NPA Data Science

Learners' Guide to the National Progression Award

Data Education in Schools

Level 4 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.

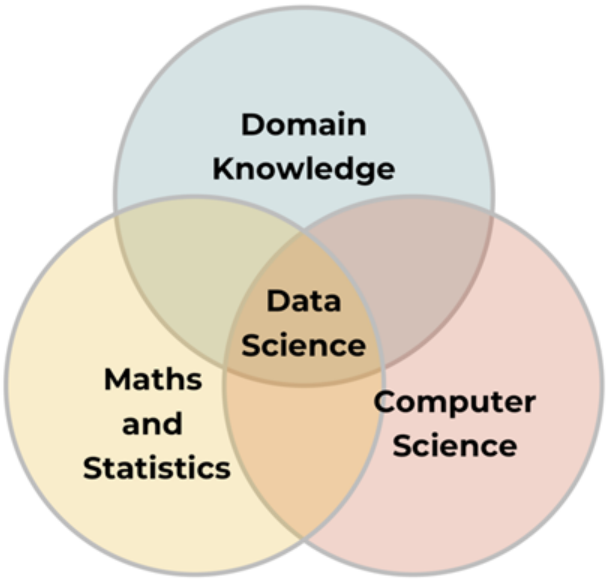
Visit www.dataschools.education for more information about support materials.

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 4 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-4-data-science-lessons/ for relevant lessons and resources for this unit, separated by outcome.



1 Outcome 1 - Describe 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 the applications of data science.

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

1b - Describe the benefits of data science.

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.

1c - State the steps in solving a problem using data science.

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.

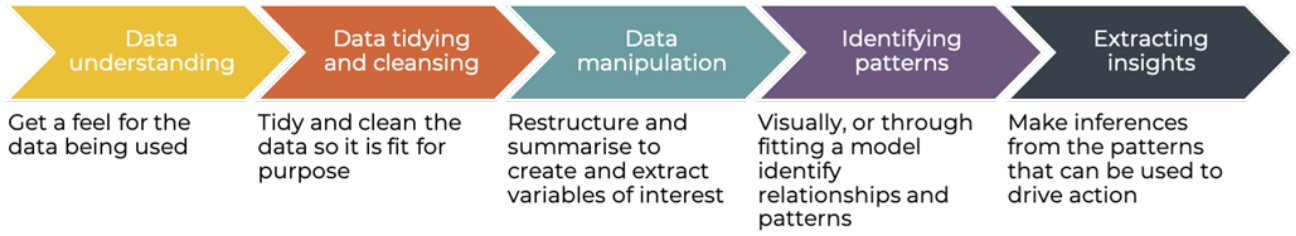


Figure 1: PPDAC's analysis step

2 Outcome 2 - State simple ways of analysing data.

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 - State common data types and data formats.

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.

2b - State simple methods of cleaning and transforming 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.

2c - State basic 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.

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

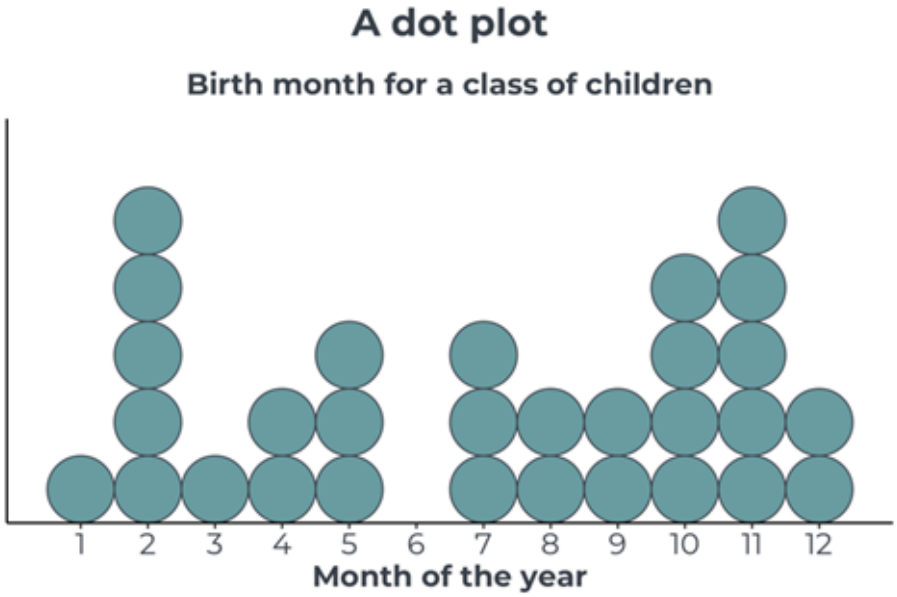


Figure 3: A dot plot.

2d - Identify types of simple data visualisations.

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.

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.

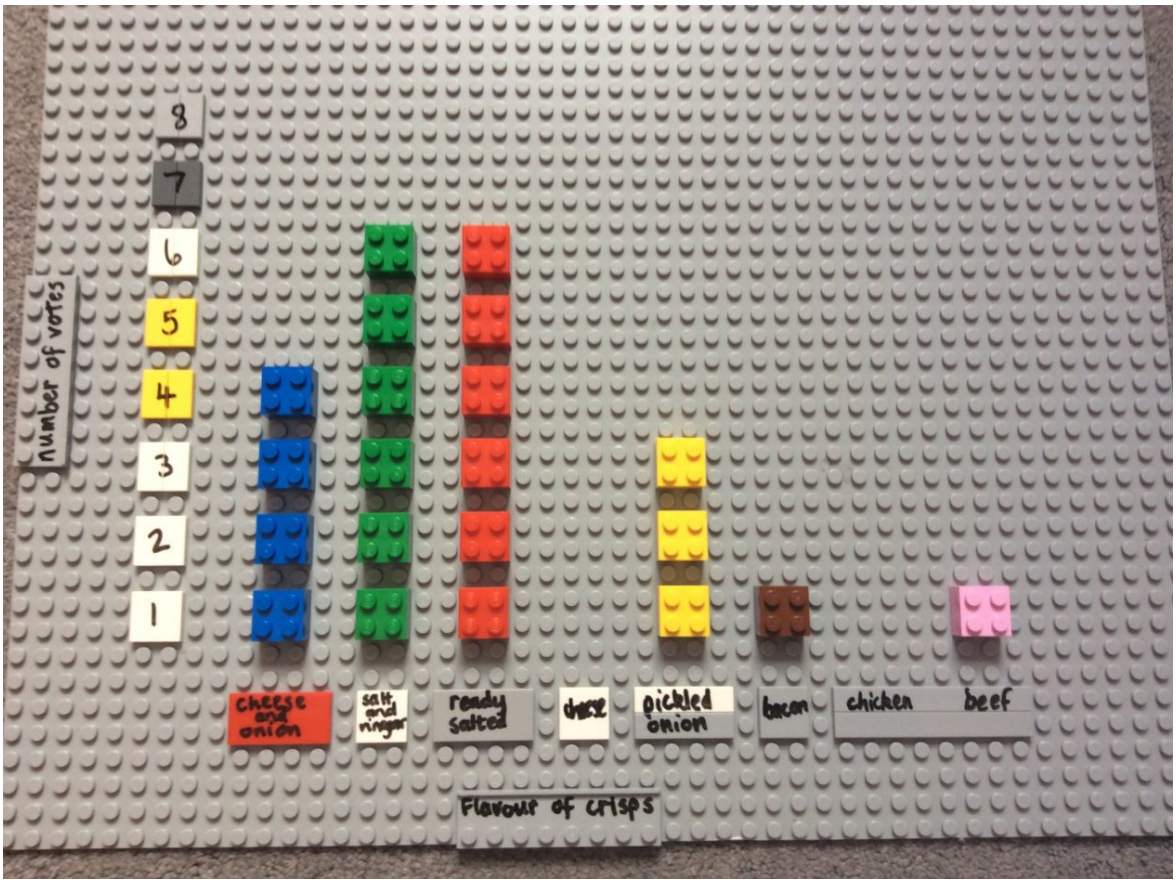


Figure 4: A dot plot made out of Lego.

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

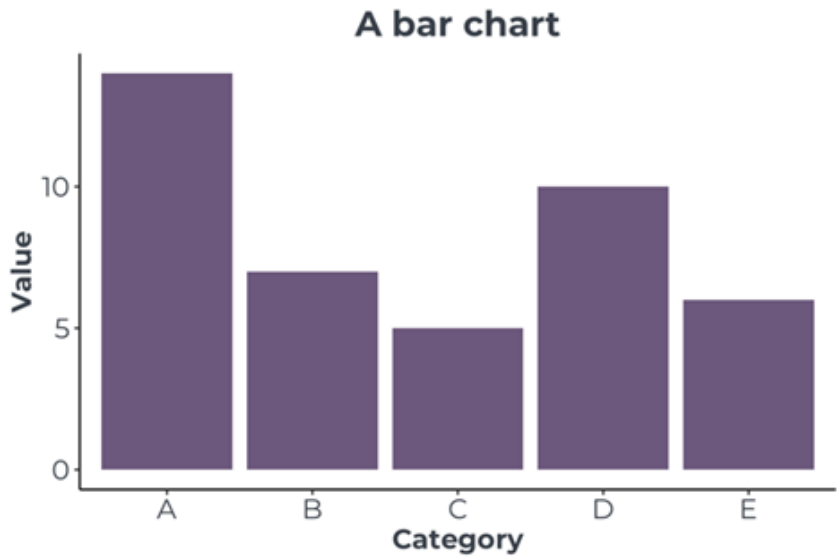


Figure 5: A bar chart.

Figure 6: A horizontal bar chart.

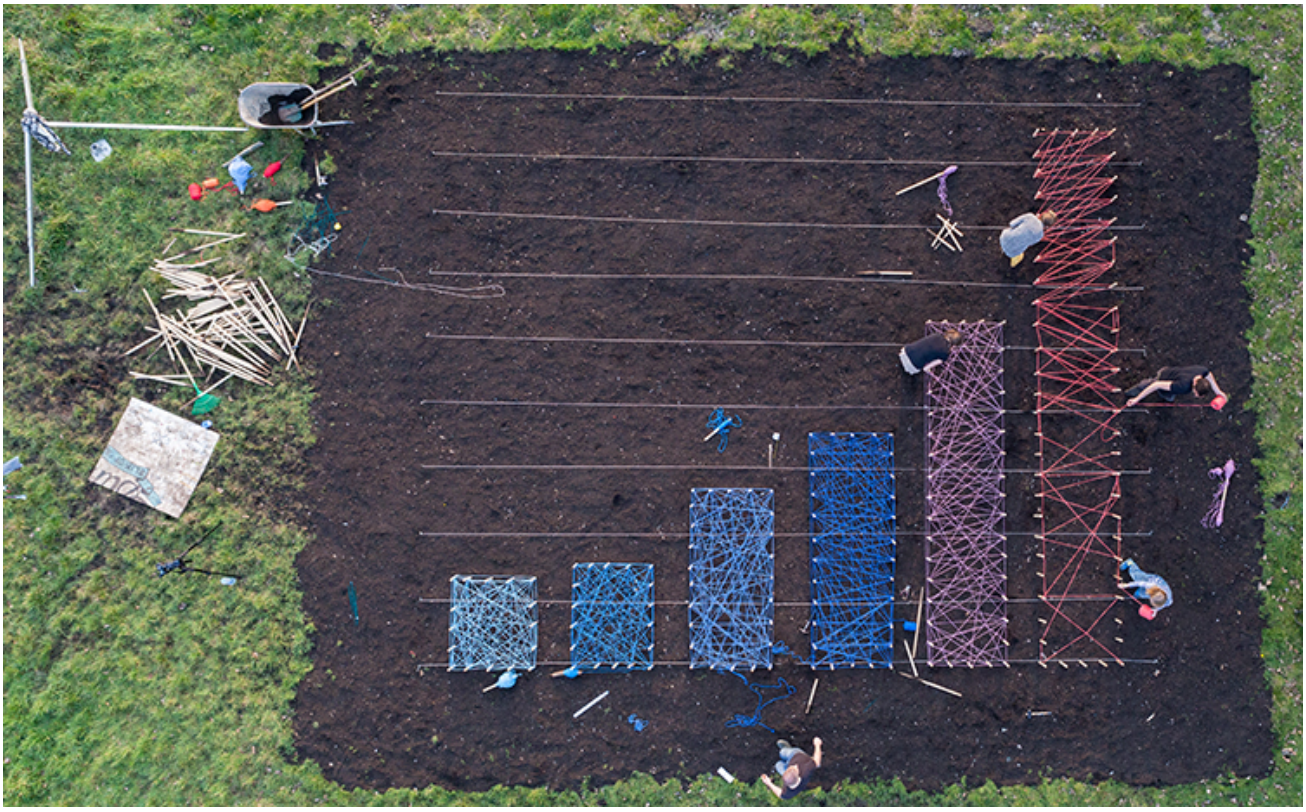


Figure 7: A vertical bar chart being made in a garden.

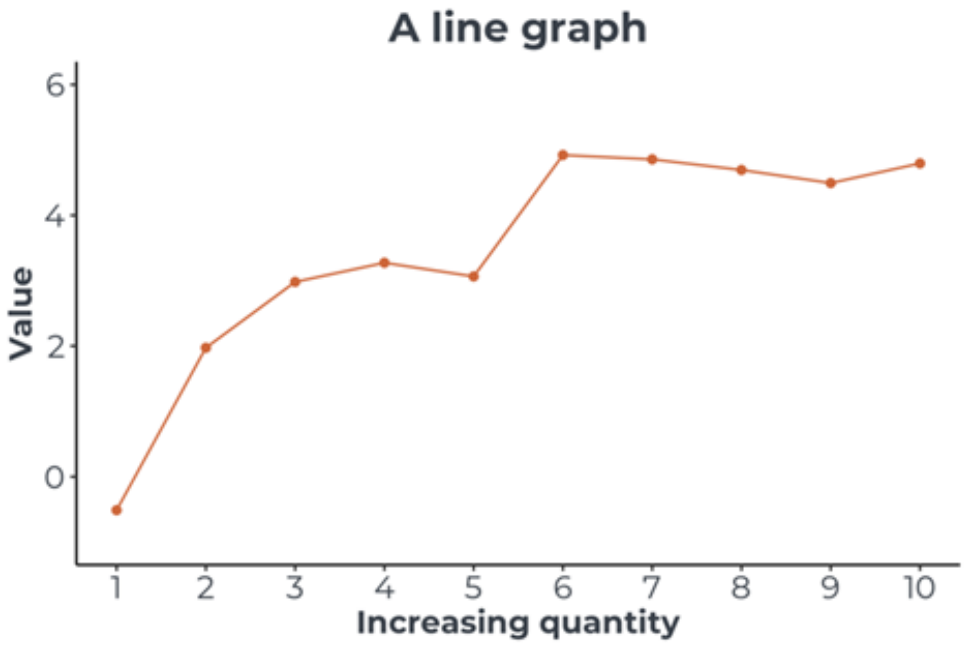


Figure 8: A line graph.

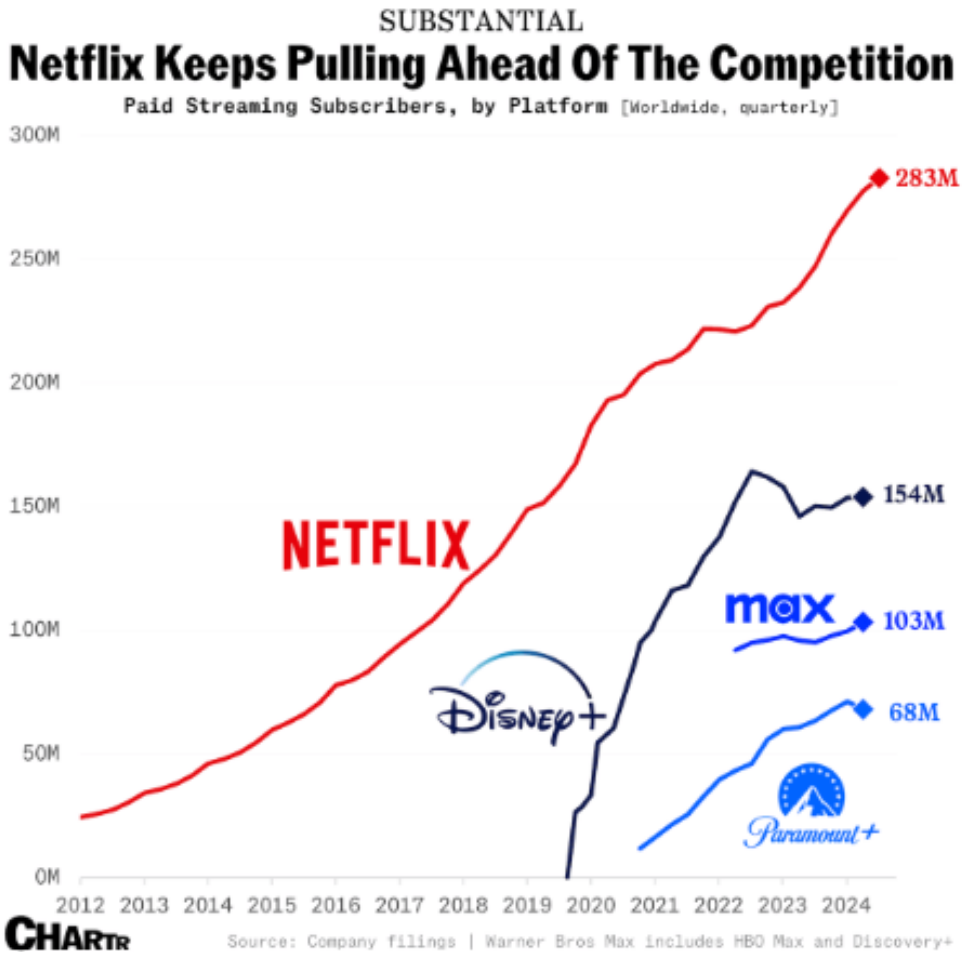


Figure 9: Line chart showing streaming service subscribers by platform.

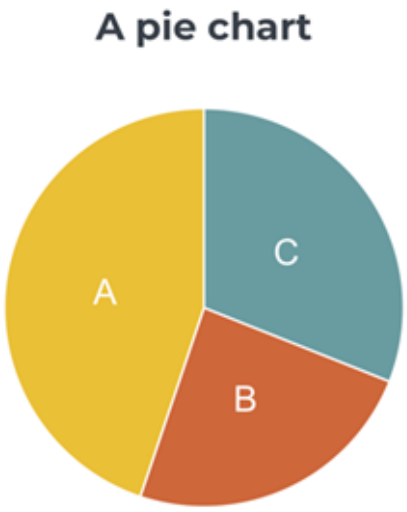


Figure 10: A pie chart.

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

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.

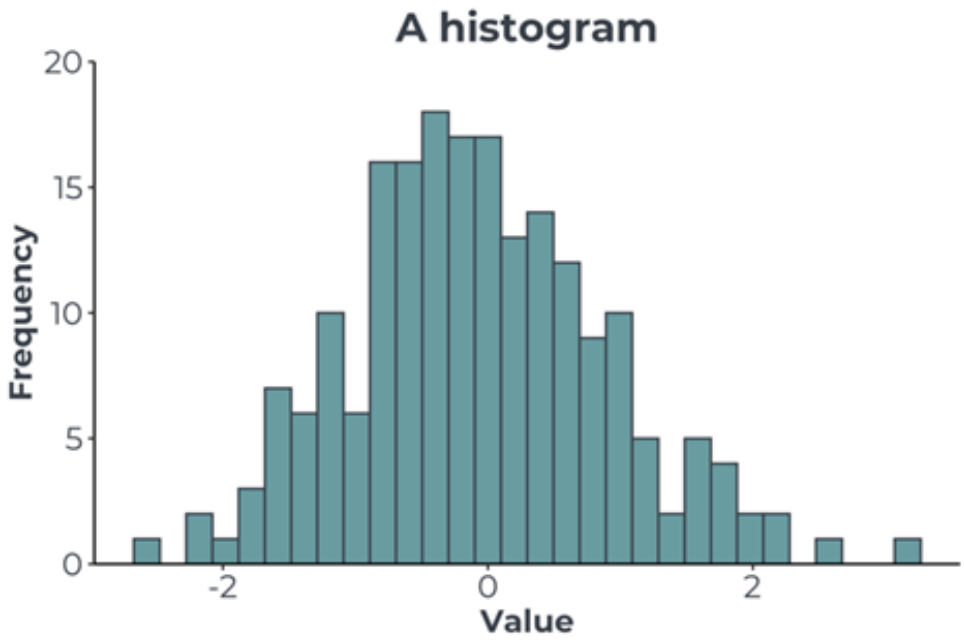


Figure 12: A histogram.

3 Outcome 3 - Analyse simple data to communicate basic insights.

3a - Perform simple data cleaning and structuring.

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

3b - Perform basic analyses including sort, filter, and summarise.

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.

3c - Visualise data to communicate basic insights.

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.