NPA Data Citizenship

Learners' Guide to the National Progression Award

Data Education in Schools

Level 4 NPA Data Citizenship 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 Citizenship unit in particular. There are separate documents available for other levels.

Scan the QR code or go to https://dataschools.education/level-4-data-citizenship-lessons/ for relevant lessons and resources for this unit, separated by Outcome.

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.

1 Outcome 1 - State the use of data in society.

This outcome focuses on the current use of data in society. This includes the use of data for social benefit, and the ethical (and unethical) use of data, such as false or deliberately misleading data. Scan the QR code on the right for relevant lessons and resources for this section.

1a - State how data is used and misused, and its impact on individuals, organisations, and society.

The role of data in decision-making.

Data plays an important part in lots of areas of society. Below are some key areas **where data plays a part in decision-making**:

1. **Home**
    - **Energy Usage:** Smart meters collect data on electricity and gas usage, helping families save money by identifying when to use appliances more efficiently.
    - **Shopping Decisions:** Online retailers track buying habits and suggest products based on past purchases or popular trends.
    - **Health Monitoring:** Smartwatches or fitness trackers collect data about steps, sleep, and heart rate, enabling people to improve their health routines.
2. **Community**
    - **Transportation:** Data from buses, trains, and traffic sensors can show where services need improvement, such as adding more buses on busy routes.
    - **Public Safety:** Communities use data from crime reports to decide where to increase patrols or install security cameras.
    - **Planning Events:** Surveys or attendance data from past events help organizers plan better events tailored to the interests of the community.
3. **Sport**
    - **Athlete Performance:** Wearable devices track speed, endurance, and recovery, helping athletes train smarter and avoid injuries.
    - **Team Strategies:** Coaches analyze data from matches, like player positions or time spent with the ball, to improve tactics.
    - **Fan Engagement:** Sports organizations use ticket sales and social media data to improve fan experiences, like offering discounts or organizing popular events.

Success Stories and Challenges in Data Usage

Below are some examples of success stories from the use of data, as well as some examples of where the use of data has caused harm.

Success Stories

1. **Using COVID-19 Data to Allocate Health Resources**

- **Case:** Hospitals use patient data to identify health risks and personalize treatments. During the COVID-19 pandemic, data was used to track infection rates, allocate vaccines, and manage hospital capacity. (*Article on Gov Website*)
- **Impact:** Lives were saved, and resources were used more efficiently.

2. **Liverpool FC Using Data to Improve Performance**
    - **Case:** In football, teams like Liverpool FC use data analytics to improve player performance and game strategies (Article). Their data-driven approach contributed to their Premier League win in 2020 after a 30-year gap. (Article)
    - **Impact:** Better team strategies and higher chances of success.
3. **Successful Recommendations for Shoppers by Amazon**
    - **Case:** Amazon uses customer data to recommend products, personalize the shopping experience, and optimize delivery times. (*Article*)
    - **Impact:** Increased sales and customer satisfaction, making Amazon a global leader in e-commerce.
4. **Using Data to Implement Congestion Tax in Stockholm**
    - **Case:** In Stockholm, Sweden, data from traffic sensors helped implement a congestion tax, reducing traffic volume by 25% and improving air quality. (*Article by IBM*)
    - **Impact:** Better urban living conditions.

Data Disasters

1. **Facebook and Cambridge Analytica Scandal (2018)**
    - **Case:** Data from millions of Facebook users was harvested without proper consent and used for political campaigns. (*Article*)
    - **Impact:** Public trust was damaged, and Facebook faced fines and stricter regulations.
2. **UK COVID-19 Data Loss (2020)**
    - **Case:** The UK government lost around 16,000 COVID-19 test results due to a spreadsheet error. The system couldn't handle the large dataset. (*Article*)
    - **Impact:** Delayed contact tracing and potential spread of the virus.
3. **Target's Pregnancy Prediction Backlash (2012)**
    - **Case:** Target used customer data to predict pregnancies and sent related ads to customers. This accidentally revealed a teenager's pregnancy to her family. (*Article*)
    - **Impact:** Privacy concerns and public criticism.

1b - State the ways your personal data may be used and shared.

Personal Data

Personal Data

Any information relating to an identified or identifiable natural person.

Data subject

The identified or identifiable living individual to whom the personal data relates.

There is a sub-category of personal data called **sensitive personal data**, which is required to be treated even more stringently than personal data. This includes the personal data of children (anyone under 18). Sensitive personal data should not be collected or processed except under certain conditions and with an identified lawful basis for doing so.

This table gives examples of data that would be classed as personal or sensitive personal data.

Personal Data	Sensitive Personal Data
• Names • Addresses • Phone numbers • Identification numbers • Location data • Online identifiers • A combination of identifiers that together can identify an individual	• Racial or ethnic origin • Political opinions • Religious or philosophical beliefs • Genetic data • Biometric data, where used for identification • Health data

What Data Is Shared When Going Online?

It is important to actively manage your privacy online, otherwise more information may be shared than necessary. The kind of information that is often being stored, and possibly shared, is more than just name and email addresses. It could be:

- Geographic location
- Web browsing habits
- Websites visited
- Products bought online
- Illnesses searched for online
- Devices used to connect to the internet
- Reading habits and history
- Food preferences
- Political views

Risks of Sharing Online

When sharing data online, users should consider these risks.

- **Privacy Breaches:** Personal information (like your address, phone number, or location) can be exposed to strangers or unauthorized parties.
  Example: Posting a photo with a visible address or geotag can reveal your location to others.
- **Identity Theft:** Hackers can use your shared information (e.g., full name, birth date, or photos of documents) to impersonate you and commit fraud.
  Example: Sharing a photo of your new ID or credit card can lead to theft of your identity or financial details.
- **Reputation Damage:** Old posts, even if intended as jokes, can resurface and harm your reputation in the future, especially when applying for jobs or schools.
  Example: A controversial tweet from years ago could lead to public backlash.
- **Phishing and Scams:** Scammers may use shared information to trick you into revealing sensitive details or money.
  Example: Posting about a recent purchase might make you a target for fake refund scams.

Benefits of Sharing Online

While there is risk in sharing data online, there are also benefits.

- **Connecting with Others:** Sharing updates, photos, and stories helps you stay connected with family, friends, and communities, even if they are far away.
  Example: Posting family photos can keep relatives updated on your life.
- **Sharing Knowledge and Ideas:** Online platforms allow you to share expertise, learn from others, and contribute to global conversations.
  Example: Writing blogs or creating tutorials helps others while showcasing your skills.
- **Raising Awareness and Advocacy:** Social media can amplify your voice and help bring attention to causes you care about.
  Example: Sharing information about environmental initiatives can inspire others to take action.
- **Expressing Creativity:** Sharing art, music, writing, or other creative content allows you to express yourself and gain feedback.
  Example: Posting your paintings on Instagram can help you build an audience and improve your skills.

How Long Does Data Stay Online?

Personal data shared online can remain **indefinitely** because:

- Platforms store backups even after deletion.
- Shared content can be copied or reshared.
- Search engines cache old versions of web pages.

Can It Be Deleted?

Removing personal data is sometimes possible in the following ways:

1. Delete Content: Remove posts and adjust privacy settings.
2. Request Deletion: Contact websites or use search engine tools like Google's "Remove Outdated Content."
3. Delete Old Accounts: Delete accounts that you are no longer using.

1c - State simple methods of 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.

Figure 1: The benefits of using high-quality data: improved customer experience, reduced risk, competitive advantage, increased revenue.

2 Outcome 2 - State data literacy concepts.

This outcome focuses on data literacy. Here we talk about how to use data, strategies for securing personal data, and different ways to visualise data. Scan the QR code on the right for relevant lessons and resources for this section.

2a - State the benefits of high-quality data.

What is high-quality data?

High-quality data refers to data that correctly represents the real-world constructs that it is referring to. High-quality data is fit for the analytical purpose which it is being used for.

This video from IBM describes simply the different factors that come into data quality.

The benefits of using high-quality data

All analysis is only as good as the data it is carried out on. Therefore, the quality of the underlying data is critical to any analysis. There are benefits in using high-quality data. For businesses, the benefits of high-quality data include:

- **Improved customer experience**: For example, if high-quality data has been used to train recommendation systems, customers are likely to receive better recommendations.
- **Reduced risk**: Using high-quality data reduces the risk of inaccurate predictions, which could be potentially harmful.

- **Competitive advantage**: Companies using high-quality data can make better predictions, leading to happier customers, meaning they have a competitive advantage over other companies.
- **Increased revenue**: The improved customer experience leads companies to have increased revenue as they have more paying users.

Spotting Poor-Quality Data

Poor-quality data can significantly impact decision-making and outcomes. It is important to identify and address issues such as:

- **Out of Date:** Data that is not current may lead to decisions based on outdated trends or information.
- **Inaccurate:** Errors or inconsistencies in data can result in incorrect conclusions and misguided strategies.
- **Incomplete:** Missing data points can skew analysis and provide an unreliable picture of reality.
- **Gathered from a Small Sample Size:** Limited data can lead to biased results that do not accurately represent the larger population.

Regular reviews and data validation processes are essential to ensure data quality and reliability.

2b - State types of data visualisations.

When analysing data, it's useful to have a visual representation of that data so that we can more intuitively understand it. We call a visual representation of data a **graph**. When looking at a graph, it is often easier to spot patterns in the relationships between different variables in the data.

Graphs, charts, plots, visualisations, diagrams – these terms all mean roughly the same thing and are often used interchangeably

In this section, we look at the different types of graphs that you should know about, and when they are most suitable to use.

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.

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

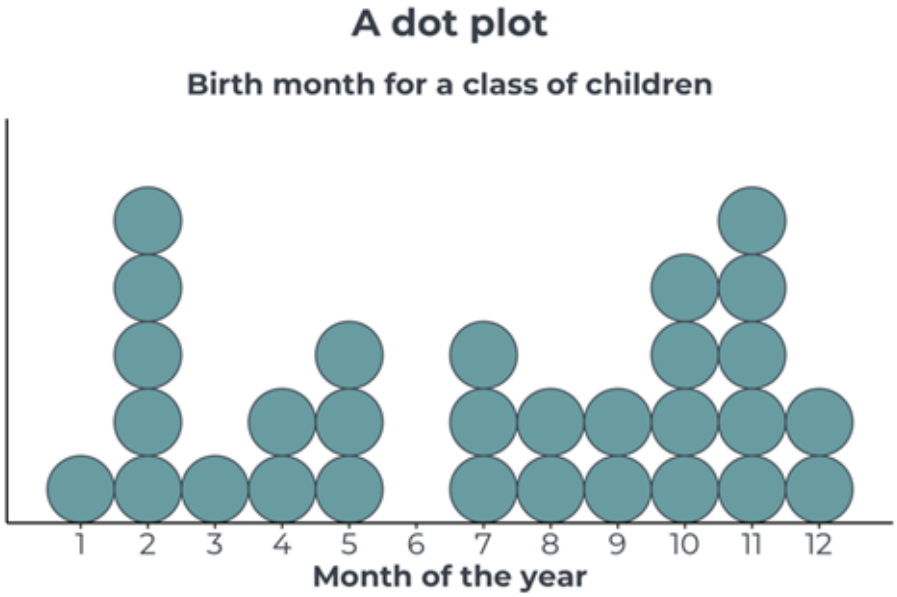

Figure 3: A dot plot.

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.

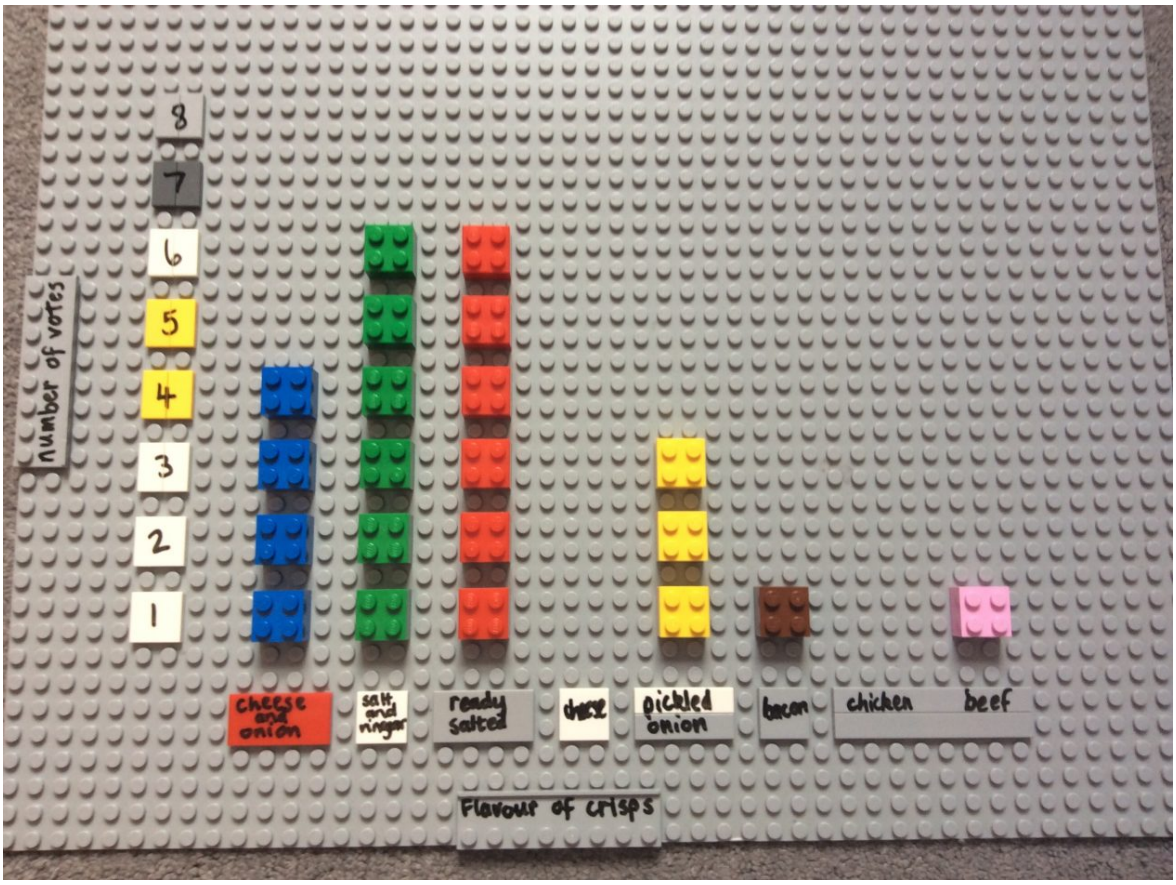

Figure 4: A dot plot made out of Lego.

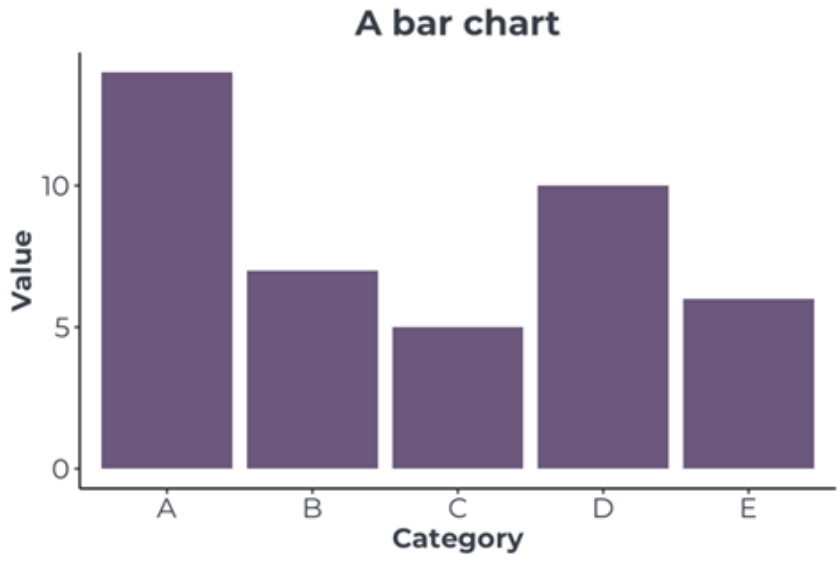

Figure 5: A bar chart.

Figure 6: A horizontal bar chart.

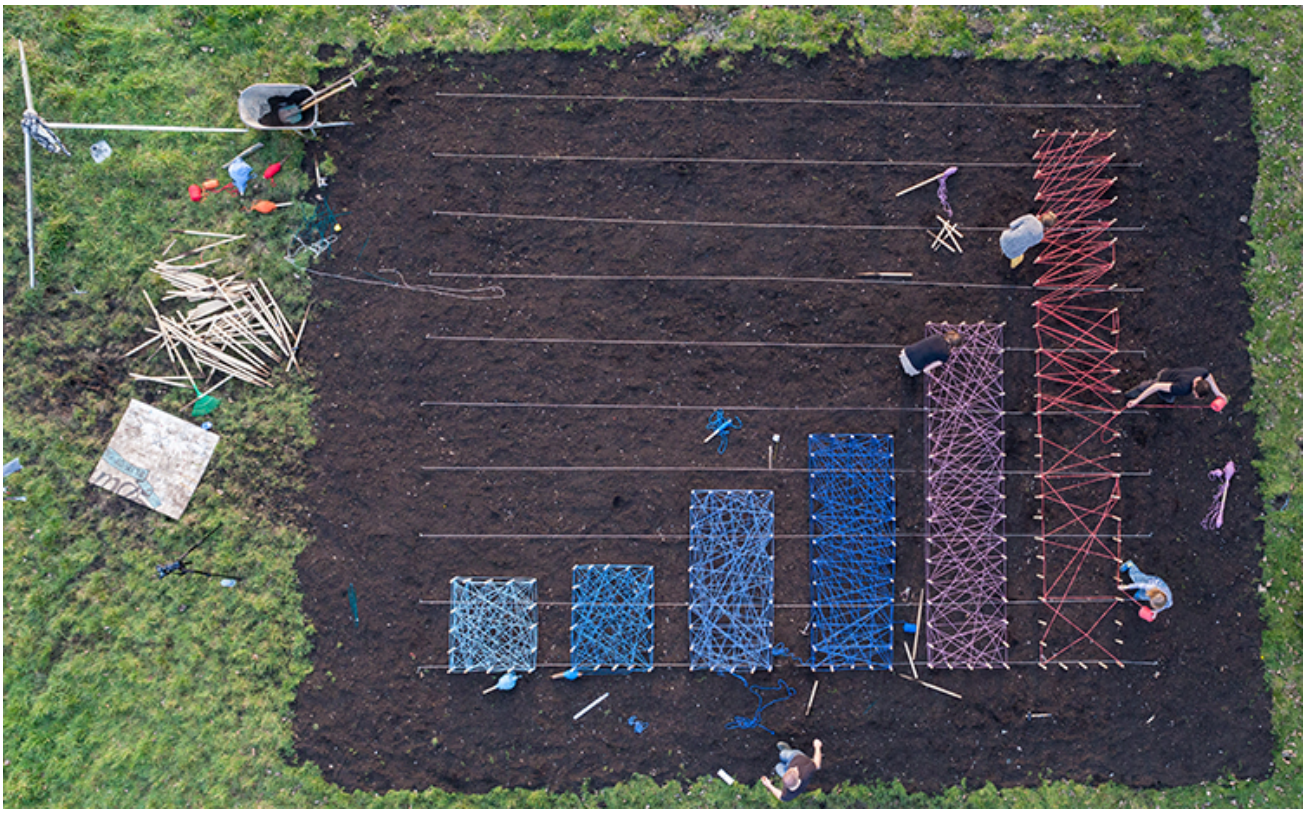

Figure 7: A vertical bar chart being made in a garden.

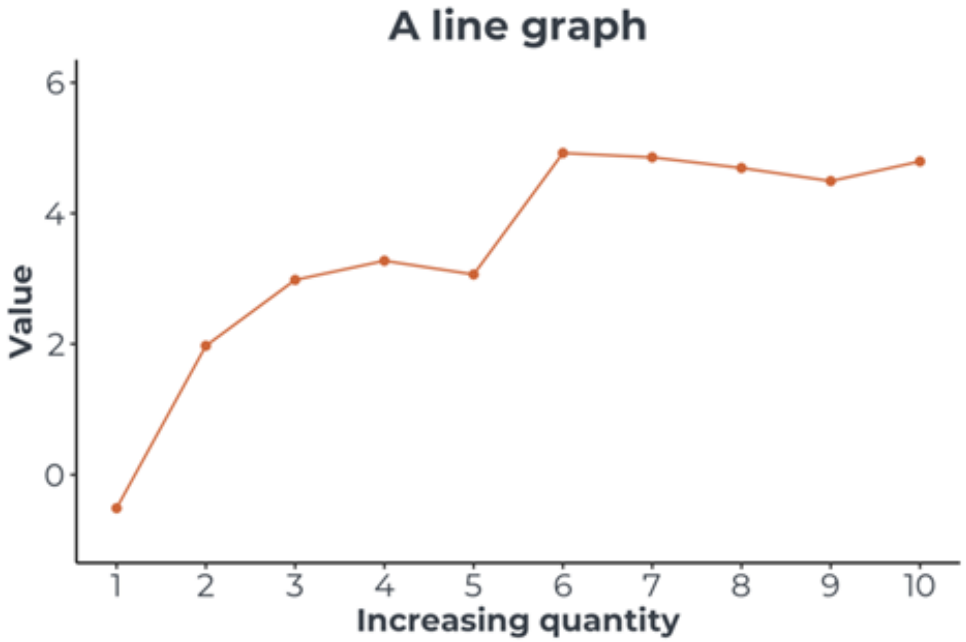

Figure 8: A line graph.

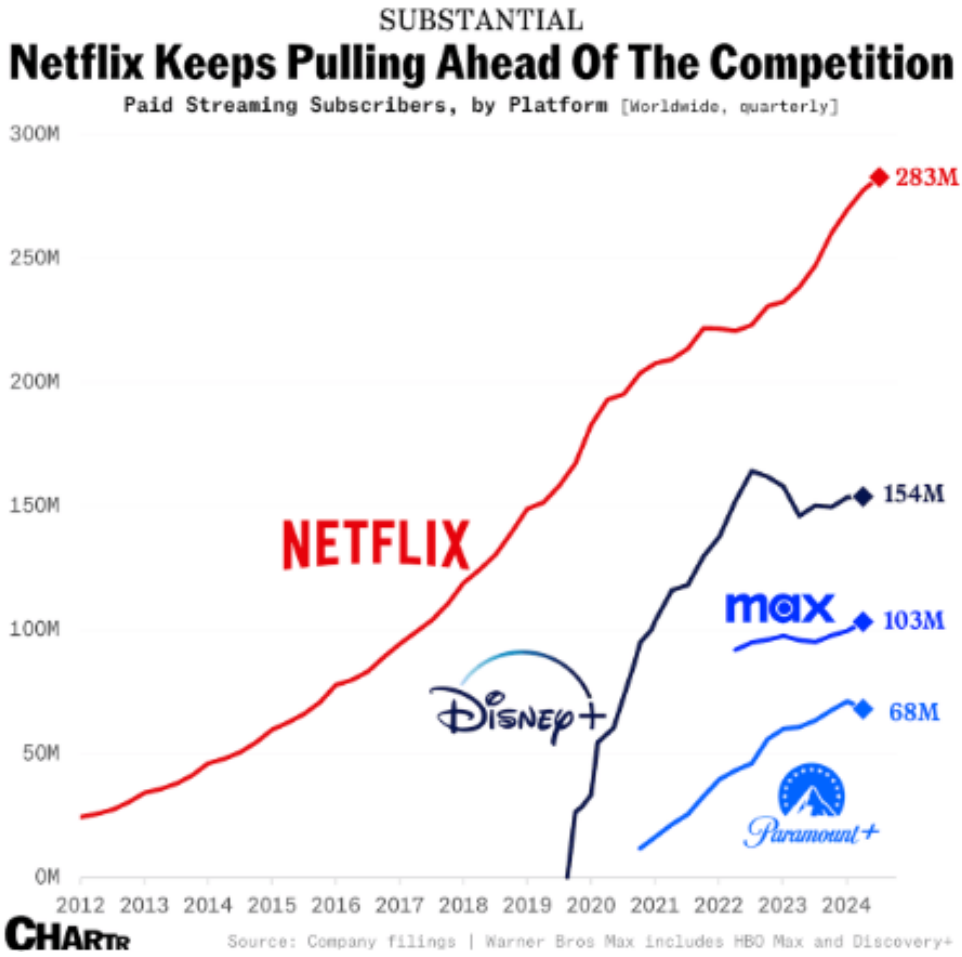

Figure 9: Line chart showing streaming service subscribers by platform.

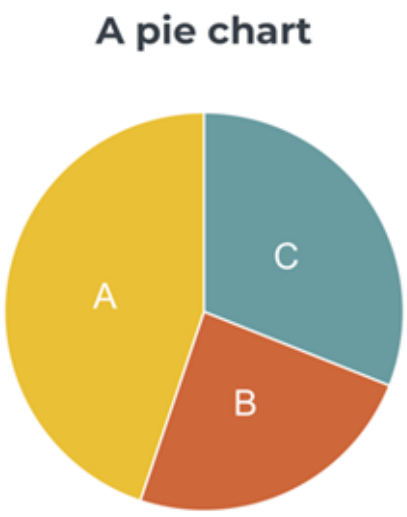

Figure 10: A pie chart.

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 compare relative segment sizes. A bar chart (Section 2.2) 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.

2c - State ways in which data visualisations can be interpreted and misinterpreted.

Some key elements of reading plots which you should be able to do include:

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

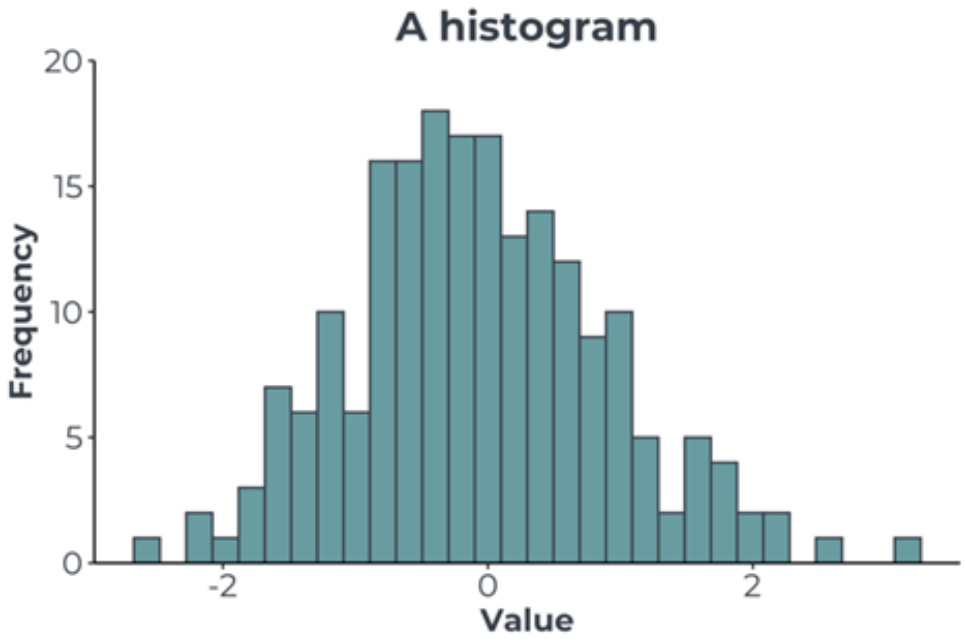

Figure 12: A histogram.

- Identify the name of the plot
- Interpret the axis
- Identify information
- Look for trends

Identifying Factors in Misleading Graphs

Sometimes data visualizations can be misleading. Some factors that could contribute to this are:

- Proportions not adding up to 100%
- Axes not starting at zero
- Missing data points
- Too many colours or segments

3 Outcome 3 - Interpret simple data.

Relevant Lessons

3a - Interpret simple data visualisations to identify patterns and trends.

In order to best describe a graph, the following should be done:

- **What is being measured?** - Identify the variables represented on the axes.
- **Quantitative description** - Use numbers and percentages to describe key data points.
- **Descriptive vocabulary** - Use terms like increase, decrease, peak, trend, correlation.
- **Consistency with the data** - Ensure your description accurately reflects the information in the visualization.

3b - Describe data visualisations in terms of quality.

Graphs are everywhere, on the news, on the internet, in reports and publications. Not all graphs are good graphs though. Good graphs convey their message at a glance, whilst bad graphs can be either deliberately misleading or just hard to decipher.

When describing a data visualisation in terms of quality, mentioning whether or not the following features are present can be useful:

- The axes are visible, labelled, and scaled correctly
- Units of measurement are given
- The data is plotted accurately
- There is a legend present
- The graph is overall neat and legible
- There is a title or caption
- There is a trend line shown, if required
- Graph helps answer the question

Examples of Graphical Crimes

Proportions not adding up to 100% (Figure 13)

When plotting proportions of a whole the numbers must always add up to the whole or 100%.

Axes not starting at zero

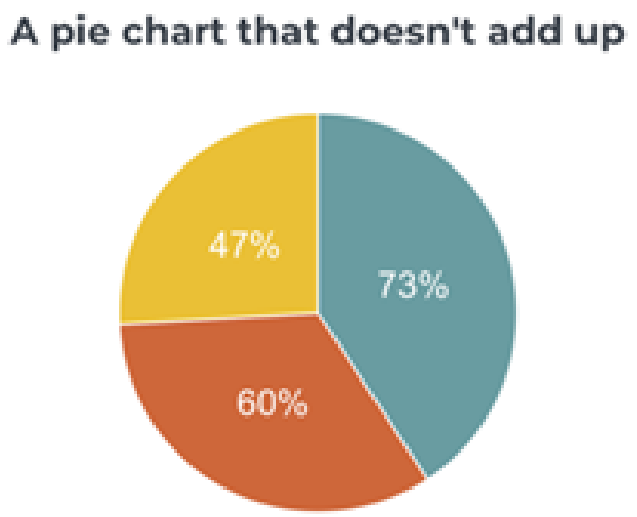

Figure 13: A pie chart that doesn't add up.

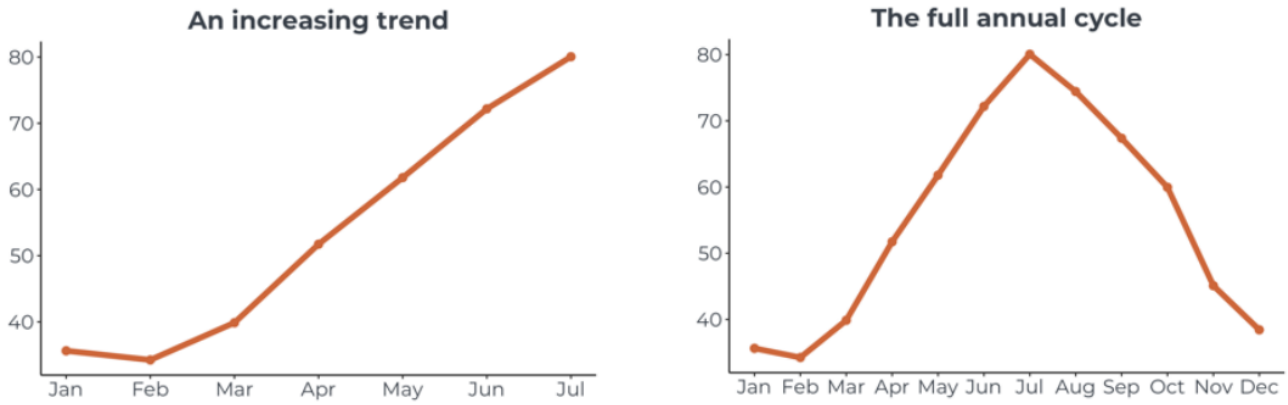

Figure 14: Two line charts. The one on the left has less datapoints than the one on the right.

Many graphics types such as bar graphs are interpreted by the reader by comparing the lengths of the different bars. If the bars do not start from zero, then the length comparison is distorted, and patterns can be made to appear that don't actually exist.

Missing data points

By choosing only data that fits the creator's objective the reader will not see the full picture. In Figure 14 only half a year of data is shown to imply a trend that doesn't exist in the second half of the year.

Too many colours and segments

Although vibrant, too many colours make a visual that is very hard to interpret. It is best to stick to one or two colours and make use of grey to de-emphasise unimportant patterns.

Pie charts should ideally be replaced with bar charts. If used, they should never have more than 2 or 3 segments. In Figure 15, all the small categories have been merged together.

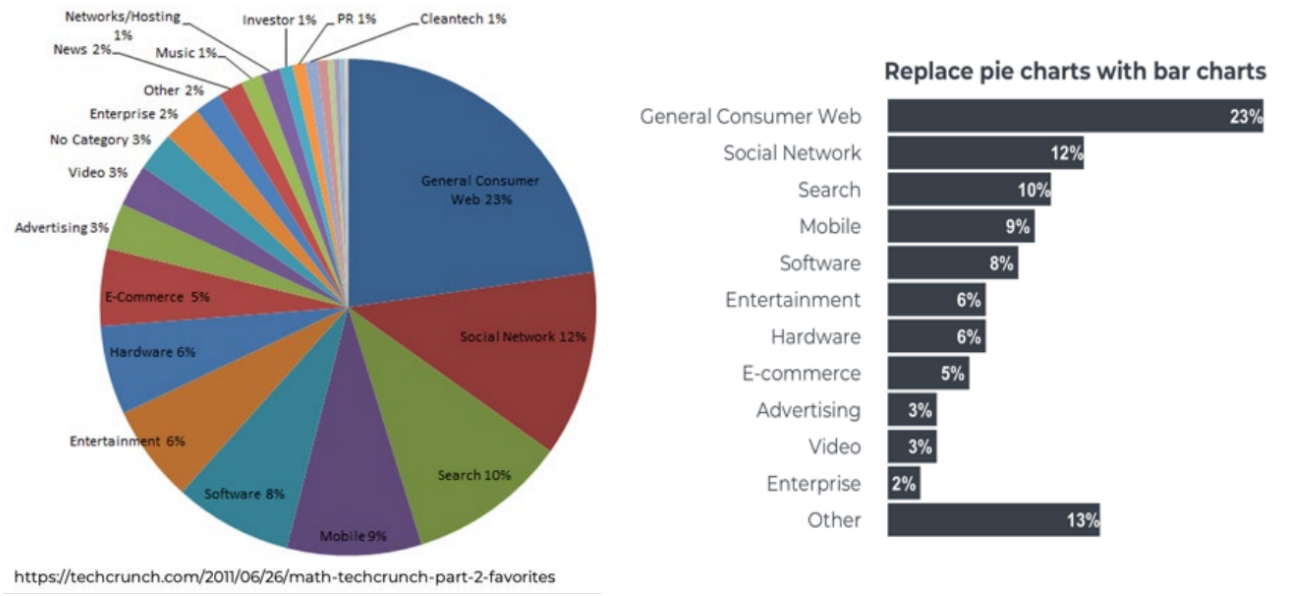

Figure 15: On the left, a pie chart with too many segments. On the right this has been replaced with a bar chart.

3c - Describe data generated by AI tools in terms of quality.

When generating output using AI, it's important to be able to think about the output critically, and to be aware that there can be mistakes and inaccuracies in it.

When describing data generated by AI, there are a number of features that can be commented on. Some of these include:

- **Inaccuracies**: For example, in images: the wrong number of fingers in people, other strange artefacts.
- **Relation to Prompt**: If the image or text is not fitting the criteria of what was requested.
- **Misinformation**: Sometimes generative AI will come out with factually wrong things. If unsure, compare the information with another trusted source such as well known news outlets.

3d - Draw conclusions from data 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?”).