas that reference other cells, enabling dynamic calculations.

Spreadsheets are interactive with users able to modify values and observe immediate changes in calculated results, facilitating "what-if" analysis. Beyond basic arithmetic, spreadsheets offer financial, statistical, and conditional functions, enhancing their analytical capabilities. Composed of rows (numbered) and columns (labelled with letters), cells are referenced by their alphanumeric coordinates (e.g., A1). Modern spreadsheet applications include multiple worksheets within a workbook, allowing for complex data management. It can also display data graphically, aiding in understanding trends and patterns.

Spreadsheets have revolutionised business processes by replacing manual systems, offering versatility across various applications where tabular data is essential. Their dynamic cell referencing system allows for efficient and flexible data manipulation, making them indispensable in both professional and personal contexts.

10.1 History of spreadsheets

10.1.1 Paper spreadsheet

The concept of organising data into tabular formats dates back to ancient times, with examples such as Babylonian clay

tablets from 1800 BCE as shown in Figure 10.1. In accounting, the term *"spread sheet"* was used by at least 1906 to describe a grid system in ledgers[1].

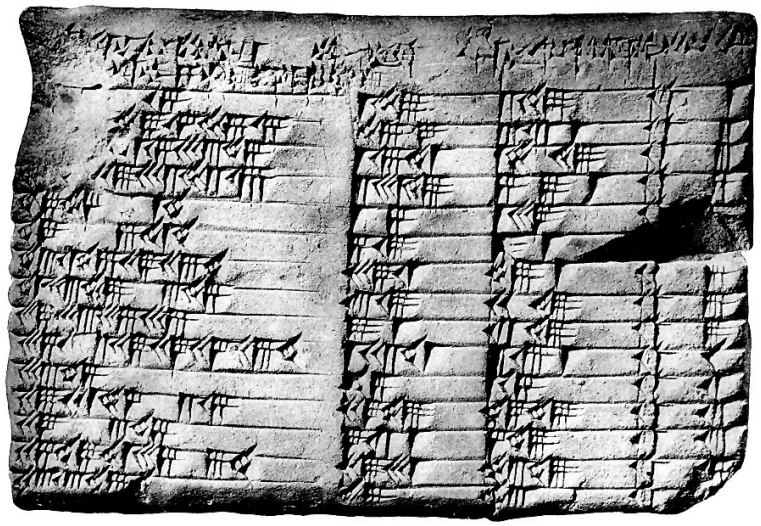


Figure 10.1: A Babylonian clay tablet believed to have been written around 1800 BC containing mathematical table written in cuneiform script

Before digital spreadsheets, *"spread"* referred to large, two-page layouts in publications. The evolution of the term *"spread-sheet"* reflects its transition from physical, oversized ledger pages with examples shown in Figure 10.2, Figure 10.3 to the digital tool we use today, emphasising their role in accounting and data organisation.

10.1.2 Electronic spreadsheet

Figure 10.4 shows key development milestones of the electronic spreadsheet.

[1]"We maintain, in our general ledger, a so-called Spread Sheet which is a long sheet with the name of each individual plant in a particular column." (from Middleton, 1933)

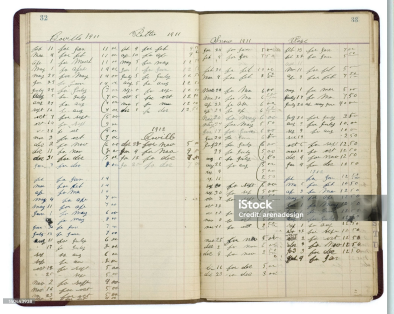


Figure 10.2: Accounting ledger from 1911

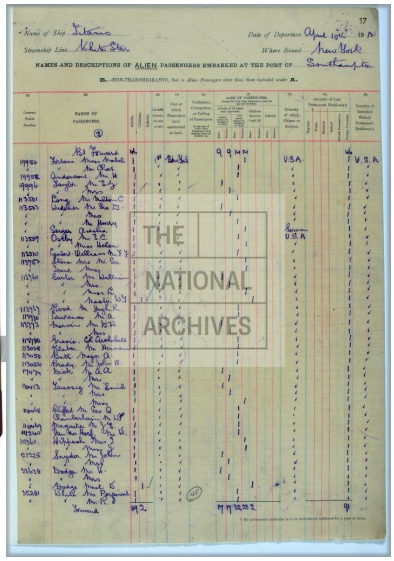


Figure 10.3: Manifest of passengers of the Titanic

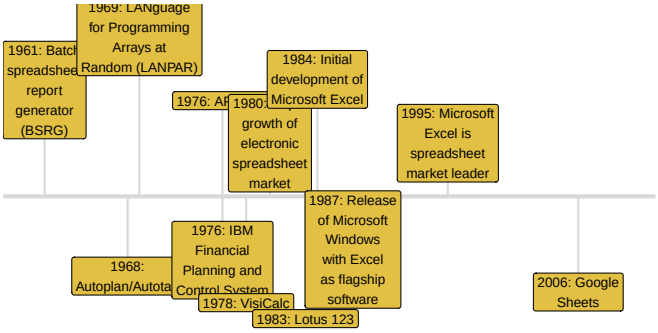


Figure 10.4: Timeline of development of the electronic spreadsheet

10.2 Spreadsheets as databases

Spreadsheets and databases share similarities but are fundamentally different. A spreadsheet is essentially a single table, while a database consists of multiple tables with machine-readable relational structures. Although a spreadsheet workbook containing multiple sheets has interacting tables, it lacks the relational complexity of a database. Spreadsheets and databases are interoperable, however, with spreadsheets being able to be converted into database tables, and database queries being able to be exported to spreadsheets for analysis.

10.3 Spreadsheets as multi-function tools

Spreadsheets are widely used software tools for every step of the data pathway - *data entry*, *storage*, *analysis*, and *visualisation*. Spreadsheets are able to implement such functionality through an *end-user development* approach.

10.3.1 End-user development

Spreadsheets are designed as *end-user development* (EUD) tools. EUD refers to techniques in which non-professional developers are able to create automated tasks and complex data objects without in-depth knowledge of a programming language. Many find using spreadsheets for calculations and analysis much easier. This most likely stems from the following key features:

Ease of use

Spreadsheets leverage spatial relationships, making it intuitive to establish program connections, unlike sequential programming which requires extensive text.

Forgiving nature

Partial results and functions can operate even if other parts are incomplete or contain errors, simplifying the development process.

Visual enhancements

Modern spreadsheets use colours, fonts, and lines to provide visual cues, aiding comprehension and organisation.

Advanced functionality

Extensions enable users to create complex functions and integrate machine learning models, expanding their capabilities beyond basic calculations.

Versatility

Beyond numerical data, spreadsheets support Boolean logic, graphical design, and even SQL queries through relational data storage and formula-based expressions.

In essence, spreadsheets offer a flexible, powerful platform that caters to diverse tasks, making them an invaluable tool for many users despite their limitations compared to traditional programming environments.

10.3.2 Limitations and shortcomings of spreadsheets

Unfortunately, the same multi-functionality and features that make spreadsheets user-friendly and easily accessible to most also make them fragile, non-robust, and prone to causing/producing errors.

In order to be able to function as a tool for the various steps in the data pathway while still being user-friendly meant combining the data interface functionality (data storage and data access) with the programming/scripting capabilities (for data cleaning/processing, analysis, and visualisation) into a single graphical user interface with no clear distinction between

them and no clear mechanism for programming/script testing. This brings about the following limitations and shortcomings:

Lack of standard mechanisms for management and quality assurance of spreadsheets produced by organisations

Given that data storage and access along with data processing/cleaning, analysis and visualisation capabilities sit side-by-side within the spreadsheet tool/software, developing routine and automated audit mechanisms for both data validity/quality and accuracy/correctness of data processing, analysis, and visualisation is nearly impossible. These audits will need to be done manually and line-by-line making them highly onerous. This is most likely the reason why a survey conducted in 2011 of nearly 1,500 people in the UK saw 72% reporting that no internal department checks their spreadsheets for accuracy, that only 13% said that internal audit reviews their spreadsheets, while a mere 1% receive checks from their risk department (Anon).

Reliability issues

An estimated 1% of all formulas in operational spreadsheets are in error (Powell, Baker & Lawson, 2009).

Practical expressiveness is limited

Whilst the graphical user interface of a spreadsheet using its *cell-at-a-time* approach is accessible and user-friendly for most users and for simple data operations, applying the same to a complex data model requiring more complicated calculations require tedious attention to detail. Users will tend to have difficulty remembering the meanings of hundreds or thousands of alphanumeric cell addresses that appear in per cell formulas.

Formulas expressed in terms of cell addresses are hard to keep straight and hard to audit

A research paper critically reviewing spreadsheet errors has shown that auditors who check both numerical results and the cell formulas find no more errors than auditors who only check numerical results (Powell, Baker & Lawson, 2008). By the nature of the *cell-at-a-time* approach, spreadsheets typically contain many copies of the same formula. Thus, when a formula needs to be edited, these edits will need to be applied to all cells containing that formula. This is in sharp contrast to a well-known principle in programming - *do not repeat yourself* or *DRY* - which emphasises the best practice of not repeating code to implement/achieve the same process or output. The *DRY* approach makes code implementation much more efficient and code auditing much more feasible and robust.

Maintenance of volumes of spreadsheets is challenging

Creating and managing a system to maintain vast amounts of spreadsheets for an individual or for an organisation is a challenging endeavour. Without built-in functionalities for proper security, version control and audit trails, and prevention of unintentional introduction of errors, it is more likely that management of spreadsheets end up being ad hoc, non-systematic, and tedious to implement.

11 Project-based workflow

As our skills as data analysts grow, we begin to understand that our ability to realise our full potential goes beyond the data tools that we have chosen to use or have been made to use. The importance of surrounding systems and infrastructure for ensuring reproducibility and long-term preservation of our work becomes increasingly significant. However, a lack of formal training or mentorship in managing these systems often leads to either feeling overwhelmed by technology or resorting to self-exploration without proper guidance.

This chapter aims to guide you gracefully into the exploration of this realm of efficient, effective, and reproducible data workflows. The concepts and practices discussed here may highlight current and existing practices that you have that are ineffectual, disorganised, and irreproducible. If so, we encourage you not to worry about these past mistakes but instead use them to raise the bar for your new work. Small but meaningful incremental changes add up over time, transforming your data quality of life.

In this chapter, we will discuss concepts and practices borrowed from the computational sciences field that use programming languages to record and automate their processes and translate them for use with the spreadsheet software that is sort of a hybrid with data processes implemented through both point-and-click steps via the mouse and through in cell commands or functions for performing calculations and operations. This translation as applied to spreadsheets is not high fidelity given the shortcomings and limitations of spreadsheets (as discussed in Section 10.3.2) but still provides enough structure and rigour compared to the typical and common ad hoc and unstructured use of spreadsheets.

11.1 Data processes as livestock rather than pets

In modern data and computing, a common analogy used to describe the management of data and computational processes is that of managing a *herd of livestock* compared to taking care of an *individual household pet*. For example, in cloud computing, individual servers are treated like "livestock" in that they can be easily destroyed and replaced via automation.

Figure 11.1: Microsoft Excel files as livestock

Figure 11.2: Microsoft excel file as a pet

It is recommended that we adopt a similar mindset when managing our data and data processes - design and develop appropriate data systems that manage data and data processes as disposable and rebuilt and re-implemented as needed rather than as precious "pets". We recommend this approach because if your workflow relies on an individual session or workspace in a non-reproducible way, it creates unnecessary risk and complexity. Instead, the focus should be on saving and relying on code and documentation to ensure reproducibility.

Applying this approach with spreadsheets is not as straightforward given the peculiarities of the tool compared to programming languages that use code to record each step of the workflow. However, steps can be done that can facilitate as much reproducibility when using spreadsheets.

11.1.1 Detailed documentation for point-and-click mouse-based steps

Point-and-click steps in a workflow implemented using a mouse can be documented either in a specific worksheet within the spreadsheet that is just meant for documentation. The documentation can also be done on a separate document either in Word document (`.docx`) format or in a text-based format such as Markdown (`.md`) or text file (`.txt`) written using a text editor (see Tip 1 for recommendations on text editors for different operating software). This separate file should be included within the directory where the spreadsheet file is located (see Figure 11.3). An example of a text file documenting steps for data cleaning is shown in Figure 11.4.

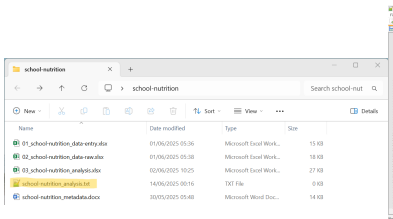


Figure 11.3: A text file for documentation notes on point-and-click mouse-based steps

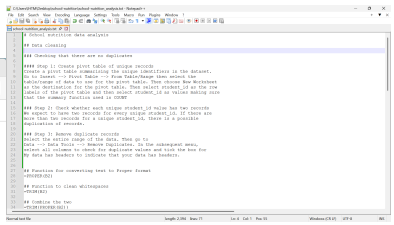


Figure 11.4: An example documentation of point-and-click mouse-based steps and in cell functions and calculations

11.1.2 Saving a source for written functions in spreadsheets

Saving a text-based source file for the syntax of in cell functions and calculations used in a spreadsheet is one way of recording the non-mouse steps of the spreadsheet workflow. A text editor (rather than a word processor) would be ideal for this as the syntax of the function or formula will be shown more appropriately. If you are already using a text editor for documenting mouse-based steps of the spreadsheet workflow,

it would be ideal to use the same text file to record in cell functions and calculations as shown in Figure 11.4.

> Tip 1: Recommended text editors
>
> Following are recommended text editors for use with different operating software:
> **For Windows**
> Notepad++ is a free source code editor and Notepad replacement that supports several languages. Running in the Microsoft Windows environment, its use is governed by GNU General Public License.
> **For Mac**
> CodeEdit is an exciting new code editor written entirely and unapologetically for macOS. Develop any project using any language at speeds like never before with increased efficiency and reliability in an editor that feels right at home on your Mac.

Create a text file to associate with every spreadsheet project that you are working on. Save the text file within the same directory as the associated spreadsheet as shown in Figure 11.3.

11.2 Organise work into projects

Organising work into projects is another best practice that provides organisational clarity for our data and data processes. Whilst this can be interpreted in many ways and that some of you may argue that you already organise your work with data in projects, the following key points give a clear indication/definition of what we mean by project-based workflows.

11.2.1 File system discipline

Simply put, this means putting all the files related to a single project in a designated folder. This applies to data, code, figures, notes, including the documentation and source files described earlier (see Figure 11.5). Depending on complexity of your project and on yours or your team's/organisation's

preferences, you might enforce further organisation into subfolders. Related and relevant file system practices are discussed in Section 11.2.3.

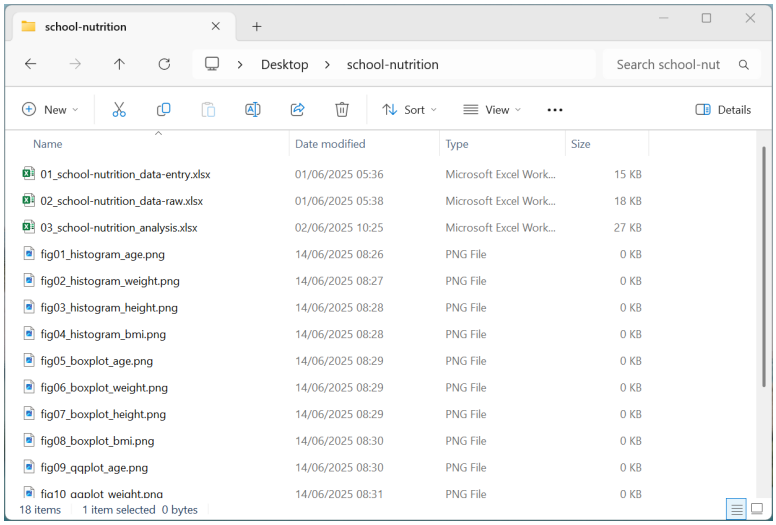


Figure 11.5: An example directory for the school nutrition project with all relevant files included

11.2.2 File path discipline

All paths are relative and, by default, relative to the project's folder. This is particularly important when you are referencing data found in one spreadsheet from within another spreadsheet for data analysis and visualisation. If files are within the same project, then relative paths make it easy to refer to associated or ancillary spreadsheets required for full analysis, visualisation, and reporting.

11.2.3 File naming

> File organisation and naming are powerful weapons against chaos.

Best practices in file naming are based on the following three principles:

Name	Size	Modified	
analysis for KA 1 PLAs.xlsx.gpg	412.6 kB	03:15	☆
Attainment result term 3.doc.gpg	11.4 kB	03:15	☆
data analysis PLAs WSI 2023.xlsx.gpg	224.7 kB	03:15	☆
data for KA 1 PGAp wsi 2023.xlsx.gpg	34.6 kB	03:15	☆
P6 National Exam Results 2022.pptx.gpg	845.5 kB	03:15	☆
S4_2021_option.xlsx.gpg	30.5 kB	Fri	☆
subject Report Term 1 2023.docx.gpg	19.4 kB	03:15	☆
TERM 1 2023 RESULTS.xlsx.gpg	46.0 kB	03:15	☆

Figure 11.6: An example of messy file names

Machine-readable

Machine-readable file names avoid spaces, punctuation, accented characters, and case sensitivity. Avoiding these makes file names easier to index and search via use of **regular expressions** and **wild card matching** or **globbing**.

A regular expression, usually shortened as *regex* or *regexp* and sometimes referred to as *rational expression*, is a sequence of characters that specifies a match pattern in text. Usually such patterns are used by string-searching algorithms for *"find"* or *"find and replace"* operations on strings, or for input validation.

```
Windows PowerShell
Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! https://aka.ms/PSWindows

PS C:\Users\IHTM> cd Desktop/school-nutrition
PS C:\Users\IHTM\Desktop\school-nutrition> Get-ChildItem | Where-Object { $_.Name -match 'fig[0-9]{2}' }


    Directory: C:\Users\IHTM\Desktop\school-nutrition


Mode                 LastWriteTime         Length Name
----                 -------------         ------ ----
-a----        14/06/2025     08:26              0 fig01_histogram_age.png
-a----        14/06/2025     08:27              0 fig02_histogram_weight.png
-a----        14/06/2025     08:28              0 fig03_histogram_height.png
-a----        14/06/2025     08:28              0 fig04_histogram_bmi.png
-a----        14/06/2025     08:29              0 fig05_boxplot_age.png
-a----        14/06/2025     08:29              0 fig06_boxplot_weight.png
-a----        14/06/2025     08:29              0 fig07_boxplot_height.png
-a----        14/06/2025     08:30              0 fig08_boxplot_bmi.png
-a----        14/06/2025     08:30              0 fig09_qqplot_age.png
-a----        14/06/2025     08:31              0 fig10_qqplot_weight.png
-a----        14/06/2025     08:31              0 fig11_qqplot_height.png
-a----        14/06/2025     08:32              0 fig12_qqplot_bmi.png


PS C:\Users\IHTM\Desktop\school-nutrition>
```

Figure 11.7: Using regular expression to find all files that start with fig followed by a two-digit number

Globbing, also known as wildcard matching, is a technique used in computer systems to match multiple files or paths based on patterns containing wildcards like * (asterisk) and

? (question mark). It's a common way to specify a set of files or paths in command-line interfaces, file managers, and programming languages. In simpler terms: globbing allows you to use patterns to find files that share a common naming structure, without having to specify each file individually.

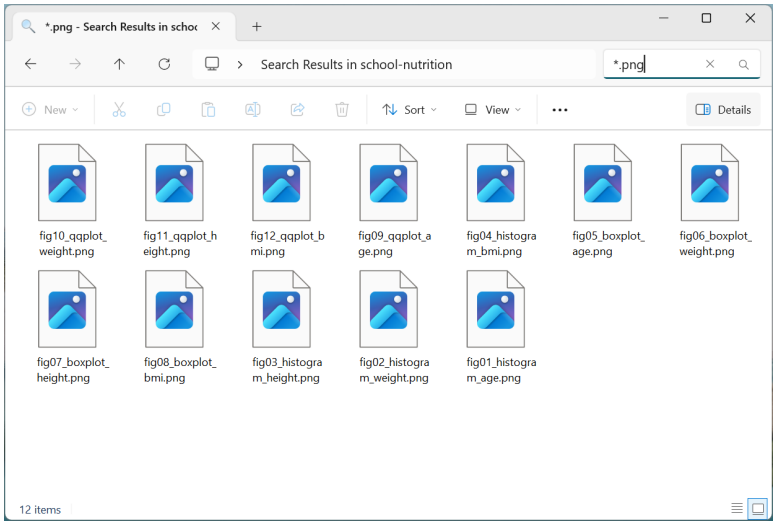


Figure 11.8: Using wildcard matching to find all PNG files

Machine-readable file names have deliberate use of delimiters/space-holders such as underscore (_) or hyphen (-). The general rule is that _ is used to delimit units of metadata while - is used to delimit words so that they are more readable. This system allows for much easier recovery of metadata from file names.

Human-readable

A file name is human-readable if it contains information on what the file contains. It should be easy to figure out what something is based on its file name. This is a similar concept to a *slug* from semantic URLs. A URL slug is the unique, identifiable portion of a web address (URL) that follows the domain name (e.g., "google.com"). It essentially acts as a "name tag" for a specific page or resource on a website, helping both users and search engines understand what the page is about.

Plays well with default ordering

File names should play well with default ordering. This is usually achieved by:

- putting something numeric first in a file name;

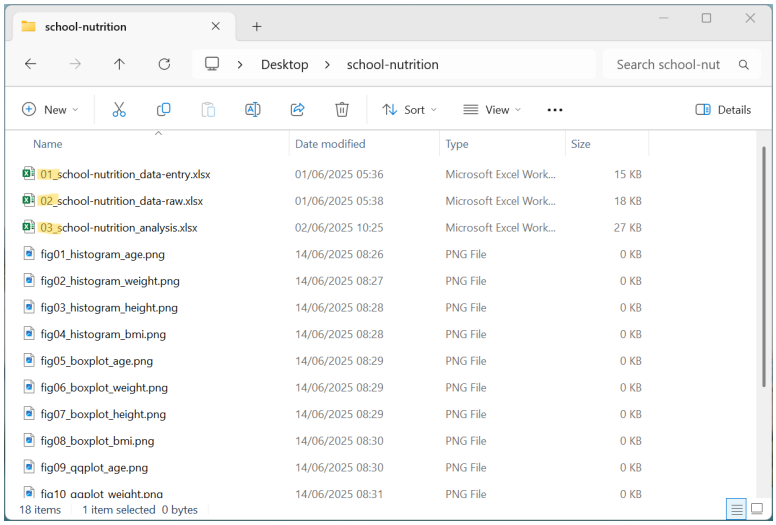


Figure 11.9: Putting something numeric first in a file name

- using the ISO 8601 standard (YYYY-MM-DD) for dates; and,
- pad the left side of other numbers with zeros.

11.3 Gains from project-based workflows

Developing the different project-based workflow habits described above collectively yields the most significant benefits. These practices ensure projects can move seamlessly across different computers or environments while maintaining reliability. Project-based workflow approach is a practical convention for achieving consistent behaviour across users and time comparable to societal norms like agreeing on traffic rules (e.g., driving on the left or right). Adhering to these conventions - whether in computing or in broader civilisation - constrains individual actions slightly for greater functionality and safety.

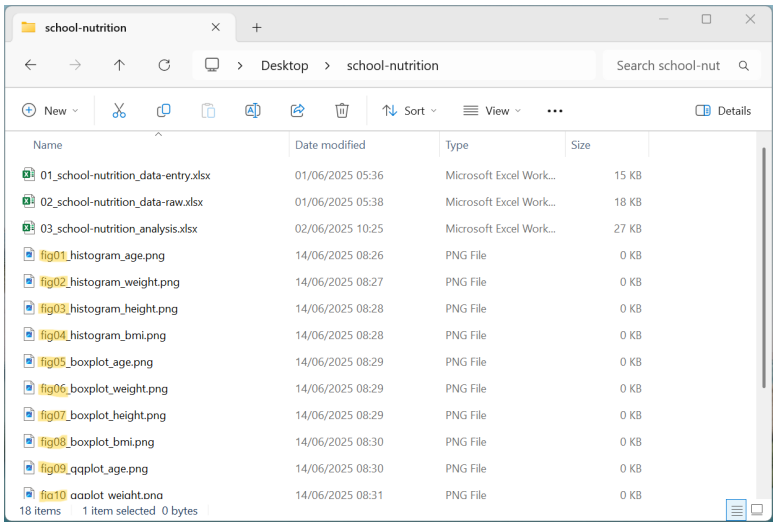


Figure 11.10: Left pad other numbers with zeros

12 Data entry/collection and storage using spreadsheets

Spreadsheets are often used as a multi-purpose tool for data entry, storage, analysis, and visualisation. Most spreadsheet software available allows users to perform all of these tasks. However, spreadsheets are best suited to data entry and storage. Analysis and visualisation should be done separately either by using other data tools or at least in a separate copy of the data file in order to reduce the risk of contaminating or accidentally changing the raw data in the spreadsheet.

Spreadsheets, by design, make humans format data to be viewed by the human eye rather than to be readable by a machine (Murrell, 2013). Data structured as such require greater amount of effort, usually in computer code, to be able to extract the information needed for analysis. However, if the initial structure is such that it is easily readable by a machine, the effort leading to analysis is much more simplified.

In this chapter, we discuss best practices in using spreadsheets as a data entry and data storage tool and provide specific recommendations for organising spreadsheet data in a way that both humans and computers can read. Following these recommendations allows the creation of spreadsheets that minimise errors, are easy for computers to process, and facilitate collaboration and public access. These well-structured spreadsheets integrate with reproducible methods, serving as a reliable foundation for robust analytic workflows.

12.1 Be consistent

Consistency is key in data organisation. Strive for uniformity in your data entry and organisation practices. By maintain-

ing this consistency from the start, you can save yourself and your collaborators from the hassle of reconciling inconsistencies later on.

Following are some examples of being consistent and why it is important. Some of these examples are also part of the recommendations listed here.

- **Use consistent codes for categorical variables.**

For a categorical variable like the sex, use a single common value for males (e.g., "male"), and a single common value for females (e.g., "female"). Do not sometimes write "M," sometimes "male," and sometimes "Male." Pick one and stick to it. In order to limit the occurrence of this inconsistency, you can enforce a data validation rule for this variable (see Section 12.13).

- **Use a consistent fixed code for any missing values.**

It is ideal to have every cell filled in so that distinguishing between truly missing values and unintentionally missing values is more straightforward. Decide right at the outset what value to use for missing values and stick with that value throughout. Do not use a note explaining why a value is missing in place of the data itself. Rather, make a separate column with such notes.

- **Use consistent variable names.**

Name variables exactly the same way throughout one file and across every other file relevant to the project. If naming is inconsistent for the same variable, those working with the data will have to work out that these are all really the same thing. See Section 12.2 for more discussion on best practices in naming things within a data file. See Section 11.2.3 for an in-depth discussion on best practices in naming files.

- **Use consistent subject identifiers.**

Create consistent and unique subject identifiers to avoid extra work in figuring out which subject/record if which.

- **Use a consistent data layout in multiple files.**

If your data are in multiple files and you use different layouts in different files, it will be extra work for the analyst to combine the files into one dataset for analysis. With a consistent structure, it will be easy to automate this process.

- **Use consistent file names.**

Have some system for naming files. Keeping a consistent file naming scheme will help ensure that your files remain well organised, and it will make it easier to batch process the files if you need to. See an in-depth discussion of file naming in Section 11.2.3.

- **Use a consistent format for all dates.**

Preferably use the standard format `YYYY-MM-DD`, for example, 2015-08-01. If sometimes you write 8/1/2015 and sometimes 8-1-15, it will be more difficult to use the dates in analyses or data visualisations. See Section 12.3 for more discussion on this.

- **Use consistent phrases in your notes.**

If you have a separate column of notes (e.g., `"dead"`), be consistent in what you write. Do not sometimes write `"dead"` and sometimes `"Dead"`.

- **Be careful about extra spaces within cells.**

A blank cell is different than a cell that contains a single space. And `"male"` is different from `" male "` (i.e., with spaces at the beginning and end).

12.2 Choose good names for things

It is important to pick good names for things such as variables. This can be hard, and so it is worth putting some time and thought into it. Section 11.2.3 provides some general principles for naming files that can also be used when naming variables in data.

12.2.1 General rules for naming

- Do not use spaces, either in variable names or file names

Spaces make programming harder. An analyst will need to surround a name that contains spaces in double quotes in order to refer to it. Use underscores (`_`) or hyphens (`-`) instead of spaces. Do not use a mixture of underscores and hyphens; pick one and be consistent.

- No extraneous spaces at the start and/or end of variable names

Be careful about extraneous spaces at the beginning or end of a variable name. `"sex"` is different from `"sex "` (with an extra space at the end) or `" sex"` (with an extra space at the start).

- Avoid special characters, except for underscores and hyphens

Other symbols (`$`, `@`, `%`, `#`, `&`, `*`, `(`, `)`, `!`, `/`, etc.) often have special meaning in programming languages, and so they can be harder to handle. They are also a bit harder to type.

The main principle in choosing names, whether for variables or for file names, is short, but meaningful. So not too short. The following table (adapted from The Data Carpentry lesson on using spreadsheets) show good and bad example variables names.

Table 12.1: Examples of good and bad variable names

Good name	Good alternative	Avoid
max_temp_c	MaxTemp	Maximum Temp (°C)
precipitation_mm	Precipitation	precmm
mean_year_growth	MeanYearGrowth	Mean growth/year
sex	sex	M/F
weight	weight	w.
cell_type	CellType	Cell type
observation_01	first_observation	1st Obs.

The first column of variable names use the *snake case* naming convention which uses an underscore to replace a space and

letters are in lower case. The second column of good alternative variable names use the *camel case* naming convention in which phrases are written without spaces or punctuation and with capitalised words.

12.3 Write dates as YYYY-MM-DD

Figure 12.1: ISO 8601 from https://xkcd.com/1179/

When entering dates, we strongly recommend using the global *ISO 8601* standard, `YYYY-MM-DD`, such as 2013-02-27.

Microsoft Excel's treatment of dates can cause problems in data. It stores them internally as a number, with different conventions on Windows and Macs (see Note 3). So, you may need to manually check the integrity of your data when they come out of Excel.

i Note 3: Date systems in Microsoft Excel

Excel supports two date systems, the *1900 date system* and the *1904 date system*. Each date system uses a unique starting date from which all other workbook dates are calculated. All newer versions of Excel calculate dates based on the 1900 date system, while older versions used the 1904 system.

When you copy dates from a workbook created in an earlier version to a workbook created in a newer version, they will be converted automatically unless the option to `"Automatically convert date system"` is disabled in `Preferences > Edit > Date` Options. If this option is disabled, you will receive a message asking whether the dates should be converted when pasted. You have two options. You can convert the dates to use the 1900 date system (recommended). This option makes the dates compatible with other dates in the workbook. Or you can keep the 1904 date system for the pasted dates only.

The 1900 date system

In the 1900 date system, dates are calculated by using *January 1, 1900*, as a starting point. When you enter a date, it is converted into a serial number that represents the number of days elapsed since January 1, 1900. For example, if you enter July 5, 2011, Excel converts the date to the serial number 40729. This is the default date system in Excel for Windows, Excel 2016 for Mac, and Excel for Mac 2011. If you choose to convert the pasted data, Excel adjusts the underlying values, and the pasted dates match the dates that you copied.

The 1904 date system

In the 1904 date system, dates are calculated by using *January 1, 1904*, as a starting point. When you enter a date, it is converted into a serial number that represents the number of days elapsed since January 1, 1904. For example, if you enter July 5, 2011, Excel converts the date to the serial number 39267. This is the default date system in earlier versions of Excel for Mac. If you choose not to convert the data and keep the 1904 date system, the pasted dates vary from the dates that you copied.

The difference between the date systems
Because the two date systems use different starting days, the same date is represented by different serial numbers in each date system. For example, July 5, 2011, can have two different serial numbers, as follows:

Table 12.2: Comparison of Excel's 1900 and 1904 date systems

Date System	Serial Number
1900	40729
1904	39267

The difference between the two date systems is *1,462 days*. This means that the serial number of a date in the 1900 date system is always 1,462 days greater than the serial number of the same date in the 1904 date system. 1,462 days is equal to *four years and one day (including one leap day)*.
*taken from Microsoft Support documentation

To avoid these issues with dates when using spreadsheets (specifically Excel), we recommend the following:

- Use a plain text format for columns in an Excel worksheet that are going to contain dates

Doing so will avoid automatic conversion of these variables into often unpredictable formats. This can be done through the following steps:

Step 1: Create a date variable (see Figure 12.2).

Step 2. Select the date variable you created (see Figure 12.3).

Step 3: In the menu bar, select *Format –> Cells –> Choose "Text"* on the left (see Figure 12.4).

This approach will only work if you are creating the date variable first and when no date values have been entered yet. If you do this on a date variable that already contain dates, Excel will convert them to a text value of their underlying numeric representation (as described in Note 3).

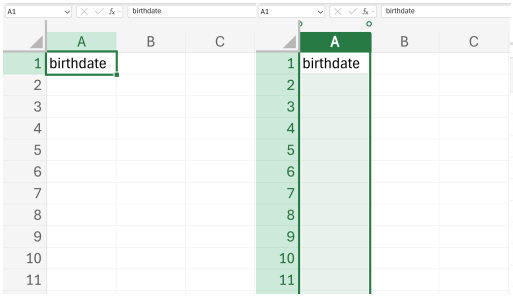


Figure 12.2: Create date variable

Figure 12.3: Select date variable

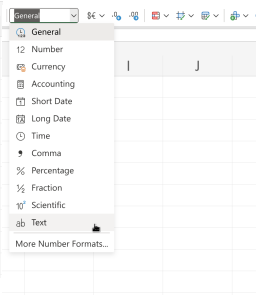


Figure 12.4: Set format to text

- Place an apostrophe at the start of a date value entry

Another way to force Excel to treat dates as text is to begin the date with an apostrophe, like this: `'2014-06-14`. Excel will treat the cells as text, but the apostrophe will not appear when you view the spreadsheet or export it to other formats. This is a handy trick, but it requires impeccable diligence and consistency.

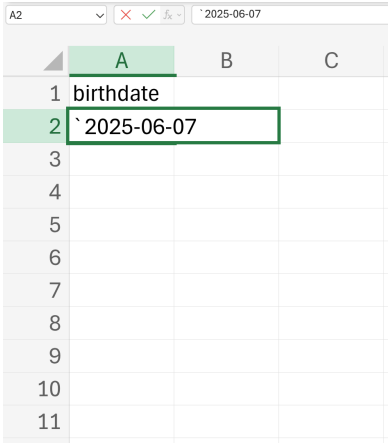


Figure 12.5: Enter an apostrophe followed by YYYY-MM-DD date format

- Create three separate columns with year, month, and day

These will be ordinary numbers, and so Excel will not mess

them up. If there is an existing date variable already, you can convert that to year, month, and day columns by using the built-in date functions in Excel that extract year (Figure 12.6), month (Figure 12.7), and day (Figure 12.8) values from a date variable.

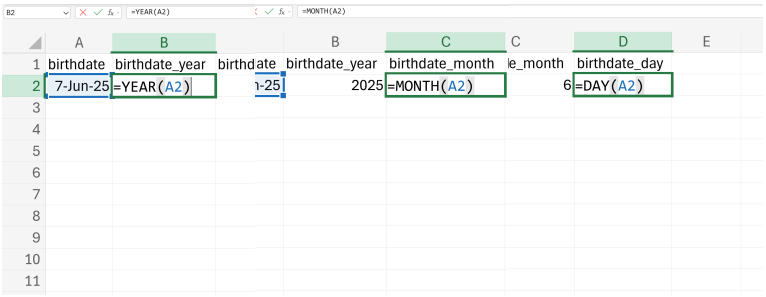


Figure 12.6: Extract year from date value

Figure 12.7: Extract month from date value

Figure 12.8: Extract year from date value

- Represent dates as an 8-digit integer of the form YYYYMMDD

For example, `20140614` for `2014-06-14` (see Figure 12.9).

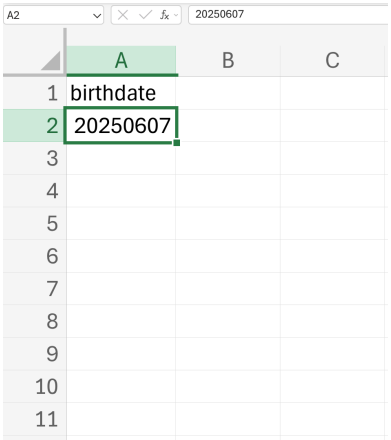


Figure 12.9: Date value as text in YYYYMMDD format

12.4 No empty cells

Fill in all cells. Use some common code for missing data to make it clear that the data are known to be missing rather than unintentionally left blank.

12.5 Put just one thing in a cell

Your spreadsheet should have cells that each contain one piece of data only. Putting more than one type of data value in a cell is not considered best practice.

For example, you might have a column with information on year (containing values of either *2022*, *2023*, or *2024*) and `sex` (*Male* or *Female*) as `year-sex` such as `2022-Male`, `2022-Female`, and so on and so forth. It would be better to separate this into `year` and `sex` columns (containing *2022* and *Male*).

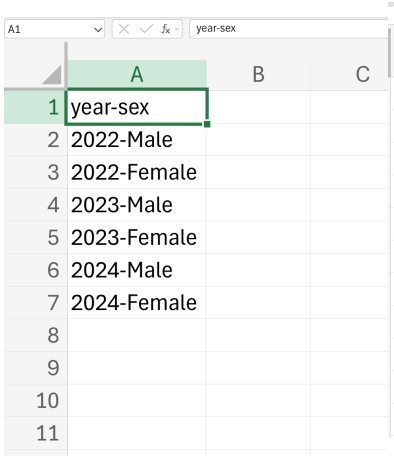


Figure 12.10: Combined year-sex variable

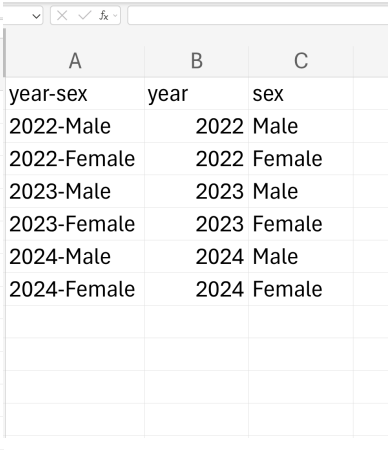
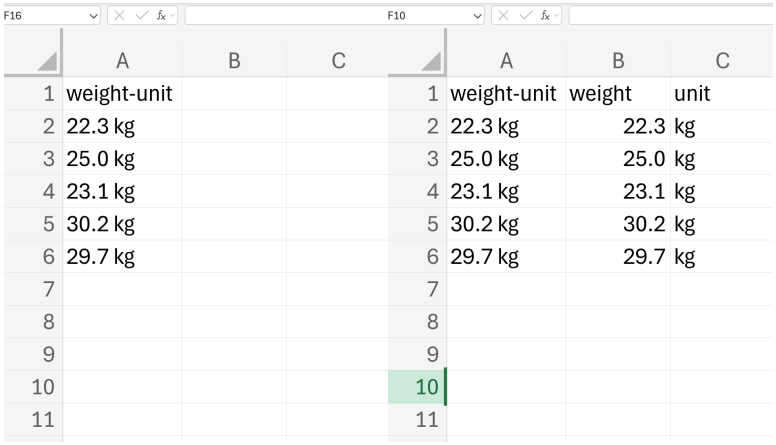


Figure 12.11: Year and sex as separate variables

Or you might include units alongside measurements such as weight. It is better to have a variable for the weight and then a separate variable for the units.



	A	B	C
1	weight-unit		
2	22.3 kg		
3	25.0 kg		
4	23.1 kg		
5	30.2 kg		
6	29.7 kg		
7			
8			
9			
10			
11			

Figure 12.12: Combined weight-unit variable

	A	B	C
1	weight-unit	weight	unit
2	22.3 kg	22.3	kg
3	25.0 kg	25.0	kg
4	23.1 kg	23.1	kg
5	30.2 kg	30.2	kg
6	29.7 kg	29.7	kg
7			
8			
9			
10			
11			

Figure 12.13: Separate weight and unit variables

It is even better to just have a variable for weight and then document the units used in a separate data dictionary (see Section 12.7 on creating a data dictionary).

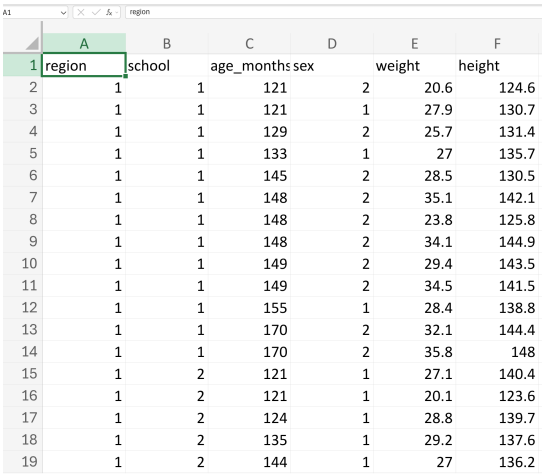
Finally, do not merge cells. It might look pretty, but you end up breaking the rule of no empty cells.

12.6 Make it a rectangle

A single big rectangle with rows corresponding to subjects and columns corresponding to variables is the best layout for data within a spreadsheet. The first row should always contain variable names. Do not use more than one row for the variable names.

12.7 Create a dictionary

Having a separate file (see Figure 12.15) that outlines all variables can be very helpful, particularly if it's organised in a structured layout so data analysts can use it effectively in their analyses. We recommend that this *data dictionary* includes the following information:



	A	B	C	D	E	F
1	region	school	age_months	sex	weight	height
2	1	1	121	2	20.6	124.6
3	1	1	121	1	27.9	130.7
4	1	1	129	2	25.7	131.4
5	1	1	133	1	27	135.7
6	1	1	145	2	28.5	130.5
7	1	1	148	2	35.1	142.1
8	1	1	148	2	23.8	125.8
9	1	1	148	2	34.1	144.9
10	1	1	149	2	29.4	143.5
11	1	1	149	2	34.5	141.5
12	1	1	155	1	28.4	138.8
13	1	1	170	2	32.1	144.4
14	1	1	170	2	35.8	148
15	1	2	121	1	27.1	140.4
16	1	2	121	1	20.1	123.6
17	1	2	124	1	28.8	139.7
18	1	2	135	1	29.2	137.6
19	1	2	144	1	27	136.2

Figure 12.14: Data in spreadsheet with rectangular layout

1. The **precise variable names** as they appear in the dataset.
2. A **detailed description** explaining what the variable represents.
3. The **measurement units** associated with each variable.
4. The **list of possible values** (for categorical variables) and/or **typical range of values expected** (for numerical variables) for that variable.

An example of this data dictionary within an accompanying Word document metadata in a project-based workflow is shown in Figure 12.16.

An alternative to a separate metadata and data dictionary file is to include this documentation as a separate worksheet within the spreadsheet containing the data.

12.8 No calculations in the raw data file

Excel files often come with various calculations and graphs included alongside the data itself. We strongly recommend keeping your primary data file free from any additional content - only raw data should be present. This is because editing the same file for calculations can lead to accidental errors. For instance, when you open a file and start typing without

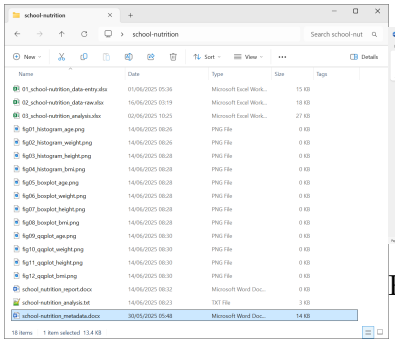


Figure 12.15: A project-based workflow structure with a separate file for metadata

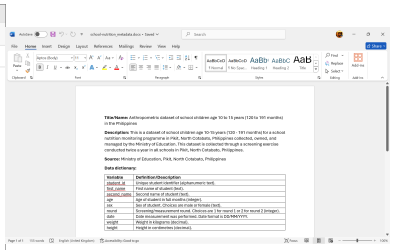


Figure 12.16: A recommended layout and contents of a project metadata containing a data dictionary

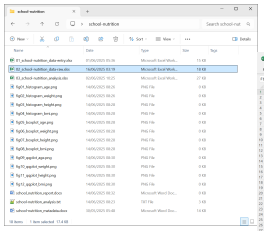


Figure 12.17: A project-based workflow structure with separate file for data

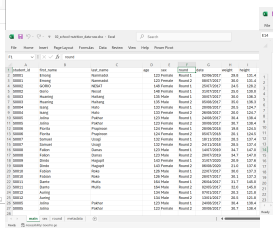


Figure 12.18: Project spreadsheet with raw data in worksheet called “main”

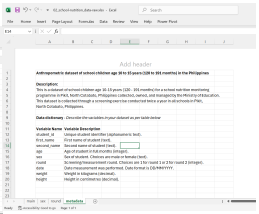


Figure 12.19: Project spreadsheet with metadata/dictionary in worksheet called “metadata”

selecting a cell, the text might end up in unexpected cells, causing problems during analysis. To prevent this, protect your main data file from changes, keep it backed up, and refrain from making edits. If you need to perform analyses or create graphs, work on a duplicate of the original file.

12.9 Do not use font colour or highlighting as data

You might be tempted to highlight particular cells with suspicious data, or rows that should be ignored. Or the font or font colour might have some meaning. Instead, add another column with an indicator variable (e.g., "trusted" with values TRUE or FALSE).

Another possible use of highlighting would be to indicate males and females in a mouse study by highlighting the corresponding rows in different colours. But rather than use highlighting to indicate sex, it is better to include a sex column, with values Male or Female.

12.10 Make backups

Regularly back up your data by storing copies in multiple locations. Consider using a formal version control system such as *git* (though it's not ideal for data files). Keep every version of your data files and label them with version numbers, like `file_v1.xlsx`, `file_v2.xlsx`, and so on and so forth. For this, the guidance on good file names discussed in Section 11.2.3 is a good reference to follow for naming your file versions. Once you've finished entering data or if taking a break, protect the file from accidental changes by setting it to *read-only* as described below.

12.11 Windows

1, Right-click the file in File Explorer or select the file and then click on the triple dot icon in File Explorer as shown in Figure 12.20.

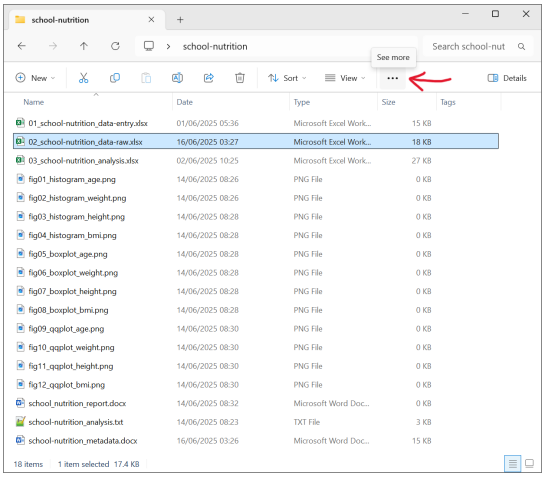


Figure 12.20: See more file options in File Explorer

2. In the drop-down menu, choose `Properties` as shown in Figure 12.21.

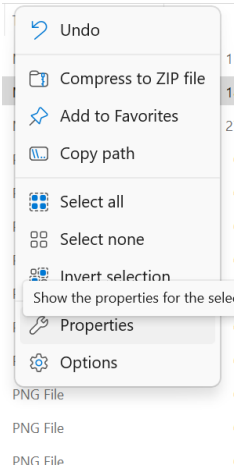


Figure 12.21: Select Properties option in the drop-down menu

3. In the pop-up menu, navigate to the `General` tab, check the `Attributes` box for `Read-only`, and confirm with `OK` as shown in Figure 12.22.

12.12 macOS

1. Open Finder and right-click on the file.

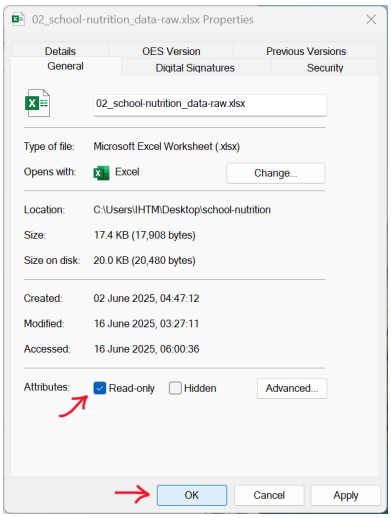


Figure 12.22: Set the file to read-only

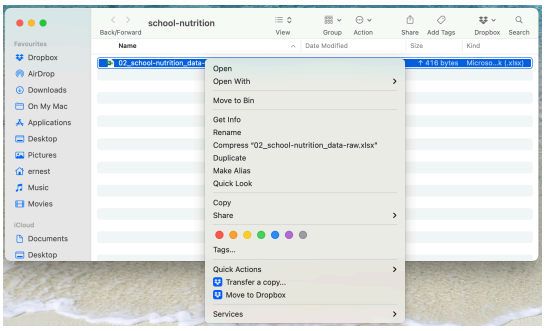


Figure 12.23: Right-click on the spreadsheet file

2. Select `Get Info` then go to `Sharing & Permissions`.

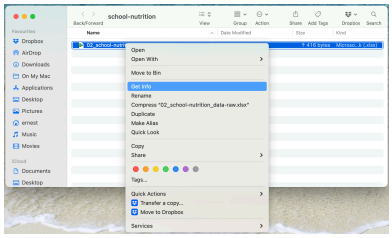


Figure 12.24: Select Get Info

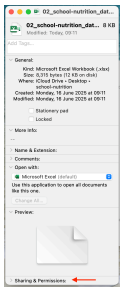


Figure 12.25: Go to Sharing & Permissions

3. Set your privileges to `Read only`.

Another option is to password protect the worksheet in the spreadsheet that has the data.

1. Select the worksheet to password-protect and then go to `Review` as shown in Figure 12.31.

2. Select `Protect Sheet` as shown in Figure 12.28.

3. Enter password to protect worksheet as shown in Figure 12.29.

4. Re-enter password to confirm protection as shown in Figure 12.30.

Password protection can also be applied to the whole spreadsheet workbook.

1. Open the spreadsheet to password-protect and then go to `Review` as shown in Figure 12.31.

2. Select `Protect Workbook` as shown in Figure 12.32.

3. Enter password to protect workbook as shown in Figure 12.33.

4. Re-enter password to confirm workbook protection as shown in Figure 12.34.

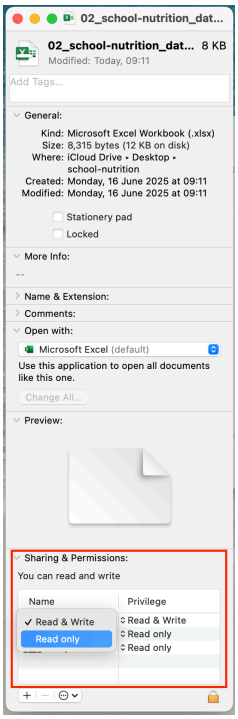


Figure 12.26: Set privileges to Read only

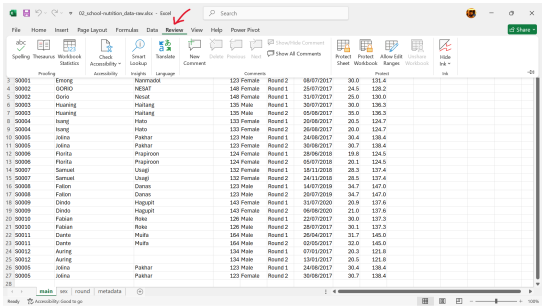


Figure 12.27: Go to Review

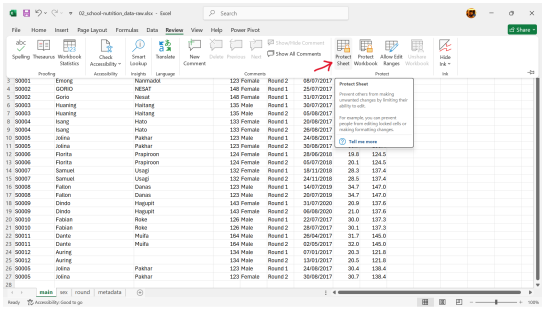


Figure 12.28: Select Protect Sheet

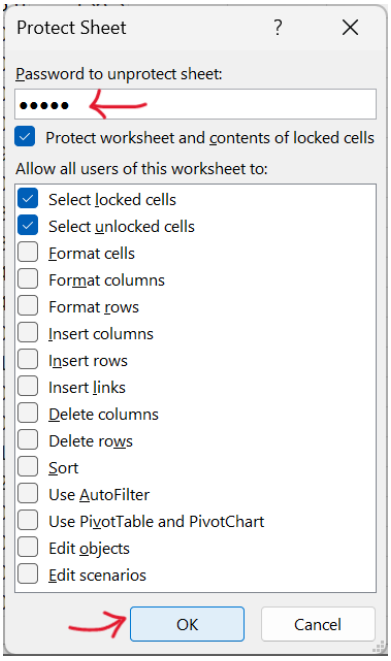


Figure 12.29: Enter password to protect worksheet

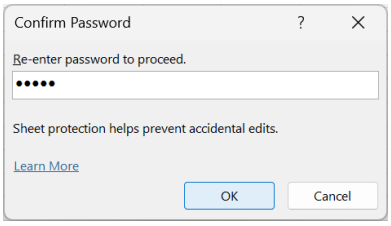


Figure 12.30: Re-enter password to confirm protection

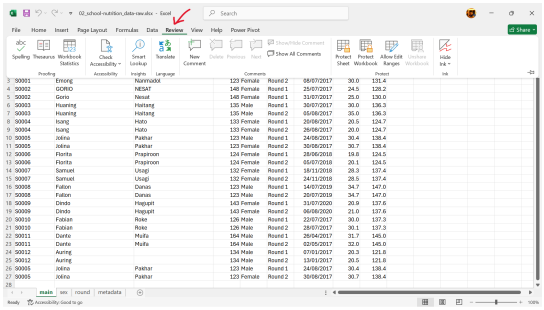


Figure 12.31: Go to Review

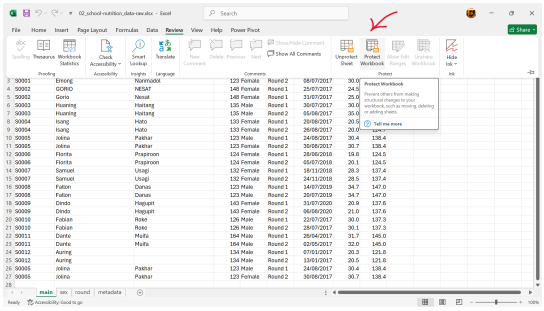


Figure 12.32: Select Protect Workbook

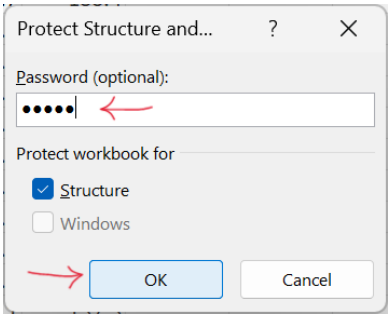


Figure 12.33: Enter password to protect workbook

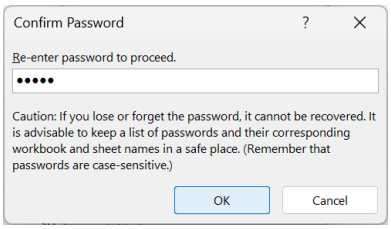


Figure 12.34: Re-enter password to confirm workbook protection

> **Important 1**
>
> Remember, always back up your data!

12.13 Use data validation to avoid errors

When handling data entry tasks using spreadsheets, it's crucial to aim for accuracy and comfort to minimise errors and reduce physical strain. Excel provides a helpful `Data Validation` feature that can prevent errors during data entry.

To use this feature:

1. Choose the column you wish to validate. In Figure 12.35, the data entry for the `age` variable is to be validated.

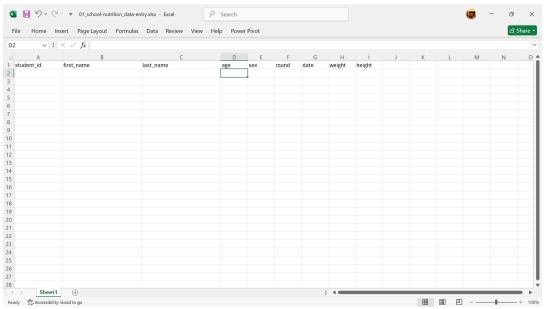


Figure 12.35: Add validation to the age column/variable

2. Go to the menu bar and select `Data --> Data Tools --> Data Validation` as shown in

3. Set up validation criteria such as:

    - Whole numbers within a specified range
    - Decimal numbers within a specified range
    - A predefined list of acceptable values
    - Text with length restrictions

In Figure 12.37, we set validation for age variable to allow only whole numbers ranging from 120 to 191 (inclusive). The `Ignore blank` option is ticked so that blank entry will be accepted.

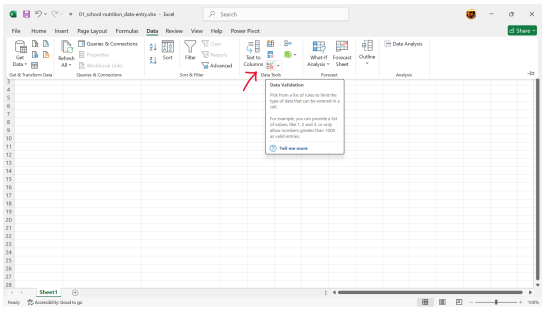


Figure 12.36: Select Data Validation in Data Tools

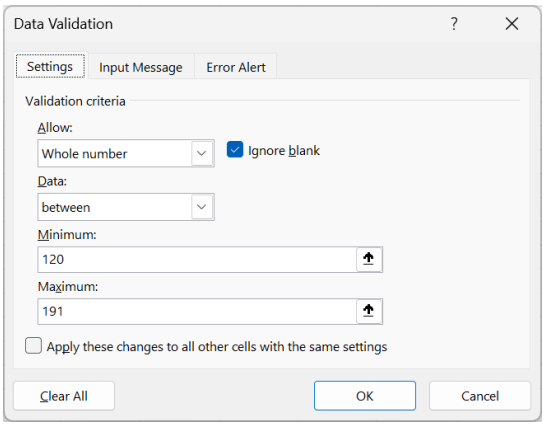


Figure 12.37: Setup validation criteria for age variable

4. Add title and message to guide data entry input as shown in Figure 12.38.

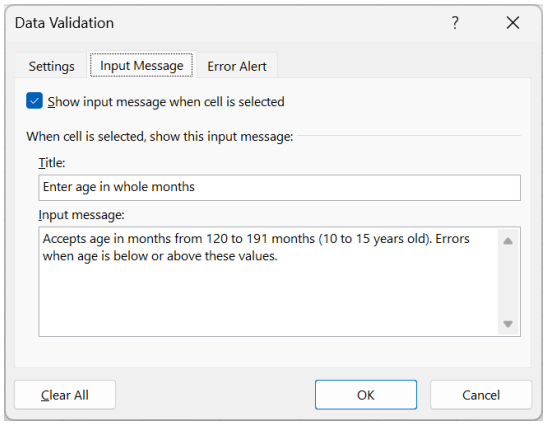


Figure 12.38: Add title and message guide for data entry

5. Add error alert to show up when incorrect data entry input is made as shown in Figure 12.39

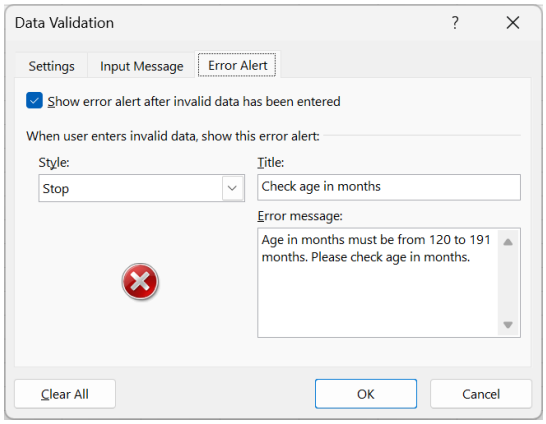


Figure 12.39: Add error alert to show up when incorrect data entry input is made

Microsoft Support has further documentation and guidance on how to apply data validation to cells here.

Additionally, formatting cells as "Text" can prevent unintended changes to data like dates or names. In Section 12.3, the steps to change formatting of cells is described and demonstrated.

While these steps may seem tedious, they are valuable in

maintaining data integrity and minimising errors during entry.

Part IV

Exploratory Data Analysis

13 Exploratory data analysis

Data analysis involves steps like cleaning, transforming, inspecting, and modelling data to extract meaningful information. This process can serve various purposes, including exploratory and confirmatory analyses, as well as descriptive or predictive tasks.

Before building models or making predictions, it's essential to explore the data to identify underlying patterns and structures. Data analysts employ both numerical and visual techniques to uncover insights that might be hidden within the dataset. However, it's crucial for analysts to avoid over-interpreting apparent patterns and to ensure that the findings are reliable for the given data and potentially applicable to new datasets as well. Exploratory data analysis fills this role.

Following are a few other definitions of exploratory data analysis (EDA).

13.1 Definitions

From Wikipedia:

> In statistics, exploratory data analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often with visual methods. A statistical model can be used or not, but primarily EDA is for seeing what the data can tell us beyond the formal modeling or hypothesis testing task.

From Wickham and Grolemund (2023):

> EDA is not a formal process with a strict set of rules. More than anything, EDA is a state of mind. During the initial phases of EDA you

> should feel free to investigate every idea that occurs to you. Some of these ideas will pan out, and some will be dead ends.

From SAS:

> EDA is necessary for the next stage of data research. If there was an analogy to exploratory data analysis, it would be that of a painter examining their tools and available time, before deciding on what best to paint.

13.2 Origins

The field of EDA got into the forefront with the publication of Tukey's Exploratory Data Analysis (Tukey, 1977). Tukey's aim in writing the book was to provide individual and isolated techniques useful to data analysts. All of Tukey's techniques in the EDA book can be done by hand with pencil and paper.

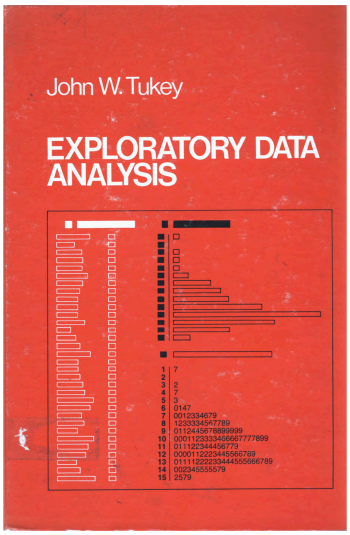


Figure 13.1: Book cover of Tukey's Exploratory Data Analysis

Following are some quotes by Tukey from the EDA book.

13.2.1 On measures

It is important to understand what you **can do** before you learn to measure how **well** you seem to have **done** it.

13.2.2 On pictures

The greatest value of a picture is when it forces us to notice **what we never expected to see**.

13.2.3 On exploration

Once upon a time, statisticians only explored.

13.2.4 On not having one right answer

There can be many ways to approach a body of data. Not all are equally good.

14 Univariate statistics

In this chapter, we will focus on exploratory data analysis of a single variable. We discuss the features of a continuous variable and how these features can be explored, elucidated, tested, and visualised. We then discuss ordinal and nominal categorical variables and the various considerations when exploring these types of variables.

14.1 Continuous variables

A *continuous variable* is a type of variable that can take on any value within a given range. It's characterised by the ability to be measured and can have an infinite number of values between any two given points. Examples include height, weight, temperature, and time.

14.1.1 Measure of central tendency

In statistics, *measure of central tendency* is a central or typical value for a probability distribution. Most common measures are *mean*, *median*, and *mode* but there are many other measures of central tendency. It is important to note that not all measures of central tendency are robust.

Mean

Mean is the sum of a set of values divided by the number of values. Mathematically, it is represented as:

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

Mean is a non-robust measure of location because it is susceptible/sensitive to a few extreme values in the data.

Median

For a dataset x of n elements ordered from smallest to greatest, if n is odd

$$\tilde{x} = x_{(n+1)/2}$$

and if n is even

$$\tilde{x} = \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$$

The median is the value that divides a dataset into two equal halves, separating the higher half from the lower half. It can be seen as the 'middle' value in a data sample, population, or probability distribution. Unlike the mean (often referred to as the "average"), the median is not influenced by extremely large or small values, making it more resistant to skewed data and providing a better representation of the dataset's center.

For instance, when analysing income distribution, the median income is often a more accurate measure of the central tendency because increases in the highest incomes do not affect the median. This characteristic makes the median particularly valuable in robust statistics, where its ability to withstand the impact of outliers is highly regarded.

> **Note 4: Mean vs Median**
>
> Consider this data of test scores of ten students.

	A	B
1	student	score
2	A	10
3	B	15
4	C	80
5	D	90
6	E	75
7	F	85
8	G	92
9	H	90
10	I	90
11	J	90

Figure 14.1: Test scores of ten students

8 out of the 10 students did really well with scores 75 and higher but two students got really low scores.
Calculating the mean scores using the spreadsheet, we use the `AVERAGE()` function:

```
=AVERAGE(B2:B11)
```

we get **71.7**.
Calculating the median scores using the spreadsheet, we use the `MEDIAN()` function:

```
=MEDIAN(B2:B11)
```

we get **87.5**.
The mean and the median can be very different from each other.
If `median > mean`, this would indicate that the continuous variable has some extremely low values
If `median < mean`, this would indicate that the continuous variable has some extremely high values
We should be using median instead of mean when performing summary measures for continuous variables

14.1.2 Measure of dispersion

In statistics, measure of dispersion describes the spread or variability of data points within a dataset. It indicates how much individual values deviate from the central tendency. Essentially, it provides insight into whether the data is tightly or loosely clustered around its center. Common measures of dispersion include *variance*, *standard deviation (SD)*, and *interquartile range (IQR)*.

Standard deviation and variance are the most popular choice for measure of dispersion but are not robust to extreme values or outliers.

Standard deviation

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

where:

n = number of observations in sample

x_i = individual data point

$\bar{x}$ = sample mean

In Excel, you can calculate standard deviation using the `STDEV()` function. Using the student scores example:

```
=STDEV(B2:B11)
```

which gives **31.6756128**.

A low standard deviation indicates that the values tend to be close to the mean of the set, while a high standard deviation indicates that the values are spread out over a wider range. The standard deviation is commonly used in the determination of what constitutes an outlier and what does not.

Interquartile range (IQR)

IQR is the difference between the 1st and 3rd quartile of the values of the continuous variable and is a more robust measure of spread.

Excel doesn't have a function to calculate IQR but it can be calculated with the `QUARTILE()` function. Using the student scores example:

```
=QUARTILE(B2:B11, 3) - QUARTILE(B2:B11, 1)
```

which gives **13.75**.

14.1.3 Distribution

Assessing distribution using boxplots

A **boxplot**, also referred to as a **box-and-whisker plot**, is a graphical tool used to summarise the distribution of a continuous variable. It provides insights into key aspects such as the median, quartiles, and any potential outliers in a clear and concise manner.

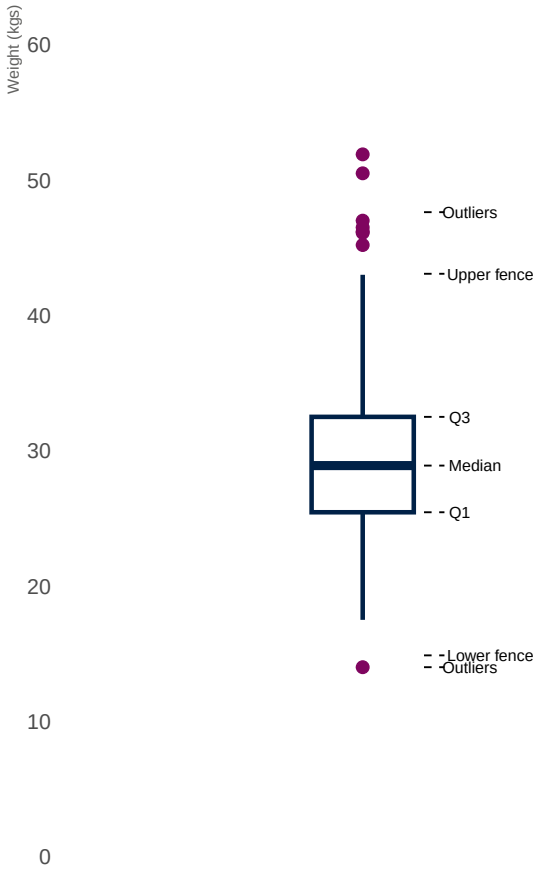


Figure 14.2: Boxplot of weight of children 10-15 years of age

Key Components of a Boxplot

1. **The Box:** The box itself illustrates the IQR, which spans the middle 50% of the data. The lower edge of the box represents the first quartile (25th percentile), while the upper edge marks the third quartile (75th percentile).
2. **The Median Line:** A vertical line within the box indicates the median, or the 50th percentile, which divides the dataset into two equal halves.
3. **The Whiskers:** Lines extending from the sides of the box, often referred to as *whiskers*, show the range of the data beyond the IQR. This is based on 1.5 times the IQR value. The lower whisker is measured out as a distance 1.5 times IQR below the lower edge of the box (lower quartile) while the upper whisker is measured out as a distance 1.5 times IQR above the upper edge of the box (upper quartile).
4. **Outliers:** Data points that lie outside the whiskers are considered outliers and are usually plotted as individual points separate from the main plot.

Plotting boxplots

14.2 Excel 2016 and later

Starting with Excel 2016, Microsoft has included a Box and Whiskers chart capability to Excel.

1. Go to `Insert` –> `Insert Statistics Chart` –> `Box and Whisker`
2. Select base chart –> `Chart Design` –> `Select Data`
3. Edit range of values to use for boxplot.
4. Specify range of weight values for boxplot.

The data is in Sheet 1 and found in range `E2:E268`.

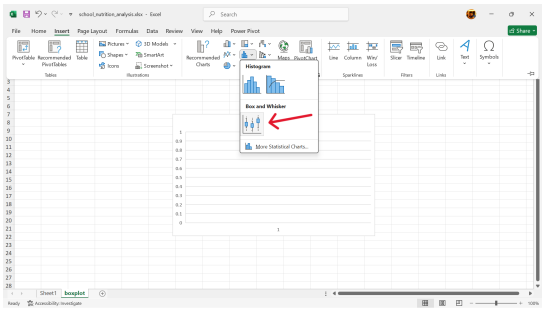


Figure 14.3: Insert Box and Whisker chart

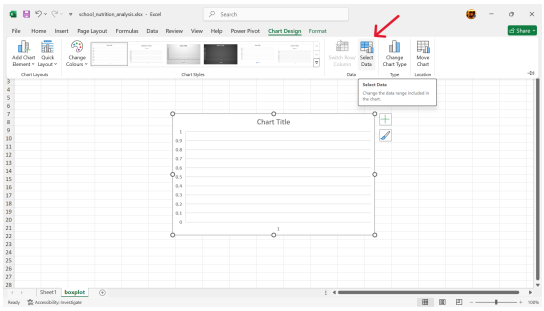


Figure 14.4: Select data to use for boxplot

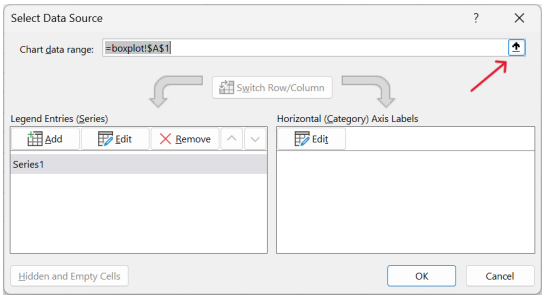


Figure 14.5: Edit range of values to use for boxplot

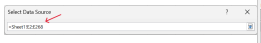


Figure 14.6: Specify range of weight values for box-plot

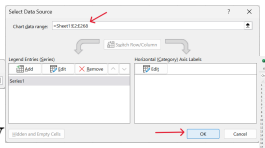


Figure 14.7: Click OK to confirm range of weight values

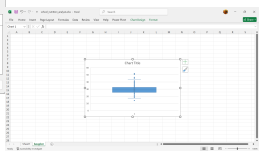


Figure 14.8: Box-plot is now created

14.3 Pre-Excel 2016

Before Excel 2016 version, there was no boxplot chart type available but boxplots can be made through the following steps:

1. Calculate the lower whisker, quartile 1, median, quartile 3, and upper whisker summary values of the weight variable.

2. Create a base stacked column chart.

`Insert` –> `Insert Column or Bar Chart` –> `Stacked Column`

3. Select data to use for stacked column chart.

Select base chart –> `Chart Design` –> `Select Data`

4. Edit range of values to use for stacked column chart.

5. Reverse stacked column chart axis.

Click on `Switch Row/Column`

6. Hide the lower column of the stacked column chart.

Click on the lower column of the stacked column chart. Format the data series by removing its fill.

7. Hide the upper column of the stacked column chart.

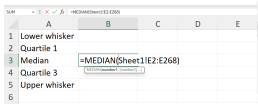


Figure 14.9: Calculate the median weight

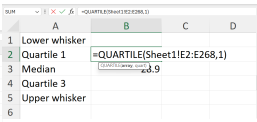


Figure 14.10: Calculate quartile 1 of weight

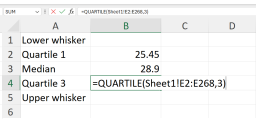


Figure 14.11: Calculate quartile 3 of weight

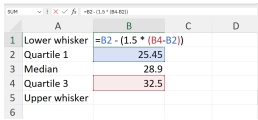


Figure 14.12: Calculate the lower whisker

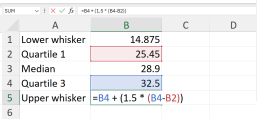


Figure 14.13: Calculate the upper whisker

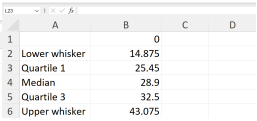


Figure 14.14: Add a starting point value of 0

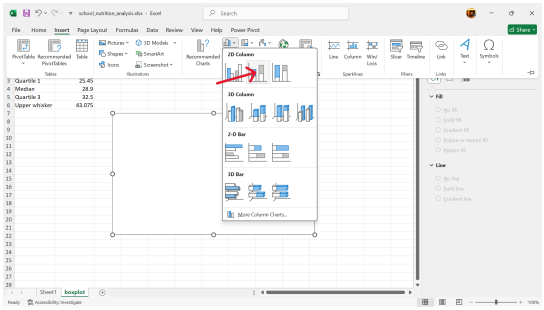


Figure 14.15: Create a base stacked column chart

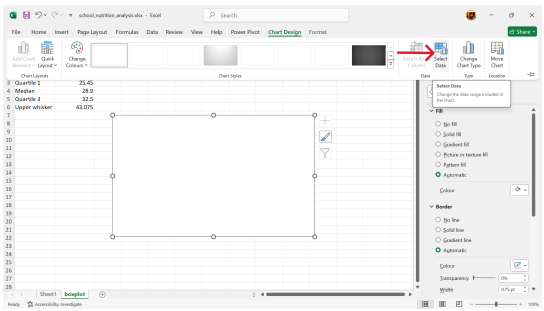


Figure 14.16: Select data to use for stacked column chart

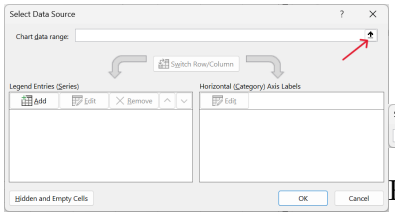


Figure 14.17: Tap to select range of values to use for stacked column chart

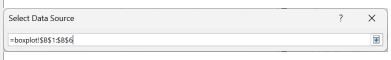


Figure 14.18: Specify range of values to use for stacked column chart

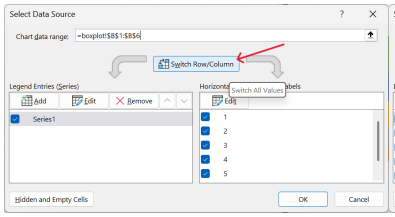


Figure 14.19: Click on Switch Row/Column

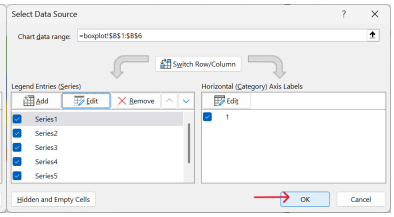


Figure 14.20: Confirm chart settings

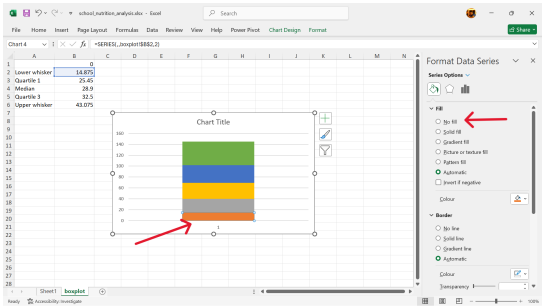


Figure 14.21: Hide lower column of the stacked column chart

Click on the upper column of the stacked column chart. Format the data series by removing its fill.

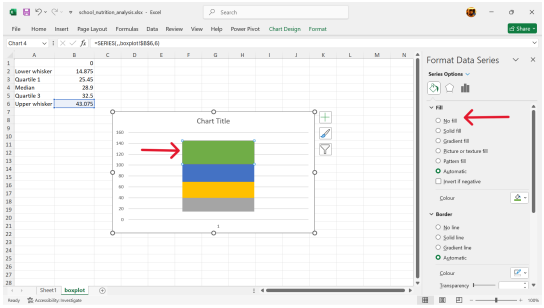


Figure 14.22: Hide upper column of the stacked column chart

8. Add error bars to replace the upper column.

Click on `Chart Elements` –> `Error Bars` –> `More Options`

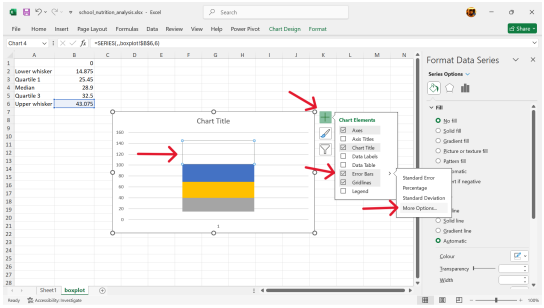


Figure 14.23: Add error bars to replace the upper column

9. Format error bars to create the upper whisker.

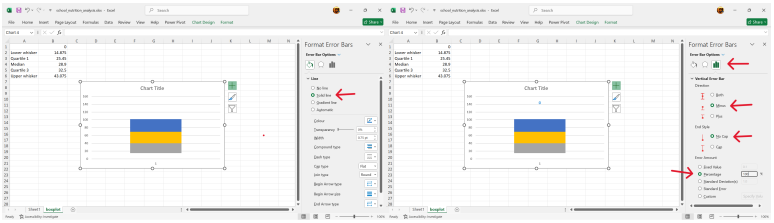


Figure 14.24: Set error bars to solid line

Figure 14.25: Format error bars

10. Hide the second lower column of the stacked column chart.

Click on the second lower column of the stacked column chart. Format the data series by removing its fill.

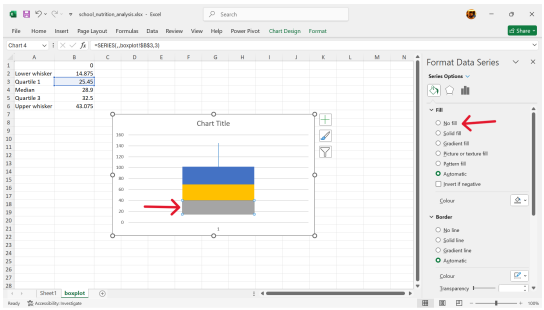


Figure 14.26: Hide second lower column of the stacked column chart

11. Add error bars to replace the second lower column.

Click on `Chart Elements` -> `Error Bars` -> `More Options`

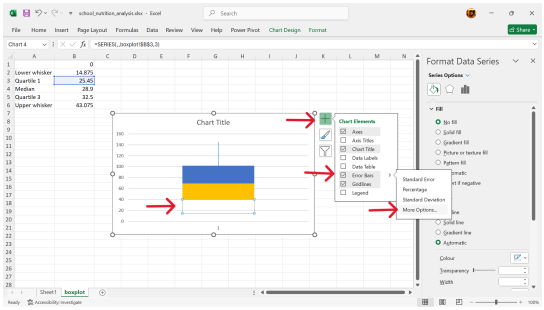


Figure 14.27: Add error bars to replace the second lower column

12. Format error bars to create the lower whisker.

13. Change fill and outline colours of the box.

14. Boxplot is now created.

*Based on Microsoft Support documentation

Interpreting boxplots

The median, represented by the line within the box, indicates the central value of the data. The height of the box and the

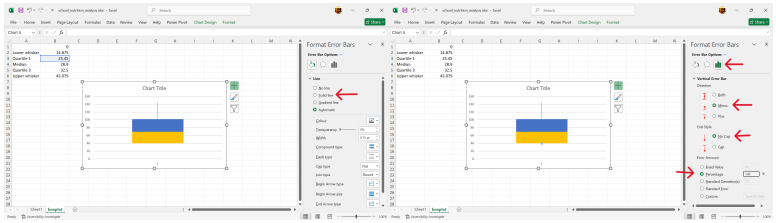


Figure 14.28: Set error bars to solid line

Figure 14.29: Format error bars

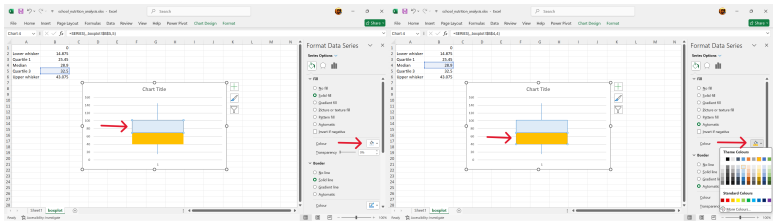


Figure 14.30: Change fill and outline colours of the upper box

Figure 14.31: Change fill and outline colours of the lower box

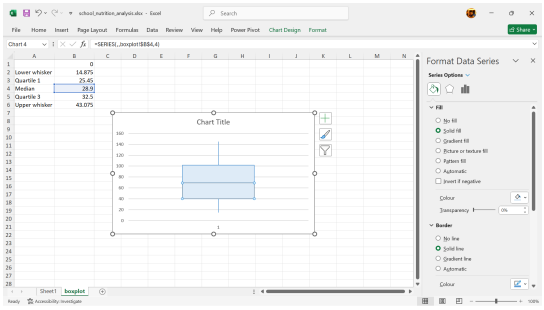


Figure 14.32: Boxplot is now created

length of the whiskers indicate the spread and variability of the data. A wider box or longer whiskers suggest greater variability. The position of the median line within the box can suggest whether the data is symmetrical, skewed left (negative skew), or skewed right (positive skew). Outliers represent data points that are significantly different from the rest of the data and may warrant further investigation. Comparing box plots for different datasets can help reveal differences in their central tendency, spread, and distribution.

Assessing distribution using histograms

A histogram is a graphical representation that displays the distribution of numerical data. It uses adjacent bars to show the frequency or count of data points within specified ranges, known as bins. Histograms provide insights into the data distribution, such as whether it is normally distributed (bell-shaped), skewed, or has multiple peaks (multimodal). They help identify patterns like clusters, gaps, and outliers in the data.

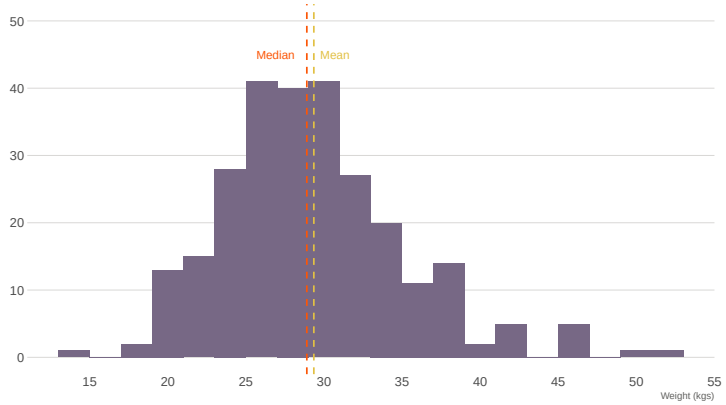


Figure 14.33: Histogram of weight of children 10-15 years of age

Plotting the histogram

On a graph, the x-axis (horizontal) represents the bins or intervals of the numerical data. The y-axis (vertical) represents the frequency or count of data points in each bin. Bars

are drawn adjacent to each other without gaps, as the bins represent continuous ranges.

In Excel, the histogram can be plotted as follows:

1. Create a base histogram plot.

Go to `Insert` –> `Insert Statistics Chart` –> `Histogram`

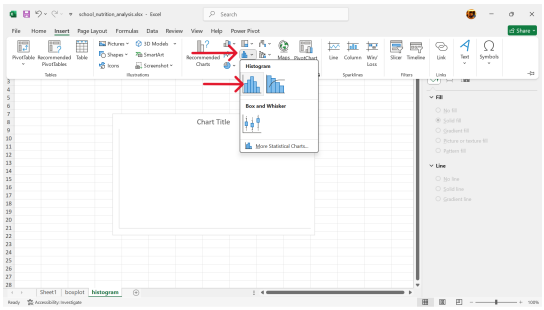


Figure 14.34: Create a base histogram plot

2. Select data to use for histogram.

Select base chart –> `Chart Design` –> `Select Data`

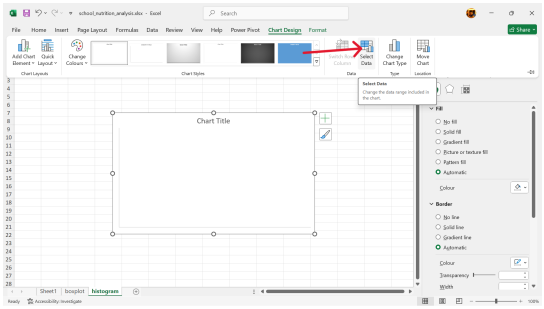


Figure 14.35: Select data to use for histogram

3. Edit range of values to use for histogram.

4. Confirm range of values to use for histogram.

Click on `OK`

5. Histogram is now created.

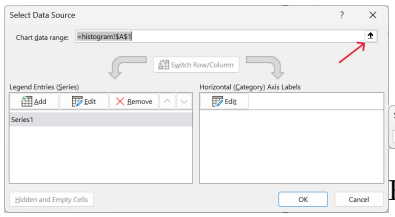


Figure 14.36: Tap to select range of values to use for histogram

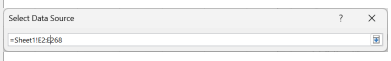


Figure 14.37: Specify range of values to use for histogram

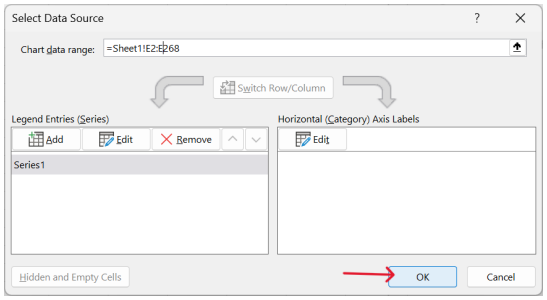


Figure 14.38: Confirm range of values to use for histogram

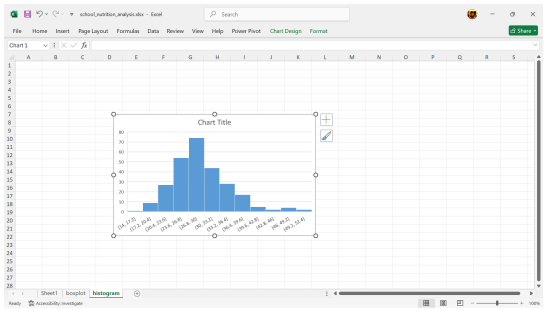


Figure 14.39: Histogram is now created

Interpretation

Histograms provide insights into the data distribution, such as whether it is normally distributed (bell-shaped), skewed, or has multiple peaks (multimodal). Histograms help identify patterns like clusters, gaps, and outliers in the data.

Assessing distribution using QQ plots

A QQ plot, or Quantile-Quantile plot, is a graphical tool used to assess whether a dataset follows a specific theoretical distribution (like a normal distribution) or to compare the distributions of two datasets. It does this by plotting the quantiles of the observed data against the quantiles of the theoretical distribution (or the other dataset).

Plotting QQ plots to test for normality

There is no built-in chart for QQ plots in Excel. To create a QQ plot to test normal distribution in Excel, we recommend creating a new worksheet and then importing the raw dataset into this worksheet before proceeding with the steps below. We use the school nutrition dataset for this demonstration.

1. Sort weight values in ascending order.

We recommend doing this step on the worksheet where the data is and create a new variable for the sorted values of weight. The values can be sorted using the `SORT()` function in Excel as follows (Figure 14.40):

```
=SORT(E2:E268)
```

This function will fill the new column/variable with the weight values sorted in ascending order without changing the order of the reference/original data. We recommend doing it this way instead of sorting the whole dataset based on weight so that you can perform the same operation for creating QQ plots for other appropriate variable in the dataset (i.e., height).

2. Give a rank to each value from 1 to the total number of rows.

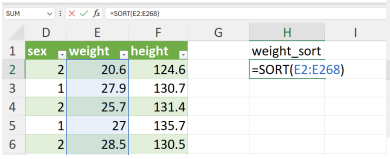


Figure 14.40: Sort weight values in ascending order

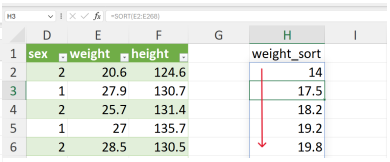


Figure 14.41: Sorted weight values in ascending order in new column

This should be created in a new column/variable. This can be done easily using the following formula (Figure 14.42):

```
=ROW() - 1
```

Copy or drag this formula from the first cell to the following cells in the new column/variable to get the rank for all weight records (Figure 14.43).

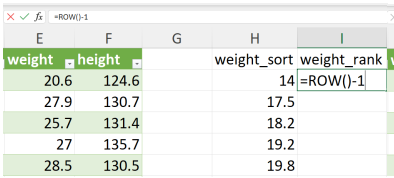


Figure 14.42: Formula for assigning a rank to the sorted weight records

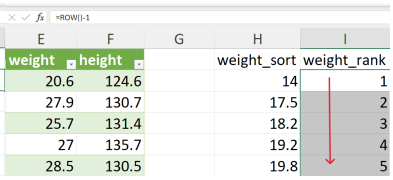


Figure 14.43: Copy or drag the formula to the rest of the rows

3. Calculate the empirical/observed cumulative probabilities.

The empirical cumulative probabilities are calculated as follows:

$$F(x) = \frac{rank - 0.5}{n}$$

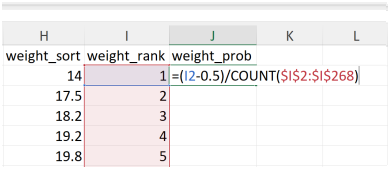
where:

$rank$ = rank of the sorted value of the variable

n = number of records/values of the variable

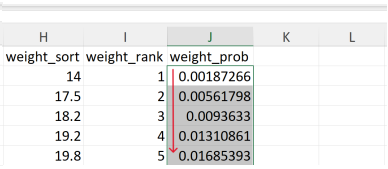
In Excel, it can be calculated as follows (Figure 14.44):

```
=(I2-0.5)/COUNT($I$2:$I$268)
```



	H	I	J	K	L
1	weight_sort	weight_rank	weight_prob		
2	14	1	=(I2-0.5)/COUNT($I$2:$I$268)		
3	17.5	2			
4	18.2	3			
5	19.2	4			
6	19.8	5			

Figure 14.44: Calculate the empirical cumulative distribution



	H	I	J	K	L
1	weight_sort	weight_rank	weight_prob		
2	14	1	0.00187266		
3	17.5	2	0.00561798		
4	18.2	3	0.0093633		
5	19.2	4	0.01310861		
6	19.8	5	0.01685393		

Figure 14.45: Copy or drag the formula to the rest of the rows

4. Calculate the theoretical quantiles.

The theoretical quantiles for each of the empirical probabilities are calculated as follows:

$$q = F^{-1} \times p$$

where:

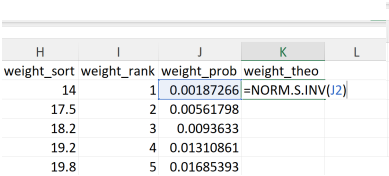
q = theoretical quantile

F^{-1} = inverse of the cumulative distribution function

p = cumulative probability

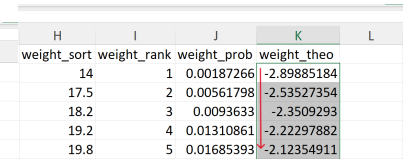
In Excel, it can be calculated as follows (Figure 14.46):

```
=NORM.S.INV(J2)
```



	H	I	J	K	L
1	weight_sort	weight_rank	weight_prob	weight_theo	
2	14	1	0.00187266	=NORM.S.INV(J2)	
3	17.5	2	0.00561798		
4	18.2	3	0.0093633		
5	19.2	4	0.01310861		
6	19.8	5	0.01685393		

Figure 14.46: Calculate the theortetical quantiles



	H	I	J	K	L
1	weight_sort	weight_rank	weight_prob	weight_theo	
2	14	1	0.00187266	-2.89885184	
3	17.5	2	0.00561798	-2.53527354	
4	18.2	3	0.0093633	-2.3509293	
5	19.2	4	0.01310861	-2.22297882	
6	19.8	5	0.01685393	-2.12354911	

Figure 14.47: Copy or drag the formula to the rest of the rows

5. Calculate the slope and intercept of the line of agreement (QQ line).

In a QQ (quantile-quantile) plot, the *line of agreement*, also known as the 45-degree line or reference line, represents the expected relationship if two datasets being compared have identical distributions. If the points on the QQ plot fall close to this line, it suggests the distributions are similar. Deviations from the line indicate differences in the distributions.

We recommend creating a new worksheet for these calculations (Figure 14.48) and then creating layout in which the various values can be calculated as shown in Figure 14.49.

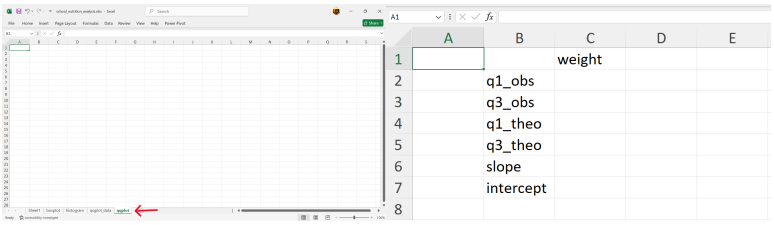


Figure 14.48: Create new worksheet for QQ line calculations

Figure 14.49: Create a layout for the calculations of the various values

a. Calculate the empirical 25% and 75% quantiles of the weight values (Figure 14.50).

In Excel, these can be calculated as follows:

```
=QUARTILE(qqplot_data!E2:E268,1)      ## 25% quantile
=QUARTILE(qqplot_data!E2:E268,3)      ## 75% quantile
```

b. Calculate the theoretical 25% and 75% quantile of the normal distribution (Figure 14.52).

In Excel, these can be calculated as follows:

```
=NORM.INV(0.25,0,1)      ## 25% quantile
=NORM.INV(0.75,0,1)      ## 75% quantile
```

c. Calculate the slope and intercept of the QQ line.

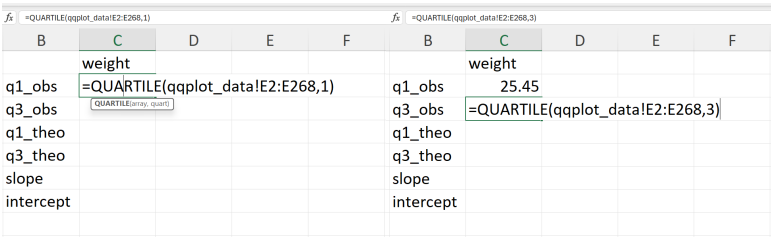


Figure 14.50: Calculate the empirical 25% quantile

Figure 14.51: Calculate the empirical 75% quantile

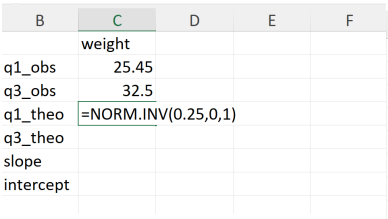


Figure 14.52: Calculate the theoretical 25% quantile of the normal distribution

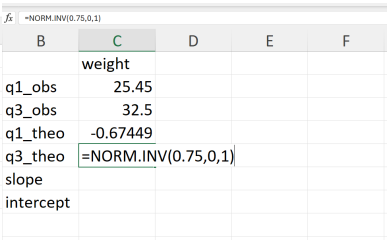


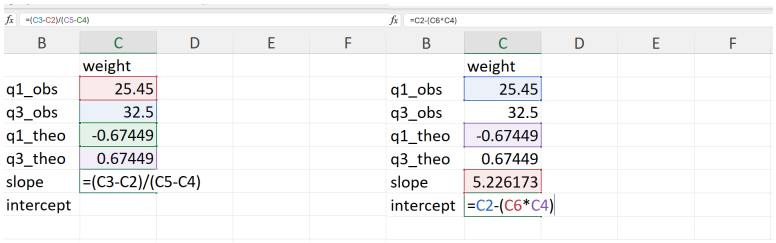
Figure 14.53: Calculate the theoretical 75% quantile of the normal distribution

The slope of the QQ line is calculated as follows:

$$slope = \frac{q75_{observed} - q25_{observed}}{q75_{theoretical} - q25_{theoretical}}$$

The intercept of the QQ line is calculated as follows:

$$intercept = q25_{observed} - slope \times q25_{theoretical}$$



	B	C	D	E	F
		weight			
2	q1_obs	25.45			
3	q3_obs	32.5			
4	q1_theo	-0.67449			
5	q3_theo	0.67449			
6	slope	=(C3-C2)/(C5-C4)			
7	intercept				

Figure 14.54: Calculate the slope of the QQ line

	B	C	D	E	F
		weight			
2	q1_obs	25.45			
3	q3_obs	32.5			
4	q1_theo	-0.67449			
5	q3_theo	0.67449			
6	slope	5.226173			
7	intercept	=C2-(C6*C4)			

Figure 14.55: Calculate the intercept of the QQ line

6. Create a base scatter plot.

In the same worksheet as the calculations, create a base scatter plot as follows (Figure 14.56):

`Insert` –> `Charts` –> `Insert Scatter (x, Y) or Bubble Chart` –> `Scatter`

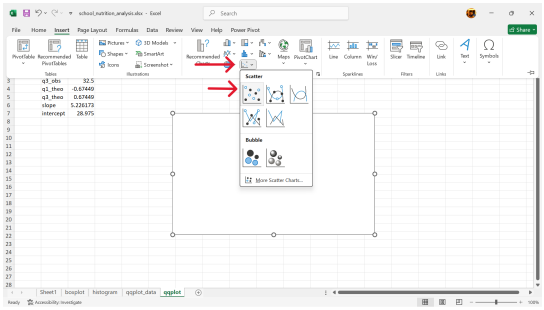


Figure 14.56: Create a base scatter plot

7. Select data to use for scatter plot.

Select base chart –> `Chart Design` –> `Select Data`

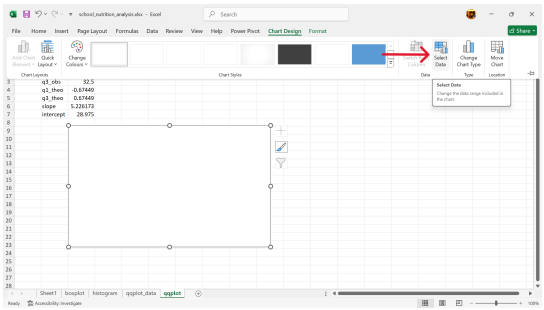


Figure 14.57: Select data to use for scatter plot

8. Add a data series to base scatter plot.

Click on `Add` (Figure 14.58)

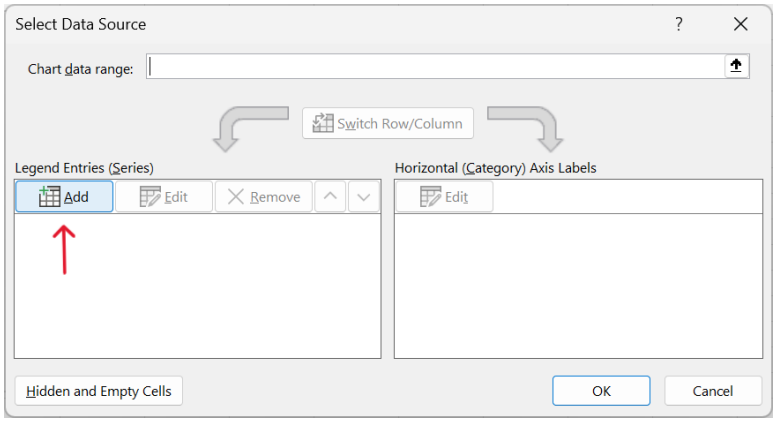


Figure 14.58: Add a data series to base scatter plot

9. Specify values for x-axis of the scatter plot.

The x-axis should use the range of values for the theoretical probabilities.

10. Specify values for y-axis of the scatter plot.

The y-axis should use the range of values for the ordered values for weight.

11. Confirm selection of values for scatter plot.

12. Add the QQ line to the QQ plot.

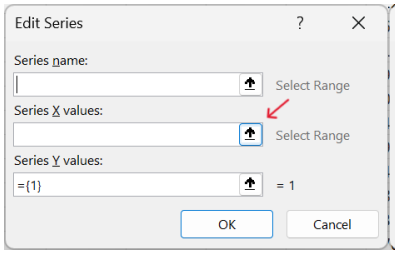


Figure 14.59: Select x-axis for editing

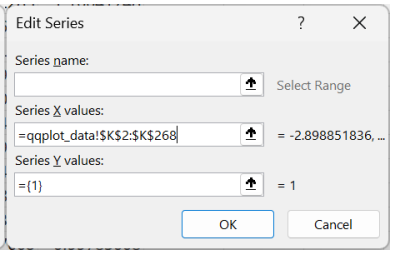


Figure 14.60: Select range of values for x-axis

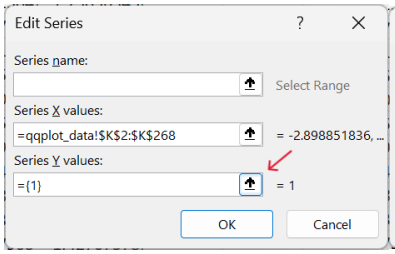


Figure 14.61: Select y-axis for editing

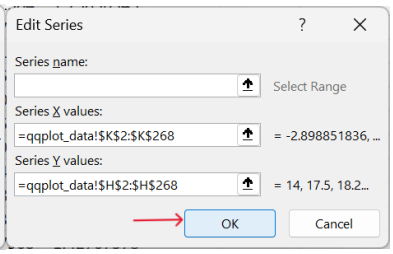


Figure 14.62: Select range of values for y-axis

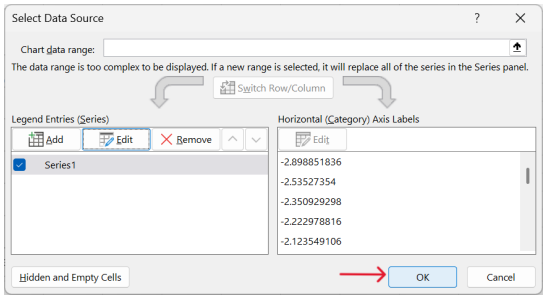


Figure 14.63: Confirm selection of values for scatter plot

`Chart Elements` –> `Trendline` –> `More Options...`

Select `Linear` for trendline and tick `Set Intercept` and then input the value for intercept that we calculated earlier (Figure 14.65).

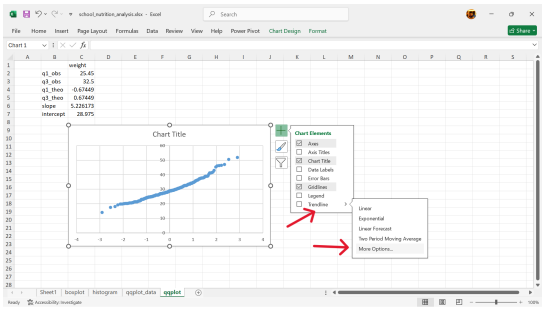


Figure 14.64: Add trendline to scatter plot

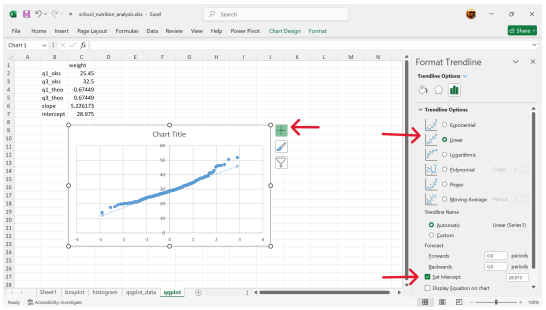


Figure 14.65: Edit settings for trendline

Interpreting the plot

If the data follows the theoretical distribution, the points in the QQ plot will roughly form a straight diagonal line. Deviations from this line indicate that the data does not follow the theoretical distribution, and the nature of the deviations can provide insights into how the data differs from the expected distribution (e.g., skewness, outliers). In simpler terms: A QQ plot is a visual check to see if your data "looks like" it came from a specific distribution. It helps you determine if your data is normally distributed, or if it has some other pattern.

14.3.1 Age Heaping

Age heaping is the tendency to report children's ages to the nearest year or adults' ages to the nearest multiple of five or ten years. Age heaping is very common that is why most reported national statistics use broad age groups.

Testing for age heaping

There is no built-in function in Excel that tests for age heaping. The following steps can be performed in Excel to test whether there is significant age heaping in a dataset with age values. These steps use the school nutrition data for demonstration. We recommend that these steps be done in a separate worksheet from the raw data to avoid contamination of original dataset.

1. Determine an appropriate divisor.

The school nutrition data records age in months. A useful way of looking at age heaping when age is recorded in months is to examine the remainders when the ages are divided by 12. So, we set the divisor to 12.

2. Create a new variable for the remainder when age variable is divided by 12.

Create a new worksheet containing the raw dataset and create a new variable as shown in Figure 14.66.

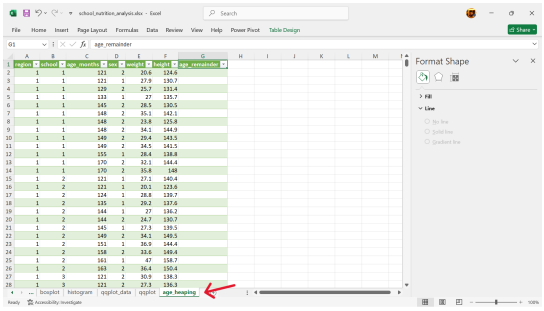
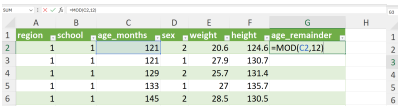


Figure 14.66: Create a new variable for the remainder value

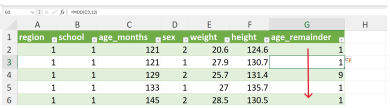
Apply the `modulo` operator to the age variable using 12 as the divisor. This can be done using the `MOD()` function is Excel as follows (Figure 14.67):

```
=MOD(C2,12)
```



region	school	age_months	sex	weight	height	age_remainder
1	1	121	2	20.6	124.6	=MOD(C2,12)
1	1	121	1	27.9	130.7	
1	1	129	2	25.7	131.4	
1	1	133	1	27	135.7	
1	1	145	2	28.5	130.5	

Figure 14.67: Get the remainder value when age is divided by 12



region	school	age_months	sex	weight	height	age_remainder
1	1	121	2	20.6	124.6	1
1	1	121	1	27.9	130.7	1
1	1	129	2	25.7	131.4	9
1	1	133	1	27	135.7	1
1	1	145	2	28.5	130.5	1

Figure 14.68: Copy or drag the formula to the rest of the rows

3. Create a summary table of the counts per remainder values.

a. In a separate worksheet (see Figure 14.69), create a summary table using Excel's pivot table functionality.

`Insert` –> `Pivot Table` –> `From table/range`

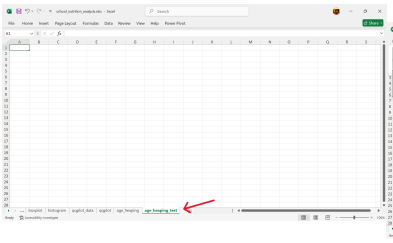


Figure 14.69: Create a new worksheet for the summary table

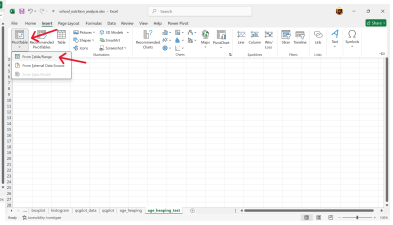


Figure 14.70: Initiate pivot table functionality in Excel

b. Select the range of data to summarise via pivot table and insert this pivot table into the new worksheet.

c. Select the rows of the summary table.

d. Select the values of the summary table.

e. The summary table is now completed.

4. Calculate the expected counts of the remainder values if there was no age heaping.

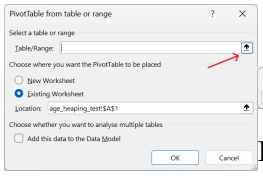


Figure 14.71: Select the range of data to summarise

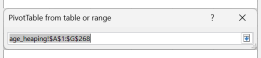


Figure 14.72: Select the range of data to summarise

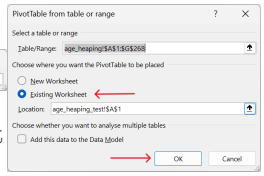


Figure 14.73: Select to insert into new worksheet

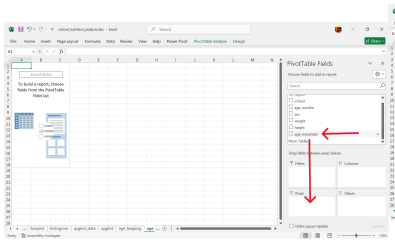


Figure 14.74: Select the variable for the rows of the summary table

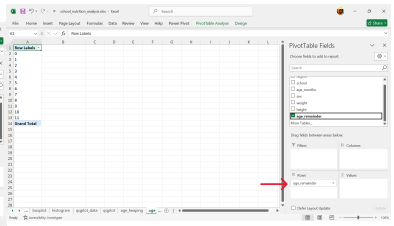


Figure 14.75: Drag the variable for the rows to the row option of the summary table

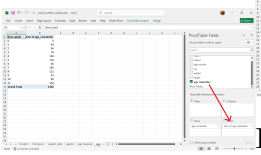


Figure 14.76: Select the values of the summary table

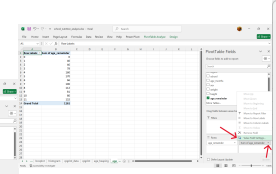


Figure 14.77: Select the value field settings of the summary table

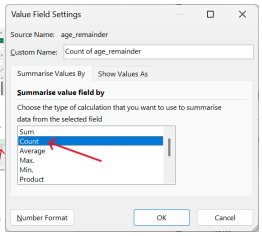


Figure 14.78: Select the appropriate summary measure

	A	B	C	D	E
1	**Row Labels**	**Count of age_remainder**			
2	0	22			
3	1	40			
4	2	23			
5	3	26			
6	4	40			
7	5	34			
8	6	11			
9	7	28			
10	8	14			
11	9	9			
12	10	8			
13	11	12			
14	**Grand Total**	**267**			

Figure 14.79: Summary table completed

The expected counts can be calculated as follows:

$$\text{expected counts} \;=\; \frac{n}{d-1}$$

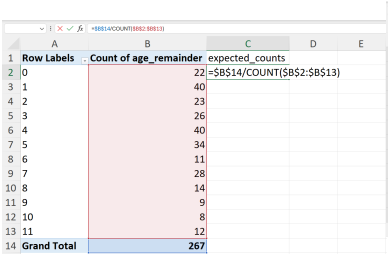
where:

n = number of records

d = divisor

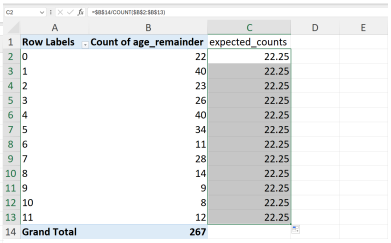
In Excel, this can be calculated using the following formula (Figure 14.80):

```
=$B$14/COUNT($B$2:$B$13)
```



	A	B	C	D	E
1	Row Labels	Count of age_remainder	expected_counts		
2	0	22	=$B$14/COUNT($B$2:$B$13)		
3	1	40			
4	2	23			
5	3	26			
6	4	40			
7	5	34			
8	6	11			
9	7	28			
10	8	14			
11	9	9			
12	10	8			
13	11	12			
14	Grand Total	267			

Figure 14.80: Calculate expected counts if no age heaping



	A	B	C	D	E
1	Row Labels	Count of age_remainder	expected_counts		
2	0	22	22.25		
3	1	40	22.25		
4	2	23	22.25		
5	3	26	22.25		
6	4	40	22.25		
7	5	34	22.25		
8	6	11	22.25		
9	7	28	22.25		
10	8	14	22.25		
11	9	9	22.25		
12	10	8	22.25		
13	11	12	22.25		
14	Grand Total	267			

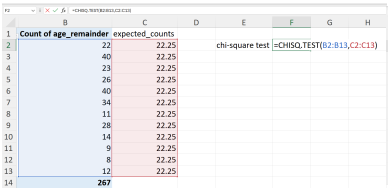
Figure 14.81: Copy or drag the formula to the rest of the rows

5. Perform a chi-square test on the actual vs expected remainder counts.

A chi-square test is performed to test whether the actual remainder values are significantly different from the expected remainder values. This test can be performed in Excel as follows (Figure 14.82):

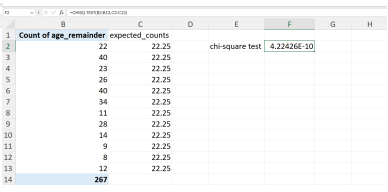
```
=CHISQ.TEST(B2:B13,C2:C13)
```

The resulting value when the `CHISQ.TEST()` is performed in Excel is the `p-value` of the chi-square test. A `p-value` of less than 0.05 indicates that there is a significant difference between the actual and expected remainder values for the age in the school nutrition dataset. This points to significant age heaping in the dataset.



	B	C	D	E	F	G	H
1	Count of age_remainder	expected_counts					
2	22	22.25		chi-square test	=CHISQ.TEST(B2:B13,C2:C13)		
3	40	22.25					
4	23	22.25					
5	26	22.25					
6	40	22.25					
7	34	22.25					
8	11	22.25					
9	28	22.25					
10	14	22.25					
11	9	22.25					
12	8	22.25					
13	12	22.25					
14	267						

Figure 14.82: Perform chi-square test in Excel



	B	C	D	E	F	G	H
1	Count of age_remainder	expected_counts					
2	22	22.25		chi-square test	4.22426E-10		
3	40	22.25					
4	23	22.25					
5	26	22.25					
6	40	22.25					
7	34	22.25					
8	11	22.25					
9	28	22.25					
10	14	22.25					
11	9	22.25					
12	8	22.25					
13	12	22.25					
14	267						

Figure 14.83: Result is p-value of chi-square test

14.3.2 Digit preference

Digit preference is the observation that the final number in a measurement occurs with a greater frequency than is expected by chance. This can occur because of rounding, the practice of increasing or decreasing the value in a measurement to the nearest whole or half unit, or because data are made up.

Testing for digit preference

There is no built-in function in Excel that tests for digit preference. The following steps can be performed in Excel to test whether there is significant digit preference in a continuous variable. These steps use the weight variable in the school nutrition data for demonstration. We recommend that these steps be done in a separate worksheet from the raw data to avoid contamination of original dataset.

1. Create a new variable for the last digit of the weight variable.

Create a new worksheet and then create a new variable for the last digit of the weight variable as shown in Figure 14.84.

The last digit of the weight variable can be extracted in Excel using the `RIGHT()` function as follows (see Figure 14.86):

```
=RIGHT(E2,1)
```

2. Create a summary table of the counts of the last digits.

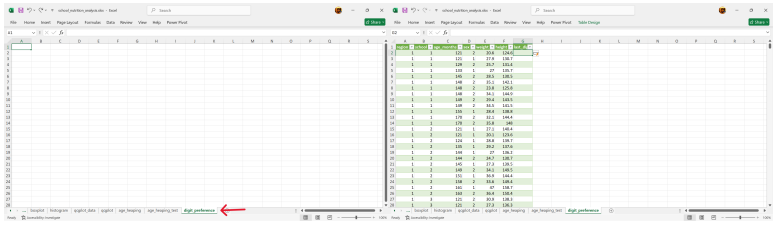


Figure 14.84: Create a new worksheet for the digit preference testing

Figure 14.85: Create a new variable for the last digit of the weight variable

E	F	G	H	I
weight	height	last_digit		
20.6	124.6	=RIGHT(E2,1)		
27.9	130.7			
25.7	131.4			
27	135.7			
28.5	130.5			

Figure 14.86: Get the last digit of the weight variable

E	F	G	H
weight	height	last_digit	
20.6	124.6	6	
27.9	130.7	9	
25.7	131.4	7	
27	135.7	7	
28.5	130.5	5	

Figure 14.87: Copy or drag the formula to the rest of the rows

Using pivot table in Excel, create a summary table of the counts of the last digits. We recommend that the summary table be inserted into a new worksheet.

a. `Insert` –> `Pivot Table` –> `From Table/Range`

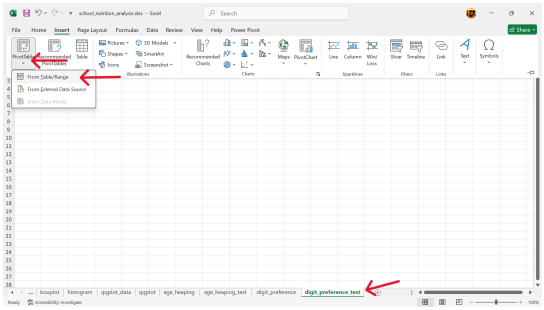


Figure 14.88: Initiate pivot table functionality in Excel

b. Select the range of data to summarise via pivot table and insert this pivot table into the new worksheet.

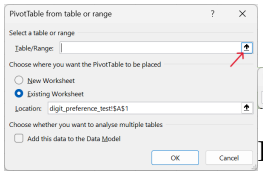


Figure 14.89: Select the range of data to summarise

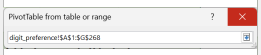


Figure 14.90: Select the range of data to summarise

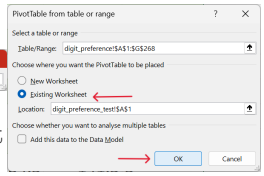


Figure 14.91: Select to insert into new worksheet

c. Select the rows of the summary table.

d. Select the values of the summary table.

3. Calculate the expected counts of the last digits if there was no digit preference.

This can be calculated as:

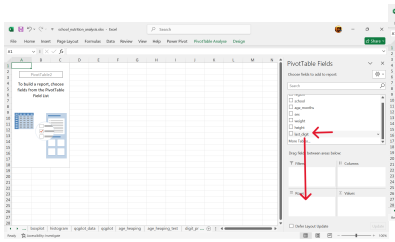


Figure 14.92: Select the variable for the rows of the summary table

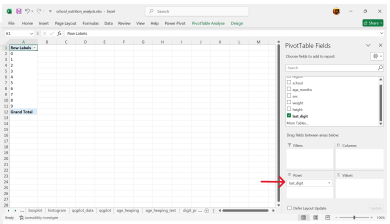


Figure 14.93: Drag the variable for the rows to the row option of the summary table

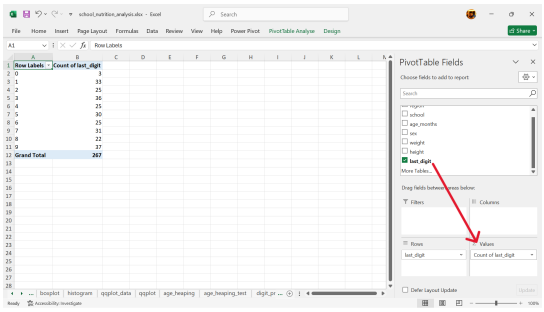


Figure 14.94: Select the values of the summary table

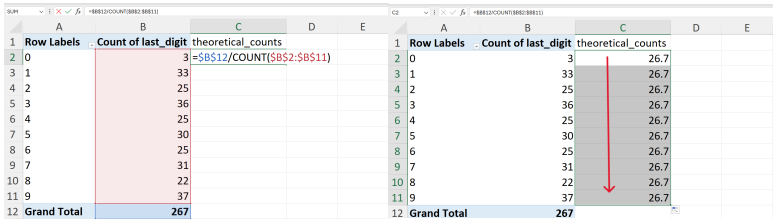
$$\text{expected counts} = \frac{n}{10}$$

where:

n = number of records

In Excel, this can be calculated as follows (Figure 14.95):

```
=B12/10
```



	A	B	C	D	E
1	Row Labels	Count of last_digit	theoretical_counts		
2	0	3	=$B$12/COUNT($B$2:$B$11)		
3	1	33			
4	2	25			
5	3	36			
6	4	25			
7	5	30			
8	6	25			
9	7	31			
10	8	22			
11	9	37			
12	Grand Total	267			

Figure 14.95: Calculate expected counts if no digit preference

	A	B	C	D	E
1	Row Labels	Count of last_digit	theoretical_counts		
2	0	3	26.7		
3	1	33	26.7		
4	2	25	26.7		
5	3	36	26.7		
6	4	25	26.7		
7	5	30	26.7		
8	6	25	26.7		
9	7	31	26.7		
10	8	22	26.7		
11	9	37	26.7		
12	Grand Total	267			

Figure 14.96: Copy or drag the formula to the rest of the rows

4. Calculate the chi-square (χ^2) statistic.

The formula to calculate the chi-square (χ^2) statistic is:

$$\chi^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}$$

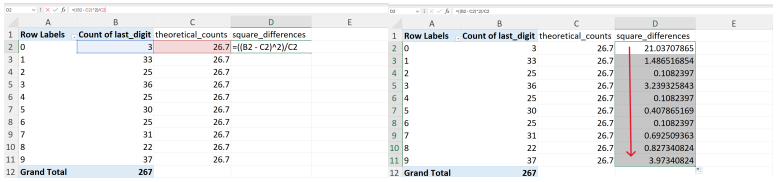
In Excel, this can be calculated by first calculating the square of per digit difference in observed and expected counts divided by the expected counts (Figure 14.97).

```
=((B2-C2) ^ 2) / C2
```

Then, these differences are summed (Figure 14.99).

```
=SUM(D2:D11)
```

5. Calculate the digit preference score (DPS).



	A	B	C	D	E
1	Row Labels	Count of last_digit	theoretical_counts	square_differences	
2	0	3	26.7	=((B2 - C2)^2)/C2	
3	1	33	26.7		
4	2	25	26.7		
5	3	36	26.7		
6	4	25	26.7		
7	5	30	26.7		
8	6	25	26.7		
9	7	31	26.7		
10	8	22	26.7		
11	9	37	26.7		
12	Grand Total	267			

Figure 14.97: Calculate the difference in observed and expected counts

	A	B	C	D	E
1	Row Labels	Count of last_digit	theoretical_counts	square_differences	
2	0	3	26.7	21.03707865	
3	1	33	26.7	1.486516854	
4	2	25	26.7	0.1082397	
5	3	36	26.7	3.239325843	
6	4	25	26.7	0.1082397	
7	5	30	26.7	0.407865169	
8	6	25	26.7	0.1082397	
9	7	31	26.7	0.692509363	
10	8	22	26.7	0.827340824	
11	9	37	26.7	3.97340824	
12	Grand Total	267			

Figure 14.98: Copy or drag the formula to the rest of the rows

D	E	F	G
square_differences			
21.03707865	chi-square statistic	=SUM(D2:D12)	
1.486516854			
0.1082397			
3.239325843			
0.1082397			
0.407865169			
0.1082397			
0.692509363			
0.827340824			
3.97340824			

Figure 14.99: Sum the square of per digit difference in observed and expected counts divided by expected counts

The digit preference score (DPS) is a summary measure that reduces the bias in digit preference testing because of sample size. The DPS takes into account sample size as shown in this formula:

$$DPS = \sqrt{\frac{\chi^2}{\sum_{i=1}^{n} O_i \times (n_{digits} - 1)}} * 100$$

In Excel, this can be calculated as follows (Figure 14.100):

```
=SQRT(F2/(SUM(B2:B11)*(COUNT($B$2:$B$11)-1)))*100
```

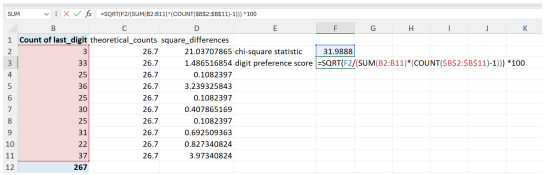


Figure 14.100: Calculate the digit preference score

Interpreting the digit preference score

The following table shows how to interpret the DPS.

DPS	Classification
< 8	Excellent
from 8 to < 12	Good
from 12 to < 20	Acceptable
20 or higher	Problematic

14.4 Categorical variables

Categorical variables represent attributes or categories instead of numerical values. They organise data into specific groups or labels, and each observation is placed into one group. Categorical variables don't have an inherent numerical order or ranking. They consist of a finite number of distinct categories or groups.

Categorical variables can be further classified as either *nominal*, *ordinal*, or *binary*.

- Nominal - categories with no inherent order such as colours or types of fruit.
- Ordinal - categories with a meaningful order such as education level - primary, secondary, college.
- Binary - a special case of categorical variable with only two categories such as yes or no.

14.4.1 Some considerations when dealing with categorical variables

Use the inherent order of ordinal variables

Order nominal variables meaningfully

Use colours for categories appropriately

15 Bivariate statistics

Bivariate means "involving two variables" in statistics. It's a method used to analyze relationships between two variables, studying how their values might connect or influence each other.

One crucial element of bivariate analysis is evaluating the correlation between two variables, which can be positive (both rise together), negative (one rises while the other falls), or non-existent (no apparent connection).

Bivariate data is often visualised using scatter plots (see Section 15.1), which can help reveal patterns or trends between the two variables.

15.1 Scatter plots

> Scatter plots are glorious. Of all the major chart types, they are by far the most powerful. They allow us to quickly understand relationships that would be nearly impossible to recognize in a table or a different type of chart… Michael Friendly and Daniel Denis, psychologists and historians of graphics, call the scatter plot the most "generally useful invention in the history of statistical graphics."

Dan Kopf

Scatter plots are graphs that show how two numerical datasets relate. Each data point is represented by a dot, positioned based on the values of the two variables being compared. They effectively demonstrate both the strength and direction of connections between these variables.

Scatter plots primarily help visualise and determine potential relationships or correlations between two variables. They

reveal if the variables trend upward together (positive correlation), downward together (negative correlation), or exhibit no noticeable connection.

A scatter plot features two axes—the horizontal *x-axis* and vertical *y-axis*. Data points are represented as dots, each positioned according to their respective x and y values.

Examining the pattern of dots in a scatter plot offers insights into the relationship between variables. For example, dots forming an upward-sloping line indicate a positive correlation, whereas a downward slope suggests a negative correlation.

15.1.1 Creating scatter plots

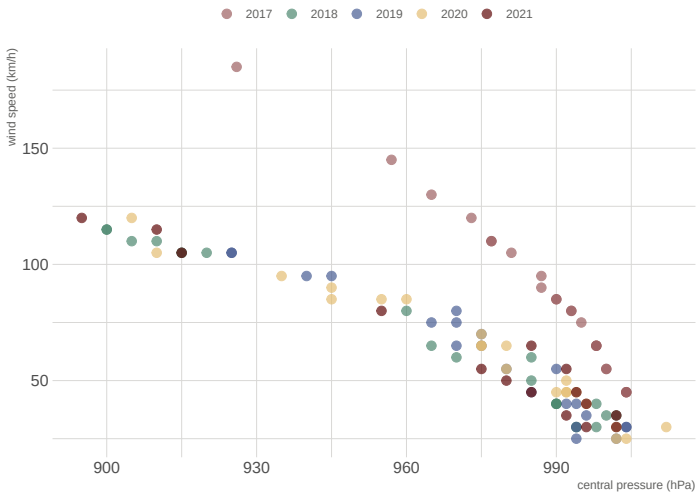


Figure 15.1: Scatterplot of pressure and speed of cyclones

Excel has a built-in functionality to create scatterplots. Following are the steps to create a scatterplot in Excel. For this demonstration, we use the cyclones dataset. We recommend creating a new Excel workbook and import the raw cyclones dataset into this workbook to avoid contamination of the original data.

1. Create a base scatterplot.

In a new worksheet, go to `Insert` –> `Insert Scatter (X, Y) or Bubble Chart` –> `Scatter`

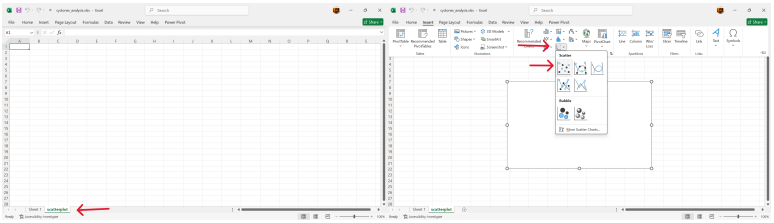


Figure 15.2: Create new worksheet

Figure 15.3: Create base scatterplot

2. Select data to use for the scatterplot.

a. Select base scatterplot –> `Chart Design` –> `Select Data`

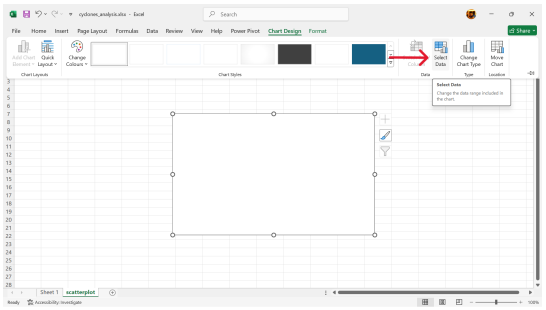


Figure 15.4: Go into Chart Design options to select data

b. Click on `Add` to add a new data series (Figure 15.5).

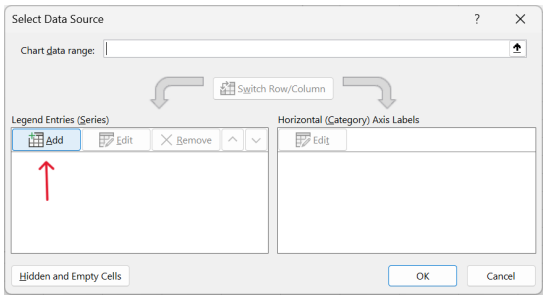


Figure 15.5: Click on Add to add new data series

c. Select the appropriate x-axis and y-axis values. For this demonstration, we will use pressure as the x-axis variable and speed as the y-axis variable.

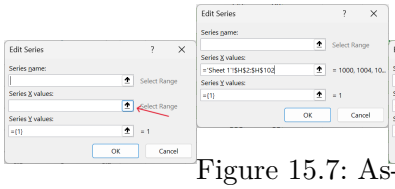


Figure 15.6: Edit x-axis values

Figure 15.7: Assign pressure to x-axis

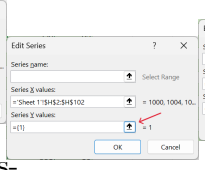


Figure 15.8: Edit y-axis values

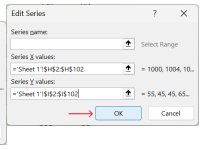


Figure 15.9: Assign speed to y-axis

d. Add a trendline (Figure 15.10).

`Chart Elements` –> `Trendline` –> `Linear`

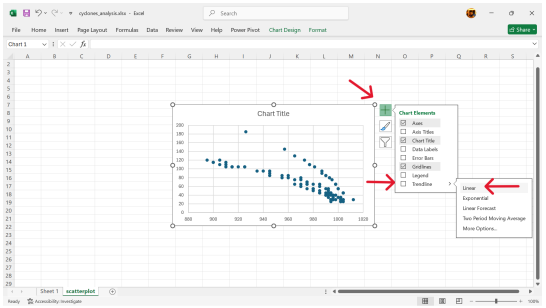


Figure 15.10: Add a trendline

e. The scatterplot is now created (Figure 15.11).

15.2 Numerical measures of association

15.2.1 Correlation

Correlation is a statistical measure indicating how strongly and in what direction two variables are connected. While it reveals the extent and nature of a linear relationship, it does not imply that one variable causes the other - it only shows how often they move together without explaining why.

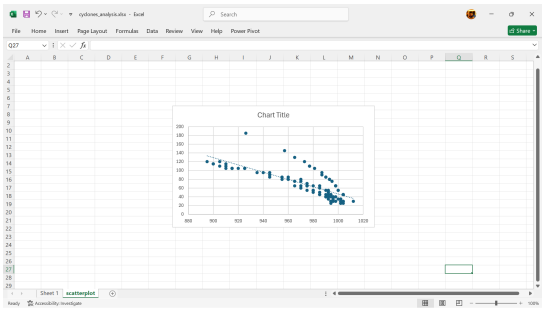


Figure 15.11: The scatterplot is now created

15.2.2 Correlation measures

In this section, we discuss the three most common numerical measures of correlation - *Pearson's correlation coefficient*, *Spearman's rank correlation coefficient*, and *Kendall's tau rank correlation coefficient*.

Of these three, Spearman's and Kendall's are non-parametric and are considered more robust than Pearson's.

Pearson's correlation coefficient

Pearson's correlation coefficient, commonly referred to as Pearson's ρ, is a statistical tool used to evaluate both the strength and direction of a linear association between two continuous variables. It indicates how closely data points align with a line of best fit. The value of ρ can range from -1 to +1.

The absolute value of Pearson's ρ reflects the strength of the linear relationship. A score of 1 or -1 signifies a perfect positive or negative correlation, respectively, implying all data points lie perfectly on a line. Conversely, a score of 0 suggests no linear relationship exists.

Pearson's ρ can be calculated as follows:

$$\rho = \frac{covariance(x_1, x_2)}{(n-1)sd_{x_1}sd_{x_2}}$$

where

x_1, x_2 = continuous variables to test correlation of

n = number of data pairs for x_1, x_2

sd_{x_1}, sd_{x_2} = standard deviation for x_1, x_2

In Excel, either the `PEARSON()` or the `CORREL()` function is used to get Pearson's ρ. To get Pearson's ρ for the correlation between pressure and speed from the cyclones dataset, we use the following calculation:

```
=PEARSON(H2:H102,I2:I102)
```

or

```
=CORREL(H2:H102,I2:I102)
```

with both giving a Pearson's ρ of **-0.7886634.**

Table 15.1 summarises how to interpret the range of Pearson's ρ values.

Table 15.1: Interpretation of various Pearson's ρ values.

Pearson's ρ	**Interpretation**
+1	Perfect positive correlation
-1	Perfect negative correlation
0	No correlation
+/- 0.1 to +/- 0.3	Weak correlation
+/- 0.4 to +/- 0.6	Moderate correlation
+/- 0.7 to +/- 0.9	Strong correlation

It's important to note that Pearson's ρ only measures linear relationships. It may not accurately reflect the relationship between variables if the relationship is non-linear. Pearson's correlation is typically used when dealing with normally distributed data that are measured on interval or ratio scales.

Spearman's rank correlation coefficient

Spearman's rank correlation coefficient (Spearman's ρ) is a statistical measure that assesses the strength and direction of a monotonic relationship between two ranked variables. It's a non-parametric test, meaning it doesn't assume data follows a normal distribution and is often used when data is ordinal or when a linear relationship isn't assumed. The coefficient

ranges from -1 to +1, with -1 indicating a perfect negative correlation, +1 indicating a perfect positive correlation, and 0 indicating no correlation.

Compared to Pearson's ρ, Spearman's ρ performs the correlation test on the rank of the values of the two variables rather than on the values themselves.

Hence, the values of the two variables are ranked first and then the Spearman's ρ is calculated based on these ranks as follows:

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$$

where:

d = difference in ranks

n = number of data pairs

In Excel, we can calculate Spearman's ρ as follows (using the pressure and speed variables in the cyclones dataset):

1. Rank each of the two variables you are testing for correlations using the `RANK.AVG()` function.

For this step, we recommend creating a new worksheet and importing the raw dataset to this worksheet to avoid contamination of the raw data (Figure 15.12).

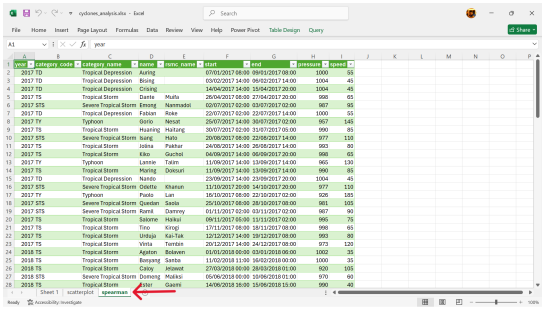


Figure 15.12: Import raw data to new worksheet

Create a new variable for the ranking of the pressure variable and then rank using `RANK.AVG()` function (Figure 15.13).

Create a new variable for the ranking of the speed variable and then rank using `RANK.AVG()` function (Figure 15.14).

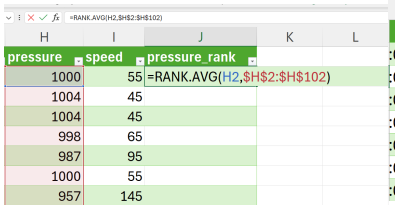


Figure 15.13: Rank the pressure variable

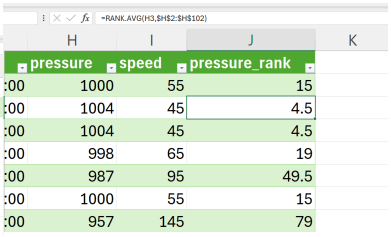


Figure 15.14: Copy/drag formula to rest of rows

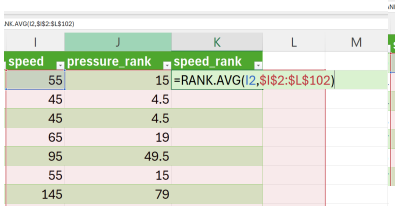


Figure 15.15: Rank the speed variable

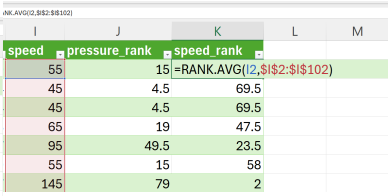


Figure 15.16: Copy/drag formula to rest of rows

2. Calculate Spearman's ρ.

a. Calculate using the `CORREL()` function in Excel.

The `CORREL()` function should be applied to the ranks for pressure and speed (see Figure 15.17) instead of the actual pressure and speed values (as used in the Pearson's calculation).

```
=CORREL(J2:J102,K2:K102)
```

Using this method, the Spearman's ρ for pressure and speed variable in the cyclones dataset is **-0.8289627**

b. Calculate using the formula.

First, get the square differences of the pressure and speed ranks (Figure 15.18).

```
=(J2-K2)^2
```

Then apply the Spearman's ρ formula as follows:

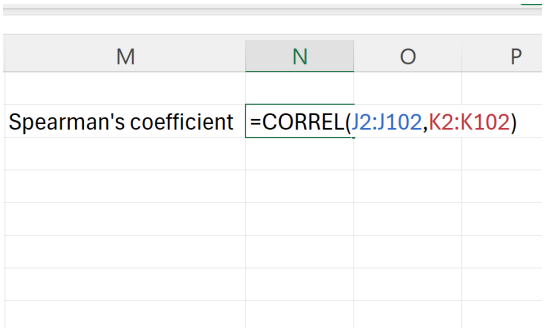
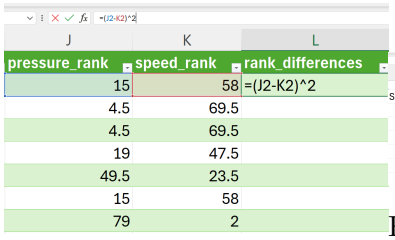


Figure 15.17: Calculate Spearman's ρ using built-in function

```
=1-((6*SUM(L2:L102))/(COUNT(L2:L102)*(COUNT(L2:L102)^2-1)))
```



pressure_rank	speed_rank	rank_differences
15	58	=(J2-K2)^2
4.5	69.5	
4.5	69.5	
19	47.5	
49.5	23.5	
15	58	
79	2	

Figure 15.18: Get squared differences of ranks

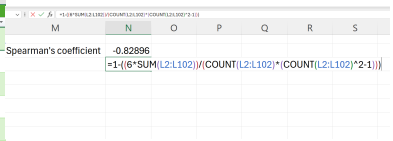


Figure 15.19: Calculate using the formula

Using this method, the Spearman's ρ for pressure and speed variable in the cyclones dataset is **-0.8226878**

> **Note 5: Difference between built-in function result and the calculated result**
>
> There is a very small difference (`-0.0063`) between the result when `CORREL()` function is used compared to when the formula is used. This is likely due to some differences in the way ranking is performed by the `CORREL()` function compared to the `RANK.AVG()` function approach used in the calculation approach.

The reference table in Table 15.1 for the Pearson's ρ can also be used to interpret Spearman's ρ.

Kendall's rank correlation coefficient

Kendall's τ, alternatively referred to as Kendall's tau rank correlation coefficient, serves as a statistical tool for evaluating the ordinal relationship between two variables. It examines how closely the rankings of data points align across two datasets, irrespective of their specific values. This measure is non-parametric, implying it does not rely on assumptions about data distribution and remains effective even in the presence of outliers.

Kendall's τ assesses how closely the rankings of two variables align. If one variable's ranking rises, does the other also tend to rise (positive relationship), fall (negative relationship), or show little pattern (nearly zero correlation)?

The method involves comparing concordant pairs - where both variables' rankings follow the same order - and discordant pairs - where their rankings are opposite in order.

Kendall's τ can be calculated as follows:

$$\tau = \frac{n_{concordant} - n_{discordant}}{\frac{n(n-1)}{2}}$$

where:

n = number of data pairs

There is no built-in function in Excel that calculates Kendall's τ. Following are steps on how to arrive at this value for pressure and speed variables in the cyclones dataset using Excel.

1. Create a new worksheet and import cyclones dataset.

We recommend creating a new worksheet and importing the raw dataset to this worksheet (see Figure 15.20) to avoid contamination of the raw data.

2. Rank pressure values.

Using the `RANK.AVG()` function in Excel, rank the pressure values as follows (Figure 15.21):

```
=RANK.AVG(H2,$H$:$H$102)
```

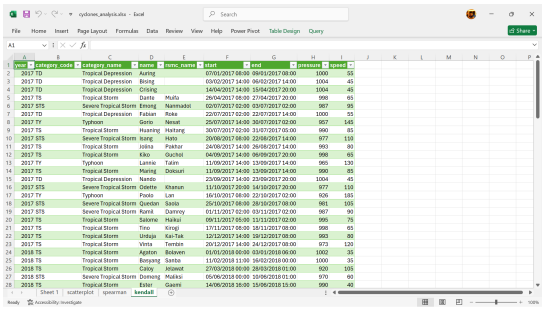


Figure 15.20: Create new worksheet and import cyclones dataset for Kendall's tau calculations

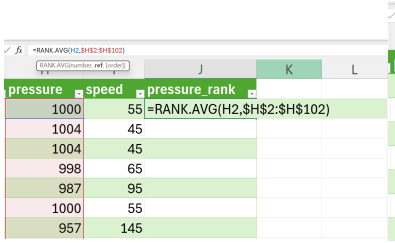


Figure 15.21: Rank pressure values

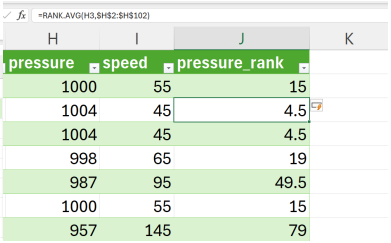
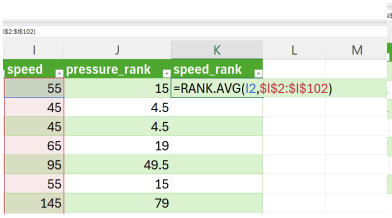


Figure 15.22: Copy/drag formula to rest of rows

2. Rank speed values.

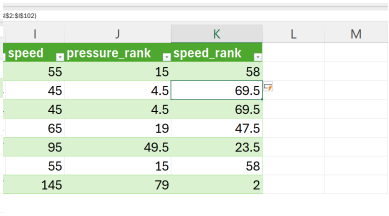
Using the `RANK.AVG()` function in Excel, rank the speed values as follows (Figure 15.23):

```
=RANK.AVG(I2,$I$:$I$102)
```



speed	pressure_rank	speed_rank
55	15	=RANK.AVG(I2,$I$2:$I$102)
45	4.5	
45	4.5	
65	19	
95	49.5	
55	15	
145	79	

Figure 15.23: Rank speed values



speed	pressure_rank	speed_rank
55	15	58
45	4.5	69.5
45	4.5	69.5
65	19	47.5
95	49.5	23.5
55	15	58
145	79	2

Figure 15.24: Copy/drag formula to rest of rows

3. Sort the table by ascending order based on the pressure rank.

Click on the sort functionality at the pressure column if available (Figure 15.25) or go to `Data` –> `Sort` to sort the table by pressure rank in ascending order.

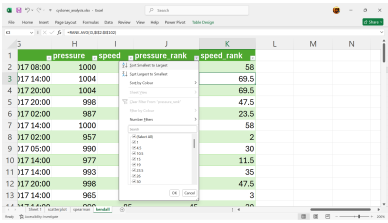


Figure 15.25: Sort table by pressure rank

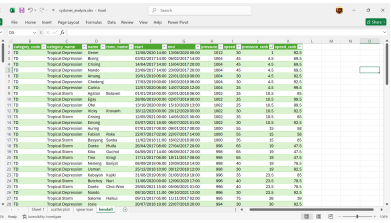


Figure 15.26: Table sorted by pressure rank

4. Create new variable for counts of concordant pairs.

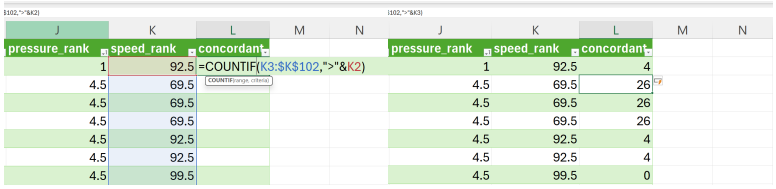
Concordant pairs are items that are ranked higher in variable X and also ranked higher in variable Y, or items ranked lower in X and also ranked lower in Y.

Since our cyclones data is now sorted in ascending order based on the pressure ranking, we can go down the array of ranks for speed to check for concordance. For example, in the cyclones

dataset, if the value of the speed rank in `K2` cell is higher than the value of the rank in the `K3` cell, then these two are counted as concordant pairs. We make this comparison for the speed rank in `K2` for every cell after it and tally the number of concordant pairs. The sum of the counts of concordant pairs for `K2` cell becomes the value for the concordant variable for row 2. We then do the same for the next row until we have counts of concordant pairs for each row of data.

This can be implemented in Excel using the following formula (Figure 15.27):

```
=COUNTIF(K3:$K$102,">"&K2)
```



J	K	L	M	N
pressure_rank	speed_rank	concordant		
1	92.5	=COUNTIF(K3:$K$102,">"&K2)		
4.5	69.5			
4.5	69.5			
4.5	69.5			
4.5	92.5			
4.5	92.5			
4.5	99.5			

Figure 15.27: Count number of concordant pairs

J	K	L	M	N
pressure_rank	speed_rank	concordant		
1	92.5	4		
4.5	69.5	26		
4.5	69.5	26		
4.5	69.5	26		
4.5	92.5	4		
4.5	92.5	4		
4.5	99.5	0		

Figure 15.28: Copy/drag formula to rest of rows

4. Create new variable for counts of discordant pairs.

Discordant pairs are items that are ranked higher in variable X and are ranked lower in variable Y, or vice versa.

Since our cyclones data is now sorted in ascending order based on the pressure ranking, we can go down the array of ranks for speed to check for concordance. For example, in the cyclones dataset, if the value of the speed rank in `K2` cell is lower than the value of the rank in the `K3` cell, then these two are counted as discordant pairs. We make this comparison for the speed rank in `K2` for every cell after it and tally the number of discordant pairs. The sum of the counts of discordant pairs for `K2` cell becomes the value for the discordant variable for row 2. We then do the same for the next row until we have counts of discordant pairs for each row of data.

This can be implemented in Excel using the following formula (Figure 15.29):

```
=COUNTIF(K3:$K$102,"<"&K2)
```

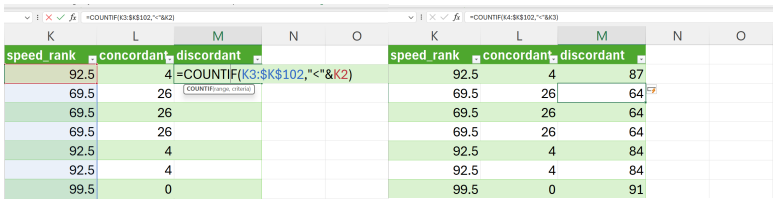


Figure 15.29: Count number of discordant pairs

Figure 15.30: Copy/drag formula to rest of rows

5. Calculate Kendall's τ.

Now that we have the counts of concordant and discordant pairs, we can calculate Kendall's τ in Excel as follows (Figure 15.31):

```
=(SUM(L2:L102)-SUM(M2:M102))/((COUNT(L2:L102)*(COUNT(L2:L102)-1))/2)
```

This results in a Kendall's τ of **-0.642178218**.

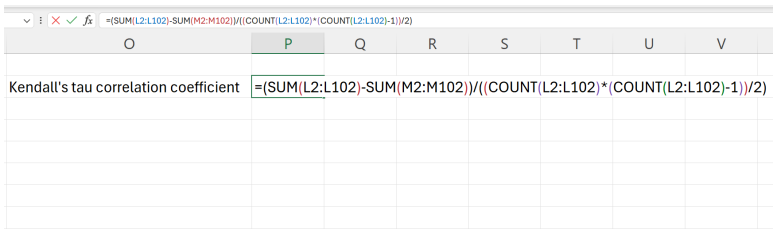


Figure 15.31: Calculate Kendall's tau

The reference table in Table 15.1 for the Pearson's ρ can also be used to interpret Kendall's τ.

16 Epidemiological statistics

In this chapter, we will cover topics and techniques on statistical methods commonly used by epidemiologists in public health investigations or studies that have wider uses and applications to other fields. These methods are considered *bread-and-butter* techniques for all epidemiologists and are generally easy to implement.

We will cover topics on *contingency tables*, *relative risk ratio*, *odds ratio*, and *t-test*. These methods can be considered bivariate statistics as they are applied on two variables but with binary categorical variables. To demonstrate these techniques, we will use the fem dataset.

16.1 Contingency tables

A contingency table, also known as a cross-tabulation, is used in statistics to display the relationship between two or more categorical variables. It organises data by showing the frequency of observations that fall into various combinations of the categories of the variables being examined. It is also usually called a *two-by-two table* as its common use is for comparing two categories per group. However, contingency tables can also be created with more than two categories per group.

For this topic, we will focus on *two-by-two* tables for simplicity and to be consistent with exploratory data analysis of bivariates.

Figure 16.1 demonstrates the structure of a two-by-two contingency table with the exposure variable on the rows and the outcome variable on the columns.

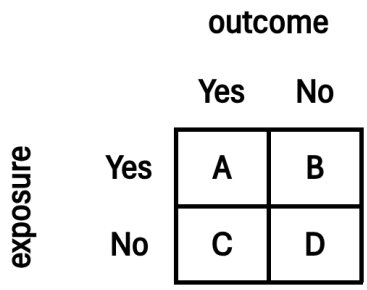


Figure 16.1: A diagram of a two-by-two contingency table

> **Note 6: Understanding exposure**
>
> Since contingency tables were developed for disease epidemiology, the term exposure has been used which usually pertains to exposure to a risk factor or known causative agent of a particular disease outcome.
> However, exposure in a general sense can also mean exposure to a factor or a condition that is known to be associated to a certain outcome which doesn't have to be a disease. For example, exposure to being female for an outcome of good grades; exposure to being married for an outcome of owning your house, etc.

16.1.1 Creating two-by-two contingency tables

In Excel, a contingency table can be easily created using pivot tables. Using the fem dataset, we can create a contingency table for the exposure variable of lost interest in sex (`SEX` variable) and the outcome variable of considered suicide (`LIFE` variable) through the following steps.

1. Create a new worksheet for the contingency table.

2. Setup pivot table.

a. `Insert` –> `Pivot Table` –> `From table/range`

b. Select table/range to pivot and insert into current worksheet.

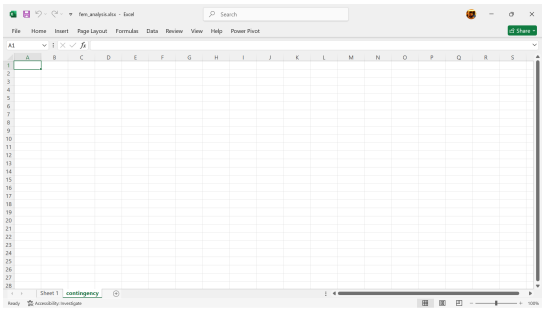


Figure 16.2: Create a new worksheet for the contingency table

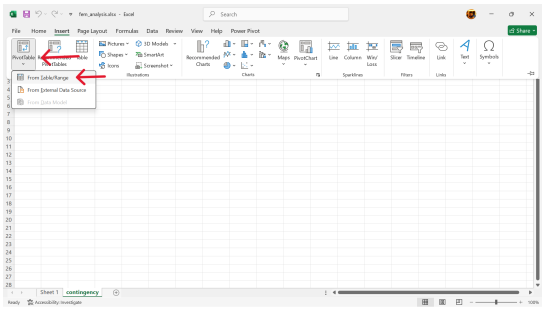


Figure 16.3: Initiate pivot table

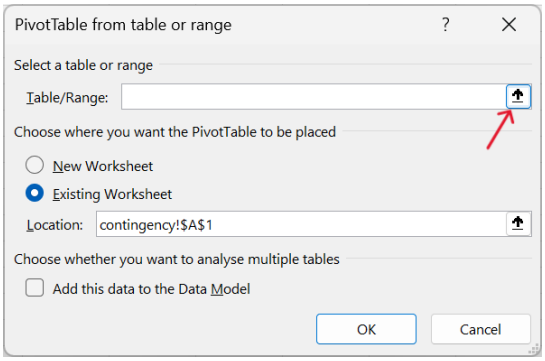


Figure 16.4: Select table/range

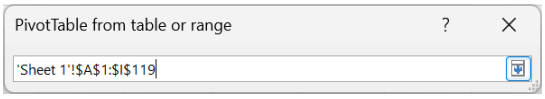


Figure 16.5: Select fem raw data table

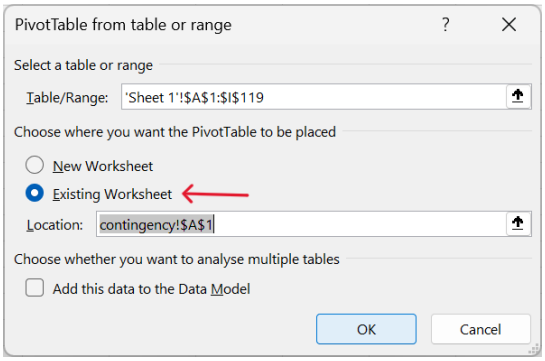


Figure 16.6: Insert into current worksheet

3. Select exposure variable of contingency table.

The variable for no interest in sex (`SEX`) is the exposure variable. Select and drag to the rows setting.

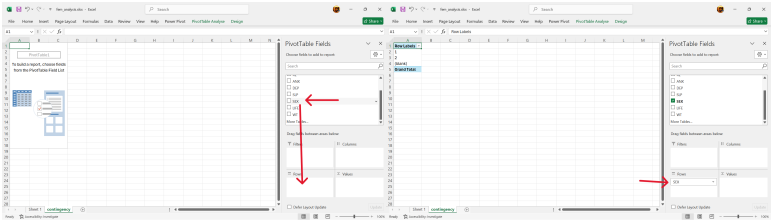


Figure 16.7: Select SEX as exposure variable

Figure 16.8: Drag SEX as row of table

4. Select outcome variable of contingency table.

The variable for considered suicide (`LIFE`) is the outcome variable. Select and drag to the columns setting.

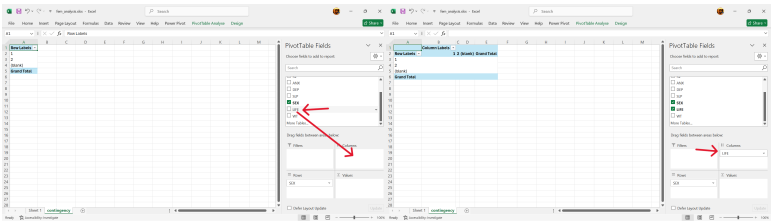


Figure 16.9: Select LIFE as outcome variable

Figure 16.10: Drag LIFE as column of table

5. Select values for the contingency table.

a. Drag the `LIFE` variable into the values setting.

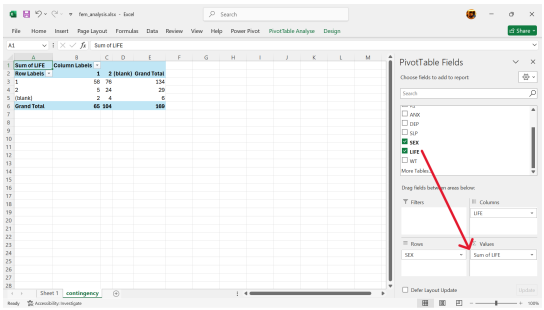


Figure 16.11: Drag LIFE into the values setting

b. Change the value setting to the COUNT summary measure.

Tap on the settings arrow on the value variable –> `Value Edit Settings` –> Select `COUNT`

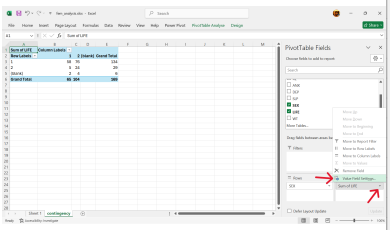


Figure 16.12: Go to value edit settings

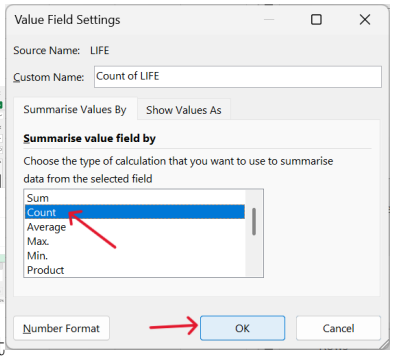


Figure 16.13: Select COUNT as summary measure

6. Remove empty values in exposure variable.

a. Click on settings button for exposure labels and untick blank.

7. Remove empty values in outcome variable.

a. Click on settings button for outcome labels and untick blank.

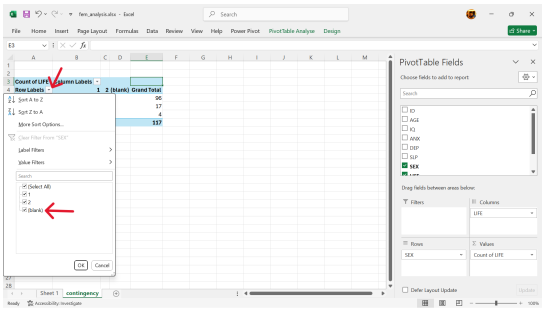


Figure 16.14: Click on settings for exposure

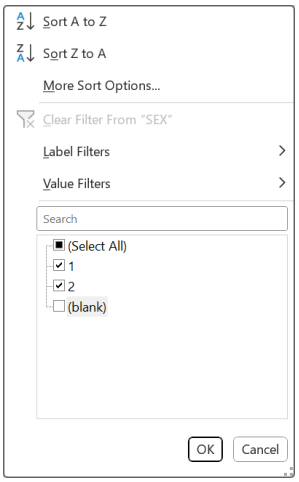


Figure 16.15: Untick blank exposure label

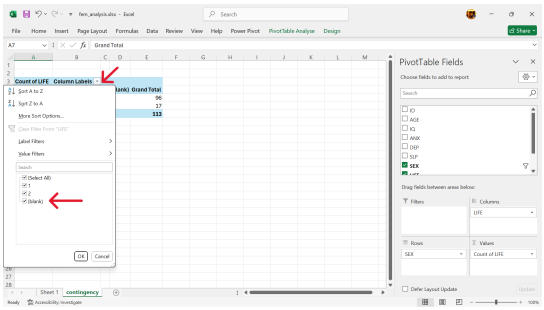


Figure 16.16: Click settings for outcome

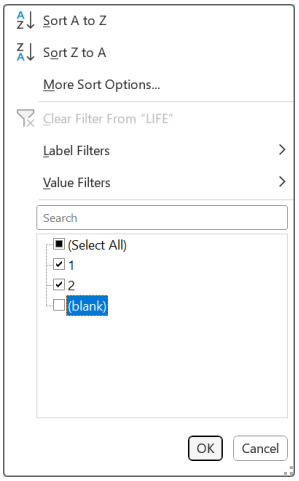
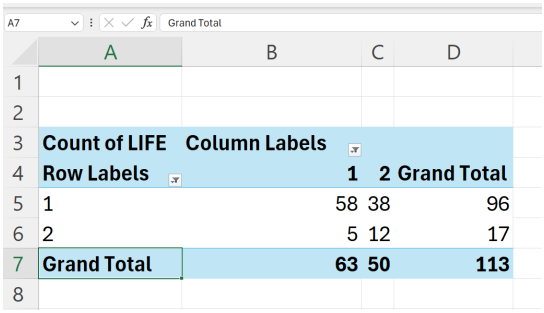


Figure 16.17: Untick blank outcome label

8. Contingency table is now complete.



	A	B	C	D
1				
2				
3	**Count of LIFE**	**Column Labels**		
4	**Row Labels**	**1**	**2**	**Grand Total**
5	1	58	38	96
6	2	5	12	17
7	**Grand Total**	**63**	**50**	**113**
8				

Figure 16.18: Contingency table is now complete

16.2 Relative risk ratio

Relative risk ratio (RRR) is a measure of the risk of a certain event happening in one group (usually called the exposed group) compared to the risk of the same event happening in another group (usually called the unexposed group). It indicates how much more likely the outcome is in the exposed group compared to the unexposed group.

16.2.1 Calculating relative risk ratio

Using the schema of a two-by-two table in Figure 16.1, the relative risk ratio is calculated as follows:

$$RRR = \frac{\frac{A}{A+B}}{\frac{C}{C+D}} = \frac{A \times (C+D)}{C \times (A+B)}$$

where:

A = exposed with outcome

B = exposed with no outcome

C = not exposed with outcome

D = not exposed with no outcome

Using the pivot table we created in Section 16.1.1, we can calculate the relative risk ratio as follows:

$$RRR = \frac{A \times (C+D)}{C \times (A+B)}$$

$$RRR = \frac{58 \times (5+12)}{5 \times (58+38)} = \frac{58 \times 17}{5 \times 96} = \frac{986}{480} = 2.054167$$

We can also perform this calculation in Excel using the pivot table we created.

```
=(B5*D6)/(B6*D5)
```

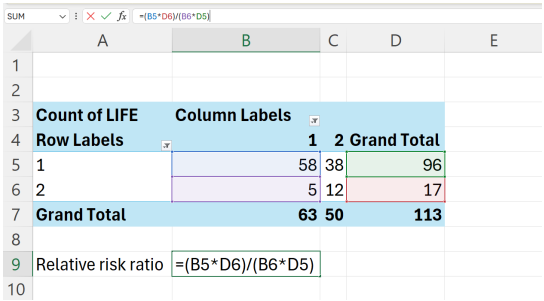


Figure 16.19: Calculate relative risk ratio

Calculating the confidence interval of the relative risk ratio

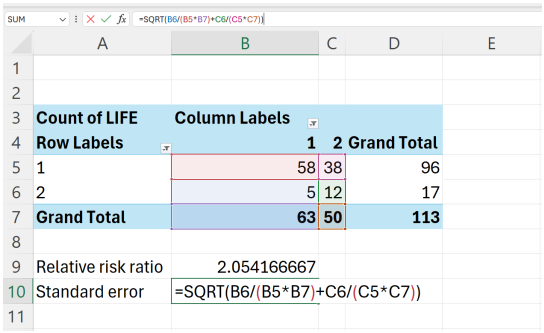
Following are the steps to calculating the confidence interval[1] of the relative risk ratio.

1. Calculate the standard error of the natural logarithm of relative risk ratio

$$SE_{log(RRR)} = \sqrt{\frac{C}{A(A+C)} + \frac{D}{B(B+D)}}$$

In Excel, this can be calculated as follows:

```
=SQRT(B6/(B5*B7)+C6/(C5*C7))
```



	A	B	C	D	E
1					
2					
3	**Count of LIFE**	**Column Labels**			
4	**Row Labels**	**1**	**2**	**Grand Total**	
5	1	58	38	96	
6	2	5	12	17	
7	**Grand Total**	**63**	**50**	**113**	
8					
9	Relative risk ratio	2.054166667			
10	Standard error	=SQRT(B6/(B5*B7)+C6/(C5*C7))			
11					

Figure 16.20: Calculate standard error of natural logarithm of relative risk ratio

2. Calculate the 95% confidence interval of relative risk ratio.

$$95\%\ CI = e^{\log(RRR)\ \pm\ 1.96\ \times\ SE_{\log(RRR)}}$$

In Excel, this can be calculated as follows:

[1] A confidence interval is a specific range of values, determined using sample data, which probably includes the actual value of an unknown population parameter. It shows how much uncertainty about a sample statistic and provides a likely interval for the corresponding population parameter. For instance, a 95% confidence interval means that if we repeated the sampling and calculation process numerous times, 95% of those intervals would include the true population parameter.

```
=EXP(LN(B9)-1.96*B10)      ## 95% LCI

=EXP(LN(B9)+1.96*B10)      ## 95% UCI
```

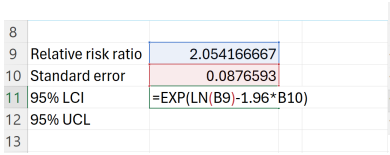


Figure 16.21: Calculate 95% lower confidence interval of relative risk ratio

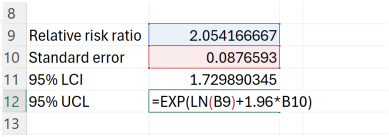


Figure 16.22: Calculate 95% upper confidence interval of relative risk ratio

The risk of suicidal ideation for those with no interest in sex is **2.05 (95% CI: 1.7298903-2.4392302)** times higher than those who have interest in sex.

16.2.2 Interpreting the relative risk ratio and its confidence interval

Table 16.1 provides guidance on how to interpret relative risk ratio.

Table 16.1: Interpretation of relative risk ratio values

Risk ratio	Interpretation
`RRR = 1`	Exposure does not affect outcome
`RRR < 1`	Risk of outcome is decreased by the exposure (protective factor)
`RRR > 1`	Risk of outcome is increased by the exposure (risk factor)

If the 95% confidence interval doesn't contain 1, this means that the risk of the outcome given the exposure is significant.

16.3 Odds ratio

Odds ratio (OR) is a measure of association between an exposure and an outcome. It represents the odds that an outcome will occur given a particular exposure compared to the odds of the outcome occurring in the absence of the exposure.

16.3.1 Calculating odds ratio

Using the schema of a two-by-two table in Figure 16.1, the odds ratio is calculated as follows:

$$OR = \frac{A/B}{C/D} = \frac{A \times D}{B \times C}$$

$$OR = \frac{58 \times 12}{38 \times 5} = \frac{696}{190} = 3.663158$$

We can also perform this calculation in Excel using the pivot table we created.

```
=(B5*D6)/(B6*D5)
```

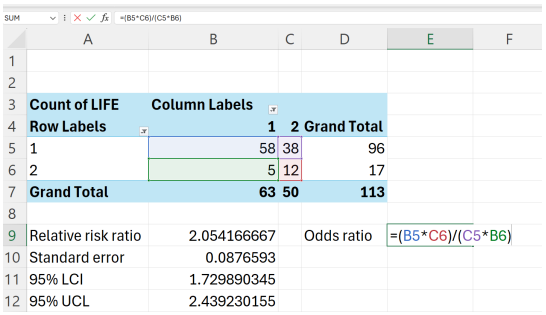


Figure 16.23: Calculate odds ratio

Calculating the confidence interval of the odds ratio

The 95% confidence interval is calculated as follows:

$$95\%\ CI = e^{\log(OR) \pm 1.96 \times \sqrt{\frac{1}{A}+\frac{1}{B}+\frac{1}{C}+\frac{1}{D}}}$$

In Excel, this can be calculated as follows:

```
=EXP(LN(E9)-1.96*SQRT(1/B5+1/C5+1/B6+1/C6))      ## 95% LCI

=EXP(LN(E9)+1.96*SQRT(1/B5+1/C5+1/B6+1/C6))      ## 95% UCI
```

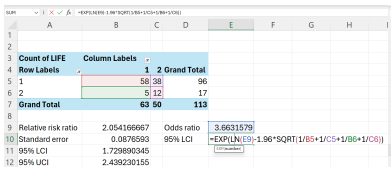


Figure 16.24: Calculate 95% lower confidence interval of odds ratio

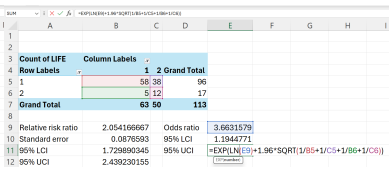


Figure 16.25: Calculate 95% upper confidence interval of odds ratio

The odds of suicidal ideation for those with no interest in sex is **3.66 (95% CI: 11.2339746-2.4392302)** times higher than those who have interest in sex.

16.3.2 Interpreting the odds ratio and its confidence interval

Table 16.2 provides guidance on how to interpret odds ratio.

Table 16.2: Interpretation of odds ratio values

Odds ratio	Interpretation
`OR = 1`	Exposure does not affect odds of outcome
`OR > 1`	Exposure associated with higher odds of outcome
`OR < 1`	Exposure associated with lower odds of outcome

If the 95% confidence interval doesn't contain 1, this means that the odds of the outcome given the exposure is significant.

16.4 Difference between relative risk ratio and odds ratio

Relative risk ratio approximates odds ratio for outcomes that are rare ($< 10\%$) and as such can be reported interchangeably.

In non-rare outcomes, odds ratio will tend to have greater magnitude than relative risk ratio but always in the same direction (negative or positive). In specific study designs, the total population-at-risk is not known hence relative risk ratio cannot be calculated.

16.5 Student t-test

Sometimes, we want to compare summary numerical values between one group and another. Unlike a contingency table that summarises the counts of the variables, this summary table will usually have the mean or median of the numerical values. We can use the *t-test* (also known as the *Student t-test*) to compare whether the mean of the values for one group is different from another group.

16.5.1 Calculating the t-test

Using the fem dataset, let's say for example we wanted to compare the mean age of those who have had thoughts of suicide to those who haven't had thoughts of suicide. We can use the t-test to compare their mean age. In Excel, there is a built in function that performs the t-test, the `T.TEST()` function. Following are the steps on how to get the mean age for each group and then how to test if there is a difference between the mean age of the two groups.

1. Sort the fem dataset by the values of the `LIFE` variable.

We recommend doing this step on a new worksheet with a fresh instance of the fem dataset imported in (Figure 16.26). Then sort the whole table based on the values of the `LIFE` variable (Figure 16.27).

2. Get the mean age for the each group value of `LIFE` variable.

```
=AVERAGE(B2:B66)      ## Average age of those who thought of suicide

=AVERAGE(B67:B118)    ## Average age of those who have not thought of suicide
```

3. Perform t-test on `AGE` variable between the two groups.

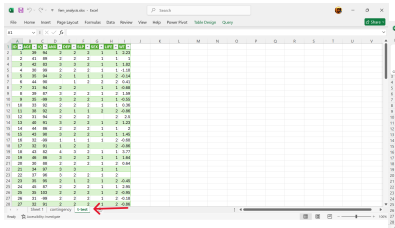


Figure 16.26: Create a new worksheet

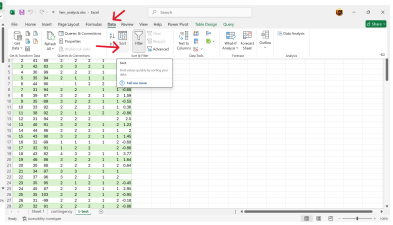


Figure 16.27: Setup sort

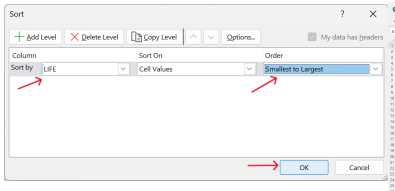


Figure 16.28: Sort by LIFE from smallest to largest

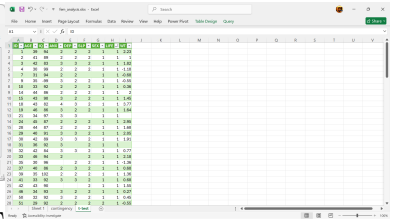


Figure 16.29: Table is now sorted by LIFE

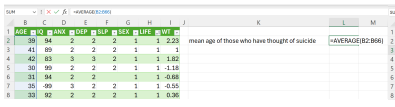


Figure 16.30: Get the mean age for those who have thought of suicide

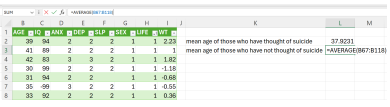


Figure 16.31: Get the mean age for those who have not thought of suicide

Using the `T.TEST()` function:

```
=T.TEST(B2:B66,B67:B118,2,2)
```

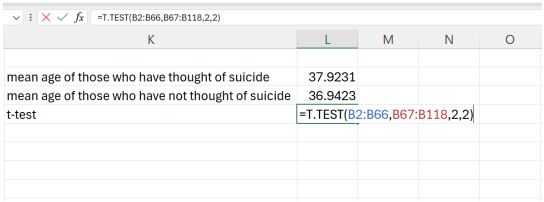


Figure 16.32: Perform t-test

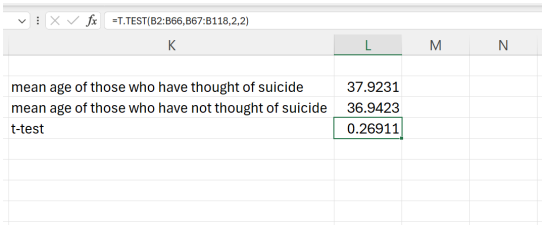


Figure 16.33: Results of the t-test

The result of the t-test is the p-value for the test. The result is **0.2691091**. These is no significant difference between the mean ages of those who thought of suicide and to those who had no thoughts of suicide.

References

Anon. Spreadsheet risk management within UK organisations. Actuarial Post: For the Modern Actuary. (https://www.actuarialpost.co.uk/article/spreadsheet-risk-management-within-uk-organisations-351.htm, accessed 12 June 2025).

Bean R (2017). How companies say they're using big data. Harvard Business Review. (https://hbr.org/2017/04/how-companies-say-theyre-using-big-data).

Carroll SR, Rodriguez-Lonebear D, Martinez A (2019). Indigenous data governance: Strategies from united states native nations. CODATA, 18(1):31. doi:10.5334/dsj-2019-031.

Choi Y, Gil-Garcia J, Burke GB, Costello J, Werthmuller D, Aranay O (2021). Towards data-driven decision-making in government: Identifying opportunities and challenges for data use and analytics. *Hawaii international conference on system sciences*. doi:10.24251/HICSS.2021.268.

Ivacko TM, Horner D, Crawford MQ (2013). Data-driven decision-making in michigan local government. SSRN Journal. doi:10.2139/ssrn.2351916.

Middleton JH (1933). Baking costs. National Association of Cost Accountants Bulletin, 14(10).

Murrell P (2013). Data intended for human consumption, not machine consumption. *Bad data handbook*. Sebastopol, CA: O'Reilly Media; 2013:31–51.

Onunga J, Odongo P (2025). Digital transformation in public administration and data-driven decision-making: A review of turkana county government. IJRIAS, IX:234–240. doi:10.51584/IJRIAS.2024.912022.

Powell S, Baker K, Lawson B (2009). Errors in operational spreadsheets. Journal of Organizational and End User Computing, 21(3):24–36.

Powell SG, Baker KR, Lawson B (2008). A critical review of the literature on spreadsheet errors. Decision Support Systems, 46(1):128–138. doi:10.1016/j.dss.2008.06.001.

Sayogo DS, Yuli SBC, Amalia FA (2024). Data-driven decision-making challenges of local government in indonesia. TG, 18(1):145–156. doi:10.1108/TG-05-2023-0058.

Stobierski T (2019). The advantages of data-driven decision-making. Business insights [web site]. (https://online.hbs.edu/blog/post/data-driven-decision-making, accessed 12 May 2025).

Tukey JW (1977). *Exploratory data analysis*. Reading (Mass.) Menlo Park (Calif.) London [etc.]: Addison-Wesley publ Addison-wesley series in behavioral science.

United Nations General Assembly (2007). United nations declaration on the rights of indigenous peoples : Resolution / adopted by the general assembly. (https://www.refworld.org/legal/resolution/unga/2007/en/49353).

Wickham H, Çetinkaya-Rundel M, Grolemund G (2023). *R for data science: Import, tidy, transform, visualize, and model data*, Second Edition. Bejing: O'Reilly.

Index