NPA Data Science

Learners' Guide to the National Progression Award

Data Education in Schools

Level 6 NPA Data Science Notes 2025

Data Education in Schools

Contents

Introduction

Welcome to the NPA Data Science Notes for 2025! These notes are designed to guide you through the content for your NPA Data Science qualification.

These notes have been written for the updated (2024) NPA Data Science specification.

This document is a summary document covering the core concepts that you will need to know in order to learn the content and undertake the assessments. It can be used by educators to introduce each topic, or for learners as they go through the course as a support resource.

Throughout the guides, you will come across links to videos, and lessons which relate to the content.

These notes are organised by learning outcome. At the beginning of each Outcome section, you will find links to the lessons related to that Outcome.

Support and Resources

These guides have been written with the support of the University of Edinburgh's Data Education in Schools team. The Data Education in Schools project aims to work with schools and colleges that are delivering this course. To date, they have worked with every school delivering this qualification, providing professional learning, facilitating sharing of resources, and working together to review materials and share the development workload.

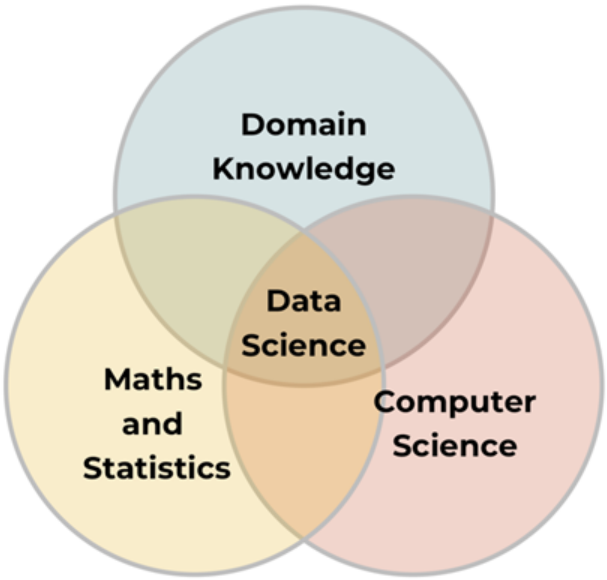
Visit www.dataschools.education for more information about support materials.

For the NPA Lessons which were developed for the previous version of this course, visit www.dataed.in/learndata. These lesson materials are also linked to throughout this guide in relevant sections.

Visit dataed.in/NPADS for more information about the qualification on the SQA site.

This document covers the Level 6 Data Science unit in particular. There are separate documents available for other levels. [Insert link to other documents here]

Scan the QR code or go to dataschools.education/level-6-data-science-lessons/ for relevant lessons and resources for this unit, separated by outcome.



1 Outcome 1 - Describe the applications of data science.

What is data?

Definition

Data: facts that can be analysed or used in an effort to gain knowledge or make decisions; information.

Data facts are distinct pieces of information that are stored and formatted so that they can be automatically interpreted by a computer. Data allows visibility of what has been happening and supports good decisions to be made for the future.

Data on its own is not valuable. Data is raw, unorganised facts that need to be processed, organised, interpreted, structured and presented before it can be turned into information. This information can then be actioned or used to create value.

What is Data Science?

The field of **data science** combines computer science, specific knowledge about a particular topic or subject area, and mathematical skills to extract insights and knowledge from data. The ability to identify the problem to solve, the correct data to use, carry out the analysis, and then implement the outcome requires all three areas to be brought together. If any one of these areas is missing, it is not possible to extract value effectively from data.

The terms **data science** and **data analytics** are often used interchangeably; however, analytics is more focused on finding insights in the data, rather than just the tools and techniques for dealing with large amounts of raw data. A data insight is an “a-ha” moment from data. Ideally it is actionable, so that once that nugget of information is known, a tangible action can be carried out.

1a - Describe contemporary applications of data science and the types of problem that data science can address.

Applications of data science

Personal

• Translation apps and websites • Image recognition • Speech recognition devices • Personalised medical treatment plans • Self-driving cars • Movie recommendations • Website sort-order and recommendations

Business

• Fraud detection • Financial risk estimation • Preventing customer attrition • Delivery logistics • Testing marketing approaches • Airline route planning • Real-time pricing optimisation • Personalised advertising

Government

• Detecting tax evasion • Preventing cyber attacks • Detecting terrorism threats • Improving national security • Improving health services • Coordinating responses to emergencies such as floods, terrorist attacks or pandemics across multiple services

Benefits of using data science

- Better / faster decision-making.
- Improved operations and processes.
- Creation of a data product.
- Understanding customer trends.
- Creating innovative products and services.

Challenges of data science

- **Data Quality**: Inconsistent, incomplete, or inaccurate data can affect model outcomes.
- **Data Privacy and Security**: Ensuring compliance with regulations and protecting sensitive information is crucial.
- **Bias and Fairness**: Avoiding bias in data and models is vital to ensure equitable outcomes.
- **Cost**: Resources required for storage, processing, and talent can be expensive.

Types of problems data science can address

1. **Predictive Analytics:** Data science excels in forecasting future events based on historical data. This is commonly used in finance for stock market predictions, in healthcare for patient outcomes, and in retail for sales forecasting.

2. **Classification Problems:** These involve categorizing data into predefined classes. For example, in healthcare, patient data can be classified to predict the likelihood of diseases, while in the finance sector, transactions can be classified as legitimate or fraudulent.
3. **Regression Analysis:** This type of problem aims to understand relationships between variables and predict continuous outcomes. Businesses use regression analysis to forecast sales, revenue growth, or to estimate customer lifetime value based on various input factors.
4. **Recommendation Systems:** Data science powers recommendation engines that analyze user behavior to suggest products, services, or content. Companies like Amazon and Netflix use these systems to enhance user experience and increase customer satisfaction.
5. **Natural Language Processing (NLP):** With the growth of text data, NLP helps in understanding and generating human language. It's used in sentiment analysis, chatbots, and language translation services, allowing businesses to grasp customer sentiment and automate responses.
6. **Image and Video Analysis:** In fields like security and healthcare, data science is used to analyze visual data. This includes facial recognition technologies and medical image diagnosis, improving both safety and health outcomes.
7. **Anomaly Detection:** Identifying unusual patterns or outliers in data is crucial for fraud detection in banking, network security, and quality control in manufacturing. Data science methods can pinpoint potential issues before they escalate.
8. **Optimisation Problems:** Data science can optimize operations in various industries, from supply chain logistics to production processes, ensuring resources are used efficiently and costs are minimized.

1b - Explain the steps in solving a problem using data science, including the potential for bias at each stage.

PPDAC

The **PPDAC** cycle is a framework to follow when asking and answering real-world problems using data.

Problem:

Identify the question being solved.

1. How much, or how many?
2. Which category or group does this belong to?
3. Is this weird?
4. Which is the best option to choose?

Plan:

Decide how to answer the question.

1. What data is needed to solve this question?
2. Where will this data come from?
3. Is there access to the data, or will it need to be collected?
4. Will there be sufficient volume of data to provide a robust answer?
5. Where and how will the data be stored?
6. Are there any ethical implications of collecting and using this data?

Data:

Collect and store the data securely.

1. Data quality checks
2. Data understanding
3. Data dictionary creation

Analysis:

1. Preparation
2. Manipulation
3. Visualisation
4. Data modelling
5. Validation of feelings

Conclusion:

Summarise and communicate

1. How does the data answer the original statement?
2. Are there any aspects of the original question that has not been addressed?
3. What are the conclusions?
4. What could be done differently next time?
5. What should happen next?

The planning phase of a data science project is critical, as often similar analyses may have been carried out in the past. Not only can time be saved by not repeating existing work, but previous analyses can also be built upon and extended.

Bias and Ethics in the PPDAC cycle

Ensuring ethical standards and minimizing bias at each stage is essential for the integrity and validity of research findings. Here, we explore considerations for bias and ethics within each phase of the PPDAC cycle:

Problem:

- **Ethical Question Formulation**: Define research questions that respect the dignity and rights of all participants. Ensure the intended research does not perpetuate stereotypes or cause harm.
- **Bias Awareness**: Be aware of any preconceived notions that may influence the framing of the problem.

Plan:

- **Transparent Planning**: Clearly document the study design and methodology. Consider potential ethical issues related to participant selection and data collection methods.
- **Minimizing Selection Bias**: Ensure a representative sample by considering inclusion and exclusion criteria carefully. Strive for diversity to avoid skewed results.

Data:

- **Ethical Data Collection**: Obtain informed consent from participants. Ensure confidentiality and privacy of data, adhering to data protection regulations.
- **Bias Detection**: Be vigilant about potential sources of bias in data collection, such as non-response or measurement errors. Implement strategies to address and mitigate these biases.

Analysis:

- **Objective Analysis**: Analyze data objectively, avoiding manipulation or selective reporting of results. Use appropriate statistical methods to ensure validity.
- **Awareness of Analytical Bias**: Recognize and address any biases introduced during analysis, such as confirmation bias or cherry-picking data that supports a hypothesis.

Conclusion:

- **Ethical Reporting**: Report findings honestly and transparently, acknowledging limitations and potential biases. Discuss the implications of the results responsibly.
- **Avoiding Overgeneralization**: Be cautious of making broad generalizations from the data, especially if the sample is not representative of the larger population.

Identify sources of public and private datasets

Data can be categorized into public and private sources. Public datasets are accessible to everyone and often provided by government bodies, organizations, and educational institutions. These resources support transparency and innovation across various sectors. On the other hand, private datasets are restricted and require special permissions, offering detailed and proprietary information typically used for business and research purposes.

Public Datasets

- **Government Portals:**
  - Data.gov.uk
  - European Union Open Data Portal
- **Organizations and International Bodies:**
  - World Bank Open Data
  - United Nations Data
- **Educational Institutions:**
  - UCI Machine Learning Repository
  - Kaggle Datasets
- **Research and Science:**
  - CERN Open Data
  - NASA Open Data
- **Social Media and Internet:**
  - Twitter API
  - Reddit Datasets

The Data Education in Schools team has a collection of publicly accessible datasets. You can view the Trello Board to see them.

Private Datasets

- **Corporate Data:**
  - Internal company databases
  - CRM systems (e.g., Salesforce)
- **Subscription Services:**
  - Bloomberg Terminal (financial data)
  - Nielsen (consumer and market data)
- **Partnership Datasets:**
  - Collaborative projects between companies
- **Proprietary Research:**
  - Market research firms like Gartner or Forrester
- **Industry-specific Sources:**
  - Healthcare records (accessible with appropriate permissions)
  - Retail transaction data

1d - Explain techniques for keeping data secure.

If data is private, it is critically important to both individuals and businesses to keep it secure. This will stop it falling into the wrong hands.

Keeping data safe is everybody's responsibility. Human beings are often unknowingly the weakest link in keeping data secure.

Personal data

information that relates to an identified or identifiable individual

Strategies for keeping personal data secure might include methods such as:

- **Strong passwords**: a combination of letters, numbers and special characters that are difficult to guess by a person or program.
- **Password manager**: a software that securely stores passwords that a user has for online accounts.
- **Anti-virus software**: Software designed to detect and destroy computer viruses.
- **Using encryption**: A way of scrambling data so that it can only be decoded by the intended recipient.

The data itself should also be secured by being encrypted both whilst being stored and when in transit. It is also good practice to ensure it contains a minimum amount of sensitive information in the first place.

- **Physical security**: Ensuring that the physical location of data is well secured, allowing only those who should access it physically.

- **Backups**: To avoid accidental loss of data by deletion or corruption, regular backups should be taken.
- **Access Limitation**: Only users with a valid reason to access the data should be able to. Permissions should be time limited and removed when no longer required.
- **Testing & Monitoring**: Regular ethical hacking tests to identify weaknesses should be carried out. Monitoring of systems access can also identify data breaches.

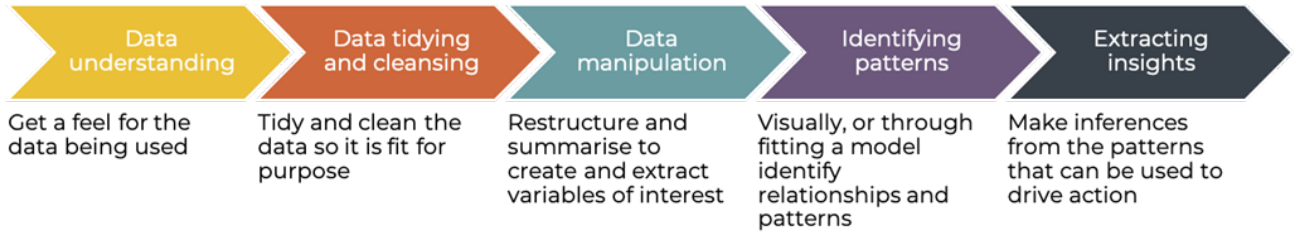


Figure 1: PPDAC's analysis step

2 Outcome 2 - Explain techniques in analysing a dataset.

Analysis Steps

Data analysis involves the transformation of raw data into useful information or insights in a structured and organised way.

It is generally accepted that around 80% of any data analyst's time is spent cleaning and manipulating data. These activities are both important and time consuming. The other activities involve detailed understanding of the dataset before any manipulation and ensuring that conclusions are drawn, and actions are taken at the end of the process.

There is a structured approach to carrying out a data analysis, which if followed will minimise mistakes and maximise the validity of the conclusions or insights extracted from the data. This is what would be done within PPDAC's analysis step as seen in Figure 1.

2a - Describe common data types and data formats, including structured and unstructured data.

Data Types

Data type

How data is stored internally to the computer.

Examples of data types:

- **Integers:** Whole numbers with no decimal or fractional parts.
- **Floating point:** Numbers that can contain a decimal or fractional part.
- **Character:** A single text character which can be a letter, number, or symbol.
- **Boolean:** Can take two possible values, such as true/false or yes/no. Often stored as 0 and 1.
- **Date and time:** The number of days or seconds passed since the 'epoch' date, normally 1/1/1970.

Changing the data type will affect the precision of the value stored.

Data Formats

Examples of different data types displayed using a variety of formats are given below:

Data type	Display format	Stored value	Displayed value
Floating point	1 decimal place	22.6176470588235	22.6
Floating point	percent	0.4893	48.9%
Date	%d-%m-%y	18496	22-08-20
Date	%B %d, %Y	18496	August 22, 2020
Date	%b %Y	-2000	Jul 1964
Time	%Y-%m-%d %H:%M:%S	1584801002	2020-03-21 14:30:02
Time	%r	1584801002	02:30:02 pm
Time	%c	-1613826000	Mon 11 Nov 11:00:00 1918
Boolean	TRUE/FALSE	1	TRUE
Floating point	£	24.99	£24.99

Structured and Unstructured Data

We can separate data into structured and unstructured.

Structured data is organised in a specific format. For example, a table with rows and columns, where each column has a defined data type like integers, strings, or dates, allowing for easy storage and analysis.

Unstructured data lacks a predefined structure - for example, documents, images, or social media content.

Types of Data

- **Nominal:** Data that represents categories with no intrinsic order.
  - Examples: Colours (red, green, blue), types of fruit (apple, banana, cherry).
- **Ordinal:** Categorical data with specific order or ranking.
  - Examples: Survey ratings (satisfied, neutral, dissatisfied), education levels (high school, bachelor's, master's).
- **Categorical:** Data that can be divided into specific groups of categories.
  - Examples: Yes/no questions, types of vehicles (car, truck, motorcycle).
- **Discrete:** Quantitative data with distinct, separate values that can be counted.
  - Examples: Number of students in a class, number of cars in a parking lot.
- **Continuous:** Qualitative data with an infinite number of possible values within a range.
  - Examples: Height, weight, temperature and time.

2b - Explain techniques for data capture, cleaning and manipulation.

Capturing data

Methods for capturing data:

- Surveys and Forms: Gather data directly from people.
- Sensors and IoT Devices: Capture real-time data from physical environments.
- Web Scraping: Extract data from websites using automated tools.
- APIs: Gather data from online services and platforms which allow access to their data.

Cleaning data

Why Do We Need to Clean Data?

Cleaning data is essential in data science because raw data is often messy, containing errors, missing values, or inconsistencies that can lead to incorrect conclusions. By cleaning data—removing duplicates, filling in gaps, and correcting mistakes—we ensure that analysis is accurate and reliable. This helps us make better decisions based on trustworthy information. For example, if a dataset has spelling mistakes in category names, it could lead to misleading results when sorting or filtering data. Cleaning data improves the quality of insights we can gain, making it a crucial step in any data science project.

The Steps in Data Tidying.

The first step in preparing data is to tidy it up. There are a few activities that this could involve, including:

- **Naming or renaming columns** so that the data is easily understood, and the names are informative.
- **Dropping unnecessary columns** so that only those required for the analysis remain. This saves disk space and speeds up processing.
- **Reformatting columns** so that numbers and dates/times are stored correctly and not as strings.
- **Fixing strings** so that the case (upper/lower) and spaces are consistent, allowing comparison.

- **Fix Missing Data** – Handle missing values by filling them in with appropriate estimates (e.g., averages or placeholders) or removing incomplete rows if necessary.
- **Fix Errors** – Identify and correct mistakes like typos, incorrect entries, or inconsistent capitalisation to improve data accuracy.
- **Convert Between Different Data Types** – Ensure values are stored in the correct format, such as changing numeric strings to actual numbers or dates to a standard date format, to enable proper calculations and analysis.

Manipulating Data

Data manipulation enables the transformation of raw data into a structured format ready for analysis. Effective data manipulation helps ensure accuracy and relevance in your findings. This section explores essential techniques and tools for manipulating data, providing practical insights for both beginners and advanced users.

Manipulating Data in Excel

Excel is a widely used tool for data manipulation with various capabilities:

- **Sorting and Filtering:**
  - Sort your data by selecting a column and clicking on "Sort A to Z" or "Sort Z to A."
  - Use the "Filter" option in the Data tab to display specific entries. [Excel Support]
- **Formulas and Functions:**
  - Use `=SUM(A1:A10)` to find the sum of a range or `=VLOOKUP(value, table, col_index, FALSE)` to search for a value.
  - Explore more functions [here].
- **Pivot Tables:**
  - Insert a Pivot Table via the "Insert" tab, drag fields to areas for a summary, and analyze complex data easily.
  - Learn more about Pivot Tables [here].
- **Data Validation:**
  - Go to the "Data" tab, click "Data Validation" to set rules for data entry, ensuring quality and consistency.
  - More on data validation [here].

Manipulating Data in Python

Python provides powerful libraries for advanced data manipulation:

- **Pandas:**
  - Use `DataFrame` for data manipulation. E.g., `df['column'].mean()` calculates the mean.
  - Pandas documentation [here].
- **NumPy:**
  - Perform mathematical operations on arrays, e.g., `np.sum(arr)`.
  - Learn more at [NumPy Documentation].
- **Matplotlib and Seaborn:**
  - Create plots with `plt.plot()` for visualization. Seaborn provides high-level interface for drawing statistical graphics.
  - More on Matplotlib [here] and Seaborn [here].

- **Data Cleaning:**
  - Use `df.dropna()` to handle missing values or `df.drop_duplicates()` to remove duplicates.
  - Explore data cleaning techniques [here].

2c - Explain the use of descriptive statistics used to summarise a dataset.

Descriptive analytics focuses on summarizing historical data to identify patterns and trends. It answers the ”what happened” question and provides a clear picture of the past.

- **Sum / Totals:** The total value of a set of numbers, found by adding them together.
- **Averages (Mean, Median, and Mode):**
  - *Mean:* The sum of all values divided by the number of values, giving the central value of a dataset.
  - *Median:* The middle value when data is arranged in order, useful when there are extreme values (outliers).
  - *Mode:* The most frequently occurring value in a dataset, useful for identifying common trends.
- **Range:** The difference between the highest and lowest values in a dataset, showing how spread out the data is.
- **Percentiles:** Values that divide a dataset into 100 equal parts, helping to compare individual data points to the overall distribution (e.g., the 90th percentile means a value is higher than 90% of the other values).
- **Frequency Tables:** A table that shows how often each value or group of values appears in a dataset, making it easier to identify patterns and trends.
- **Standard Deviation:** A measure of the amount of variation or dispersion in a dataset. A low standard deviation indicates that data points are close to the mean, while a high standard deviation indicates more spread out data.

Predictive Analytics

Predictive analytics uses statistical models and machine learning techniques to predict future outcomes based on historical data. It answers the ”what could happen” question, helping to anticipate trends and behaviors. You are not expected to use these, but you are expected to know the difference between predictive and descriptive analytics. Some examples of predictive analytics are:

- **Regression Analysis:** Predicts a dependent variable based on one or more independent variables.
- **Classification:** Categorizes data into predefined classes or categories.
- **Time Series Analysis:** Analyzes data points collected or sequenced over time to forecast future events.

Mark	Tally	Frequency
4	II	2
5	II	2
6	IIII	4
7	𝍸	5
8	IIII	4
9	II	2
10	I	1

Figure 2: A frequency table

Describe the Composition of a Structured Dataset

Structured datasets are organized in a way that makes them easily accessible and analyzable. They typically consist of specific data structures which store and manage data efficiently.

Strings: A collection of characters combined to create alphanumeric text. Used for storing words, sentences, or identifiers such as names and codes.

Array: A fixed-size structure for storing items of the same data type. Efficient for managing static collections of numeric data or other simple data types (such as strings).

Vector: A one-dimensional array, commonly used in mathematical computations. Useful for storing operations that require linear data processing like statistical analysis.

List: A dynamically sized structure which can contain different data types. Flexible for storing collections of varied data types, such as numbers and strings.

Data Frame: A two-dimensional structure designed for holding datasets. Each column can hold different data types, but all must contain the same number of items. Essential for data analysis, allowing for efficient manipulation of complex datasets, typically seen in tools like R or Pandas (Python).

2d - Explain techniques for data visualisation and data storytelling, including accessibility considerations.

Frequency tables

Frequency tables (Figure 2) provide a structured way to display how often each value in a dataset occurs. They are particularly useful for summarizing categorical data and identifying patterns.

A typical frequency table lists categories alongside their corresponding counts or frequencies. They offer a clear, concise overview of data distribution, making it easier to spot trends and outliers. Frequency tables are often used as a preliminary step before creating more complex visualizations like bar charts or histograms.

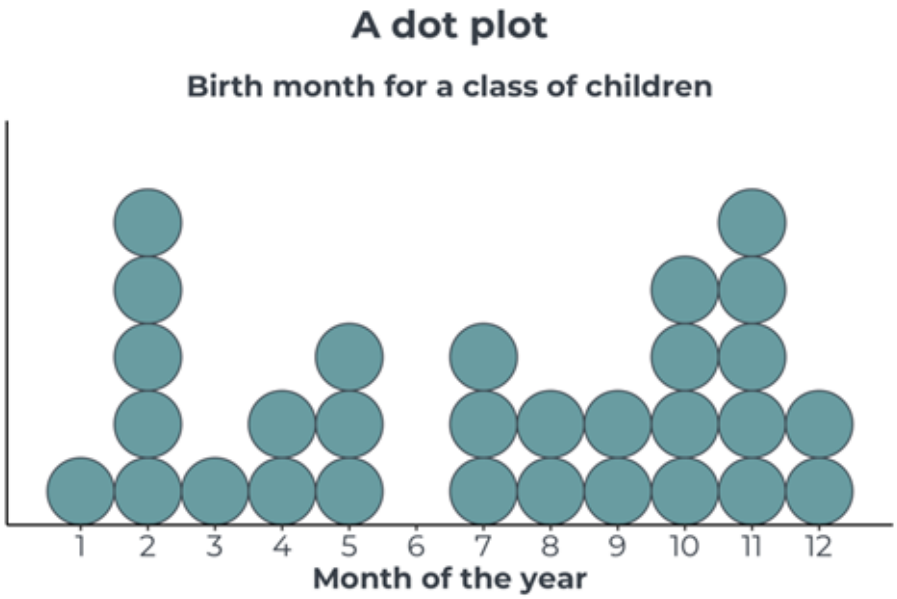


Figure 3: A dot plot.

Dot plot

In a dot plot (Figure 3, 4), each dot represents a single observation. For example, this dot plot records the month each child in a class of children was born. The dots can also be swapped for icons or images for a more visually appealing graphics.

Bar Chart

Bar charts (Figure 5) use rectangular bars to compare values in different categories. The bars normally show the counts or sizes of categorical data. Since there is no connection between the bars, they are normally shown not touching.

A horizontal bar graph (Figure 6) is often a good option when there are many categories, or the category labels are long. It is also possible to reorder the bars, which makes it easier to see the smallest and largest categories. These can also be highlighted by using different colours.

Line graph

Line graphs (Figure 8) are used to show the change, or evolution of a numerical variable as another quantity varies. Both the x-axis and y-axis are numeric, with the x-axis containing the varying quantity. This is often time but could be another varying quantity such as temperature or distance. The data points in a line graph are joined sequentially by lines.

Pie Chart

Pie charts (Figure 10) show the proportion of a whole. The total of the pie must add up to 100%. Although popular, pie charts are often not the best choice of graph to use, since it is much more difficult for human brains to estimate relative angles, or segments of the chart.

When they are used with more than two or three segments, it isn't easy to pick out slivers or

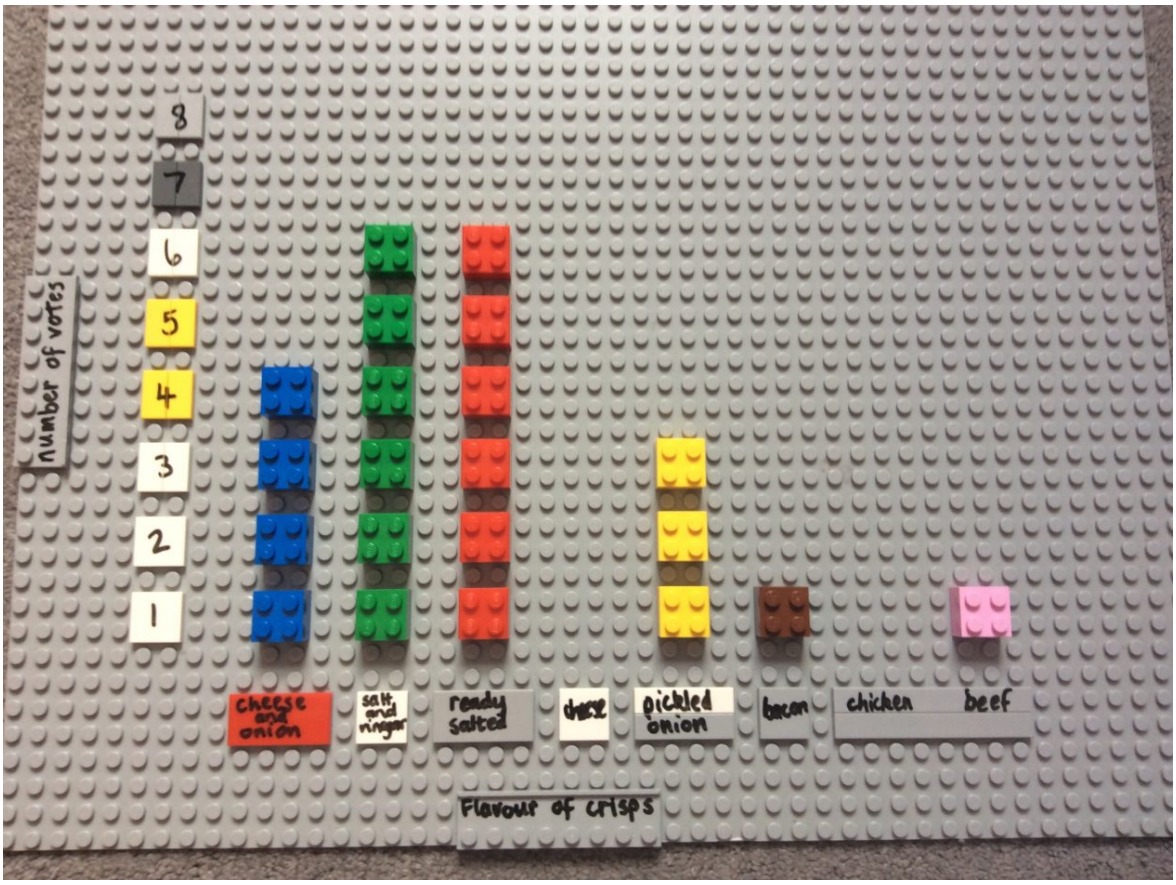


Figure 4: A dot plot made out of Lego.

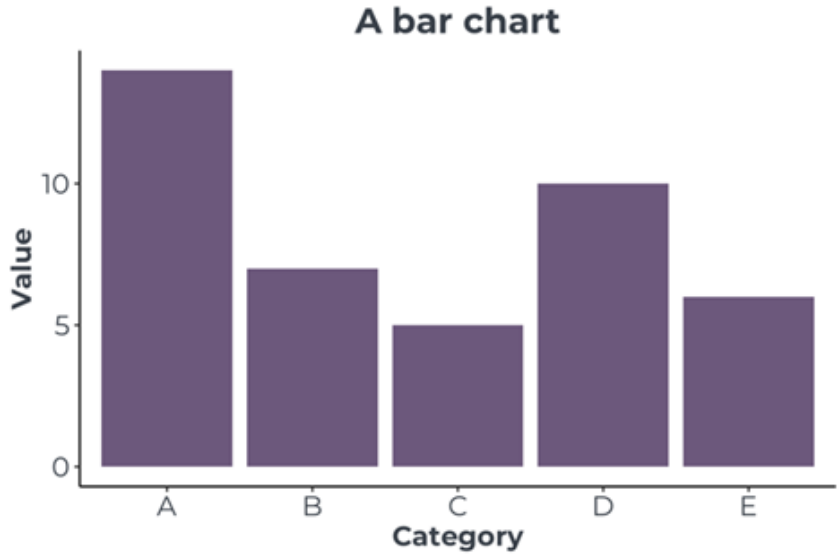


Figure 5: A bar chart.

Figure 6: A horizontal bar chart.

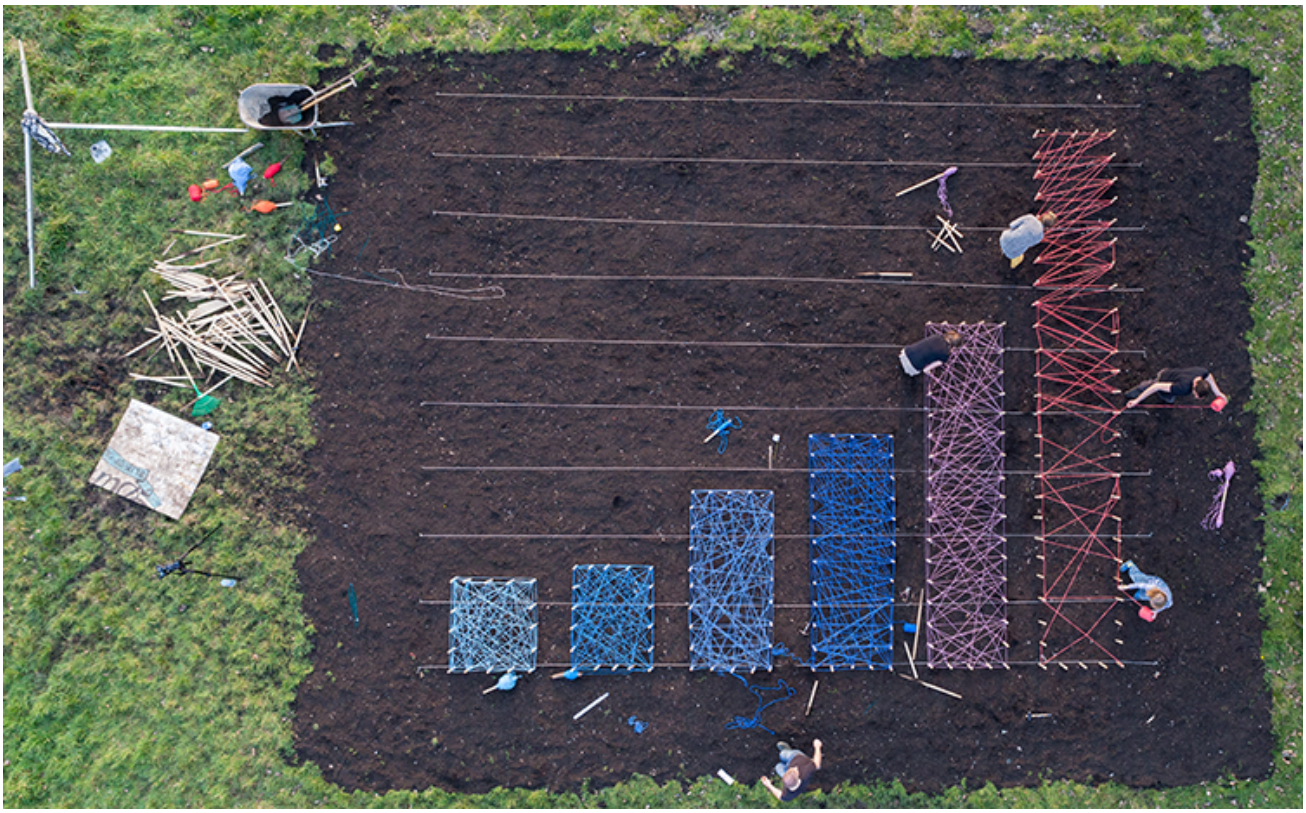


Figure 7: A vertical bar chart being made in a garden.

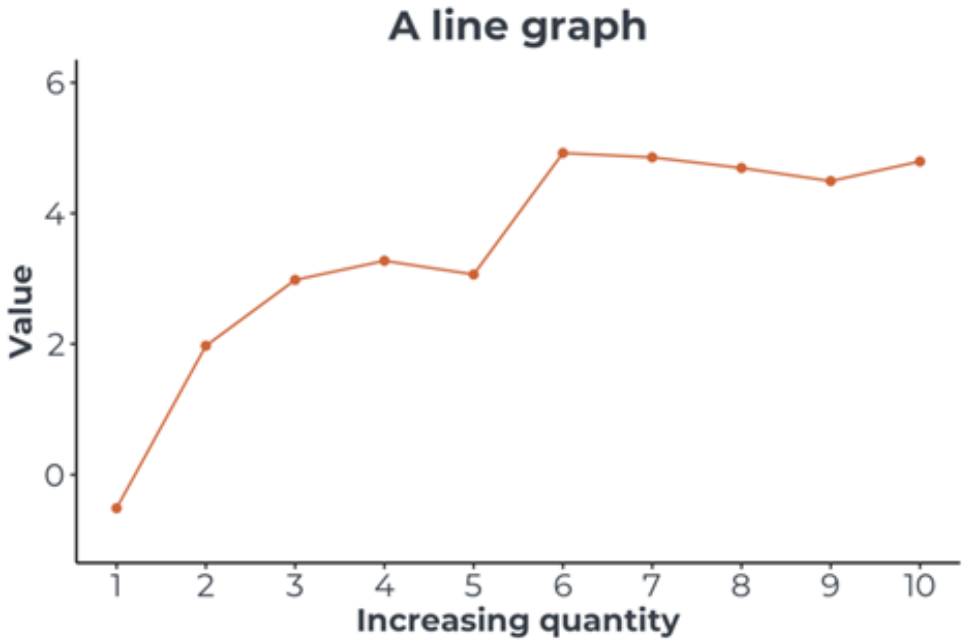


Figure 8: A line graph.

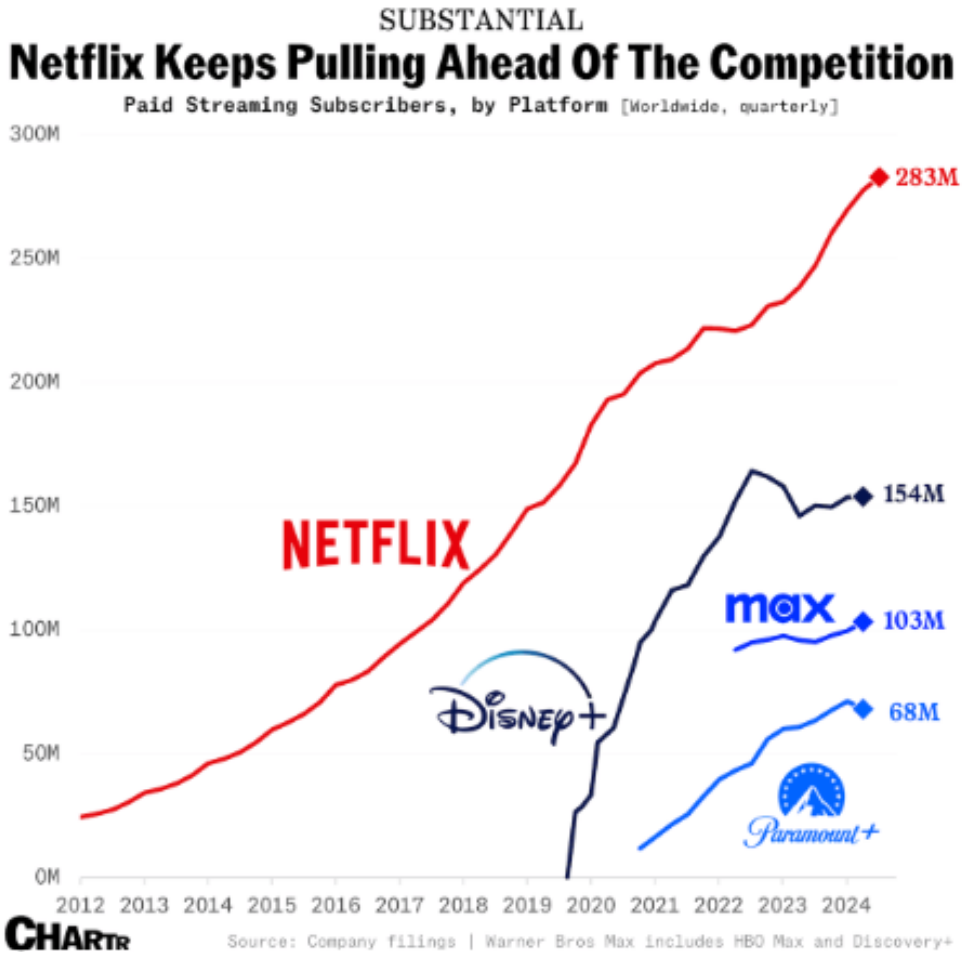


Figure 9: Line chart showing streaming service subscribers by platform.

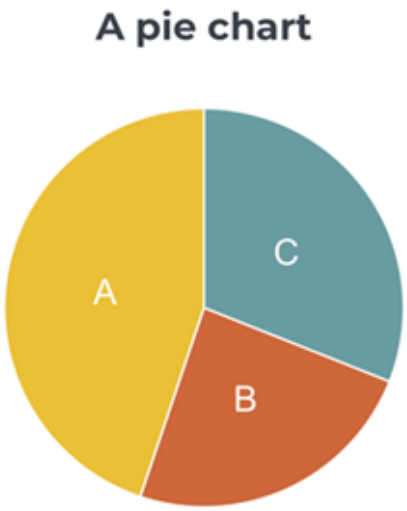


Figure 10: A pie chart.

Figure 11: Pie chart showing the amount of pie eaten versus the amount not (yet) eaten.

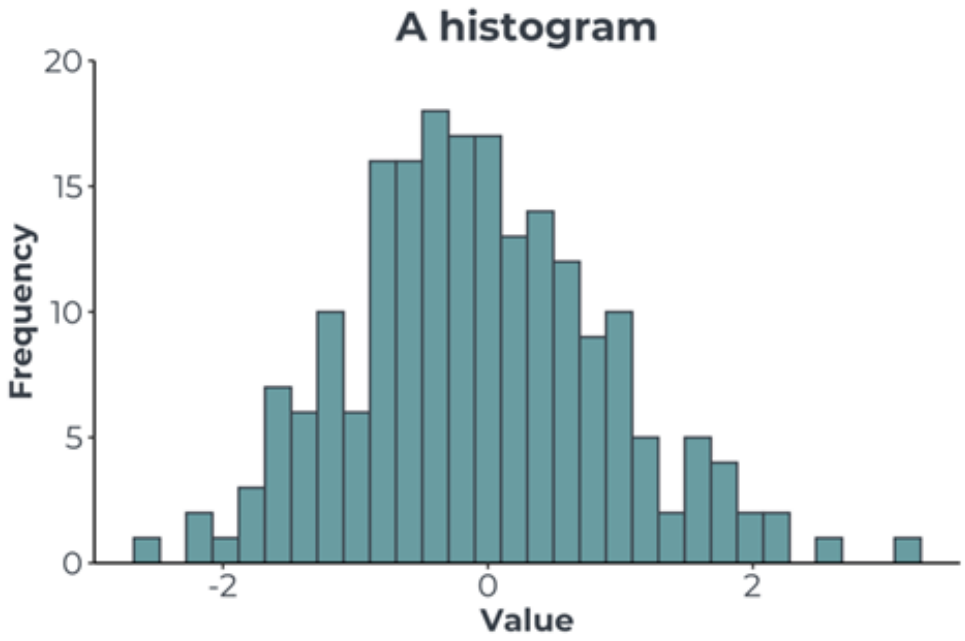


Figure 12: A histogram.

compare relative segment sizes. A bar chart (Section 2.5) can always be used in place of a pie chart and is much clearer to read.

Histogram

A histogram might look very similar to a bar chart, but it is fundamentally different since it is plotting numerical rather than categorical data.

Histograms (Figure 12) are used to examine the distribution of a numerical variable. The x-axis contains the value of the numerical variable, which is then binned into ranges, and the frequency of points in the range is displayed on the y-axis. The bars on a histogram should always be displayed as touching, since the variable is continuous.

Scatterplot

Scatterplots (Figure 13) are used to show the relationship between two numerical variables. Both the x-axis and y-axis contain numerical quantities. There is often a line of best fit added to demonstrate the relationship between the two variables.

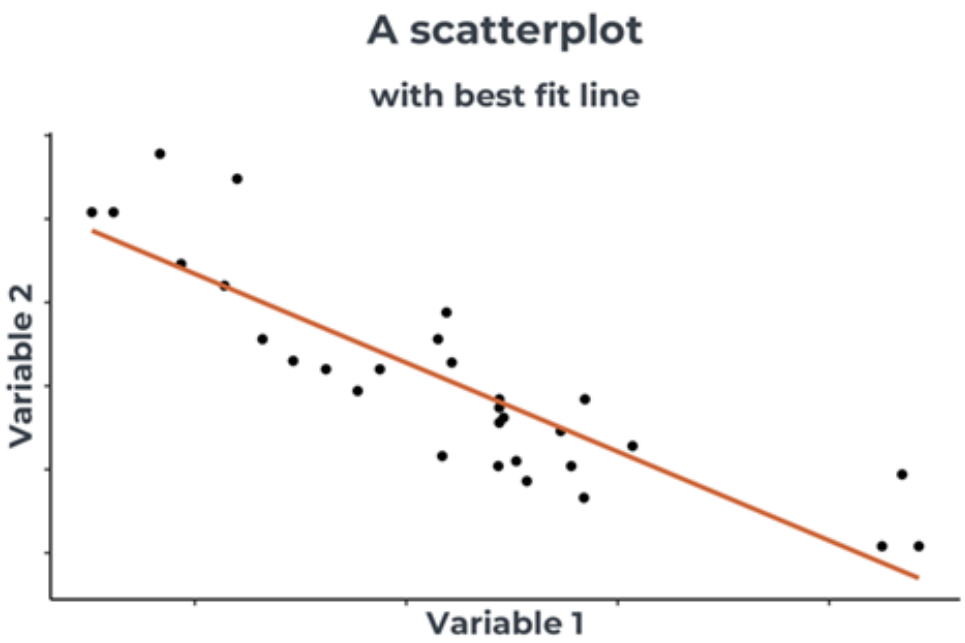


Figure 13: A scatterplot

3 Outcome 3 - Analyse data to extract insights.

In this outcome, you will apply the knowledge that you have gained in Outcome 1 and Outcome 2. You will perform analyses on datasets!

Spreadsheets

Spreadsheet tools are suitable for quick ad-hoc data analysis and visualisation. The two most popular tools are Microsoft's **Excel** and **Google Sheets**. Both tools have scripting languages which can be used to automate repetitive tasks. Challenges with all spreadsheet tools include auditability, error checking and reproducibility of analysis.

Excel (Figure 14)

- Available at office.com
- Wide array of graphical choices
- Maximum dataset size of 17 billion items
- Version control not automatic

This video by Kevin Stratvert gives a helpful introduction to Excel: Video.

Google Sheets (Figure 15)

- Available at sheets.google.com
- More limited set of graphical choices
- Maximum dataset size of 5 million items
- Automatic version control

Google Sheets works similarly to Excel, with some differences in syntax and capabilities.

Programming Languages for Data Analysis

The advantages of using a programming language for analysis and visualisation are flexibility and reproducibility.

The main open source languages that data scientists use are R and Python; however, there are commercial languages available too. R and Python both have special data types designed for manipulating tidy tabular data. They all have multiple packages for creating visualisations and more complex dashboards that can be implemented in web applications.

Python

- A object-oriented capable programming language
- A recommended IDE to use is Spyder, available as part of the Anaconda suite

Popular Python packages for data science:

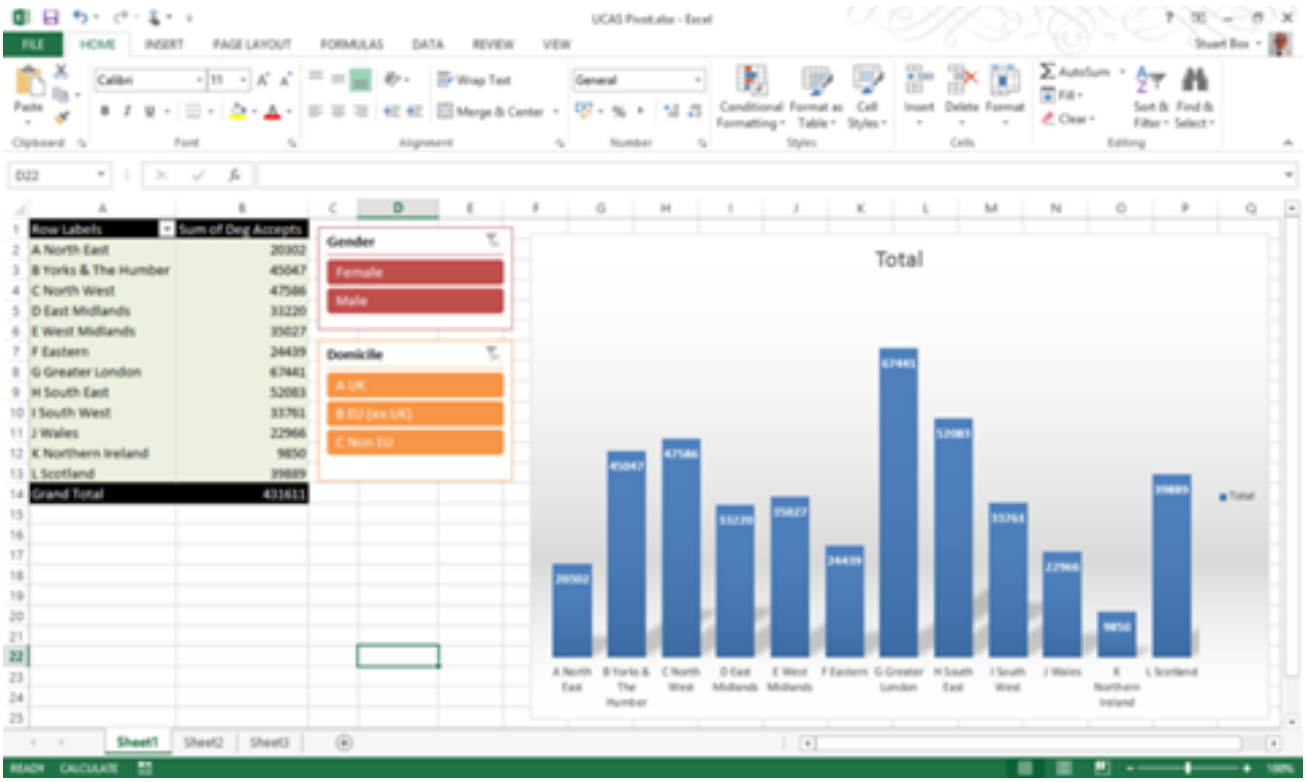
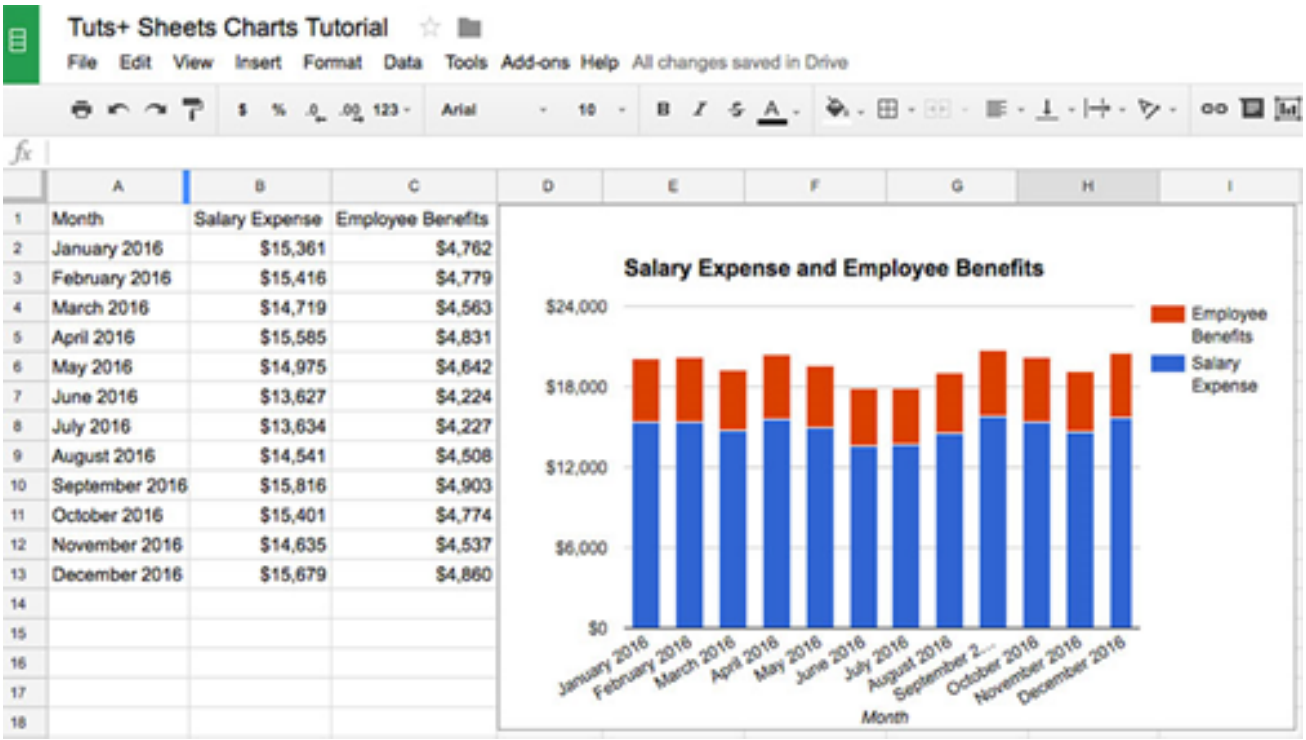


Figure 14: A screenshot of Microsoft Excel.



	A	B	C
1	Month	Salary Expense	Employee Benefits
2	January 2016	$15,361	$4,762
3	February 2016	$15,416	$4,779
4	March 2016	$14,719	$4,563
5	April 2016	$15,585	$4,831
6	May 2016	$14,975	$4,642
7	June 2016	$13,627	$4,224
8	July 2016	$13,634	$4,227
9	August 2016	$14,541	$4,508
10	September 2016	$15,816	$4,903
11	October 2016	$15,401	$4,774
12	November 2016	$14,635	$4,537
13	December 2016	$15,679	$4,860

Figure 15: A screenshot of Google Sheets.

- Pandas for data manipulation
- Numpy for data functions
- Matplotlib or Seaborn for visualisation
- Plotly for interactive plotting

3a - Plan an analysis of a dataset to solve a problem.

In the planning phase of a data analysis, we decide how we will answer the question that we have posed and the steps that we will follow in order to gain our insights.

In this example, let's investigate the question "How does student satisfaction relate to academic performance?"

In order to plan this analysis, let's list the steps that are required in order to carry in out.

1. **Data collection**: In order to carry out our analysis we require data. We might collect this from our fellow students through a survey. Which data are we interested in for this study? We are interested in student satisfaction and academic performance. We might ask the students to provide a rating from 1 to 10 stating their satisfaction at school, and their average grade.
    - How satisfied are you at school from 1-10?
    - What is your average grade (percentage) over the last year?

    If we are interested in any other data items, we could ask for those too. For example, it might be interesting to investigate how satisfaction relates to year group.

    - What year group are you in?
2. **Analysis**: Although we have not yet collected our data, it is helpful to plan how we will analyse our data. Visualisations will help us to understand our data further, and present our findings to others. We might decide that suitable visualisations are:
    - A scatterplot with satisfaction on the y axis and average grade on the x axis.
    - A bar chart showing satisfaction, coloured by year group.

Now that we know how we will collect and analyse our data, we can begin our investigation, collecting data and creating visualisations. [Link to section which explains this]

3b - Identify potential sources of bias in a dataset.

You should be able to identify sources of bias in your particular dataset.

This video gives an example of a dataset of children's heights and arm span where biases are present, and defines the term "artefact". [Maybe describe this here??]

3c - Tidy, clean and manipulate a dataset.

You are required to carry out the steps to tidying and cleaning a dataset. Refer to Section 2.3 for these steps.

At level 6, you are expected to automate these processes in some way. For example, instead of manually removing rows with missing items, using the filter tool to do it automatically.

3d - Perform analyses on the data.

This section describes the practical steps you need to take to perform certain basic analysis on an Excel database.

Sort

To sort in Excel, do the following:

1. Select all the data including the headings that you need to sort.
2. In the Home ribbon, click on '*Sort Filter*' then on '*Custom Sort...*' **OR** Right click and find the '*Custom Sort...* option.
3. Choose the column header you would like to sort by (e.g. height_m)
4. Select the order you want to sort by (e.g. smallest to largest)

Filter

Filter

To choose some of the rows in a dataset based on some criteria.

When filtering data it can help to think about the following questions:

- What **data** do I have?
- What do you **need** from the data?
- What **criteria** do I need to filter my data by?

How to Filter in Excel

1. To turn on the filter options highlight all the data you want to filter.

   Press 'Ctrl' and 'Shift' and 'L' if you are using Windows or 'Command' and 'Shift' and 'L' for Mac at the same time.

   This tells Excel to make your selected cells a table. Excel now shows filter arrows next to each heading. We can use these to filter our data.
2. Click on the small arrow that has now appeared next to the column you want to filter, and select 'Number Filters'.
3. Fill in the 'Custom AutoFilter' box and press OK.

Summarise

Summarise

To condense the rows in a dataset (often to a single value) by performing a calculation on the data items within a variable.

In the similar way that you can perform calculations on columns of data, you can perform calculations on rows of data.

The most common calculations performed on rows of data are,

- Count (number of rows)
- Total
- Average

Summarising in Excel

1. Create a new dataset with the variable headings you have selected and row labels for summary types you will calculate.
2. Type in the calculation you will use to summarise the data.
3. Copy the calculation you have just typed into the first variable, and paste into the remaining variables of the new row.
4. Repeat the process for any other summary calculations you need.

Consolidate

Consolidating Data in Excel

1. Open the worksheets containing the data you wish to consolidate.
2. Create a new worksheet where you will perform the consolidation.
3. Click on the cell in the new worksheet where you want to place the consolidated data.
4. Go to the ”Data” tab in the Excel ribbon and click ”Consolidate.”
5. In the Consolidate dialog box, select the function you want to use (e.g., SUM, AVERAGE).
6. Click ”Add” to select the data range from each worksheet. Ensure that each range has the same row and column structure.
7. Check the ”Top row” and ”Left column” options if your ranges have labels that you want to use.
8. Click ”OK” to complete the consolidation process.

Group

Group allows you to combine multiple rows or columns so they can be expanded or collapsed. This is useful for improviing the readability of larger data sets.

Grouping Data in Excel

1. Select the rows or columns you want to group in your worksheet.
2. Go to the ”Data” tab in the Excel ribbon.
3. Click on ”Group” in the Outline section.
4. Choose either ”Rows” or ”Columns” depending on your selection.
5. Excel will create a group, allowing you to collapse or expand the selected data.
6. To ungroup, select the grouped data and click ”Ungroup” in the Outline section.

3e - Create accessible data visualisation to extract insights.

Creating simple visualisations in Excel

To create simple visualizations, such as bar or line charts using Excel, follow these steps for effective graph creation:

1. **Prepare Your Data:**
   - Ensure data is clean and organized in columns and rows with clear headers.
2. **Select Data Range:**
   - Highlight the data for visualization, including headers if needed.
3. **Insert Chart:**
   - Navigate to the Insert tab on the Excel ribbon.
   - Select your desired chart type, such as Bar or Line Chart.
4. **Choose Chart Style:**
   - Pick a style that best represents your data from the options available.
5. **Customize the Chart:**
   - **Title:** Add a descriptive chart title.
   - **Axes:** Label axes to clarify their meaning.
   - **Legend:** Ensure the legend is clear.
   - **Data Labels:** Add labels to show specific values if needed.
6. **Format the Chart:**
   - Use Chart Tools to adjust colors, fonts, and styles for clarity.
7. **Review and Adjust:**
   - Check the chart’s accuracy and clarity; adjust as necessary for better insight.

This section will describe to you how to make specific types of graphs in Excel.

Creating a bar chart in Excel

How to create a bar chart in Excel:

1. **Insert a bar chart**: Highlight the dataset you are going to plot. Then in the Insert tab, click on the icon that looks like a bar chart.
2. **Select the type of bar chart**: Select the type of bar chart you would like from the drop-down list. In this case we will use a horizontal bar chart.
3. **Amend your graph**: Click on the + symbol next to the graph to add elements to graph such as Axis Titles. Then right click on any element you would like to change (e.g. the Title, Axis title)

Creating a line graph in Excel

1. **Insert a line graph**: Highlight the dataset you are going to plot. Then in the Insert tab, click on the icon that looks like a line graph.
2. **Select the type of line graph**: Select the type of line graph you would like from the drop-down list. The graph will now appear in the workbook.

Making visualisations accessible

Here are some guidelines to follow when making visualisations to ensure that they are accessible to a wide audience.

1. Use high contrast colour schemes. Orange and blue are widely thought of as the most colour blind friendly colours. Using high contrast will also help the visually impaired.
2. Don't rely on colour alone to convey information. Make sure to include clear labels and potentially patterns so that people who have difficulty differentiating colours are still able to easily understand the visualisation.
3. Ensure the size of text is suitable. Make sure that when a visualisation is displayed, the text is large enough to be read easily by those viewing it.

3f - Make recommendations based on conclusions and communicate findings.

Tips for Drawing Conclusions and Communicating Findings

In order to draw conclusions from data, we should make a claim about what a graph is showing in response to a question or issue. The reasoning used to reach a claim should be clear and logical.

When communicating findings, we should present with an audience in mind (such as peers, family, school management, or community) with a purpose, such as to inform or persuade.

1. Use values (e.g., "The average test score was 75%").
2. Use visualisations to support findings (e.g., "The bar chart shows that football is the most popular sport among students").
3. Answer a question based on the data (e.g., "What is the most common age group in the survey?").

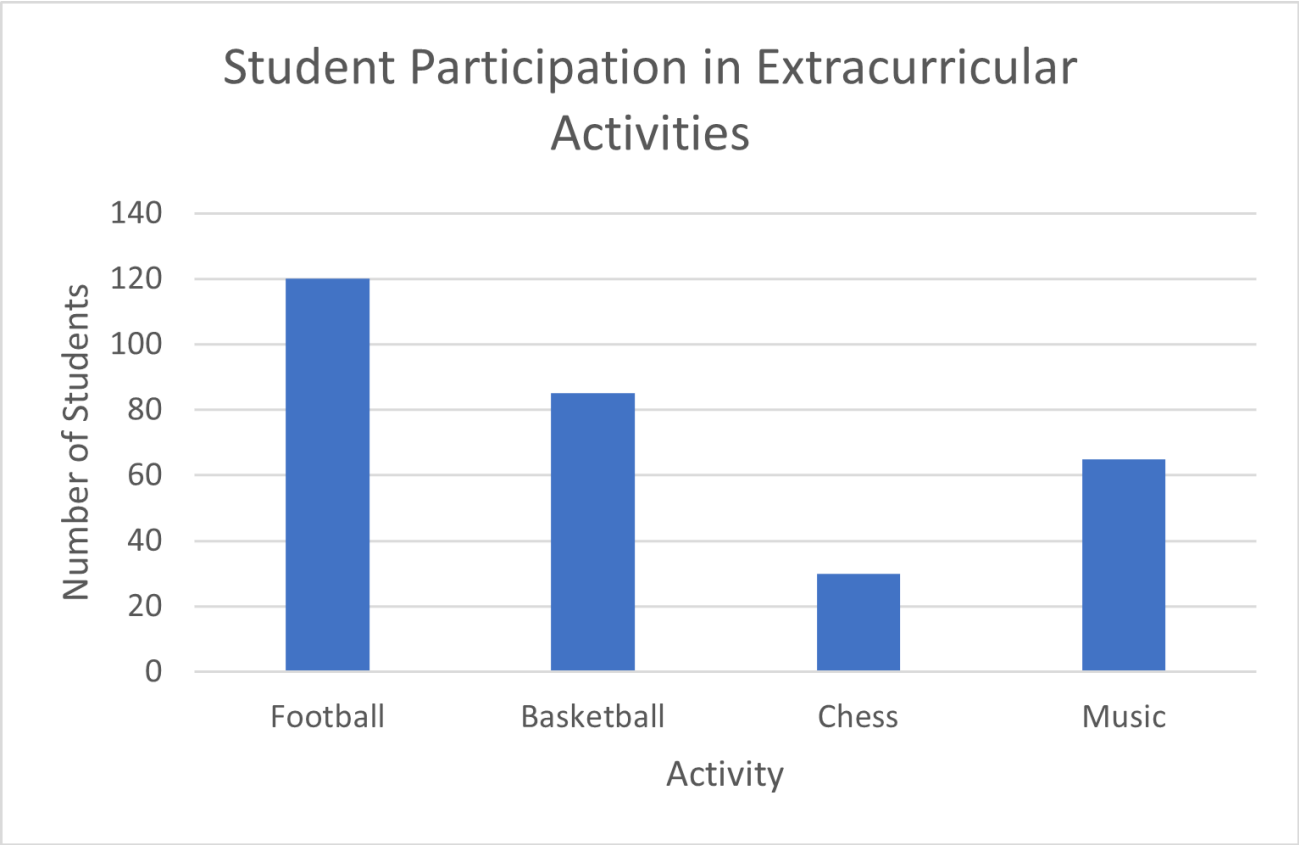


Figure 16: A bar chart showing how many students partake in different extracurricular activities.

Examples of drawing conclusions, communicating findings, and making recommendations.

Example 1: Figure 16. The bar chart shows that football has the highest participation with 120 students, while chess has the lowest with 30 students. We can conclude that football is the most popular extracurricular activity among students, indicating a strong interest in team sports. We might suggest that school resources could be allocated to support more football events, and initiatives can be developed to increase interest in less popular activities like chess.

Example 2: Figure 17. We can observe that there seems to be a positive correlation between study time and exam performance. This leads us to the potential conclusion that if students study for longer they will achieve a higher score. We might consider studying longer if we would like to achieve a higher grade.

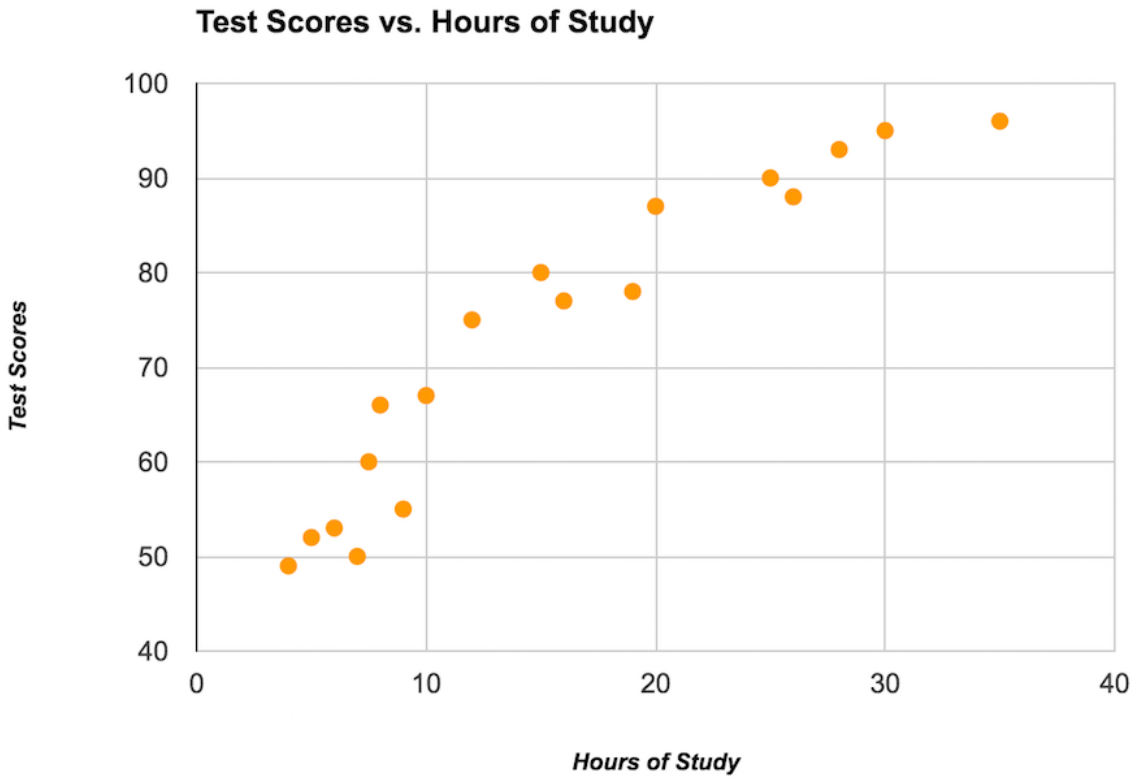


Figure 17: Graph of time studied against exam performance.