STC 137 Class Notes

Department of Statistics, University of Pretoria

Last Updated: 2025-02-02

Contents

Preface

In modern society, whether you are reading a magazine, watching TV or scrolling through social media, you encounter statistics. You can't avoid it! Data collection, analysis and communication have become part of day-to-day activities. If you want to be an educated consumer and citizen, you need to understand how statistics are used and also misused in our daily lives. In other words, you need to be *data literate*. The purpose of this book is to give students, in any field of study, a conceptual and practical introduction to the field of statistics. The main focus is on the use of basic statistical techniques for, among others, *collecting data*, *understanding the data* by exploring it, *analyzing the data*, *interpreting the data* and *communicating the data* to facilitate data-driven decision-making. The book is application oriented. The discussion of each technique is followed by application examples. This book is written with the needs of the practitioner in mind. The mathematical prerequisite is matric (Grade 12) level mathematics.

Chapter 1

An Overview of Data literacy

In modern society, we frequently see the following types of statements in the media:

- Momentum Chief Executive, Jeanette Marais, said while 80% of the [withdrawal requests from the two-pot retirement system] are from people between 30 and 49 years old, there was concern over requests from the 50-59 age group, which made up 16% of the total. (Reuters, 27 September 2024)
- Despite the ever presence of sunshine and wind, only 8% of South Africa's power comes from renewables compared to a global average of 29%. (IOL.co.za, 11 December 2023)
- BHP Billiton, the Australian-based diversified global miner, says it expects global electricity consumption for data centers to rise from around 2% of total demand today, to 9% by 2050, with copper demand in data centers increasing six-fold by then. (Mining.com, 30 September 2024)
- According to StatsSA's latest report, motor trade sales for March 2024 (measured in real terms by constant 2019 prices) decreased by 10.4% year-on-year, 7% month-on-month, and 2.9% quarter-on-quarter. (businesstech.co.za, 16 May 2024)
- Unsurprisingly, a trial found that people with heart disease who were obese or overweight reduced their risk of having a severe cardiovascular event — including death, stroke or heart attack — by 20% when they took semaglutide. (Nature.com, 25 September 2024).
- Released this week, the sixth South African HIV prevalence, incidence, and behaviour Survey (SABSSM VI) found a 7.4% HIV prevalence rate

> in the province [Western Cape] for 2022, down from 8.6% in 2017. (Mail & Guardian, 29 September 2024)

The numerical facts in the preceding statements – 80%, 16%, 8%, 29%, 2%, 9%, 10.4%, 7%, 2.9%, 20%, 7.4% and 8.6% - are referred to as **statistics or statistical information**. This type of information can help us understand the trends in economics, business, health and employment and thus enable us to make more informed decisions. Statistical information is **data** that has been recorded, classified, organized, related, or interpreted within a framework so that meaning emerges[1]. Thus, an essential skill towards effectively making use of statistical information is data literacy. In this Chapter, we give an overview of data literacy, including key data literacy skills that will be covered more broadly in the coming chapters. Thus, this chapter lays the foundation for the rest of the book.

1.1 Introduction

In modern society, **data** is everywhere. It is collected when you

- make a purchase online (such as Takealot);
- access a learning management system (such as ClickUP);
- click an advertisement;
- like or comment on someone's social media post;
- stream music or movies online (using platforms such as Spotify and Netflix);
- review your experience with a product or service online (such as our stay at an Airbnb); and
- engage in physical activity and even while you are sleeping! (when you are wearing a fitness tracker such as a Fitbit or Garmin).

This implies that data has now become more widely accessible to organizations as well as ordinary individuals. Using this **data**, organizations (and individuals) can make more data-driven decisions with **statistics** instead of intuition. This often leads to increased performance and efficiency, especially when compared to less data-driven approaches. However, access to data on its own isn't enough to ensure organizational success. The users of this data, such as the employees in an organization, have to understand and know how to leverage the data and this requires that they have the necessary data literacy skills.

[1] https://www150.statcan.gc.ca/n1/edu/power-pouvoir/ch1/definitions/5214853-eng.htm

1.1.1 Definition

Bhargava and D'Ignazio (2015), define data literacy as the ability to read, work with, analyze and argue with data[2]. Reading the data involves understanding the data and being able to interpret it. Working with data involves creating, acquiring and managing the data. Analyzing the data involves filtering, sorting and aggregation of the data. Arguing with the data means using the data to communicate.

In this book, we adopt an expanded form of the above definition of *data literacy*. We define data literacy as the ability to manage, understand, explore, analyze, interpret and communicate with data in a meaningful way. Data literacy does not require an individual to be an expert but to show an understanding of the basic data fundamentals like data sources, data types, measurement scales, types of analysis, data cleaning, data analysis tools (such as Excel)[3], concepts that will be explored in more detail in this book.

1.1.2 The importance of being data literate in modern society

In modern society, data plays a pivotal role for the proper functioning of governments, businesses, households and individuals. Being data literate: (1) can lead to timely response to socio-economic or health related issues, (2) it can create sustained economic value, (3) it can assist with making informed decisions, (4) it can improve communication and (5) lead to professional or career advancement. In this section, we briefly discuss the benefits of being data literate for each of these various stakeholders.

- Governments

Information from a population census can assist a government official or department to efficiently allocate government resources by, for instance, making sure that there is an equal distribution of services such as health and education.

- Businesses

Information from a marketing survey can assist a business to reduce its costs, improve its operational efficiency into operational excellence, improve its competitive advantage and market positioning by, for instance, understanding the trends among its customer base.

[2]https://www.media.mit.edu/publications/designing-tools-and-activities-for-data-literacy-learners/

[3]https://online.hbs.edu/blog/post/data-literacy

- Households

Statistical information on the annual government budget can assist households adjust savings strategies based on changes in social welfare; identify job opportunities based on the allocation of public funds to sectors such as construction and make informed decisions through civic participation and democratic engagement.

- Individuals

Statistical information from a health survey or smart watch can assist an individual to improve their health and well-being; In modern society, misinformation spreads faster and widely often resulting in inaccurate decision-making. A data-literate public is more important today in order to anticipate and prevent the negative consequences of misinformation.

1.1.3 Data Literacy Skills

The essential skills needed in order to be data literate can be divided according to the definition of data literacy as managing, exploring, analyzing, understanding, interpreting and communicating with data. The first three skills are more technical and will be explained in the following chapters. The last three skills are more practical and will be demonstrated in the following section.

1.1.4 Data Literacy in Practice

1.1.4.1 Being able to interpret or understand and use the information contained in a chart or graph.

1. The figure below (obtained from the 2022 *South African National HIV Prevalence, Incidence, Behaviour and Communication Survey (SABSSM VI)*[4]) shows a map of South Africa (SA) as a **heatmap**[5]. For each province, the map displays the HIV prevalence, which is the percentage of the population in the province that are HIV positive.

From Figure **??**, we can see that, in 2022, the Western Cape and Northern Cape had the lowest HIV prevalence (8.2 – 11.6%).

[4] https://sahivsoc.org/Files/SABSSMVI-SUMMARY-SHEET-2023.pdf

[5] A heatmap is a two-dimensional visual representation of data using colors, where the colors all represent different values (https://www.investopedia.com/terms/h/heatmap.asp)

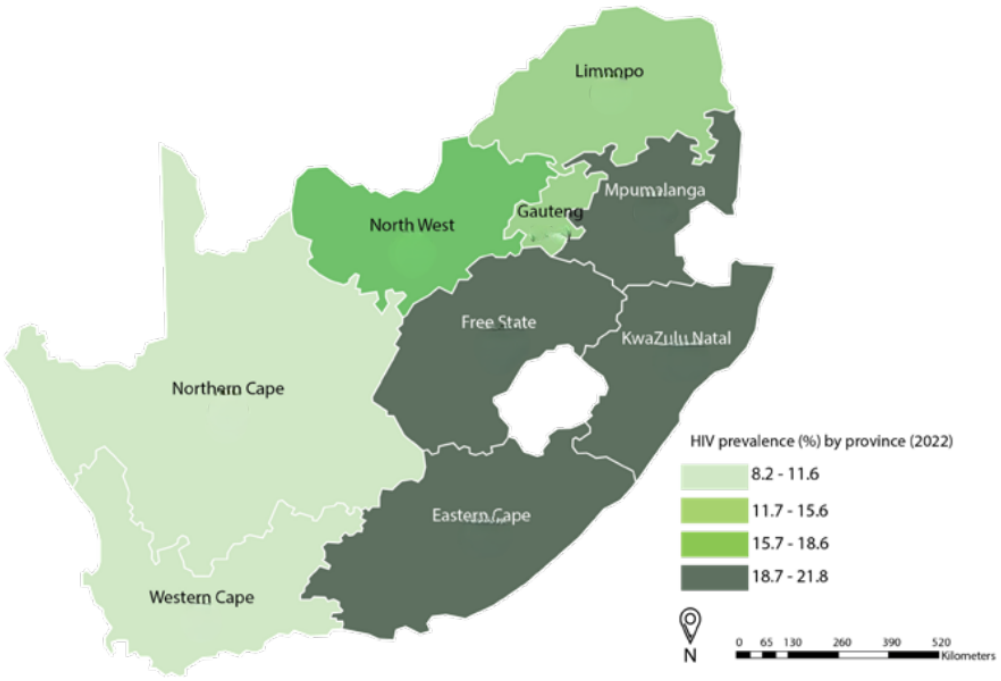


Figure 1.1: A heatmap of South Africa showing the HIV prevalence rate in each province in 2022

2. The figure below (Figure **??** shows a map (from an article in the September 30, 2024 issue of the *Wall Street Journal*) of the United States of America (USA) as a heatmap. For each state of the USA, the map displays the share of the total power consumed by data centers in 2023. The lighter the section on the map, the less the share of the power consumed by data centers.

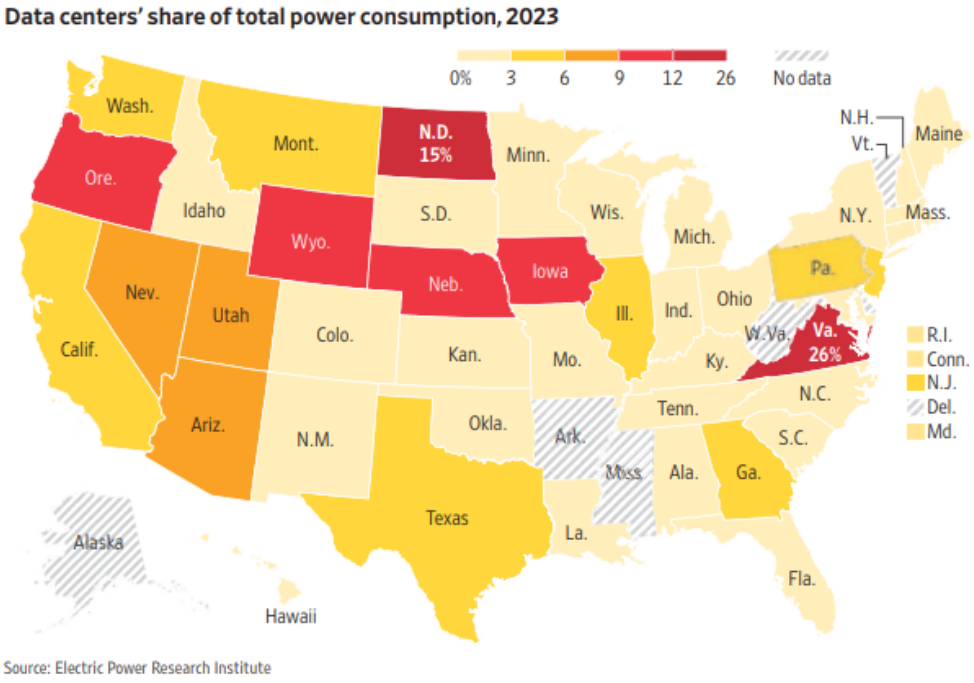


Figure 1.2: A heatmap of the United States showing the data centers' share of total power consumption by state in 2023

From Figure **??**, we can see that in the state of Michigan (Mich.), for instance, data centers consume a small share (0 – 3%) of the total power consumed by the state. Whereas, in the state of Virginia (Va.) data centers consume a large share (26%) of the total power consumed in the state.

1.1.4.2 Being able to contextualize a number given in the media:

1. The following is an extract from an article in the Financial Times UK[6]

[6] https://www.ft.com/content/da407b47-4133-470a-9574-508cee43e107

BHP warns AI growth will worsen copper shortfall

The growth of artificial intelligence will exacerbate a looming shortage of copper, a metal vital for the clean energy transition, miner BHP has warned.

The rise of data centres and AI, which requires more energy-intensive computing, could boost global copper demand by 3.4mn tonnes a year by 2050, BHP's chief financial officer Vandita Pant told the Financial Times.

"Today, data centres are less than 1 per cent of copper demand, but that is expected to be 6 to 7 per cent by 2050," she said. "There is a lot of copper in data centres."

The article talks about the growth in future copper demand from data centers as a result of Artificial Intelligence (AI). Data centers are now being used to run AI models and this process is energy intensive. Copper is used in various aspects of a data center[7]. The growth in data centers means the growth in copper demand. Today, of all the things copper is used for in the world, data centers account for less than 1%. This number can grow to 6% or 7% by 2050 as a result of an increase of an additional 3.4 million tonnes of global copper demand.

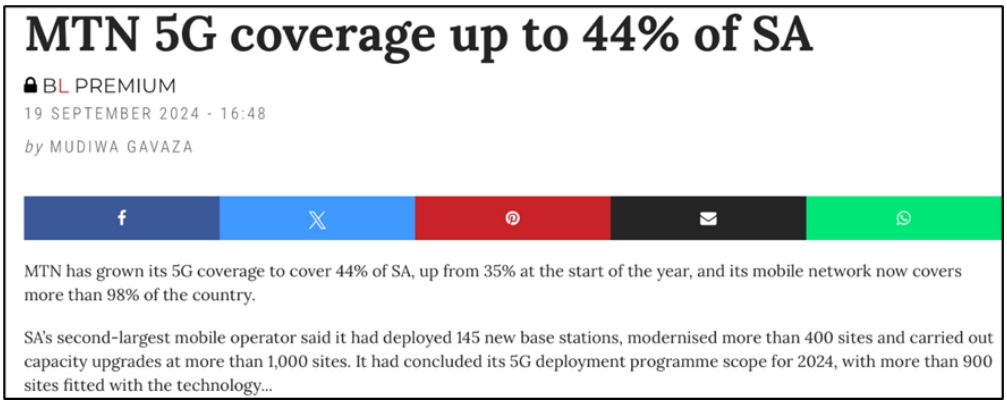
Possible implications of the above statistics: Increased job growth in the copper value chain and construction of data centers; Increased demand for AI and data center expertise; Increased demand for mining engineers with expertise in copper mining.

1.1.4.3 Recognising and being able to identify and avoid wrong interpretations of statistical information in the media.

1. Consider the following extract of an article from the Business Day[8]

[7] https://www.visualcapitalist.com/sp/copper-the-critical-mineral-powering-data-centers/

[8] https://www.businesslive.co.za/bd/companies/telecoms-and-technology/2024-09-19-mtn-5g-coverage-up-to-44-of-sa/#:~:text=Mobile%20operator%20concludes%20deployment%20scope,sites%20fitted%20with%20the%20technology&text=MTN%20has%20grown%20its%205G,than%2098%25%20of%20the%20country



MTN 5G coverage up to 44% of SA

BL PREMIUM

19 SEPTEMBER 2024 - 16:48

by MUDIWA GAVAZA

MTN has grown its 5G coverage to cover 44% of SA, up from 35% at the start of the year, and its mobile network now covers more than 98% of the country.

SA's second-largest mobile operator said it had deployed 145 new base stations, modernised more than 400 sites and carried out capacity upgrades at more than 1,000 sites. It had concluded its 5G deployment programme scope for 2024, with more than 900 sites fitted with the technology...

How can we interpret the statistical information of 44% and what are its implications?

What it means:

44% of the total land area in South Africa has access to MTN's 5G network signal.

What it doesn't mean:

44% of South Africans use MTN's 5G mobile network.

2. The United States' (US) Center for Disease Control (CDC) reported that 99% of the monkeypox cases in the US occurred in men.

   This means that:

   Of all the reported monkeypox cases in the US, 99% of them were men.

   It doesn't mean that:

   All the monkeypox cases in the US are men. Thus, falsely, implying that men are more likely to get monkeypox (99% chance!).

3. A study is conducted to assess the effectiveness of a new weight loss drug across two age groups: young adults (20-30 years) and older adults (50-60 years). The results are as follows:

Combined Data

	Experiment Group	
	Drug	Placebo
Average Weight loss (kg)	4.5	2.5

Separate Data

	Experiment Group	
Age group	Drug	Placebo
Young	7	2
Older	2	3

Based on the combined results, we can conclude that the drug was effective across the board. However, if we consider Young adults separate from Older adults, we can see that the drug was only effective for younger adults but not for Older adults. Thus, if we only look at the combined data without considering the age groups, we make the wrong conclusion that the drug is equally effective for all ages, while in reality the effectiveness varies between age groups. This reversal of conclusion based on combined and separate data is called **Simpson's Paradox**. For another example of Simpson's Paradox, watch the following video:

4. A study showed that 90% of drought-tolerant plants have a certain gene X. Moreover, 20% of plants with gene X are drought tolerant.

- Incorrect interpretation

  Menzi, a plant physiologist at the University of Cape Town, concludes that, since 90% of the drought-tolerant plants have gene X, having gene X causes or is a strong indicator that a plant is drought-tolerant.

- Correct interpretation

  Hlengiwe, a plant physiologist at the Univeristy of Pretoria, concludes that although gene X is common among drought-tolerant plants (shown by a large percentage 90%!), it is not a strong indicator of drought tolerance on its own. Other factors, such as environmental conditions or additional genes, may play a significant role.

5. Suppose that a researcher observes that ice cream sales and drowning incidents are positively correlated over time. As ice cream sales increase, the number of drowning incidents also increases.

- Incorrect interpretation

  Eating ice cream causes drowning. This is an example of confusing correlation with causation.

- Correct interpretation

  There is no causal relationship between ice cream sales and drowning incidents. Instead both variables are influenced by a third variable **hot weather**. Hot weather increases the likelihood of people swimming in pools, lakes, or the ocean, which in turn increases the risk of drowning. At the very same time, hot weather also increases the demand for ice cream, as people seek cold treats to cool down.

  Thus, the correlation between ice cream sales and drowning incidents is **spurious**. Spurious correlation arises when two variables are coincidentally related or influenced by the same third variable, known as a confounding variable, such that they appear to be causally related.

1.1.5 Exercises to Section 1.1

Question 1

What is data literacy?

Question 2

In each of the following scenarios, identify the statistical information:

a. A recent poll showed that 30% of South African wealth is held by the top 1% of the wealthy individuals in the country.

b. An environmental survey revealed that females were more environmentally conscious than males.

c. A health study showed that people with nausea are more likely to have headaches.

d. In a retirement planning survey, the majority of people who responded were under the age of 35.

Question 3

Answer the following as True or False.

a. Data literacy is only useful to data analysts, statisticians and data scientists.

b. It is not necessary to have an advanced knowledge in Statistics in order to be data literate.

c. One of the skills of data literacy is the ability to argue with data.

d. Data literacy can improve my communication skills.

Question 4

Answer the following questions based on Figure **??**.

a. What is the HIV prevalence rate for the North West province?

b. Which of the nine provinces have the largest HIV prevalence?

c. Which one of the following provinces has the highest HIV prevalence?

   i. North West
   ii. Limpopo
   iii. Gauteng

d. How many coastal provinces have the largest HIV prevalence (18.7 – 21.8%) and how many inland provinces have the largest HIV prevalence?

Question 5

Answer the following questions based on Figure **??**.

a. What share of the total power consumption in the state of Washington (Wash.) is used by data centers?

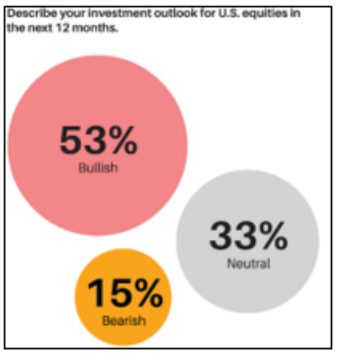
b. Which of the following states has the largest share of its total power consumed by data centers?

   i. Texas

   ii. Nebraska (Neb.)

Question 6

Consider the following figures (from the May 6, 2024 issue of *Barrons' Magazine*) which summarizes the responses by professional investment managers on their outlook on: (***left figure***) public USA companies (Bullish means they have a positive outlook, bearish means they have a negative outlook and neutral means they are neither positive nor negative.) and (***right figure***) which global public equity market they think will perform the best.



a. What can you say about the outlook of the majority of managers on USA public companies?

b. Which global public equity market is the least favored?

c. Something is wrong with these figures, can you identify it?

Question 7

For each of the following statements, state whether or not the interpretation of the statement is accurate. If not, motivate your answer.

a. Of all the people born in Gauteng, 25% of them speak Sotho. Therefore, for anyone who speaks Sotho, there is a 25% chance/probability that they were born in Gauteng.

b. The potato yield in the Limpopo province was found to be positively correlated with the coal production in the Mpumalanga province. Therefore, high potato yields lead to increases in coal production.

c. About 75% of the South African population has access to the internet. In other words, this implies that 3 in 4 South African citizens have some form of access to the internet.

1.2 Data

1.2.1 What is data?

Data are the raw facts and figures, about objects or events, that are collected, analyzed and summarized for presentation and interpretation in order to make informed decisions. Data typically arises as a result of a **study**. For instance, suppose the South African Reserve Bank (SARB) wants to forecast inflation in the next 12 months (the study). They will collect data about the inflation rate in the previous years and other factors influencing inflation. All the data collected by the SARB is referred to as a **data set**.

Table 1 gives a data set containing information about the 9 provinces of South Africa. The data was obtained from the 2022 South African census. The South African census is conducted every 10 years. A census provides information on the demographic, socioeconomic and geographic characteristics of the entire population, as well as household characteristics.

	A	B	C	D	E	F	G	H	I
1	**Province**	**Coastal (1) or Inland (2)**	**Population size**	**% of households with no internet access**	**Sex ratio**	**Median age**	**Province where most of the population migrate**	**% of population with no schooling**	**% of homeless persons**
2	Eastern Cape	1	7230204	34.3	90	27	WC	7.2	7.2
3	Free State	2	2964412	20.8	90.4	28	GP	5	6
4	Gauteng	2	15099422	13.6	101.8	30	WC	3.9	45.6
5	Kwa-Zulu Natal	1	12423907	18	91	28	GP	8.3	13.9
6	Limpopo	2	6572720	31.9	89.2	26	GP	14.1	3.4
7	Mpumalanga	2	5143324	22	92.4	27	GP	11.7	2.3
8	Northern Cape	1	1355946	28.8	93	27	WC	6.6	1.1
9	North West	2	3804548	27.6	98.2	27	GP	7.9	3.1
10	Western Cape	1	7433019	16.1	94	31	EC	2.3	17.5
11	**Source**: Census 2022, Statistics South Africa								

Figure 1.3: Table 1: Data set on information about the 9 provinces of South Africa

1.2.2 Properties of a data set

Any data set has three essential properties, namely

1.2.2.1 Elements

These are the entities or objects on which the data are collected. Each of the nine provinces in Table 1 are the elements of the data.

1.2.2.2 Variables

These are the characteristics of interest about the elements. The data set in Table 1 has eight variables:

Coastal or Inland: The province's area of location in South Africa; It can be either coastal or inland.

Population size: The number of people in a province.

% of households with no internet access: The number of households that have no access to the internet in a province as a percentage of the total population in a province.

Sex ratio: The number of males for every 100 females in a province. A value above 100 indicates that there are more males than females.

Median age

Province where most of the population migrate: A province in South Africa where most of the population in another province migrate to.

% of population with no schooling: The number of persons in a province with no formal schooling as a percentage of the total number of persons in the province.

% of homeless persons: The number of homeless persons in a province as a percentage of the total number of homeless persons in South Africa.

1.2.2.3 Observations

These are the sets of measurements obtained from each element. From Table 1, the first element (Eastern Cape) has the following measurements:

1 7230204 34.3 90 27 WC 7.2 7.2

The second element (Free State) has the following measurements:

2 2964412 20.8 90.4 28 GP 5 6

and so on. The data set with 9 elements has 9 observations.

1.2.3 Observational and experimental data

Data can come from an observational study or an experimental study. **Observational data** are records of what is actually taking place in a particular situation. The data in Table 1 are an example of observational data. As another example of observational data, a bank might observe client visits at one of their branches to collect data on variables such as the length of time a typical client spends at the branch, the number of clients visiting the branch on a given day

(e.g. Monday or last day of the month), the age of the clients and so on. Statistical analysis of this data may, for instance, help bank management decide whether or not to close the branch in order to reduce operational costs.

Experimental data is data that is obtained under controlled conditions. It is typically used to test a hypothetical statement. For instance, suppose a pharmaceutical company would like to test whether a new drug they developed is effective for weight loss. To obtain the data, researchers select a sample of individuals. The individuals are instructed to follow the same diet. One group (treatment group) of individuals is given a dose of the new drug and another group (control group) is not given the new drug. After two months, we collect data on the weight of individuals in each group and compare it with the weight data collected before the experiment. Statistical analysis of the data can help determine whether the average weight loss of the treatment group is significantly greater than that of the control group.

1.2.4 Cross-sectional and time-series data

For appropriate analysis, interpretation and communication with the collected data, a distinction must be made between **cross-sectional data** and **time series data**. Cross-sectional data are data collected at the same point in time about two or more elements. The data in Table 1 are cross-sectional because they describe eight variables about the nine provinces (elements) at the same point in time (2022). Time series data are data collected over several time periods. For instance, the figure below shows the estimated Life Expectancy at birth in South Africa for the period 2002 to 2024. Time series plots, such as the figure below, are very useful in understanding what happened in the past, identifying whether there are any trends over time and, often, forecast future values of the time series.

A time series plot is typically easy to understand and interpret. For example, from the figure, we can see that life expectancy at birth declined between 2002 and 2006. Incidentally, this was a period of increasing HIV prevalence and lack of awareness and information about prevention measures. However, due to increased awareness of HIV, expansion programmes to prevent mother-to-child transmission coupled with access to ARVs, life expectancy at birth increased from 2006 before declining in 2020 due to COVID-19.

1.2.5 Population and sample

Typically, the elements in a study consist of a large number of objects or events. For instance, suppose an accounting firm, such as KPMG, wants to determine whether the amount of accounts receivable reflected on a client's financial statement fairly represents the actual amount of accounts receivable. Usually large industrial or manufacturing businesses, such as Bidvest Group, will have a large

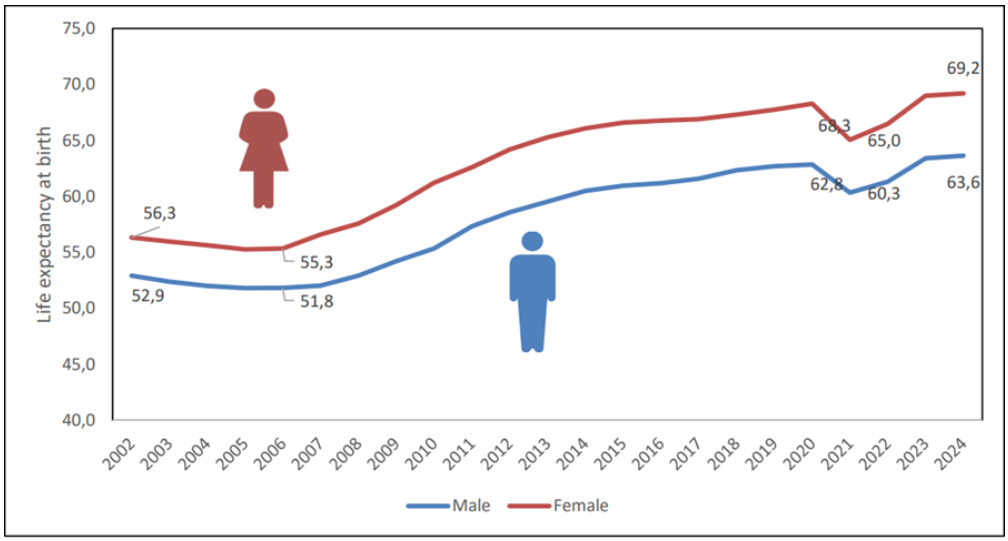


Figure 1.4: Figure: Life expectancy at birth in South Africa over time 2002 – 2024

number of accounts receivables which will make reviewing and validating every account too costly and time-consuming. As a common practice, the audit staff selects a subset of all the accounts. This subset of accounts is known as a **sample** and all of the accounts receivable are known as a **population**. The process of collecting the data is called **sampling**.

Note that, although the elements of the data in Table 1 are the provinces, the data itself was collected from individuals from the population of South Africa. Therefore, the elements in a census survey or study are a population.

1.2.6 Exercises to Section 1.2

Question 1

What is a data set?

Question 2

Name the properties of a data set.

Question 3

In 2023, FNB commissioned a retirement survey to find out how prepared South African consumers were for retirement. Is the data obtained from the survey respondents experimental or observational?

Question 4

Suppose a study is conducted to find out whether vaping (or e-cigarette smoking) is less harmful compared to tobacco smoking. Is the data obtained from the study experimental or observational?

Question 5

Figure **??** gives a bar chart showing the annual revenue of Shoprite over a ten-year period from 2014 to 2023 (The data were obtained from the annual financial statements of Shoprite[9] and the graph was plotted using Excel).

Figure 1.5: Shoprite Revenue for the period 2014 to 2023

a. What is the variable of interest?

b. Are the data cross-sectional or time-series?

c. What can you say about the trend in Shoprite's revenue overtime?

Question 6

Figure **??** shows the quarterly FNB/BER Consumer Confidence Index for the period 2014, 1st quarter, to 2024, 2nd quarter[10]. The index measures confidence

[9] https://www.shopriteholdings.co.za/shareholders-investors/financial-results-archive.html?p=1

[10] https://www.fnb.co.za/blog/investments/articles/EconomicsWeekly-20240913/?blog=investments&category=Economics&articleName=EconomicsWeekly-20240913

among South African consumers based on their outlook on the economy and their household financial position. A higher value indicates high confidence[11].

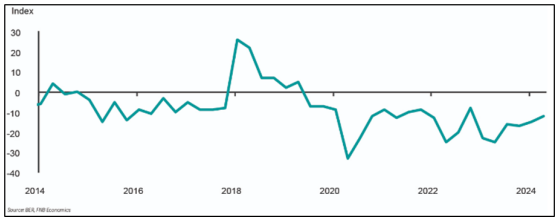


Figure 1.6: The Quarterly FNB/BER Consumer Confidence Index for the period 2014 (1st Quarter) - 2024 (2nd Quarter)

a. Are the data cross-sectional or time-series?

b. Comment on the confidence of the South African consumer over time.

Question 7

The South African Bureau of Economic Research collects data from South African adults living in predominantly urban areas to measure consumer confidence in South Africa. Survey respondents are asked about

- their expectation about the performance of the economy;
- their expectation about the financial position of households;
- their rating of the present time to buy durable goods (e.g. electronic appliances).

a. What is the population being studied?

b. Is surveying mostly people in the urban areas a good way to get a good picture of consumer confidence in a country like South Africa?

Question 8

The 2024 Annual FNB Retirement survey collected data from a sample of 1072 South African consumers to, amongst other things, uncover their preparedness for retirement. Almost half of the respondents indicated that they do not have a retirement plan in place and 20% of the respondents indicated an annual income of more than R 850 000.

a. What is the population of interest in this survey?

[11] https://www.ber.ac.za/Documents/Index/FNBBER-Consumer-Confidence-Index

b. Is the data collected from the survey observational or experimental?

c. Does this survey involve cross-sectional data or time-series data?

d. Describe any useful insights for FNB that can be obtained from the collected data.

1.3 Data Management - Foundations and Concepts

Data management is the practice of collecting, organizing, preparing, protecting and storing data so that it can be used efficiently, securely and cost-effectively in the decision-making process. In modern society, data of different types is generated in large volumes from a variety of sources at an unprecedented speed, thus a robust data management solution is important to extract meaningful and enduring value from the data. Over and above the latter, data management is important to facilitate ease of data migration and transformation and also for regulatory compliance.

1.3.1 Data management process

The data management process is made up of the following components:

- **Data collection** is the process of gathering the necessary data from the various data sources about the variables of interest for a particular study. This process typically involves a process referred to as sampling (see Chapter **??**).
- **Data organization** involves integrating different types of data, such as structured and unstructured data (see Chapter **??**). This process is also referred to as data warehousing.
- **Data preparation** involves cleaning and transforming raw data into a form that is suitable for further processing and analysis. This process is important for identifying and removing errors and duplicates in the data and also filling in missing data. This increases the accuracy and quality of the data. Data preparation is also known as data wrangling (see Chapter **??**).
- **Data governance** involves, amongst others, the processes and practices used to ensure data protection, security and privacy. **Data protection** includes safeguarding the data and restoring important information in the event of say, a data breach. **Data security** refers to safeguarding the data against theft, corruption and unauthorized access. **Data privacy** refers to safeguarding the collection, use and disclosure of personal and sensitive data to comply with policy and regulation such as the Protection of Personal Information Act (POPIA) of South Africa.
- **Data storage** involves the retention of the data for future access. In modern society, data is usually stored in a digital format using an SQL database or a spreadsheet. The data files are kept on a personal computer or, in the case of large volumes of data (so-called Big Data), on servers, also known as cloud storage.

1.3.2 The benefits of data management in modern society

At its core, the benefits of an effective data management system include:

- **Availability and visibility** – effective data management increases the visibility of the data by making it easily accessible. This in turn leads to high frequency data-driven decision-making.
- **Reliability** – a good data management system leads to accurate decision-making by making sure that the data is reliable and up to date.
- **Security** – a good data management system protects the data against loss and ransom-ware type data breaches. Moreover, it ensures that the data are used within the bounds of policy and regulation in an ethical manner.
- **Scalability** – a good data management system can allow repeatable data queries that build upon each other and thus keep the data up to date. Moreover, this mitigates inconsistencies and duplications of queries.

1.3.3 The challenges of data management in modern society

As is the case with any useful strategy, there are challenges towards effective data management. These includes, among others,

- The size (or volume) of the data collected

  As mentioned in Section **??**, data is everywhere. Given the size of the data generated today, traditional storage devices with storage capacity of up to gigabytes (GB) are no longer enough. We need data storage infrastructure with capacity up to terabytes (TB: ~1000 GB) and petabytes (PB: ~1000 TB).

- The speed (or velocity) at which the data is generated

  Since data is collected at every second of every minute, we need sophisticated infrastructure to quickly effect the changes and keep the data up-to-date for future analysis.

- The variety (or integration) of the data

  Data can come from multiple sources (e.g. social media and drone), types (e.g. structured and unstructured data) and formats (e.g. text and videos). This requires sophisticated infrastructure for data integration.

- The veracity (or quality) of the data

  Given the variety of the data coupled with the speed at which data is generated, it raises a concern over the accuracy and consistency of the information. This can lead to duplication and errors in the data.

- Changing rules and regulations

  The storage and use of data must comply with personal data protection rules and regulations while preventing cyber-attacks.

- Data security and privacy

  Protecting sensitive data while ensuring compliance with data regulations and preventing unauthorized access while ensuring data accessibility for rapid data-driven decision-making.

1.3.4 Strategies for data management

The following strategies can address some of the major data management challenges that organizations face in modern society:

- Data security and access control

  Develop a multi-layered data security system that has a robust role-based access control and data encryption system with clear audit trails.

- Data integration improvement

  Develop a robust ETL (Extract, Transform and Load) process to extract data from various sources and transform it into a standardized format which can be loaded into a central storage system.

- Data quality improvement

  Develop data validation rules, data profiling processes (such as analyzing and assessing the data to gain insights into its consistency and completeness) and error detection and correction procedures.

- Data storage and cost optimization

  Implement a tiered data storage solution with a balance between on-premises and cloud in order to optimize the use of storage and reduce data processing costs.

- Data speed optimization

  Develop a data indexing system for quick data access. Implement caching[12] strategies to reduce the number of database queries and improve the response times.

[12] Caching is to temporarily store data so that future requests for the data can be accessed faster.

1.3.5 Exercises to Section 1.3

Question 1

What is data management?

Question 2

The National Health Laboratory Services experienced a data breach which led to delays in processing laboratory tests across public health facilities in Gauteng.

a. Which one of the following actions should be taken to ensure that the data can be quickly recovered should this happen in the future?

   i. Improve data privacy
   ii. Improve data protection
   iii. Improve data storage
   iv. Improve data security
   v. Optimize data integration

b. Which one of the following actions should be taken to ensure that patient's personal information is not compromised should this happen in the future?

c. Improve data privacy

ii. Improve data protection

iii. Improve data security

iv. Improve data storage

v. Optimize data integration

Question 3

For each of the following scenarios, what do you think is likely to become a challenge in data management? Motivate your answer.

a. TymeBank, a digital bank, is reportedly on-boarding up to 5000 customers every day (that is, about 150000 customers every month).

b. A startup bank insurance (bancassurance) company collects its client's data from cameras in their homes, smart watches and banking transactions.

c. A major South African bank and a life insurance company decided to merge their businesses. When it came to combining their client's data, they failed to notice duplicates because of different data formats.

d. The University of Pretoria and University of Johannesburg are collaborating on a clinical trial for a new cancer drug. Researchers on both sides tend to store

sensitive data on their personal computers and they use different names for the data files.

e. Researchers at the United Nations Convention on Climate Change are conducting an environmental monitoring study. Their sensors are collecting data in different formats, they have no backup system for their data and there is no access control on the data.

1.4 Data Ethics

Data ethics refers to the principles or practices that seek to preserve the trust of the owners of data, from how the data is collected to how it is stored and used. In essence, data ethics concerns the measures put in place to ensure that the data are handled appropriately throughout the data management process. This is an important issue in modern society given value and ubiquitous nature of data.

1.4.1 The importance of data ethics

The following real-world examples are meant to demonstrate the consequences of unethical data management thus highlighting the importance of embedding ethical data management principles:

- In June 2018, Liberty Holdings, Africa's largest life insurance company, suffered a major data breach in which hackers access confidential client information. This raised the concern of how organizations protect and store their client's data.
- In September 2021, the South Africa Department of Justice experienced a ransomware attack in which the department's IT systems were compromised affecting all electronic services such as bail services, email and the departmental website. This attack raised concerns over whether government systems are adequately secured, given that they manage sensitive citizen data.
- In October 2017, the Master Deeds experienced a massive data leak that exposed approximately 60 million South African citizens' personally identifiable information (PII) such as ID numbers, contact details and addresses. The data was later found on a public and unsecured server. This incident raised concerned over people's right to privacy. Furthermore, this revealed a poor or no strategy for data governance.

These examples are a small snapshot of the poor management of data and its consequences for the owners of the data. Thus, ethical data management principles are important in order to:

- protect customer and, in general, human rights.
- protect customer or client loyalty to your business and society as a whole.
- ensure regulatory compliance and avoid penalty costs.
- reduce and, at best, prevent data breaches.

Regardless of who you are, the significance of data ethics is apparent. Although data is a powerful asset that can be used to drive innovation and improve lives, without ethical safeguards, it can also be misused and thus leading to harm.

Understanding data ethics enables us to better navigate how our information is used, ensuring that we can protect ourselves while still engaging in a data-drive society.

1.4.2 Data ethics principles

The following are some of the fundamental principles that can guide ethical data management:

- **Ownership:** Each individual's personal information is owned by themselves. It is therefore unlawful and unethical to collect information about an individual without their consent. It can in fact be considered stealing. Consent can be obtained from individuals through written agreements, agreeing toterms and conditions and accepting cookies on websites.
- **Transparency:** The individual whose data is collected has the right to know how it will be stored and used. It is therefore important for a company to publish a data policy documentation that will explain to the individuals how the data will be stored, why it is collected and how will it be used.
- **Privacy:** It is important that the company collecting and using personal information of individuals ensure that the information is kept private. Just because the individual gave the company consent to collect personal information, does not mean they want the information to be made public. Such information includes names, surnames, home address, contact information etc. The company should ensure that the data is securely stored so that it cannot end up in the wrong hands through hacking. When working with the data it can also be anonymised by removing the personal information so that an individual cannot be identified through the data.
- **Intention:** Before collecting data, one should clearly state why they need the data, what they will gain from using it and what possible changes, if any, will they make after making use of the data. If your intentions are to use the data to cause harm or for any other bad reason, it is unethical. Therefore, when collecting data it is important to do so with good intentions. Also, do not collect any data that is not necessary for the end goal.
- **Accountability:** The company collecting the data must take responsibility for the data collected including protecting it from data breaches and misuse. This important for maintaining trust between the company and its clients.

- **Bias:** The data as well as the algorithms used for the analysis should not have any inherent biases that will skew the results. Such biases can include amongst others racial, gender and socioeconomic biases.

1.4.3 Exercises to Section 1.4

Question 1

Which of the following is NOT a key principle of ethical data use?

a. Privacy

b. Bias

c. Profit maximization

d. Transparency

Question 2

What does "informed consent" mean in the context of data collection?

a. Data subjects must be informed about the specific use of their data and must voluntarily agree to it.

b. Data subjects must be forced to share data if it benefits society.

c. Data subjects should be informed only after data collection has taken place.

d. Data can be collected without consent if it is anonymised.

Question 3

What is "data minimization" in data ethics?

a. Limiting data collection to the minimum amount needed to achieve the stated purpose.

b. Deleting data after analysis to reduce storage costs.

c. Sharing data only with third parties who minimize its use.

d. Using smaller datasets for faster processing.

Question 4

Explain the concept of "bias" in data and why it can lead to unethical outcomes.

Question 5

What are the potential ethical concerns with using personal data collected for one purpose (e.g., marketing) for a different purpose (e.g., medical research)?

1.5 Data Exploration – Foundations and Concepts

Data exploration is an important first step in the data analysis process. It involves uncovering the characteristics, patterns and hidden insights in the observed data set in an effort to gain a deeper understanding of the data. This step is usually referred to as exploratory data analysis (EDA). During EDA, statistical techniques are used to look for similarities, patterns, anomalies and also identify relationships between the different variables in the raw data. This step is also important for investigating the quality of the data by identifying missing values, duplicate data entries and\or inconsistencies in the data. Data exploration can be seen as part of, or as following, the data preparation step in the data management process.

1.5.1 Approaches to exploratory data analysis

- **Descriptive statistics** are the main numerical features of the data such as the central tendency of the data (using the mean or median) and the spread of the data (using the variance or standard deviation). For instance, suppose you are a quantitative analyst working for a bank and you have a spreadsheet of the loan amounts given out during the current financial period. A descriptive analysis, such as calculating the mean and standard deviation of the loan amount, can give you the average loan amount and the dispersion in the loan amounts, respectively. As another example, suppose we want to understand the relationship between wind speed and air temperature, we can calculate the covariance to identify whether there is a relationship between these variables, and if so, is it positive or negative. Moreover, we can calculate the correlation coefficient to quantify the strength of the relationship.

- **Tabular and graphical statistics** are summaries and interpretations of the data using graphs and tables. This is useful for identifying trends and relationships between numerical variables (using scatter plots) and categorical variables (using cross-tabulations), distributional patterns (using histograms, frequency distributions and the ogive) and outliers (using a box-and-whisker plot). For instance, the figure below shows a box and whisker plot of a sample of data. It can be seen from the figure that the data point marked “Outlier” might be unusual in relation to the rest of the data.

- **Inferential statistics** uses sampled data to draw conclusions about the population. This includes testing various hypotheses statements you may have about the population where the data was obtained such as whether

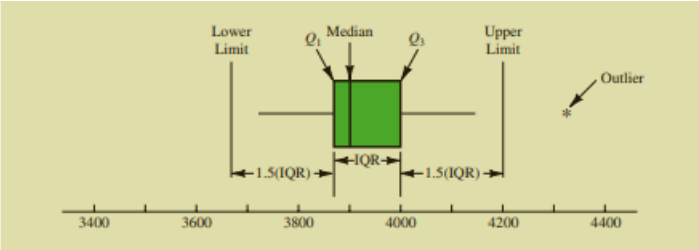


Figure 1.7: A box-and-whisker plot showing an outlier

certain variables follow a specific distribution or whether the observed relationship between two or more variables, from a graph, is significant. For instance, suppose that you are a quantitative analyst for a bank and you want to test the hypothesis that the average loan amount for male clients is large than that of female clients. This type of test can be easily performed using the Data Analysis tool in Microsoft Excel.

1.5.2 Advantages and disadvantages of data exploration

Data exploration has several benefits, including:

- Offering a comprehensive understanding of the data set before conducting the actual analysis.
- Enhancing the quality of the data.
- Highlighting important features and potential issues
- Providing insight into appropriate analysis techniques
- Guiding future research questions and directions

Limitations of data exploration may include:

- Difficulty visualizing the high-dimensional data
- May become complex for complicated data structures
- It can be time consuming and subjective.
- Misrepresenting data by choosing the wrong summary indicators

1.5.3 Stages in the exploratory data analysis process

The data exploration process involves a series of stages or steps which are aimed at obtaining a comprehensive understanding of the data and the implications thereafter. This includes

1. Understand the problem understudy

   A crucial first step in the EDA process is to clearly state the problem that led to the collection of the data. This will help in various ways, such as formulating relevant questions, choosing the appropriate analytical tools and identifying inconsistencies or anomalies in the data.

2. Examine the structure of the data

   Once you have a clear understanding of the problem, the next step is to familiarise yourself with the data by examining its structure which includes the size of the data, the number of variables and their data types. Check for inconsistencies and anomalies in the data and also missing values or data entries.

3. Handle the inconsistencies, anomalies and missing values

   The next crucial step is to have a strategy to handle any inconsistencies, anomalies and\or missing data entries identified in the second step.

4. Examine the statistical aspects of the data

   After addressing the issues identified in the second step, the next step involves the examination of the distribution, central tendency and variability of all the numerical variables. Moreover, various assumptions about the data can be tested. Among other things, this exercise will assist in identifying variables that deviate from expected patterns and may need further processing.

5. Transform the data for analysis

   The next step is to prepare the data to be ready for data analysis. Based on your insights from the previous step, you may have to apply certain transformations to the data to make them conform to expectations. This involves, among others, standardizing or normalizing numerical variables, taking the log or square root to correct for skewness in numerical variables and dummy coding categorical variables.

6. Visualize the data

   After transforming the data, you can now visualize it using tables and graphs such as frequency distributions and histograms. Data visualization is covered in some detail in Section **??**.

7. Communicate the findings and use them for further data analysis.

1.5.4 Exercises to Section 1.5

Question 1

What is data exploration?

Question 2

For each of the following scenarios, identify the approach used for data exploration:

a. A research team analysed the health data of City of Tshwane residents. They found that 12% of the residents have at least one non-communicable disease.

b. An environmental scientist analysed data on monthly carbon dioxide (CO_2) emissions and the monthly number of people hospitalized for respiratory issues in Pretoria. A scatter plot showed that there is a positive relationship between these two variables.

c. A fitness instructor analysed data on the exercise patterns of people in the district of the Cape Winelands in the Western Cape. She plotted the data on a histogram and found a bimodal distribution of exercise habits, suggesting that there are two distinct behavioural groups.

d. A marine biologist analysed data collected over 10 years by sensors from the bottom of the Indian ocean. She found that the average temperature was 24 degrees Celsius.

e.A hospital administrator at Hatfield General hospital claims that less than 10% of the daily hospital admissions are for serious injuries. In an analysis of hospital data, he was able to confirm his claim.

1.6 Data visualization – Foundations and Concepts

According to the Harvard Business Review, data visualization is a must-have data literacy skill for junior and even senior management. This is because it provides the only way for them to make sense of the work they do. As the world becomes more complex, most problems are increasingly hard to understand, much less fix, if they cannot be visualized. In modern society, organizations and individuals can use data visualization to generate and illustrate ideas or discover patterns, trends and outliers in the data.

Data visualization is the graphical representation of data using charts, graphs, maps, animations and infographics. The goal of visualizing data is to clearly and effectively communicate the key characteristics of a data set in a way that is easy to understand.

1.6.1 Common uses of data visualization in modern society

- **Comparison and benchmarking** – visualizing data can allow us to compare observations, variables or time periods. For instance, a meteorologist might want to compare the amount of rainfall before and after the first industrial revolution.
- **Monitoring and tracking** – Data dashboards (such as Tableau® and Microsoft's Power BI) can be used to monitor key performance indicators (KPIs) in a company or organization in a manner that is easy to read, understand and interpret. For instance, in order for the manager of a local Shoprite store to ensure enough merchandise on hand, the manager can refer to a real-time data dashboard showing the hourly sales volumes, inventory on hand, hourly number of transactions processed.
- **Data exploration** – as already mentioned in Section **??**, data visualization is a core step in EDA.
- **Hypothesis testing** – results from visualizing data can be used as an informal approach to formulate and test hypotheses. For instance, the plot in the figure below can lead us to hypothesize that there is a positive linear relationship between the volumes of electric vehicle sales (in R millions) and the number of television advertisements.

- **Educational and knowledge sharing** – visualizations can simplify complex topics and make information more easily accessible to the general public in order to support education and training.

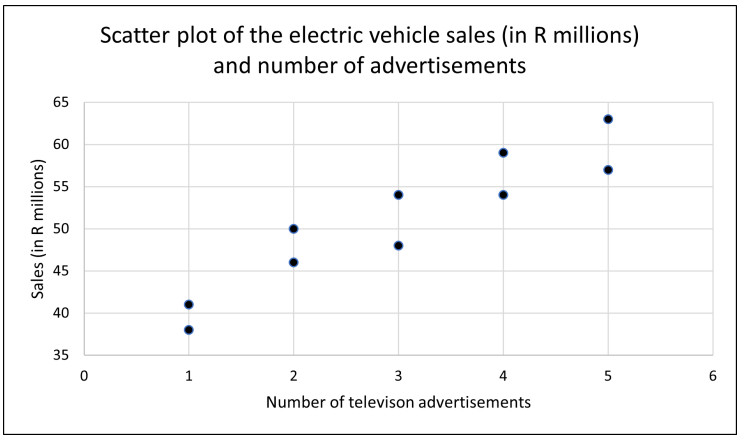


Figure 1.8: Scatter plot of electric vehicle sales (in R millions) and the number of advertisements

1.6.2 Types of data visualizations

The choice of a data visualization tool depends on the purpose of the study and the type of data available.

1. For the purpose of visualizing the distribution of the data, we can make use of a
    a. **Frequency distribution** – show the number of observations in each of several non-overlapping categories or classes for both numerical and categorical data.
    b. **Bar chart** – to show the frequency distribution for a categorical variable.
    c. **Pie chart** – to show the relative frequency distribution for a categorical variable.
    d. **Dot plot** – to show the distribution of a numerical variable over the entire range of the data.
    e. **Histogram** – to show the frequency distribution of a numerical variable over a set of class intervals.
    f. **Stem-and-Leaf** – to show the rank order and shape of the distribution of numerical data.
    g. **Box-and-Whisker plot** – to show the distribution of numerical data using five numbers calculated from the data.

2. For the purpose of identifying whether or not the data has a trend overtime, we make use of a

   a. **Line chart** which can be used to visualize time series data.

3. For the purpose of making comparisons between two or more variables in a data set, we can make us of a

   a. **Multiple boxplots** – to compare the distributions of two or more numerical variables.

   b. **Side-by-side or stacked bar charts** – to compare two categorical variables.

   c. **Pivot-tables (crosstabulation)** – to compare the frequency distribution of two categorical distributions.

4. For the purpose of describing the relationship between numerical variables we can make use of

   a. **Scatter plots** to (1) represent the relationship between two numerical variables, (2) identify the type of pattern represented in the scatter plot (linear constant upward or downward trend, curved pattern or no apparent pattern at all); (3), if there is a pattern, determine how strong is the pattern (do all the points follow the pattern exactly or not? In Section **??**, we will define a numerical measure that can quantify the strength of a linear pattern) and (4) identify if there are any unusual observations (points that are far from the cluster of the majority of the points).

5. For the purpose of displaying geographical data, we can make use of

   a. **Heat map** – to show the intensity or density of a variable across a geographical area using a color gradient. These maps are effective for visualizing, for instance, crime hotspots.

   b. **Flow map** – to show the direction and magnitude of the movement, migration or flow of people, goods or information between different locations.

   c. **Choropleth map** – to represent quantitative or qualitative data associated with geographic regions such as countries. In contrast to a heat map, in a choropleth map, the geographic regions are not based on the variable of interest but are chosen based on known spatial information. These maps are effective for visualizing data on, for instance, population density or income levels.

   d. **Cartograms** – to show the relative influence or importance of different geographical regions based on some quantitative variable such as population size. These visual tools work by distorting the size of the geographical regions by making them proportional to the numerical value of the variable of interest. For instance, geographic regions with larger population sizes will be larger on the map.

e. **Point maps** – to show the location of a specific outcome, such as the site of a wildfire, on the map.

f. **Bubble maps** – to show the relative magnitude of a specific outcome, such as the impact of a wildfire at a specific location, on the map. The larger the bubble, the higher the impact.

1.6.3 Components of a data visualization tool

In order for a visual summary to effectively communicate the message behind the data, it must possess the following core features:

- An appropriate descriptive title that explains the data being shown.
- Clear labels with the units of measurements and appropriate scales for both the x-axis (horizontal) and y-axis (vertical).
- A legend identifying the different data series, where appropriate.

The figure below shows a graph of the 14-day Covid-19 infection rate in the South Africa provinces of Gauteng and Kwa-Zulu Natal for the period 1 December 2020 to 28 February 2021.

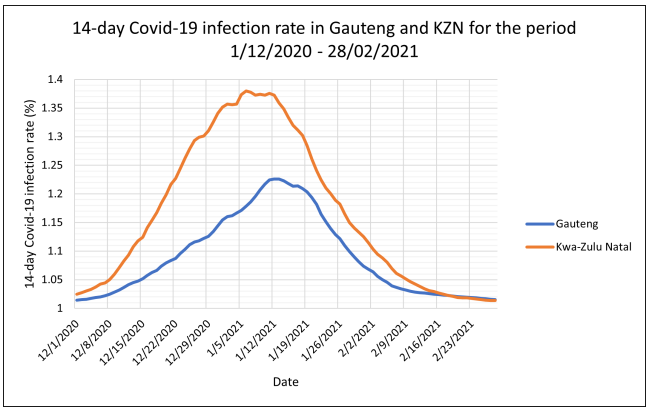


Figure 1.9: The 14-day Covid-19 infection rate in the South African provinces of Gauteng and Kwa-Zulu Natal for the period 1 December 2020 to 28 February 2021

Figure **??** is a copy of Figure **??** highlighting the core features of a graph.

1.6.4 Advantages and Disadvantages of data visualization

Data visualization can be an advantageous skill to have because

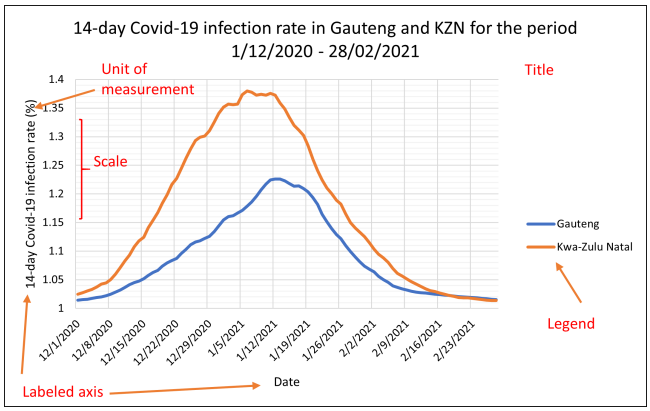


Figure 1.10: Same as Figurereffig:covid-1

- it appeals to any audience which means you can communicate the data to anyone;
- it allows you to communicate efficiently and effectively about data;
- quickly detect patterns or anomalies in the data;
- it facilitates timely decision-making;

While there are many obvious advantages to data visualization, there are also some less obvious disadvantages to data visualization such as:

- using the wrong visualization tool;
- concluding from a scatter plot that the observed correlation implies causation;
- making biased conclusions.
- Due to their ease of apprehension, they can often be used to spread misinformation.

1.6.5 Exercises to Section 1.6

Question 1

What is data visualization?

Question 2

For each of the following scenarios, specify the appropriate visualization technique (s):

a. The City of Tshwane is concerned about income inequality, the mayor wants to investigate the distribution of income.

b. In an effort to efficiently allocate police personnel, the City of Johannesburg wants to know which areas have the most criminal incidents.

c. The forestry, fisheries and the environment ministry of South Africa wants to compare the greenhouse gas emissions across industries in the primary, secondary and tertiary sector.

d. The Western Cape province has noted an increase in the number of people from other parts of South Africa, the administration wants to know what is the province of origin for most of the migrants?

e. A botanist wants to understand the effect of temperature on the rate of growth of a plant.

Question 3

For each of the following scenarios, specify whether or not the chosen visualization technique is appropriate, if not, state the reason and specify an appropriate technique:

a. A financial analyst uses a pie chart to identify the trend in the rand-dollar exchange rate for the period 2000-2020.

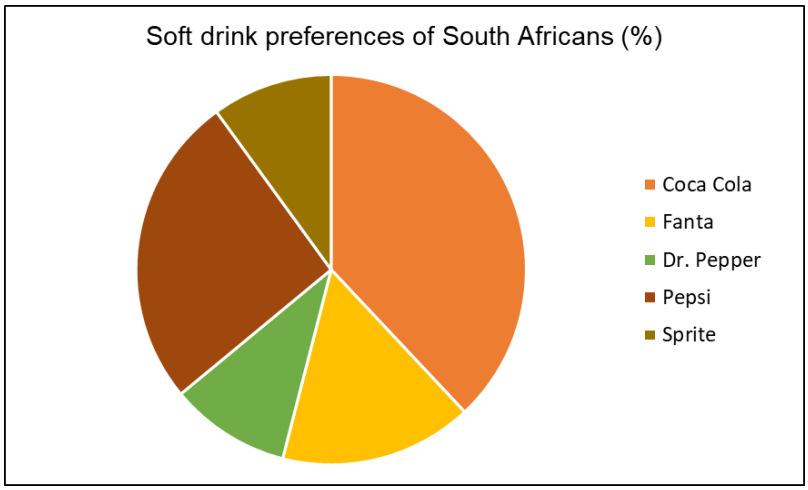
b. A psychologist uses a box-and-whisker diagram to identify outliers in terms of IQ.

c. A zoologist uses a line chart to compare the territory sizes between male tigers and female tigers.

d. A sociologist uses a bar chart to study the relationship between an individual's years of education and their voting frequency.

Question 4

The following is a pie chart of the soft-drink preferences of 50 individuals in South Africa:



Identify what is wrong with the above visualization.

1.7 Data analysis – Foundations and Concepts

Data analysis is the process of systematically using statistical techniques to explore, clean, transform and model data with the purpose of discovering useful information which can then be used to support decision-making.

1.7.1 The Data Analysis process

Data analysis can be viewed as a sequence of steps combining some of the data literacy skills we have already discussed so far in this book. This process can be summarized as follows:

1. **Define the study**

   The first and important step in the data analysis process is to clearly define the problem that the analysis aims to address by stating the objectives and specific questions that the analysis aims to answer.

2. **Data collection**

   Once the study is defined, the next step is to collect the relevant data. This can be done using various methods to be discussed in Chapter **??**. The choice of method used to collect data depends on the nature of the problem and the questions being asked.

3. **Data cleaning**

   After collecting the data, the next step is to clean the data. This involves identifying and rectify errors, missing values and inconsistencies in the data. Data cleaning will be further discussed and practically demonstrated in Chapter **??**.

   Note: The second and third step are part of data management. See Section **??** for more details.

4. **Data exploration**

   After cleaning the data, we conduct a preliminary analysis to understand the characteristics of the data. See Section **??** for more details on how this is done in practice.

5. **Data transformation**

   Following from the results of the exploration phase of the data analysis, the data is prepared for analysis by encoding categorical variables, scaling (normalizing or standardizing) some numerical variables and, if necessary, handling outliers.

6. **Data modelling (analysis)**

   Now the data is ready for the actual analysis. This step involves using statistical and mathematical techniques on the data to discover patterns, relationships, similarities or trends.

7. **Interpretation and visualization**

   After the data analysis, the next step is to interpret the obtained results and present them in a manner that can be easily understood. This can be done through visualization.

1.7.2 Approaches to Data Analysis

There are different types of data analysis and each one serves a unique purpose. The choice of which one to use will depend on the nature of your study and what kind of questions you seek to answer.

1. **Descriptive analysis** is used to describe and summarize the collected data to understand what happened in the past. For example, a university might use descriptive analysis to find out how many first-year students passed last year.
2. **Diagnostic analysis** follows descriptive analysis by going a step further to explain or diagnose why something happened. For example, suppose the number of first years that passed last year dropped significantly, diagnostic analysis can be used to find out why this happened.
3. **Predictive analysis** is used to forecast or predict what might happen in the future based on historical data. For example, a university can use predictive analysis to predict the number of first-year students that will pass next year. In other words, based on what has happened in the past we can find out what could happen in the future.
4. **Prescriptive analysis** is used to make recommendations on what course of action to take in order to reach a desired outcome. For example, suppose a university wants to increase the pass rate for first-year students, a prescriptive analysis might suggest the best course of action towards reaching this outcome.

1.7.3 Uses of data analysis in modern society

Data analysis is a very important data literacy skill that finds application across various fields and domains of application such as:

- **Marketing research**, where it can be used to assist businesses to understand market trends, consumer preferences and help identify opportunities for product development.

- **Medical diagnosis**, where it can be used to interpret medical images (e.g. MRI scans) and also assist in early detection of a disease.
- **Medical drug discovery**, where it is used by pharmaceutical companies, such as Johnson and Johnson, to develop a drug by conducting clinical trials and testing the effectiveness of the drug.
- **Fraud detection**, where it can be use by banks to identify unusual transaction patterns and detect fraudulent activities.
- **Risk management**, where it is used by financial services companies to assess a client's credit risk (i.e. will they be able to pay back the loan) and model risks in the forex market or stock market.
- **Quality control**, where it is used to monitor and control the quality of products on the production line.
- **Social science research**, where it is used to analyze survey data to study overall human behavior and sentiment.
- **Recommendation systems**, where it used by platforms such as Spotify and Netflix to recommend music or shows that you might like based on the content that you viewed in the past.
- **Environmental monitoring**, where geographical (remote sensing) data is used to monitor ecological changes such as deforestation, water quality and air pollution.

1.7.4 Data analysis techniques

There are many techniques used in data analysis and each one serves a unique purpose and application. In section [Data exploration], we discussed data exploration. This is the most basic technique for data analysis. In this section, we will briefly discuss some of the most commonly used and emerging techniques.

1. **Correlation analysis**

   Correlation analysis is a technique used to understand the linear relationship between two or more numerical variables. A simple measure that is commonly used to describe this relationship is the **correlation coefficient**, usually denoted by r. The correlation coefficient is used to quantify the strength of the linear relationship between two numerical variables. The value of r will always lies between -1 and 1. Values close to -1 or $+1$ indicate a strong linear relationship. The closer the value of r is to zero, say less than 0.5 or more than -0.5, the weaker the linear relationship. As an example of the use of the correlation coefficient, suppose we want to know the strength of the relationship between a student's matric final

mark and their final mark at the end of their first-year of study at a university. Given a correlation coefficient of $r = 0.93$, we can say that there is a strong positive linear relationship between a student's matric final mark and their first-year final mark. In other words, a larger final matric mark is strongly associated with a larger first-year final mark. Finally, please note that correlation measures the linear association between two numerical variables and not necessarily causality. In other words, a high correlation between two variables does not mean changes in one variable will cause changes in the other variable.

2. **Regression analysis**

   Regression analysis is a statistical technique used to understand the dependence of one variable, known as a dependent variable, on one or more other variables, known as independent variables. It is commonly used for predictive analysis. A simple and widely applicable approach to regression analysis is the least squares line. Consider two numerical variables x and y which are assumed to follow a straight line pattern. The relationship between x and y can be described using a straight line given by the equation

$$y = A + Bx \tag{1.1}$$

where y is the dependent variable, assumed to depend on x known as the independent variable. The term A is the y-intercept and B is the slope which represents an increase in y for every unit-increase in x. For a given sample of (x, y) points, we can obtain an estimate of (**??**), given by,

$$\widehat{y} = a + bx \tag{1.2}$$

where a and b are estimates of A and B, respectively, obtained by the method of least squares. Hence, equation (**??**) is known as the least-squares regression line.

Equation (**??**) can be used to:

- describe the dependence of y on x allowing us to learn more about the process that produces y.
- comment on the type of linear pattern between x and y (whether its positive, $b > 0$, negative, $b < 0$, or no pattern exists, $b = 0$).
- measure the influence that x has on y based on the magnitude of the value of b.
- predict the future value of y for a given value of x.

As an example of the use of the least-squares line for regression analysis, consider the least squares line $\hat{y} = 60 + 5x$ estimated to a data set on student population (in 1000s), x, and quarterly pizza sales (R 1000s), y, for a sample of 10 restaurants located near university campuses. The following points can be made about the fitted least-squares line:

- $b = 5 > 0$ which implies that as student population, x, increases, quarterly sales increase.
- $a = 60$, which means for a restaurant that is not located close to a university (that is, $x = 0$), the quarterly sales are $R60000$.

Lastly, we can use the least-squares line to predict the quarterly sales for a given size of the student population. For $x = 16$, representing 16000 students, the quarterly sales are predicted to be $\hat{y} = 60 + 5(16) = 140$ or $R140000$.

3. **Cluster analysis**

   Cluster analysis is used to group a set of objects or entities in such a way that objects in the same group (or cluster) are more similar than those in other clusters. It is commonly used in recommendation systems and market research to find consumers with similar preferences.

4. **Dimension reduction**

   As implied by the name, this technique is used to reduce a large number of variables into few variables in such a way that the remaining variables capture the maximum possible information from the original variables. It is commonly used in conjunction with some of the already-mentioned techniques such as cluster analysis for medical diagnosis.

5. **Hypothesis testing**

   This technique is used to make inference or statements about population characteristics (such as the mean) using sample data. It is commonly used in the control and monitoring of quality in a production process.

6. **Time series analysis**

   This technique is used for the analysis of time-series data. It is commonly used for predictive analysis and understanding the trend overtime. It is commonly used in forecasting or predictive analysis.

7. **Sentiment analysis**

   This technique is used to extract the emotional tone (negative, positive or neutral) behind text data. It is commonly used to understand customer feedback.

8. **Spatial data analysis**

   This technique is used for the analysis of geographical (remote sensing) data, that is data with a spatial component (e.g. geographic location represented by coordinates). The most important use of this technique is in disease tracking to identify hotspots.

1.7.5 Computational tools and software for data analysis

Modern data analysis is carried out using sophisticated computational tools and software that cater for different needs and levels of expertise. This includes

1. **Python**

   Python is the most popular high-level general-purpose programming language that can be used for a variety of tasks including data analysis. It is relatively easy to learn and has a range of libraries (pandas, NumPy and Matplotlib) that make it a favorite data analysis and visualization tool among data analysts and data scientists.

2. **R programming language**

   R is an open source (free software) programming language developed specifically for statistical computing and visualization. Overtime, R has evolved to have a wide set of capabilities such as statistical software development and scientific text editing. It is a popular software among statisticians because it features tools such as hypothesis tests, correlation analysis, regression analysis, among many others.

3. **SQL (Structured Query Language)**

   SQL is a language for managing and manipulating data that is sored in databases.

   Note that it is not necessary to know programming in order to do data analysis. The following are the most popular non-programming data analysis tools used in industry:

4. **Excel**

   Microsoft Excel is a spreadsheet that is most widely used for data analysis because it is easier to use. It offers a range of features for data collection (using the Sampling tool), exploration (using the Descriptive Analysis tool), modelling (using the Regression Analysis tool) and visualization (using the Charts tool).

5. **SAS (Statistical Analysis System)**

   SAS is an advanced license-based software developed specifical for statistical data analysis and visualization. It is made up of procedures that can perform tasks such as data exploration (PROC MEANS), hypothesis testing (PROC TTEST) and many others.

6. **Power BI**

   Power BI is a powerful business analytics tool developed by Microsoft. It enables users to create interactive visualizations with self-service business intelligence capabilities. Power BI is used to transform raw data into useful insights that are easy to understand through dashboards and reports.

7. **Tableau**

   Tableau is a business analytics tool used to create interactive and shareable dashboards that show trends, variations and densities for important day-to-day business metrics through charts and graphs.

1.7.6 Exercises to Section 1.7

Question 1

For each of the following case studies, specify which type of data analysis is appropriate:

a. Eskom wants to reduce overall electricity waste and improve the stability of the national electricity grid. To achieve this, they want to propose energy-saving strategies to its customers.

b. Absa has noticed a decrease in their corporate clients overtime. They want to know what could be behind this.

c. In order to inform her decision on many warm clothing to bring on her trip to Essen, Germany in January, Renate wants to know what the average temperature will be in Essen, Germany in January.

d. In order to inform their decision on the interest rate at their next meeting, the South African Reserve Bank wants to know what the inflation rate will be in the next 12 months [predictive].

e. A general practitioner (GP) wants to know how many COVID-19 patients she treated between 2020 – 2022.

Question 2

For each of the following case studies, specify which data analysis tool (s) is/are appropriate:

a. Given past data on electricity consumption, Eskom wants to determine the amount of electricity that will be consumed in the next winter season.

b. A plant physiologist wants to understand dependence of plant growth on factors such as water availability, temperature and soil nutrient levels.

c. Suppose that you want to study the response of a plant to changes in temperature, drought, salinity.

d. A biochemist wants to classify proteins with similar structural characteristics.

e. The World Health Organization (WHO) wants to identify geographic hotspots for the monkeypox disease.

f. The WHO wants to understand public opinion about a new strain of virus from social media posts.

g. In order to inform their decision on the interest rate, the South African Reserve Bank (SARB) wants to forecast the average value of the rand per US dollar in the next 12 months.

Chapter 2

Sources and types of data

Add content here.

Chapter 3

Data Handling in Excel

In this chapter, you will be introduced to some of the fundamental skills needed to work with data in Excel. Excel is a powerful tool enabling us to organise, analyse and visualise data efficiently. This chapter will not only enhance your ability to manage data but also lay the foundation for the more advanced data analysis.

Topics covered in this chapter includes:

- Data importing
- Data cleaning
- Data manipulation

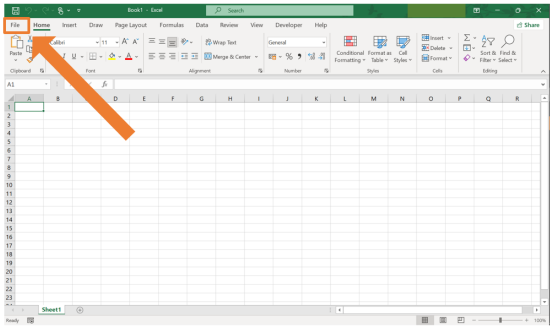
3.1 Data importing

In this subsection, the focus is on importing data into Excel from different sources such that we can start working with it. This is a crucial first step when doing data analysis.

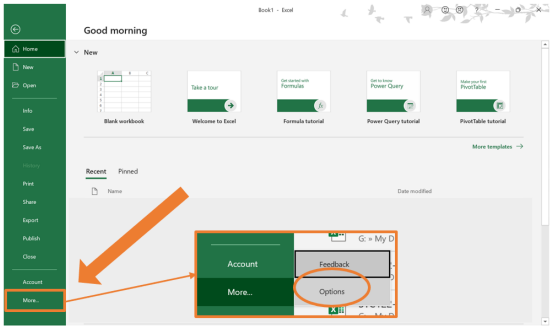
3.1.1 Change decimal separator settings in Excel

On most devices, the default setting in Excel is to use a comma as a decimal separator. This causes some difficulties especially when working with .csv (comma separated value) files.

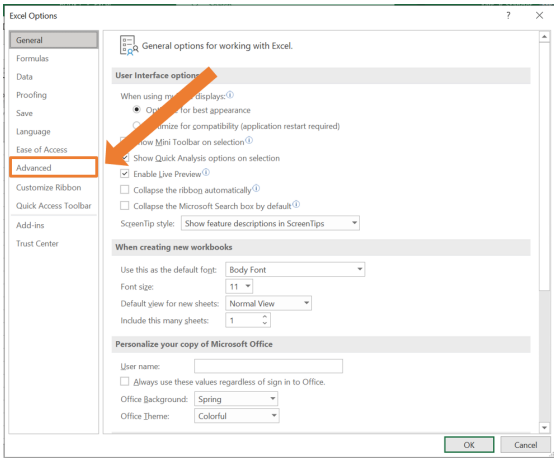
1. Go to the **File** tab.



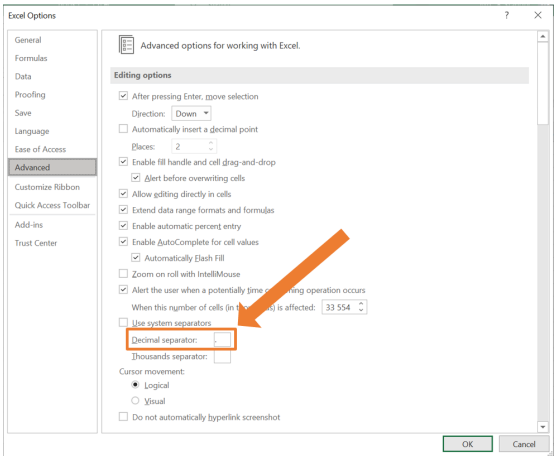
2. At the bottom of the menu at the left, select **More...** and then **Options**.



3. The Excel Options window will open. Select **Advanced** on the menu at the left.



4. Under **Editing Options** there is a you can specify the **Decimal separator**. In the box, make sure to have a full stop (`.`) instead of a comma (`,`).

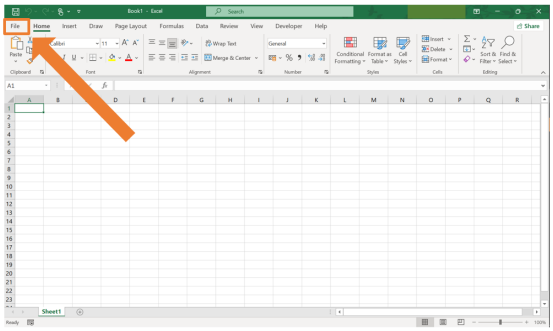


5. Click **OK** to save the changes.

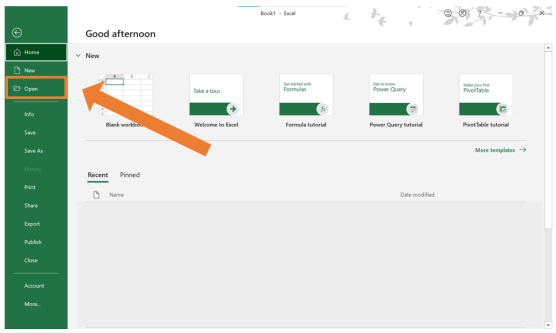
3.1.2 Opening a .xlsx file in Excel

A file with the `.xlsx` extension is a standard Excel file. Such files can be opened in Excel effortlessly. The following steps can be followed:

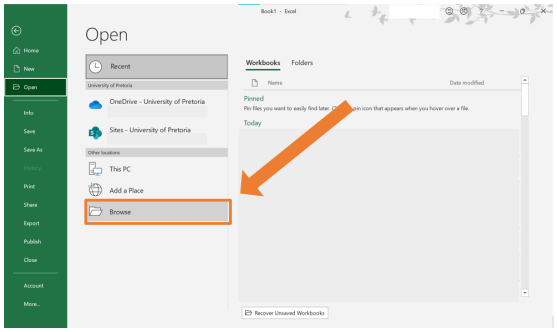
1. Go to the **File** tab.



2. Select **Open**.



3. Click on **Browse**.



4. Navigate to the location of the file you want to import and select it.

3.1.3 Importing a .txt file in Excel

Some external programs can export data in a text (`.txt`) file. If you wish to do some data manipulation or work with the data in any other way, you will need to import the text file in Excel.

When data is stored in a text file, various symbols (known as delimiters) are used to indicate the separation between columns. These delimiters help structure the data so that tools like Excel or programming languages can interpret it accurately. Here's an explanation of common delimiters used in text files:

1. Tabs

```
Name    Age     Occupation
John    28      Engineer
Alice   35      Designer
```

2. Commas

```
Name,Age,Occupation
John,28,Engineer
Alice,35,Designer
```

3. Semicolons

```
Name;Age;Occupation
John;28;Engineer
Alice;35;Designer
```

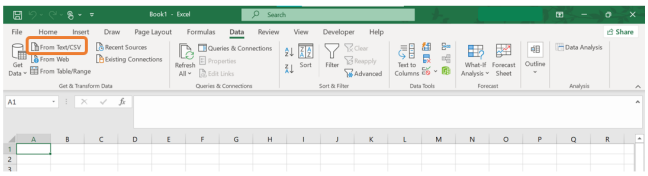
4. Pipes

```
Name|Age|Occupation
John|28|Engineer
Alice|35|Designer
```

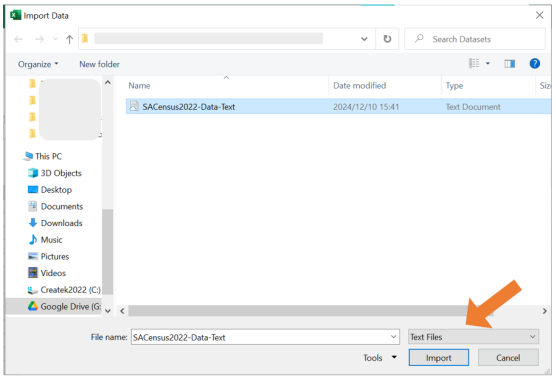
3.1.3.1 Example

You have obtained access to a data set available exclusively in text format, named `SACensus2022-Data-Text.txt`. To work with this data set, you need to import it into Excel. Follow these steps to complete the process:

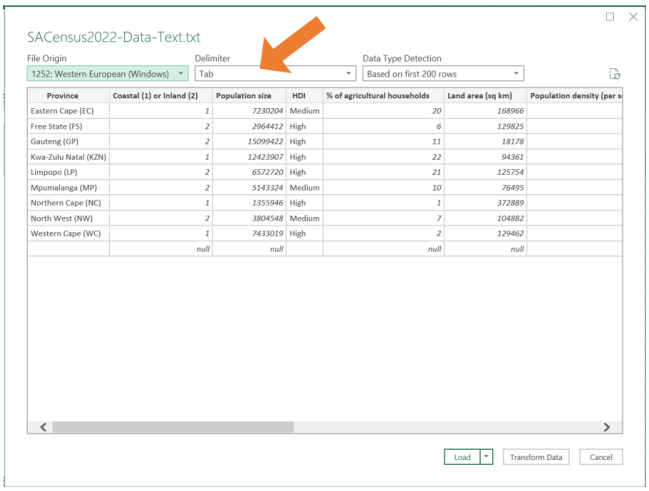
1. Open Excel and navigate to the **Data** tab. In the **Get/Transform Data** group click the **From Text/CSV** button.



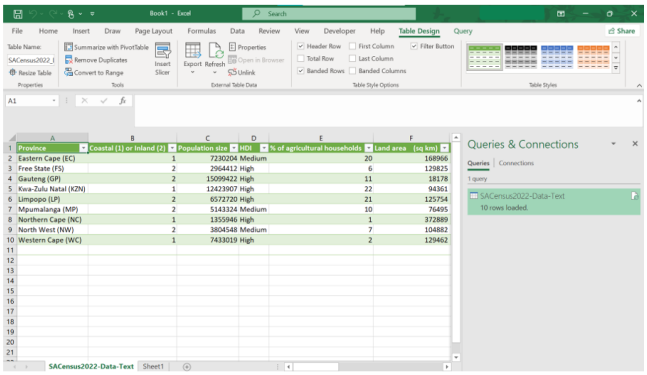
2. A file selection window will appear. Navigate to the location of the file you want to import and select it. By default, the dropdown menu at the bottom right will show "Text Files (`.txt`, `.csv`)" as the file type. Ensure this option is selected.



3. The Import Wizard will open in Excel. Check that the data columns are separated correctly. Excel typically detects the delimiter (e.g., commas, tabs) automatically, but you can change it using the dropdown menu at the top of the wizard if needed.



4. Click **Load** to import the data into a new worksheet. The data will now be displayed in Excel, ready for use.



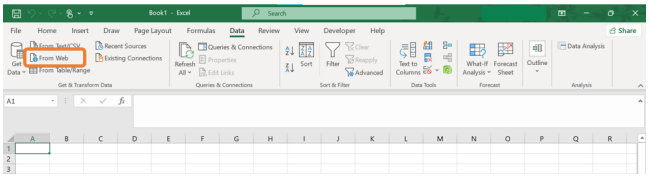
NOTE: The same process can be followed to open a `.csv` file in Excel. In the case where the settings for the decimal separator is correct, a `.csv` file can be opened using *File* > *Open* > *Browse* and selecting the desired file.

3.1.4 Importing data from a website

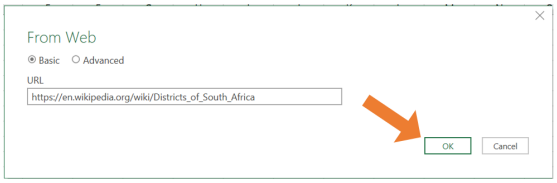
It is also possible to import a data table into Excel directly from a website, saving you the effort to manually retype the information you need.

Suppose you need to compile a list of all the districts in South Africa. You found such a list on Wikipedia and want to import it into Excel. Follow these steps:

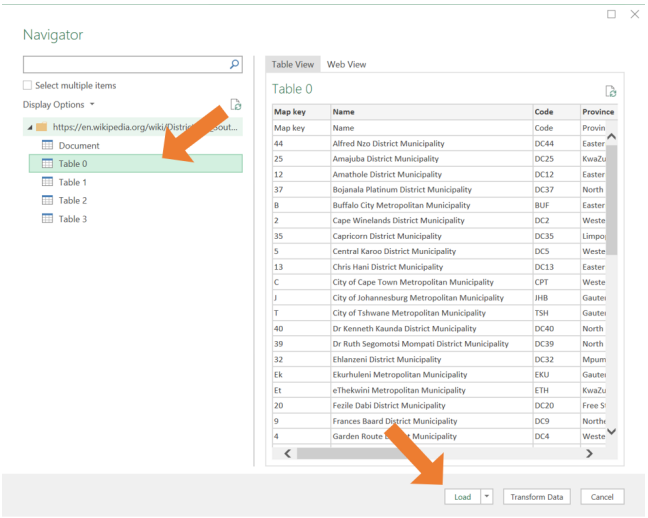
1. Open Excel and navigate to the **Data** tab. In the **Get & Transform Data** group click the **From Web** button.



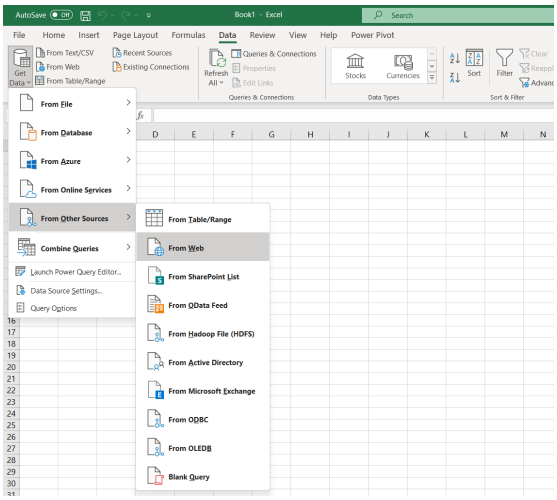
2. In the **From Web** wizard, paste the URL of the website containing the table you want to import, then click **OK**.



3. Next the **Navigator** window will open. On the left-hand side, you will see a list of tables available from the webpage. Click on a table name to preview the contents of the table on the right-hand side. Select the correct table you want to import. Click on **Load** to import the table into your Excel worksheet.



Alternatively, the **From Web** wizard can be accessed using the **GetData** button in the **Get & Transform Data** group.



3.2 Data cleaning

Data cleaning is an essential process of identifying and correcting inconsistencies and inaccuracies within a dataset. This process improves the quality, accuracy and reliability for the data. In fact, a data analyst spends a lot of time on preparing the data for analysis. It is very rare that the raw data is in the correct format and without any errors.

Data cleaning is the process of transforming raw data into consistent data that can be analysed.

To ensure that the data set is cleaned and refined before starting an analysis is crucial to assure that the analysis is accurate and therefore that better-informed decisions can be deduced from the analysis. When statistical analysis is performed on data that isn't properly cleaned, the integrity of the results and findings are compromised and not trustworthy.

Therefore, the analysis is only as reliable as the data that is used for the analysis, making data cleaning an essential step.

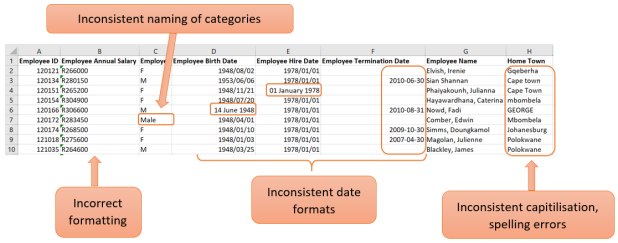
The specific data cleaning techniques that one will use always depends on the application at hand and the analysis needed. Here is a list of some data cleaning techniques:

- Removing the data that is not necessary for your analysis.
- Identifying and removing observations that are duplicated.
- Correcting typing errors, errors in capitalisations or inconsistent naming conventions.

- Removing or imputing missing data.
- Encoding categorical data either to or from a numerical format.
- Ensuring that the data is in a consistent format.

3.2.1 Example: Data cleaning in Excel

You are given a data set containing information on employees working at Orion Sales. This data set is named `orion_sales_staff.xlsx`. Once you have opened the data set in Excel, you come across some problems with the data set that requires data cleaning.

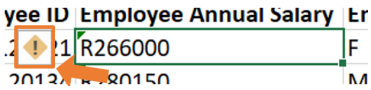


Correct the formatting of a cell

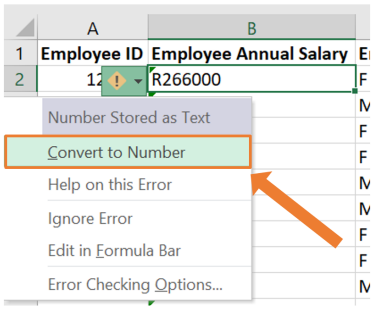
In the example above, the values in the "Employee Annual Salary" column are seen as text. This is clearly visible with the left alignment of the values in the cell. Because annual salary is a numerical value, the formatting needs to be corrected so that Excel can handle the values as numbers.

This can be corrected with the following steps:

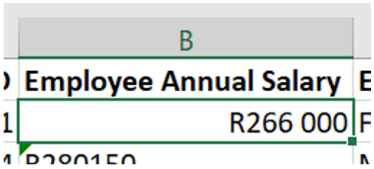
1. If you click in the cell, there will be a yellow block with an exclamation mark on the left.



2. By clicking on the warning sign, a dropdown list will open where you can choose the option "Convert to Number".



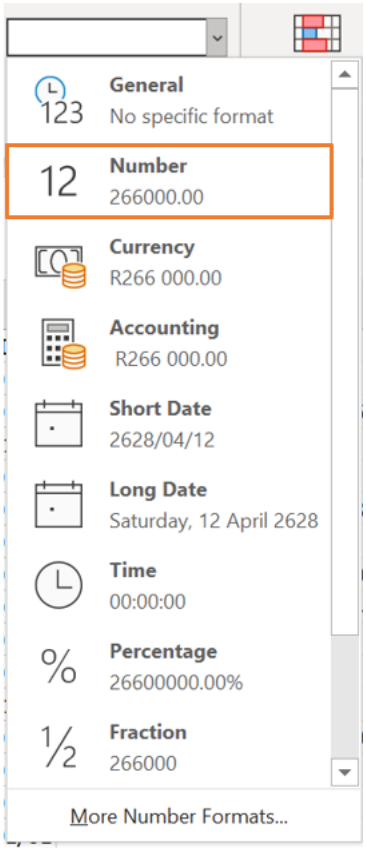
3. Now the cell's formatting will be corrected. This can be seen by the value being right aligned within the cell.



4. To do this for an entire column, The same procedure can be followed by highlighting all the cells that need to be changed.

This process will automatically change the format of the cells from text to "currency". If you work with this data set in Excel, this formatting will be sufficient as Excel knows how to work with this. If the data set is imported into another statistical program (which you will be familiarised later on), the format of this column needs to be "Number" instead of "Currency".

This can be done by navigating to the "Number" group on the Home tab in Excel and opening the dropdown list. From this dropdown list, the appropriate formatting can be selected which is "Number" in this case.

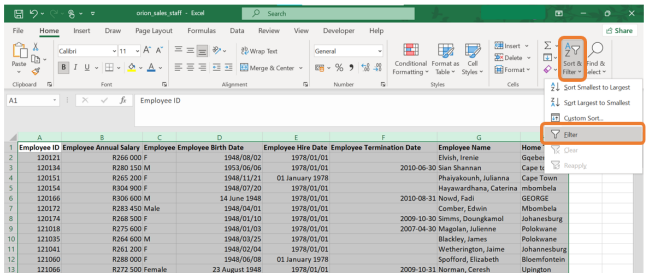


Inconsistent naming of categories

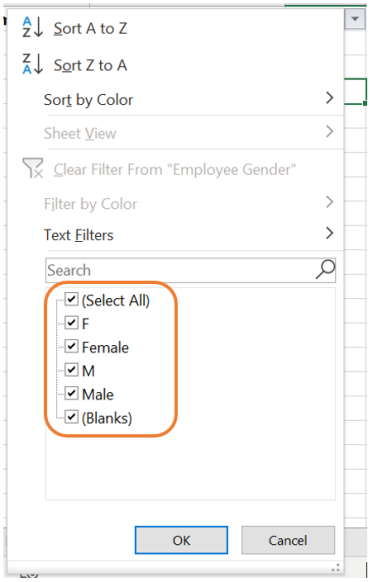
In column C of the example above, the categories are not named consistently. This column consists of the gender of each employee. By glancing at the screenshot of the first few lines of the data, it can be seen that the categories are labelled as "M" (for male) and "F" (for female). However, in line 7, the entry is "Male". There can be more occurrences like this further in the data. Having inconsistent naming of categories is not ideal as this will cause problems when you start to work with the data. For example, when constructing frequency tables or barplots.

This can be corrected with the following steps:

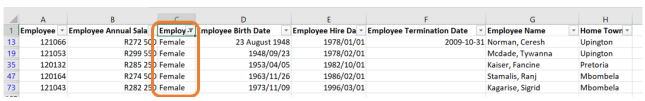
1. Select all the columns of the data set. In the "Editing" group on the "Home" tab, click on "Sort & Filter" and then on "Filter. This will allow you to filter the data by a specific column.



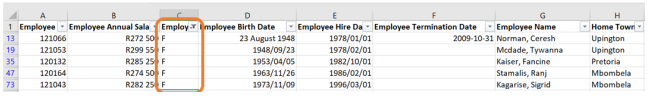
2. The filter icon will now be next to each column name. When clicking on the item next to the column name you would like to filter on, a menu will open up. On this menu, you have the ability to sort the data by this column or filter certain categories. In the screenshot below, all the categories are ticked meaning that all the observations of the data set will be displayed. In the example below, it can be seen that there are categories named "M", "F", "Male" and "Female". We would like to change all the observations with category "Female" to "F" and all the observations with category "Male" to "M".



3. When removing the tick marks of all the categories except one, only the observations from the specific category will be displayed. In our case, let us first filter all the observations where the gender is indicated as "Female".

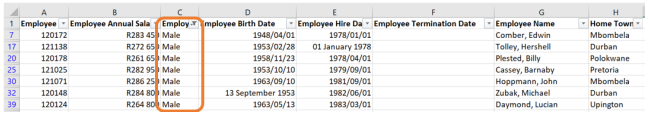


4. Having all the observations filtered out, you can manually change the categories to “F”.



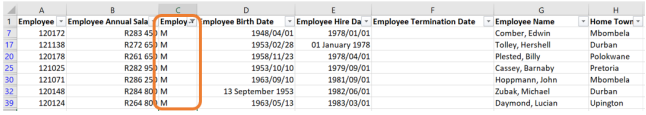
	A	B	C	D	E	F	G	H
1	Employee	Employee Annual Sala	Employ	Employee Birth Date	Employee Hire Da	Employee Termination Date	Employee Name	Home Town
13	121066	R272 50	F	23 August 1948	1978/01/01	2009-10-31	Norman, Ceresh	Upington
19	121053	R299 55	F	1948/09/23	1978/02/01		Mcdade, Tywanna	Upington
35	120132	R285 25	F	1953/04/05	1982/10/01		Kaiser, Fancine	Pretoria
47	120164	R274 50	F	1963/11/26	1986/02/01		Stamalis, Ranj	Mbombela
73	121043	R282 25	F	1973/11/09	1996/03/01		Kagarise, Sigrid	Mbombela

5. The came can be done for the males. Again, click on the filter icon next to the column name. Remove the tick marks of all the categories except the one with “Male”.



	A	B	C	D	E	F	G	H
1	Employee	Employee Annual Sala	Employ	Employee Birth Date	Employee Hire Da	Employee Termination Date	Employee Name	Home Town
7	120172	R283 45	Male	1948/04/01	1978/01/01		Comber, Edwin	Mbombela
17	121138	R272 65	Male	1953/02/28	01 January 1978		Tolley, Hershell	Durban
20	120178	R261 65	Male	1958/11/23	1978/04/01		Plested, Billy	Polokwane
25	121025	R282 95	Male	1953/10/10	1979/09/01		Cassey, Barnaby	Pretoria
30	121071	R286 25	Male	1963/09/10	1981/09/01		Hoppmann, John	Mbombela
32	120148	R284 80	Male	13 September 1953	1982/06/01		Zubak, Michael	Durban
39	120124	R264 80	Male	1963/05/13	1983/03/01		Daymond, Lucian	Upington

6. Then, manually change the categories to “M”.



	A	B	C	D	E	F	G	H
1	Employee	Employee Annual Sala	Employ	Employee Birth Date	Employee Hire Da	Employee Termination Date	Employee Name	Home Town
7	120172	R283 45	M	1948/04/01	1978/01/01		Comber, Edwin	Mbombela
17	121138	R272 65	M	1953/02/28	01 January 1978		Tolley, Hershell	Durban
20	120178	R261 65	M	1958/11/23	1978/04/01		Plested, Billy	Polokwane
25	121025	R282 95	M	1953/10/10	1979/09/01		Cassey, Barnaby	Pretoria
30	121071	R286 25	M	1963/09/10	1981/09/01		Hoppmann, John	Mbombela
32	120148	R284 80	M	13 September 1953	1982/06/01		Zubak, Michael	Durban
39	120124	R264 80	M	1963/05/13	1983/03/01		Daymond, Lucian	Upington

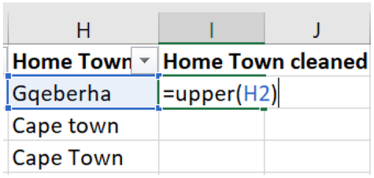
Inconsistent capitalisations

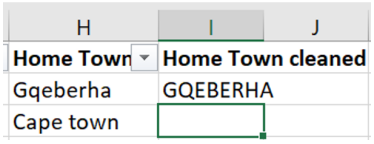
In the column named “Home Town”, there are a few observations where the name of the town is written in all capital letters, all small letters or the second word is written with a small letter instead of a capital letter.

The best and quickest way to correct this is to create a new column and use built-in Excel functions to do the correction. Let us call this column “Home Town cleaned”.

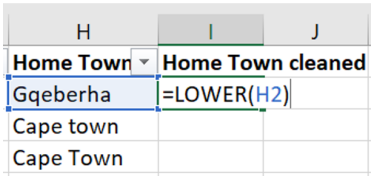
Next, we will introduce three new functions that will alter how the names of the towns are written.

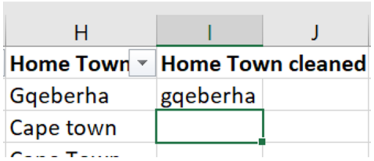
1. The first function is `=UPPER()`. This function will return the word in the selected cell written in all capital letters.



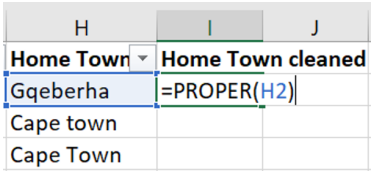


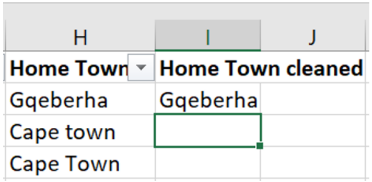
2. The second function is `=LOWER()`. This function will return the word in the selected cell written in all lower case letters.





3. The third function is `=PROPER()`. This function will return the word in the selected cell where the first letter of each word is written in a capital letter followed by lower case letters.



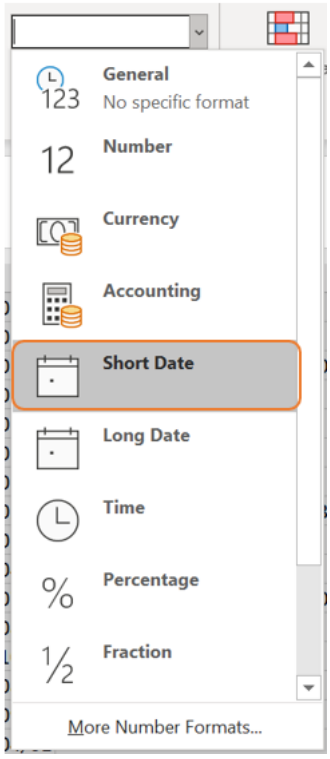


In the same column, some of the town names are misspelled. This can be corrected similarly to how you corrected the inconsistencies in the category names.

Correcting the date format

The columns "Employee Birth Date", "Employee Hire Date" and "Employee Termination Date" are examples of columns containing dates.

The dates can be corrected by selecting all the columns containing dates, then selecting the "short date" format on the drop-down list in the "Number" group on the "Home tab.



3.2.2 Missing data

Missing data occurs when there is no value for a variable of a certain observation. This is a common issue in data analysis and can arise for various reasons. Missing data can have a significant effect on data analysis and conclusions drawn from such analysis.

Missing data often results from non-response. For example, in a survey, a respondent may leave a question unanswered. This usually happens in the case of sensitive information such as salary. Another reason for missing data is caused by errors made by the researcher during data collection or entry.

Types of missing data:

- **Missing completely at random (MCAR)**: With this type of missing data, the probability of an observation being missing is entirely random and independent of any other variable in the data set. For example, at the end of a customer service call, customers might be asked to complete a satisfaction survey. Some individuals may choose not to respond which causes missing data.
- **Missing at random (MAR)**: With this type of missing data, the probability of an observation being missing depends on the values of other variables in the data set but not the missing variable itself. For example, after visiting a dermatology clinic, customers might be asked to fill in a survey on their gender and skincare routine. If females are more likely to respond, the missing data can be explained by the gender of the individual.
- **Missing not at random (MNAR)**: With this type of missing data, the probability of a data point being missing is related to the missing value itself. For example, some individuals might prefer not to answer sensitive information on a survey such as their salary.

Some methods for handling missing data:

- **Imputation**: In some cases, the missing data can be replaced with estimated values. Some common approaches include filling in missing values with the mean, median or mode of the variable.
- **Interpolation**: This method of handling missing data is to fill in missing data based on the adjacent datapoints. This is a popular method to use in time series data.
- **Deletion**: In some cases, it may be appropriate to remove the variable entirely from the analysis when the variable has a high proportion of missing values. Another method is to delete an entire observation if one or more of the variables contains missing data.
- **Model-based approaches**: With this method predictive models are used to impute the missing values in the data set based on other variables in the data set.

When working with data that contains missing values, caution should always be taken. The type of missing data as well as the analysis at hand should guide the data analyst on how to handle the missing data.

3.2.2.1 Example

You are given a data set called `Diabetes Missing Data.xlsx` which contains vital measurements of 30 patients. Some of the data is missing and you are required to explore this in Excel.

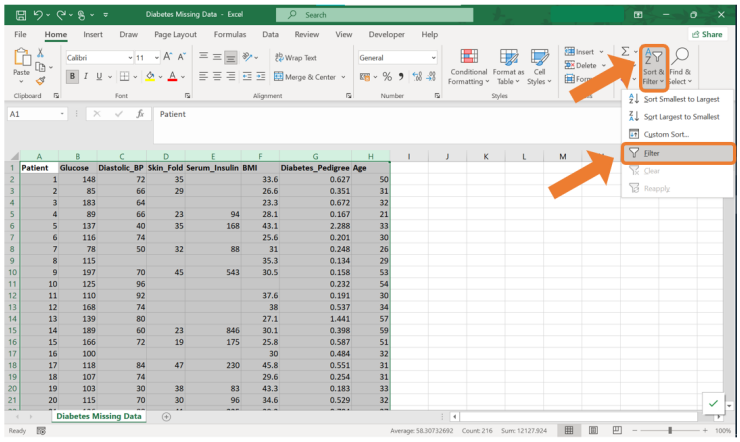
	A	B	C	D	E	F	G	H
1	Patient	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age
2	1	148	72	35		33.6	0.627	50
3	2	85	66	29		26.6	0.351	31
4	3	183	64			23.3	0.672	32
5	4	89	66	23	94	28.1	0.167	21
6	5	137	40	35	168	43.1	2.288	33
7	6	116	74			25.6	0.201	30
8	7	78	50	32	88	31	0.248	26
9	8	115				35.3	0.134	29
10	9	197	70	45	543	30.5	0.158	53
11	10	125	96				0.232	54
12	11	110	92			37.6	0.191	30
13	12	168	74			38	0.537	34
14	13	139	80			27.1	1.441	57
15	14	189	60	23	846	30.1	0.398	59
16	15	166	72	19	175	25.8	0.587	51
17	16	100				30	0.484	32
18	17	118	84	47	230	45.8	0.551	31
19	18	107	74			29.6	0.254	31
20	19	103	30	38	83	43.3	0.183	33

Filter a column for missing values

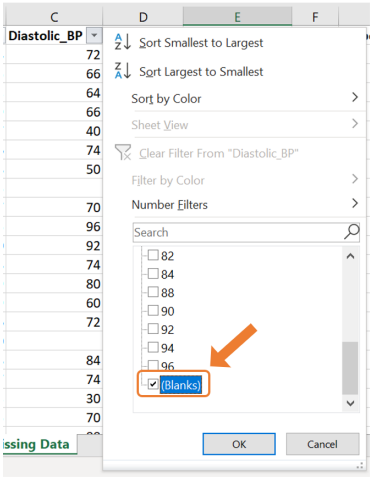
Missing values can be represented in various ways, such as `NA`, `N/A` or as a blank cell.

In this example, we will filter the `Diastolic_BP` column to display only the observations with missing values for this variable. Follow these steps:

1. Select all the columns of the data set. In the **Editing** group on the **Home** tab, click on **Sort & Filter** and then on "Filter". This will allow you to filter the data by a specific column.



2. Click on the Filter icon next to the column name which will open up a menu. Untick all the tick boxes except the one labelled "(blank)". Then click OK.



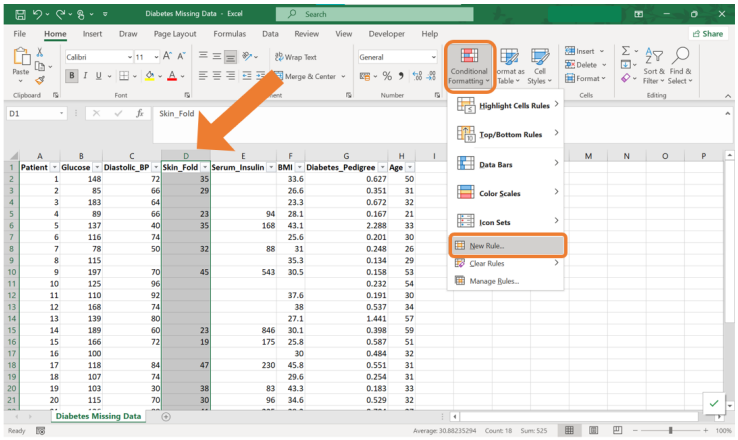
3. This will then only display the observations with a missing value for this variable.

	A	B	C	D	E	F	G	H
1	Patient	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age
9	8	115				35.3	0.134	29
17	16	100				30	0.484	32

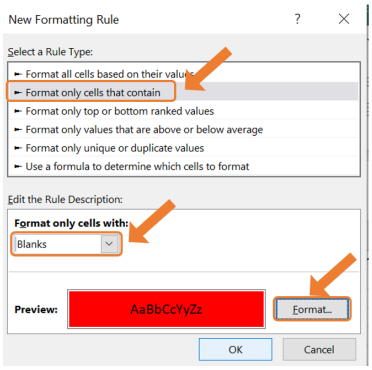
Use conditional formatting to highlight cells with missing values

In the case where you simply want to highlight the cells with missing values, conditional formatting can be used. For this example, we will highlight the missing values in the `Skin_Fold` variable. Follow these steps:

1. Select the column on which you want to apply the conditional formatting. Navigate to the **Home** tab, click on **Conditional Formatting** and choose **New Rule**.



2. In the New Formatting Rule dialog box, select **Format only cells that contain**. Then at the dropdown menu under **Format only cells with:**, select "Blanks". By clicking on the **Format** button, you can specify the formatting style for the highlighted cells.



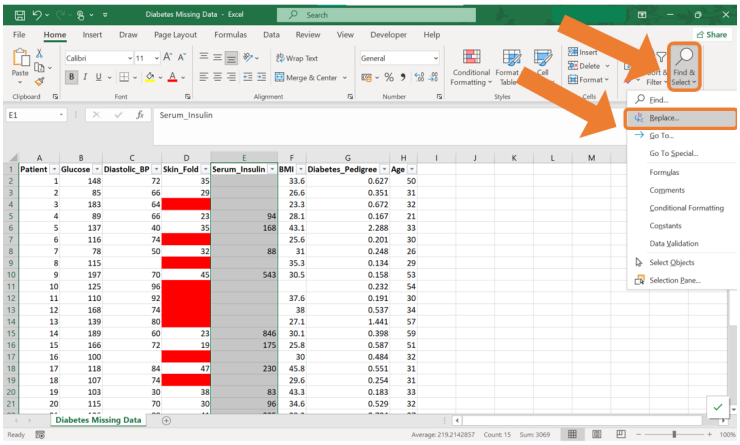
3. As a result, all the cells with missing values will be highlighted.

	A	B	C	D	E	F	G	H
1	Patient	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age
2	1	148	72	35		33.6	0.627	50
3	2	85	66	29		26.6	0.351	31
4	3	183	64			23.3	0.672	32
5	4	89	66	23	94	28.1	0.167	21
6	5	137	40	35	168	43.1	2.288	33
7	6	116	74			25.6	0.201	30
8	7	78	50	32	88	31	0.248	26
9	8	115				35.3	0.134	29
10	9	197	70	45	543	30.5	0.158	53
11	10	125	96				0.232	54
12	11	110	92			37.6	0.191	30
13	12	168	74			38	0.537	34
14	13	139	80			27.1	1.441	57
15	14	189	60	23	846	30.1	0.398	59
16	15	166	72	19	175	25.8	0.587	51

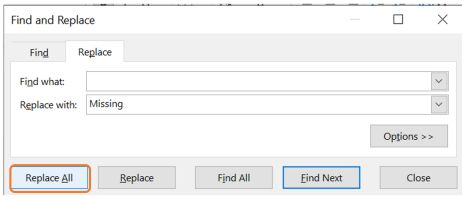
Find and replace missing values

When you want to replace all the missing values of a certain variable with the same value, **Find & Replace** can be used. In this example, replace all the missing values from the `Serum_Insulin` variable with the word "Missing". Follow the following steps:

1. Select the desired column. Navigate to the **Home** tab, click on **Find & Select** and choose **Replace**.



2. In the Find & Replace dialog box, leave the **Find what** field blank to target all blank cells. Enter the word "Missing" in the **Replace with** field. Click on **Replace All** to apply this change to the entire column.



3. As a result, all the cells that were empty previously will now contain the word "Missing".

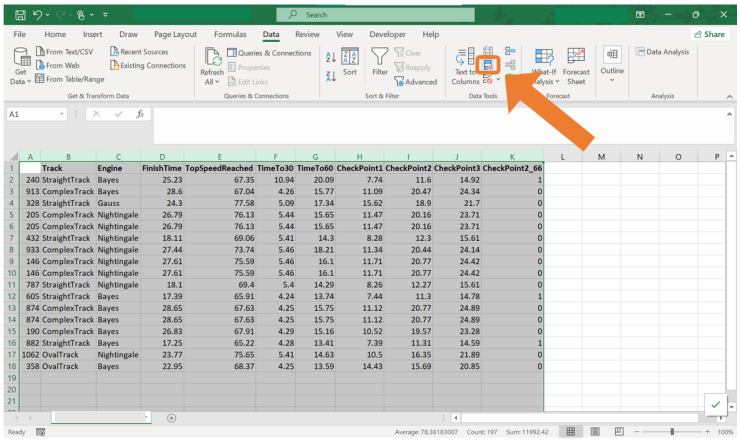
	A	B	C	D	E	F	G	H
1	Patient	Glucose	Diastolic_BP	Skin_Fold	Serum_Insulin	BMI	Diabetes_Pedigree	Age
2	1	148	72	35	Missing	33.6	0.627	50
3	2	85	66	29	Missing	26.6	0.351	31
4	3	183	64		Missing	23.3	0.672	32
5	4	89	66	23	94	28.1	0.167	21
6	5	137	40	35	168	43.1	2.288	33
7	6	116	74		Missing	25.6	0.201	30
8	7	78	50	32	88	31	0.248	26
9	8	115			Missing	35.3	0.134	29
10	9	197	70	45	543	30.5	0.158	53
11	10	125	96		Missing		0.232	54
12	11	110	92		Missing	37.6	0.191	30
13	12	168	74		Missing	38	0.537	34
14	13	139	80		Missing	27.1	1.441	57
15	14	189	60	23	846	30.1	0.398	59

3.2.3 Duplicate values

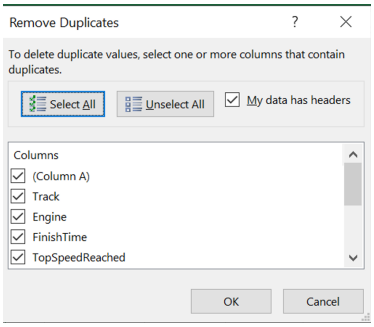
When working with a real-world data set, you may encounter duplicated records. Such records do not provide additional information and can slow down the analysis process or lead to inaccuracies. Therefore, it is best to remove such observations.

To remove duplicated rows in Excel, the following steps can be followed:

1. Select the entire data set in Excel. Navigate to the **Data** tab and in the **Data Tools** group, click on the **Remove Duplicates** icon.



2. In the Remove duplicates dialog box, you can select the columns where duplicates should be identified. In most cases, it will be appropriate to select all the columns to ensure rows are entirely unique. Click **OK** to remove all the duplicates.



3.2.4 Exercises

1. Explain why data cleaning is an essential step for the analysis.
2. What are the common types of issues that you can encounter in a real-word data set?
3. Why is it necessary to remove duplicate records?
4. List some data cleaning techniques and provide an example of where such technique might be necessary.
5. Name the functions in Excel that can be used to change the capitilisation of words in Excel.
6. Why is it important to standarise text entries (for example, converting "Yes" and "yes" to the same format) in a data set? Can you think of an example where inconsistencies in text entries can affect analysis?
7. What are some common reasons for missing data in a data set? Provide some examples of where missing data occurs.

3.3 Data manipulation

Data manipulation is the process of adjusting data so that it is easier to work with and more organised. Data manipulation is a crucial part of the analytical process because it allows statisticians to prepare data in a format that meets the requirements of specific analyses. The specific data manipulation needs depends on the application at hand as well as the statistical analysis that is required. Proper data manipulation enhances the interpretability of data, ensures accuracy in computations, and enables the effective application of statistical methods. Without it, raw data might obscure patterns, relationships, and insights that are vital for informed decision-making.

Data manipulation may include the following:

- Removing the data that is not necessary for your analysis.
- Identifying and removing rows that are duplicated.
- Encoding categorical data either to or from a numerical format.
- Conditional formatting in Excel.
- Combining data sets.
- Splitting and combining columns.
- Pivot tables in Excel to reshape the data set.

Indicator	Description
1	Critically Endangered (CR)
2	Vulnerable (VU)
3	Near Threatened (NT)
4	Least Concern (LC)

3.3.1 Example: Super Animals

A few years ago, the Super Animal Cards were available at Pick n Pay stores. These collectible cards feature illustrations and fascinating facts about various animals. The aim was to spark curiosity and environmental awareness among children. Each card highlights a different animal's unique characteristics, habitat and conservation status.

The information from these cards are collected in a data set called `super-animals.xlsx`.

The VLOOKUP function

The VLOOKUP function in Excel is a powerful tool to make data more descriptive and meaningful by referencing values from another table. In this example, the conservation status of animals is initially indicated with numeric values (1, 2, 3, and 4) in the dataset. To make this information more interpretable, we use a lookup table containing the corresponding descriptions for each number:

By replicating the lookup table in a separate sheet of the Excel file, the VLOOKUP function can map the numeric indicators to their corresponding descriptions. This method not only improves clarity but also allows for easier data analysis and reporting, making it a practical approach to handle coded information in datasets.

The syntax for the VLOOKUP function is:

```
=VLOOKUP(lookup_value, table_array, col_index_num, [range_lookup])
```

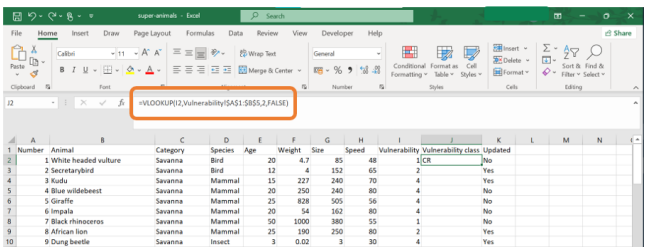
- `lookup_value`: The value you want to search for in the first column of the table.
- `table_array`: The range of cells containing the table, including the column with the lookup value and the column with the return value.
- `col_index_num`: The column number (relative to the table) from which to retrieve the result.
- `range_lookup (optional)`: Specifies whether the match should be exact (`FALSE`) or approximate (`TRUE`). By default, it's approximate.

In this example, the process of using the VLOOKUP function to map descriptive conservation statuses to numeric indicators involves the following steps:

1. Begin by entering the lookup table in a separate sheet within the Excel file. This table should contain the numeric indicators in one column and their corresponding descriptions in another.

	A	B
1	Indicator	Description
2	1	CR
3	2	VU
4	3	NT
5	4	LC

2. In a new column of your dataset, type the VLOOKUP formula to reference the lookup table. Ensure you use absolute cell references (with dollar signs, e.g., `$A$1:$B$5`) for the lookup table range. This prevents the table reference from shifting when you drag the formula down to apply it to other rows.

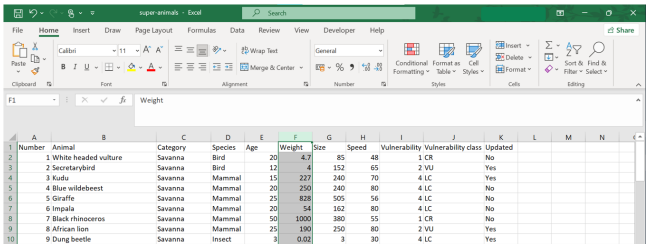


Conditional formatting

Conditional formatting is a tool in Excel that allows users to automatically apply formatting to certain cells based on specified criteria. The formatting can be colours, icons, or many others. You can create rules based on predefined options or custom formulas.

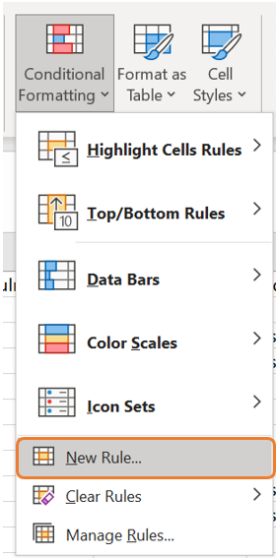
In the example of the Super Animal Cards, conditional formatting can be used to highlight the top 10 animals with the highest weights. This is particularly useful for quickly identifying the heaviest animals in the data set. To achieve this in Excel:

1. Select the column containing the weight data.

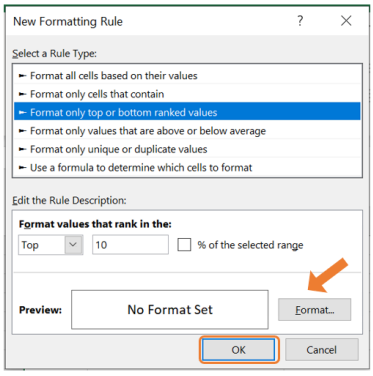


	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	Number	Animal	Category	Species	Age	Weight	Size	Speed	Vulnerability	Vulnerability class	Updated			
2	1	White headed vulture	Savanna	Bird	20	4.7	85	48	1 CR		No			
3	2	Secretarybird	Savanna	Bird	12	4	152	65	2 VU		Yes			
4	3	Kudu	Savanna	Mammal	15	227	240	70	4 LC		Yes			
5	4	Blue wildebeest	Savanna	Mammal	20	250	240	80	4 LC		No			
6	5	Giraffe	Savanna	Mammal	25	828	505	56	4 LC		No			
7	6	Impala	Savanna	Mammal	20	54	162	80	4 LC		No			
8	7	Black rhinoceros	Savanna	Mammal	50	1000	380	55	1 CR		No			
9	8	African lion	Savanna	Mammal	25	190	250	80	2 VU		Yes			
10	9	Dung beetle	Savanna	Insect	3	0.02	3	30	4 LC		Yes			

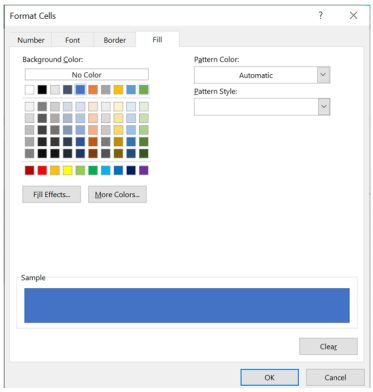
2. Navigate to the **Home** tab, click on **Conditional Formatting** and choose **New Rule**.



3. In the New Formatting Rule dialog box, select **Format only top or bottom ranked values**. This allows you to target the top 10 entries in the dataset.



4. Specify the formatting style, such as a bold font or a shaded cell color, to visually emphasise the top 10 entries.



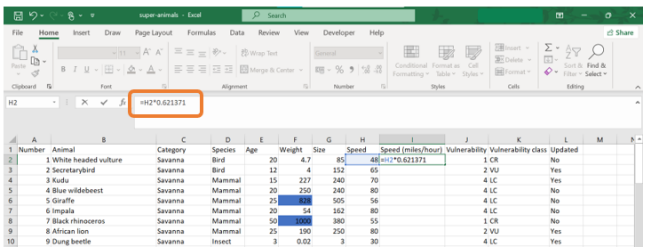
Numerical calculations in Excel

Excel can also perform basic numerical calculations, such as addition, subtraction, multiplication, and division. These operations can be executed directly in cells using simple formulas:

- Addition: `= A1 + B1` adds the values in cells A1 and B1.
- Subtraction: `= A1 - B1` subtracts the value in cell B1 from the value in cell A1.
- Multiplication: `= A1 * B1` multiplies the values in cells A1 and B1.
- Division: `= A1 / B1` divides the value in cell A1 by the value in cell B1.

These formulas can also combine multiple operations using parentheses for clarity and order of precedence. For example, `= (A1 + B1) * C1` first adds the values in cells A1 and B1, then multiplies the result by the value in cell C1.

In the example of the Super Animal Cards, we can convert the speed from kilometers per hour to miles per hour by multiplying the values in column H with 0.621371.



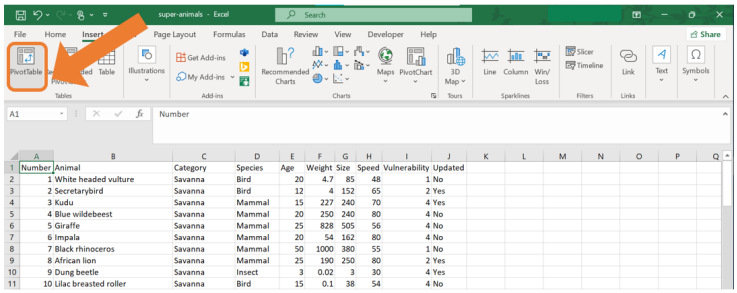
Constructing crosstabulations

The PivotTable tool in Excel is a powerful feature for constructing crosstabulations involving two or more variables. In this example, we will consider two cases:

1. When both variables are categorical.
2. When one variable is categorical and the other is numerical.

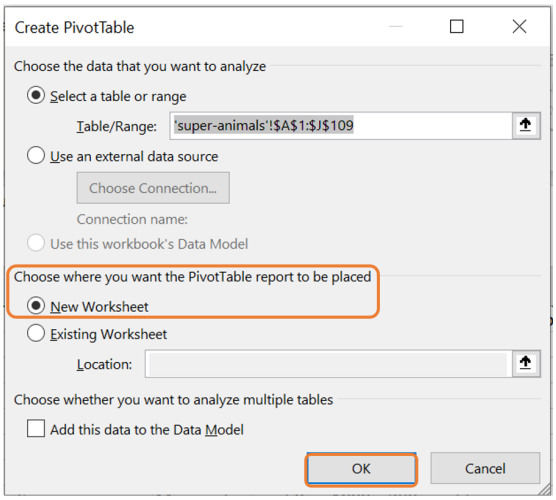
- **Case 1: Create a crosstabulation of the species and the category of the super animal card.**

1. Navigate to the **Insert** tab and select **PivotTable**.

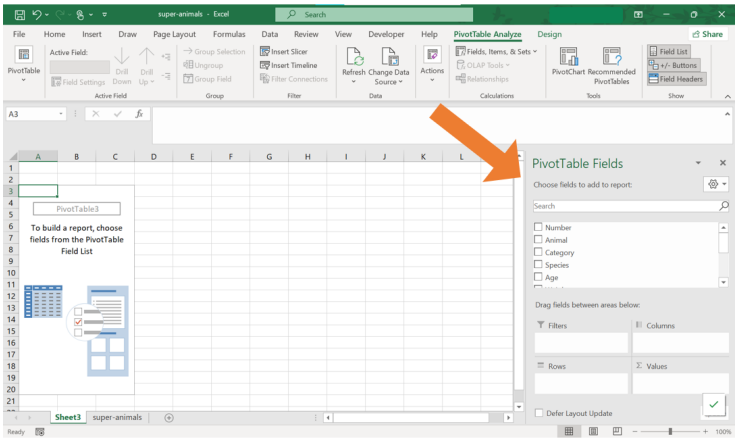


Number	Animal	Category	Species	Age	Weight	Size	Speed	Vulnerability	Updated
1	White headed vulture	Savanna	Bird	20	4.7	85	48	1	No
2	Secretarybird	Savanna	Bird	12	4	152	65	2	Yes
3	Kudu	Savanna	Mammal	15	227	240	70	4	Yes
4	Blue wildebeest	Savanna	Mammal	20	250	240	80	4	No
5	Giraffe	Savanna	Mammal	25	828	505	56	4	No
6	Impala	Savanna	Mammal	20	54	162	80	4	No
7	Black rhinoceros	Savanna	Mammal	50	1000	380	55	1	No
8	African lion	Savanna	Mammal	25	190	250	80	2	Yes
9	Dung beetle	Savanna	Insect	3	0.02	3	30	4	Yes
10	Lilac breasted roller	Savanna	Bird	15	0.1	38	54	4	No

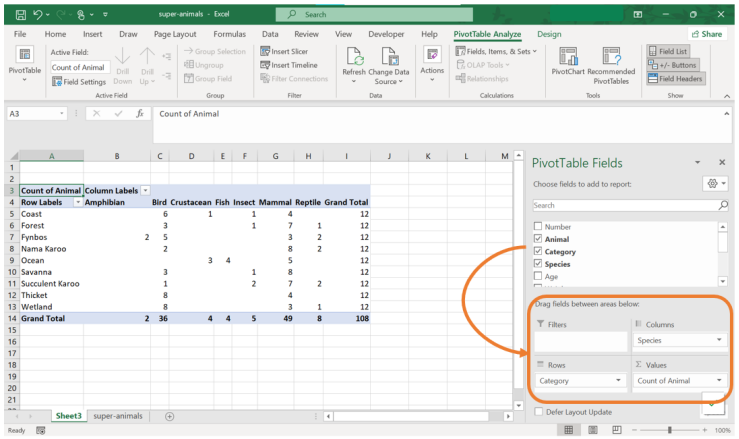
2. The Create PivotTable wizard will appear, where you can review the data range that is automatically detected and choose where the PivotTable will be created. For this example, select **New Worksheet**. Click **OK**.



3. A new worksheet will open up with the PivotTable Fields on the right.

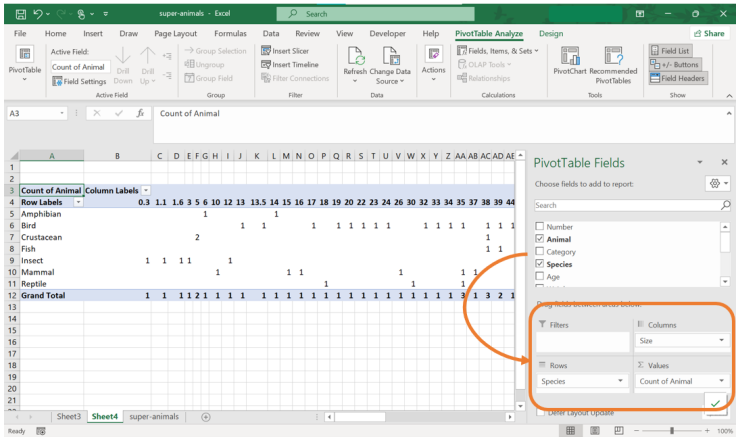


4. Using the PivotTable Fields menu, the Pivot Table can be set up. Drag the column headings to their desired positions. Drag "Category" to the Rows area. Drag "Species" to the Columns area. Drag "Animal" to the Values area to count the number of animals in each category.

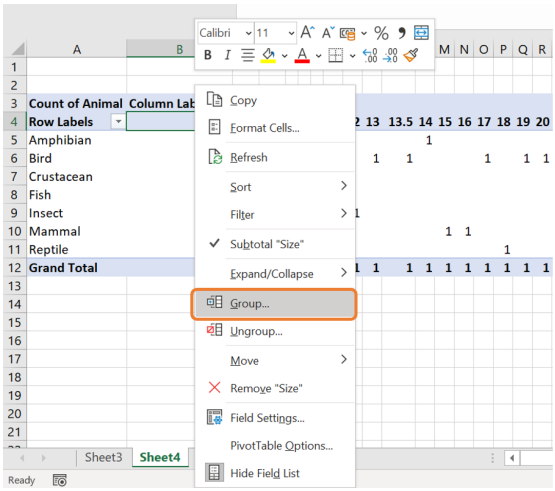


- **Case 2: Create a crosstabulation of the Species and the Size of the animal.**

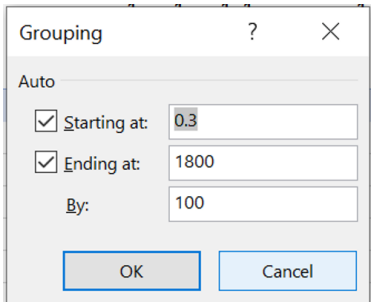
  1. Create the Pivot Table as in steps 1 and 2 above.
  2. Drag the column headings to their desired positions. Drag "Species" to the Rows area. Drag "Size" to the Columns area. Drag "Animal" to the Values area.



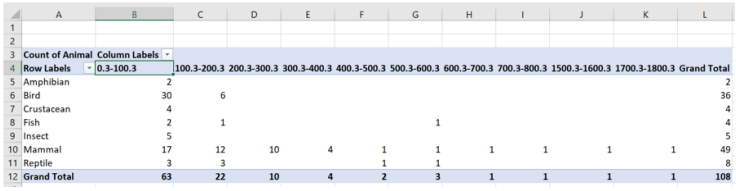
3. Since "Size" is a numerical variable, it must be grouped into ranges. This can be done by selecting the cell with the first value of the "Size" variable, right-click and select **Group** from the menu.



4. In the Grouping window, specify the desired range settings.



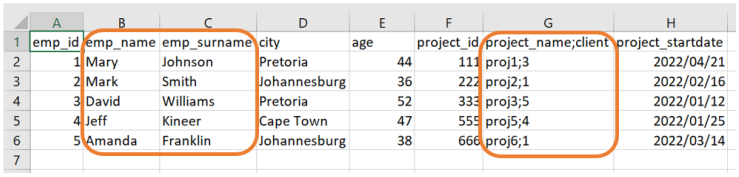
5. The PivotTable will now display a neater and more interpretable corsstabulation for species by grouped sizes.



Count of Animal	Column Labels										
Row Labels	0.3-100.3	100.3-200.3	200.3-300.3	300.3-400.3	400.3-500.3	500.3-600.3	600.3-700.3	700.3-800.3	1500.3-1600.3	1700.3-1800.3	Grand Total
Amphibian	2										2
Bird	30	6									36
Crustacean	4										4
Fish	2	1				1					4
Insect	5										5
Mammal	17	12	10	4	1	1	1	1	1	1	49
Reptile	3	3			1	1					8
Grand Total	63	22	10	4	2	3	1	1	1	1	108

3.3.2 Example: Employees

You are employed as a data analyst for a company and have been provided with a data set containing details about employees and the projects they are working on.



	A	B	C	D	E	F	G	H
1	emp_id	emp_name	emp_surname	city	age	project_id	project_name;client	project_startdate
2	1	Mary	Johnson	Pretoria	44	111	proj1;3	2022/04/21
3	2	Mark	Smith	Johannesburg	36	222	proj2;1	2022/02/16
4	3	David	Williams	Pretoria	52	333	proj3;5	2022/01/12
5	4	Jeff	Kineer	Cape Town	47	555	proj5;4	2022/01/25
6	5	Amanda	Franklin	Johannesburg	38	666	proj6;1	2022/03/14
7								

In this example, we will illustrate how to:

- Combine two columns in Excel into one.
- Splitting one column into two columns.

Combining columns

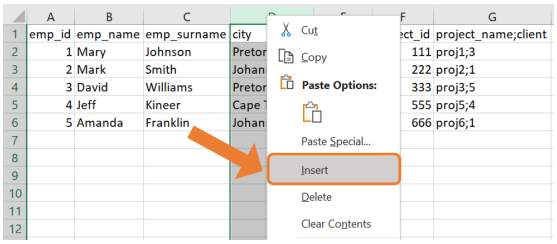
To combine the information of two columns into one, we will make use of the `CONCAT` function. This function is used in Excel to join, or concatenate, two or more text strings into a single string. In this example, we will join the name and the surname of the employees into a single column. The syntax for the `CONCAT` function is:

```
=CONCAT(text1, [text2], ...)
```

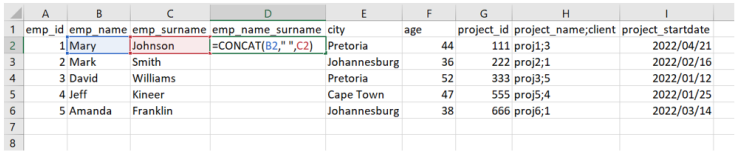
where `text1, text2,...` are text strings, cell references or ranges to be combined separated by commas. It is important to note that the `CONCAT` function does not automatically add any deliminators such as spaces or commas between the text strings. This must be explicitly added as part of the list of arguments.

For this example, if we want to combine the text in the “emp_name” and “emp_surname” variables, the following steps can be followed:

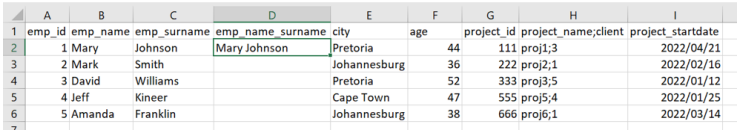
1. Insert an empty column where the combined variable will be placed. To do this, select the column where the new column should appear before, right-click and select **Insert** from the menu. This will create an empty column which you can name "emp_name_surname".



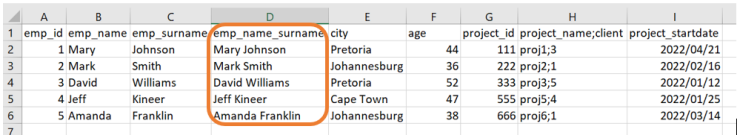
2. In the first cell of the new column, use the `CONCAT` function to combine the name of the employee (in column B) with their surname (in column C). Remember to add a space between the name and the surname!



3. The result of the `CONCAT` function is as follows:



4. Copy the formula down the rest of the columns to combine the names and the surnames for all the employees in the data set.



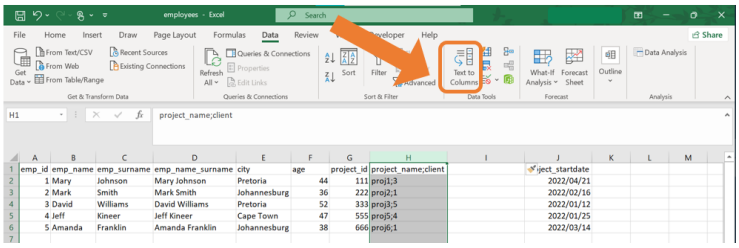
Splitting columns

To split columns in Excel, the **Text to Columns** functionality can be used. This functionality is used to split data in a single column into multiple columns

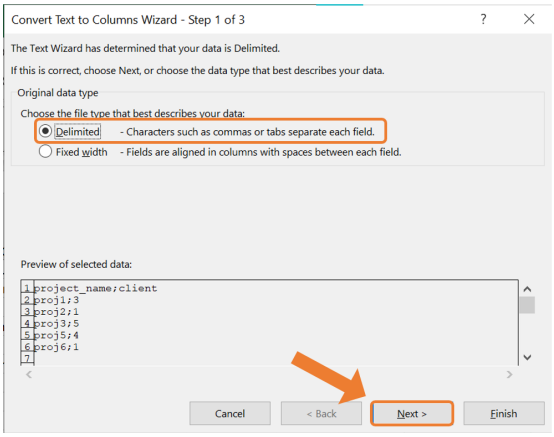
based on either a specific deliminator or a fixed width. This is particular useful when the combined data needs to be separated into distinct fields for easier analysis or formatting.

For this example, the project name and the client name are displayed in a single column, separated by a semicolon (;). To split them into two columns, follow these steps:

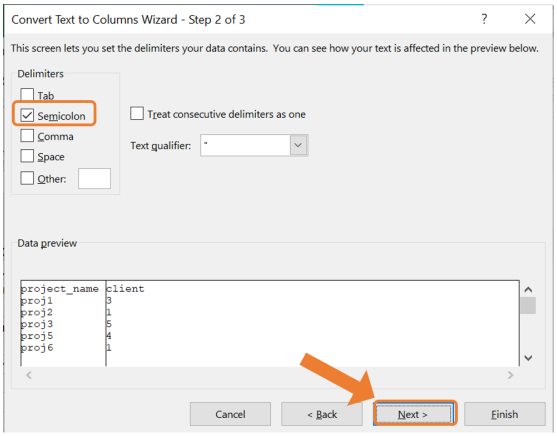
1. Again, add an empty column next to the column which requires splitting. Then, select the column with the combined information. Navigate to the **Data** tab and select **Text to columns**.



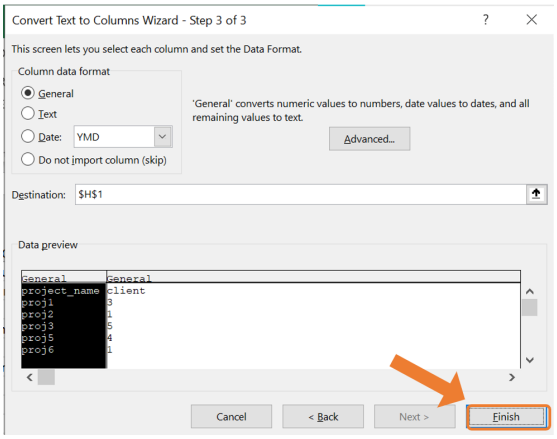
2. The Convert Text to Columns Wizard will open. In this case, the data is "Delimited" because the information is separated with a semicolon. This is often automatically detected. Click **Next**.



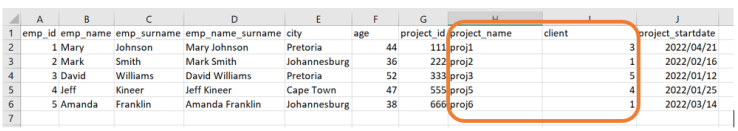
3. In the next step, select the appropriate deliminator, in this case, "semicolon". You can preview the split data in the "Data Preview" section to confirm the results. Click **Next**.



4. In the final step, choose the data format for each column. For this example, you can keep the default settings. Click **Finish**.



5. To combined data will now be split into two separate columns: one for the project name and one for the client name.



	A	B	C	D	E	F	G	H	I	J
1	emp_id	emp_name	emp_surname	emp_name_surname	city	age	project_id	project_name	client	project_startdate
2	1	Mary	Johnson	Mary Johnson	Pretoria	44	111	proj1	3	2022/04/21
3	2	Mark	Smith	Mark Smith	Johannesburg	36	222	proj2	1	2022/02/16
4	3	David	Williams	David Williams	Pretoria	52	333	proj3	5	2022/01/12
5	4	Jeff	Kineer	Jeff Kineer	Cape Town	47	555	proj5	4	2022/01/25
6	5	Amanda	Franklin	Amanda Franklin	Johannesburg	38	666	proj6	1	2022/03/14
7										

3.3.3 Basic Excel functions

Excel has many built-in functions that can assist in doing more numerical calculations, creating new variables in a data set as well as summarising data.

Summation

The `SUM` function can be used to calculate the total of a range of numbers. The syntax for the `SUM` function is:

`=SUM(number1, number2, ...)` or `=SUM(range)`

For example, `SUM(A1:A10)` will return the sum of all the numbers in cells A1 to A10.

Averaging

The arithmetic mean of a range of values can be calculated with the `AVERAGE` function. The syntax for the `AVERAGE` function is:

`=AVERAGE(number1, number2, ...)` or `=AVERAGE(range)`

For example, `AVERAGE(A1:A10)` will return the average of all the numbers in cells A1 to A10.

Counting

Counting functions can be useful for analysis the structure of the data set. Here we will consider three counting functions popularly used in Excel:

- The `COUNT` function counts the number of cells containing a numerical value. The syntax for the `COUNT` function is:

`=COUNT(value1, value2, ...)` or `=COUNT(range)`

For example, `COUNT(A1:A10)` will return the number of cells containing a numerical value in cells A1 to A10.

- The `COUNTA` function counts the number of non-empty cells. The cells can contain numerical values, text or values from any other data type. The syntax for the `COUNTA` function is:

`=COUNTA(value1, value2, ...)` or `=COUNTA(range)`

For example, `COUNTA(A1:A10)` will return the number of non-empty cells in the range A1 to A10.

- The `COUNTBLANK` function counts the number of empty cells. The syntax for the `COUNTBLANK` function is:

`=COUNTBLANK(value1, value2, ...)` or `=COUNTBLANK(range)`

For example, `COUNTBLANK(A1:A10)` will return the number of empty cells in the range A1 to A10.

Minimum and maximum values

The `MIN` and `MAX` functions can be used to identify the smallest and the largest value in a range. The syntax for the `MIN` and the `MAX` functions is:

- `=MIN(number1, number2, ...)` or `=MIN(range)`
- `=MAX(number1, number2, ...)` or `=MAX(range)`

For example, `MIN(A1:A10)` will return the smallest value in the range A1 to A10 and `MAX(A1:A10)` will return the largest value in the range A1 to A10.

IF function

The `IF` function performs a logical test and returns one value if the condition is true and another value if the condition is false. The syntax for the `IF` function is:

`=IF(logical_test, value_if_true, value_if_false)`

- `logical_test`: Any value or expression that can be evaluated to TRUE or FALSE.
- `value_if_true`: The value that is returned if `logical_test` is TRUE.
- `value_if_false`: The value that is returned if `logical_test` is FALSE.

For example, `=IF(A1>50, "Pass", "Fail")` will return “Pass” if the value in cell A1 is greater than 50; otherwise it will return “Fail”.

Countif function

The number of cells than meet a certain condition can be counted with the `COUNTIF` function. The syntax for the `COUNTIF` function is:

`COUNTIF(range, criteria)`

- `range`: The range of cells from which you want to count non-empty cells.
- `criteria`: The condition in the form of a number, expression or text that defines which cells will be counted.

For example, `=COUNTIF(A1:A10, ">10")` will count all the cells from A1 to A10 which contains a numerical value greater than 10.

Sumif function

The `SUMIF` function is used to sum the values in a range that meets a specified criteria. The syntax for the `SUMIF` function is:

```
=SUMIF(range, criteria, [sum_range])
```

- `range`: The range of cells you want to evaluate.
- `criteria`: The condition or criteria in the form of a number, expression or text that defines which cells will be added.
- `sum_range` (optional): The actual cells to sum. If this argument is not specified, the cells in `range` will be used.

For example, suppose you have a data set with two columns, where column A contains the product category and column B contains the number of sales.

- To sum the sales for "Electronics", you can use the formula:

```
=SUMIF(A1:A10, "Electronics", B1:B10)
```

where Excel will look for the word "Electronics" in the range A1 to A10 and the corresponding sales values in B1 to B10 will be summed.

- To sum all the sale values greater than 500, you can use the formula:

```
=SUMIF(B1:B10, ">500")
```

where Excel will sum all the sales values greater than 500 in the range B1 to B10.

Absolute cell referencing

In Excel, you can use absolute cell referencing to ensure that the formula keeps referencing to the same cell or range of cells regardless or where the formula is copied or moved. The syntax for absolute cell referencing is: `$Column$Row`. For example, `$A$1` refers to cell A1 and this reference will not change regardless of where the formula is copied in the Excel sheet.

Absolute cell referencing is useful in scenarios where you need to repeatedly reference the same cell or range of cells. Common examples include:

- **Using a constant value**: When applying a discount stored in a specific cell across multiple calculations.
- **Referencing a total**: Referring to a grand total cell in calculations.

3.3.4 Exercises

1. Describe a scenario where you might use the **Text to Columns** functionality in Excel.
2. Explain the difference between using a delimiter and using fixed width in the **Text to Columns** tool. Provide examples for each.
3. What is the difference between `COUNT`, `COUNTA`, and `COUNTBLANK` functions? Provide an example for each.
4. You are analysing sales data, and the column containing the product names has been combined with the client names, separated by a comma. Describe how you would use Excel to split this column into two separate columns.
5. Explain how the `CONCAT` function can be used to combine data from two columns. Provide an example of when this might be useful.
6. Explain the importance of absolute cell referencing in Excel. Provide an example of when it would be necessary.
7. Describe a situation where combining data from two separate columns would be helpful for analysis. Which function in Excel would you use, and why?

Chapter 4

Sampling

Content coming soon...