NPA Data Citizenship

Learners' Guide to the National Progression Award

Data Education in Schools

Level 6 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 6 Data Citizenship 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-citizenship-lessons/ for relevant lessons and resources for this unit, separated by outcome.

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

This outcome focuses on recognising the use and limitations of data in business and society. This includes the use of data for social benefit, the ethical (and unethical) use of data, such as the use of biased, false or deliberately misleading data.

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.

1a - Explain the technological, economic and societal reasons for the growth of data.

In recent years, data is needed more, and larger and larger amounts of data are being stored and processed. Reasons for this increase in the amount of data and how much it's used include factors such as:

Technological Reasons for Growth

- It is easier and cheaper to collect, store, process, and analyse data.
- Computers are faster and more powerful.

- Sensors are smaller and more accurate.
- Networks are wider in coverage and faster in bandwidth.
- Cloud computing enables access to and sharing of data from anywhere.
- Increased adoption of AI technologies, including Machine Learning (ML) and Generative AI.

Economic Reasons for Growth

- Data is valuable for businesses and organisations.
- It helps improve products.
- It helps improve services.
- It helps optimise processes.
- It supports better, evidence-based decisions.

Societal Reasons for Growth

- Access to information.
- Many connections and interactions between people through social media platforms.
- Growing need to address global and community issues.

Real-Life Examples Showing Reasons for the Growth of Data

- **Social Media Platforms:** The rise of platforms like Instagram and TikTok, where millions of photos and videos are uploaded daily, demonstrates the explosion of data generation.
- **Online Education:** With more educational content and lectures being hosted online, there is a substantial increase in data related to learning analytics.
- **Remote Work:** The shift to remote work has led to increased use of digital communication tools, generating large amounts of data related to productivity and collaboration.

1b - Explain how individuals, organisations and society extract value from data.

- Better and faster decision-making
- Improved operations and processes
- Creation of a data product
- Understanding customer trends
- Creating innovative products and services
- Volume, Variety, Veracity, Vulnerability, Visualisations

1c - Explain types of bias and the impact of misuse of data on individuals and society.

In data, bias is a systematic error or distortion in data that prevents it from accurately reflecting the reality it is supposed to represent. It occurs when the data is inaccurate, incomplete, or doesn't accurately represent the population it's meant to reflect, leading to skewed, misleading, or unfair outcomes.

Types of Bias

There are multiple types of bias.

Statistical Bias: A systematic error in collecting, analysing, or interpreting data that consistently skews results away from the true value.

Data Bias / Sampling Bias: Occurs when the data collected is not representative of the overall population, often due to non-random sampling methods.

Algorithmic Bias: Bias introduced by automated systems or algorithms, often reflecting or amplifying existing biases in the data or design.

Causes of Bias

Bias can be introduced in multiple ways, including the below.

Sample Bias: A type of data bias where the selected sample does not accurately reflect the population it's intended to represent.

Exclusion Bias: Bias that arises when certain groups or data points are systematically left out of the analysis, distorting the results.

Measurement Bias: Occurs when the tools or methods used to collect data introduce consistent errors, leading to inaccurate measurements.

Confirmation Bias: The tendency to favour information that confirms pre-existing beliefs, often resulting in overlooking evidence that contradicts those beliefs.

Stereotype Bias: Bias that results from applying generalised beliefs or stereotypes to interpret or analyse data about individuals or groups.

Survivorship Bias: Focus on the subjects that have "survived" a selection process while ignoring those that did not, leading to overly optimistic conclusions.

Simpson's Paradox: A phenomenon where a trend observed within several groups reverses or disappears when the groups are combined.

Correlation Bias: The mistaken assumption that because two variables are correlated, one must cause the other, ignoring other possible influencing factors.

1d - Explain the rights and responsibilities of organisations that use personal data.

Organisations that use personal data must follow GDPR rules so that individuals' privacy is protected.

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.

GDPR

GDPR is an EU-wide law that applies to the processing of personal data either for activities carried out by processors established in the EU, whether or not the processing takes place inside or outside the EU. It also covers offering goods, services, or monitoring behaviour within the EU, whether or not the processor is based in the EU.

Your Data Rights

Individuals have certain rights under GDPR. These are listed in the boxes below.

Informed

A privacy notice provides transparency about the use of their personal data. This should be in clear plain language and be age-appropriate if aimed at children.

Access

Individuals can ask to see what data is held on them. This is called a Subject Access Request. This should be free and should be dealt with within one month.

Rectification

Any data that is incorrect or incomplete should be fixed for free. This request should be dealt with within one month.

Erasure

Known as the right to be forgotten, individuals can ask for personal data to be deleted. This can be refused in certain circumstances such as crime prevention or public health reasons.

Restrict processing

Individuals can ask to limit how data is used. This could be especially relevant if awaiting rectification.

Portability

Individuals can move their data between different providers, however currently agreed standards for most data sharing does not exist.

Automated decisioning

Individuals can insist that decisions, such as applications for credit, are not made using automated algorithms and can request they be made manually.

Object

This is the right to object to their data being processed. This could be as simple as unsubscribing from marketing.

GDPR Responsibilities

GDPR states **responsibilities** for those **holding data**:

- Provide **accurate** and **up-to-date** data

- Be responsible when using their GDPR rights and consider the **impact** of their request on organisations
- Be aware of **risks of sharing personal data**; secure devices and accounts from unauthorised access

1e - Describe types of ethical risks that can be introduced through the use of data.

The use of data can introduce various ethical risks that need careful consideration:

- **Fairness:** Ensuring that data practices do not lead to discrimination or biased outcomes against any group.
- **Equality:** Maintaining equal access and opportunities for all individuals when using data-driven technologies.
- **Privacy:** Protecting individuals' personal information from unauthorized access or misuse.
- **Trust and Transparency:** Building and maintaining trust by being open about how data is collected, used, and shared.
- **Truth:** Ensuring the accuracy and honesty of data to prevent misinformation or false conclusions.
- **Health:** Considering the impact of data use on individuals' physical and mental well-being.
- **Human Rights:** Respecting and upholding fundamental rights when handling data.
- **Criminality:** Preventing and addressing the use of data for illegal activities or harmful purposes.

1f - Describe methods of data security.

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.

Some more advanced ways that users can protect their data are:

- **Multifactor Authentication:** Adds an extra layer of security by requiring multiple forms of verification before granting access.
- **Biometrics:** Uses unique physical characteristics, like fingerprints or facial recognition, to verify identity.
- **Wiping Drives Prior to Disposal:** Permanently erases data from storage devices to prevent unauthorised access.
- **Firewalls:** Monitors and controls incoming and outgoing network traffic to block unauthorised access to the network.
- **VPNs (Virtual Private Networks):** Encrypts internet connections to protect data and maintain privacy online.
- **Software Upgrades:** Keeps systems secure by updating to the latest versions, which often include security patches.

- **Limiting the Data Stored:** Only keep essential data to reduce the risk of breaches.
- **Back-ups:** Regularly save copies of data to prevent loss in case of system failure.
- **Access Limitation:** Restrict data access to authorised personnel only.
- **Testing and Monitoring:** Continuously test security measures and monitor systems for unusual activity.

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

2 Outcome 2 - Explain data literacy concepts.

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.

Figure 2: Reasons for poor-quality data.

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.

2a - Explain the ways of measuring data quality.

In order to measure the quality of data, consider the six core dimensions: accuracy, completeness, validity, uniqueness and timeliness.

Reasons for Poor Quality Data

There are a number of ways that data quality can be poor. These include:

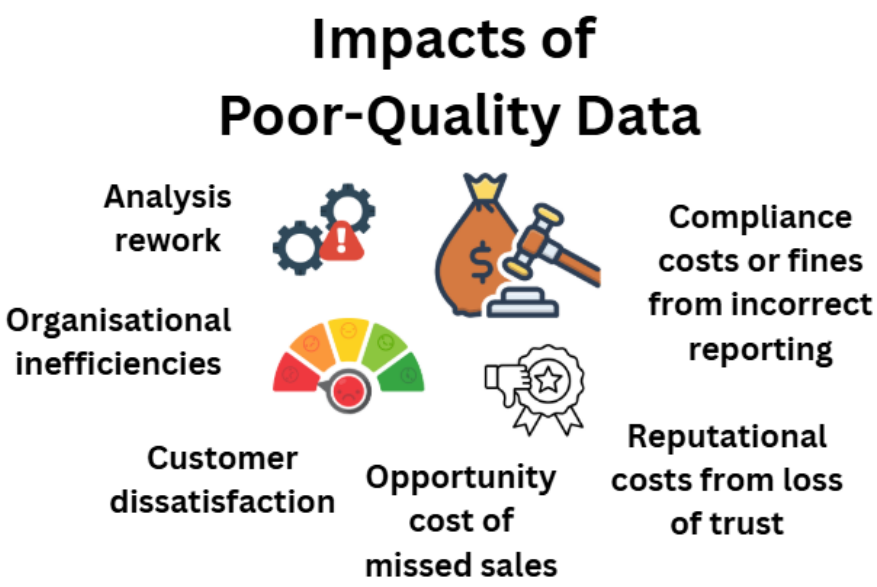

Figure 3: The impacts of poor quality data.

- **Duplicate data (Uniqueness)**: If data entries are duplicated, this might skew calculations about the distribution of data such as averages.
- **Inaccurate data (Accuracy)**: Inaccurate data means that any predictions we make will be less accurate, and any claims made about findings are less plausible.
- **Outdated information (Timeliness)**: Outdated information means that data is less applicable to any current work that is being done.
- **Missing values (Completeness)**: This makes the overall spread of our data less accurate, particularly if missing data is concentrated in one specific area.
- **Non-standardised data (Consistency)**: This means data lacks a consistent format or structure, which makes it difficult to compare or analyse.
- **Data security and privacy (Validity)**: If data security is poor, unauthorised changes could be made to it, making it inaccurate or corrupted.

Impacts of Poor-Quality Data

There are negative impacts of using poor-quality data in training or analysis for a company. These include:

- **Analysis rework**: Having to redo parts of an analysis because defects in the data were discovered too late.
- **Organisational inefficiencies**: Waste, delays and duplication that spread across teams and processes when data isn't fit for purpose.
- **Customer dissatisfaction**: Less accurate predictions and insights will lead a customer to be dissatisfied.
- **Opportunity cost of missed sales**: The value of sales that could have been achieved but weren't because of poor quality data.

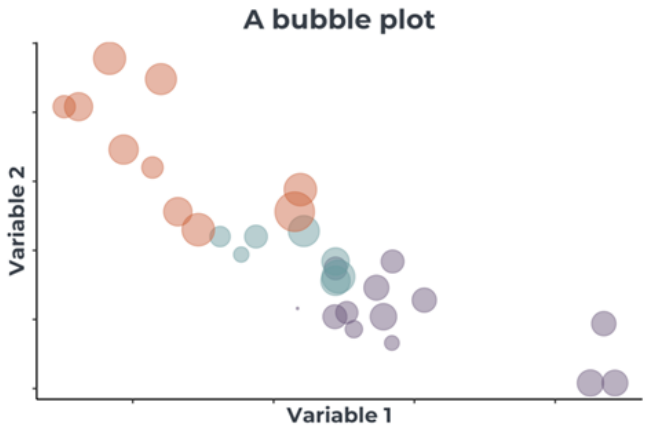

Figure 4: A bubble plot

- **Reputational costs from loss of trust**: If errors are made due to poor quality data, as a company, your reputation may be damaged.
- **Compliance costs or fines from incorrect reporting**: In some cases, if inaccurate claims are made, fines will have to be paid for inaccurate reporting.

2b - Explain ways that data can be visualised to tell a story.

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.

Bubble plot

A bubble plot (Figure 4) is like a scatter plot, but with extra information provided by the bubble size, and in this case colour as well. They are a simple way of adding an extra dimension to a chart.

Heat map

A heat map (Figure 5) is able to show patterns between three variables. The first two variables are demonstrated spatially, and the third variable utilises a colour scale. Heat maps are best for identifying spatial patterns rather than reading off accurate values.

Time series graph

A time series graph (Figure 6) is a special type of line graph, with time on the x-axis and regular repeated measurements of a variable on the y-axis. Time series are good for spotting long term trends, a regular seasonal variation, or even a cyclical variation that doesn't align with the seasons.

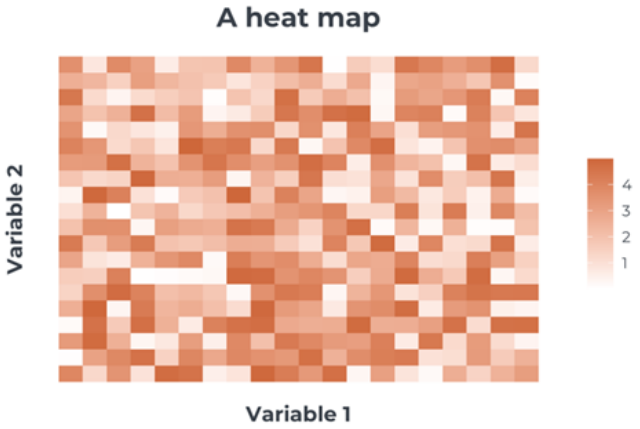

Figure 5: A heat map

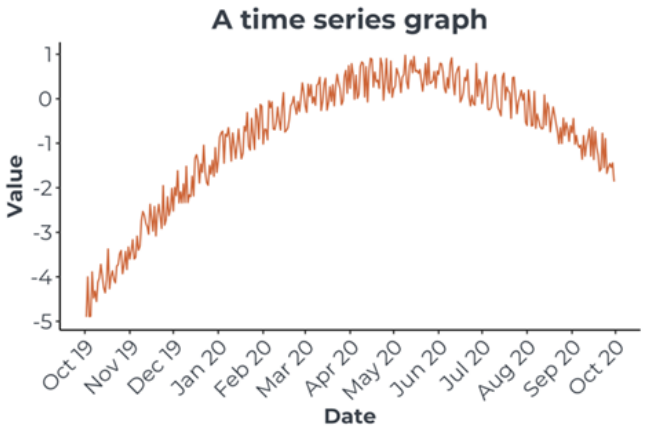

Figure 6: A time series graph.

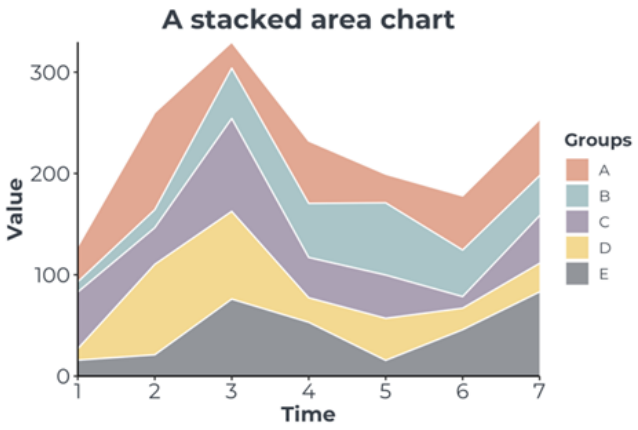

Figure 7: A stacked area chart

Stacked area chart

An area chart (Figure 7) or a stacked area chart highlights proportions changing over the varying quantity. As well as raw data volumes, this can also be done as proportions of the whole, which is useful for example in demonstrating the change in survey results over time.

Annotated box plot

Box plots (Figure 8) are a standardised way of displaying the main summary statistics of a distribution, but they require familiarisation first to be able to read them correctly.

The middle part of the box shows the interquartile range (IQR) from the 25th percentile (Q1) up to the 75th percentile (Q3), with the central line being the median of the data. The lines coming out of the box (the whiskers) are set at 1.5*IQR beyond quartiles 1 and 3. Any values outside this range are considered to be outliers and are plotted separately.

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 9) 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.

Choropleth map

Maps are very useful for demonstrating spatial patterns in data. A choropleth map (Figure 10) uses a colour scale to represent values of a quantity. In the example, it can clearly be seen that the Highland region had the highest number of road fatalities in the year. However, the size of the region is not proportional to the value, so larger regions can appear more emphasised than they should. It is better practice to plot normalised values (densities) rather than raw values.

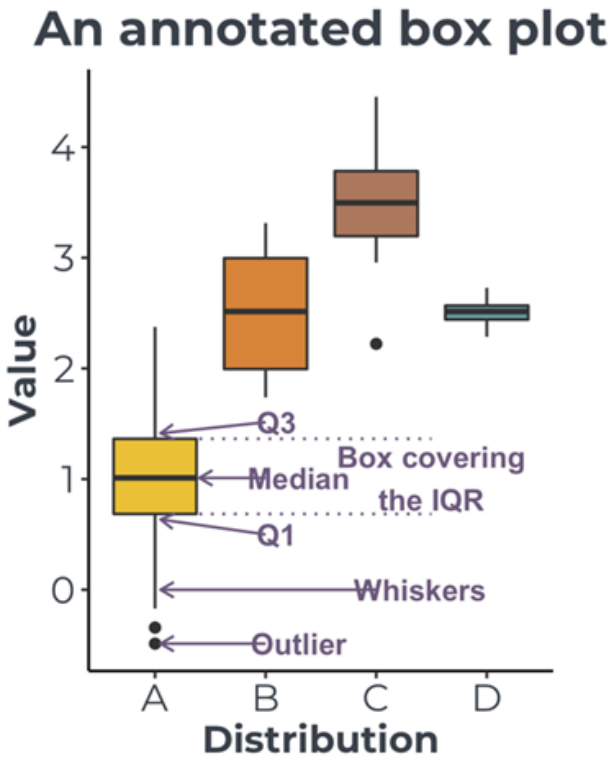

Figure 8: An annotated box plot.

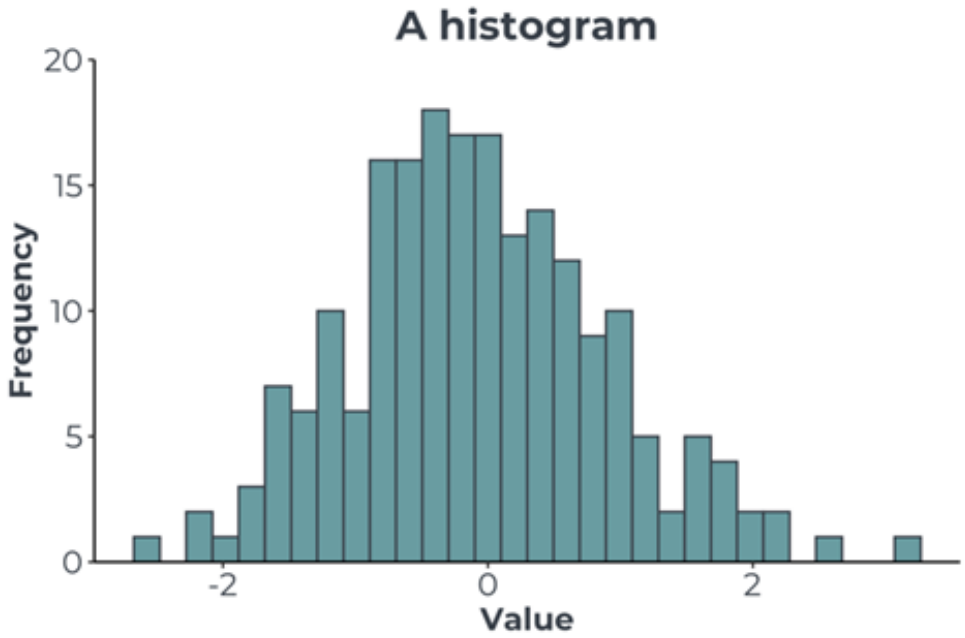

Figure 9: A histogram.

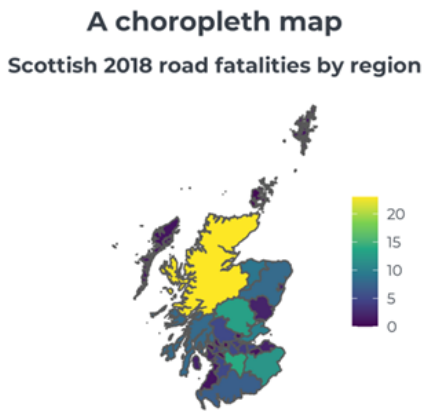

Figure 10: A choropleth map.

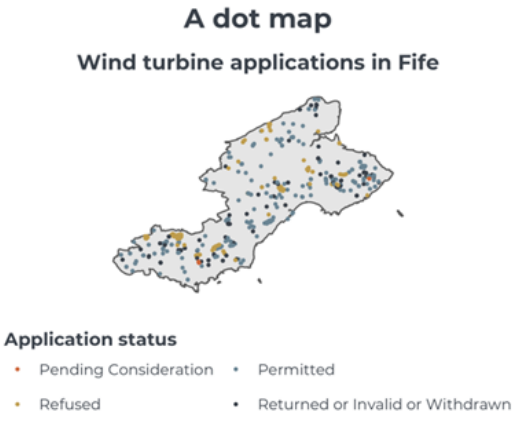

Figure 11: A dot map.

Dot map

This (Figure 11) is a dot map that shows the spatial distribution of points in a geographical area. It is good for spotting clusters, however with a large number of dots, they can be overplotted and make it more difficult to spot patterns. It is not easy to extract exact values from dot maps.

Bubble map

A bubble map (Figure 12) takes the dot map one step further and shows the magnitude of the variable as well as the spatial distribution. However, it suffers from the same problems of overplotting.

Considerations when creating visualisations

When telling a story with visualisations, you should consider the following:

- **Graph choice**: It is important to choose a graph type that is appropriate for what you are trying to convey.
- **Text**: In visualising data, text can be helpful. Adding a descriptive title or annotations of key points can be a useful way to highlight specific data points or trends.

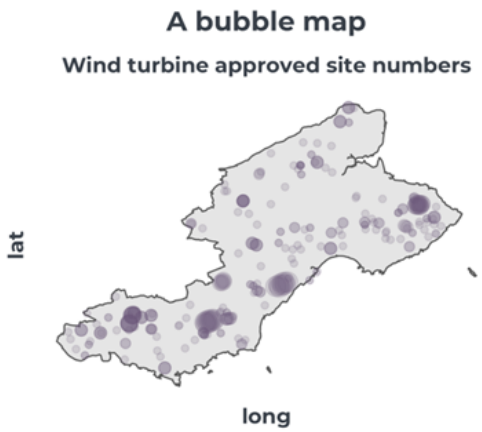

Figure 12: A bubble map.

- **Colours**: Colours can be a useful aspect of a visualisation, in grouping data in separate groups or highlighting key points. When using colour, consider colour-blind individuals and try to use the most colour-blind friendly colour combinations.
- **Position**: If using multiple graphs when telling a story, consider the positioning and order of each of them, to ensure that the story is presented in the clearest way possible.

Reading Plots

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

- 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

2c - Explain the Importance of Domain Knowledge When Solving Problems Using Data

Domain Knowledge

Specialised knowledge of a specific field, discipline or activity.

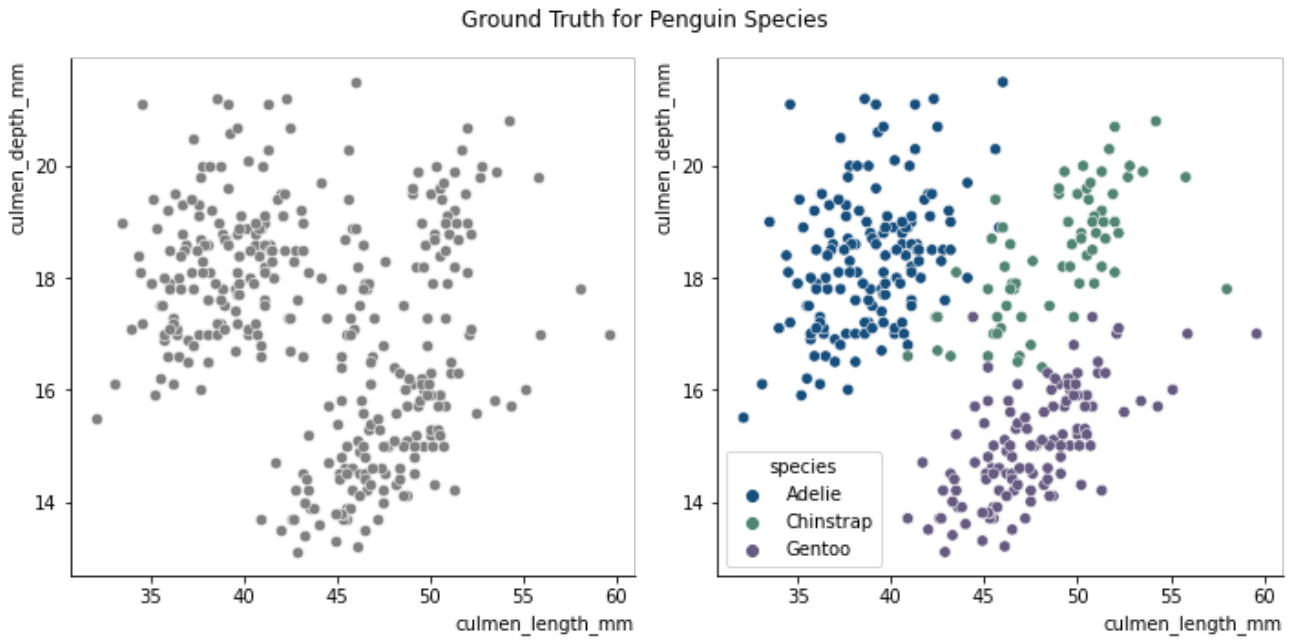

Figure 13: A scatter plot showing how penguin bill length relates to width

Domain knowledge is crucial when solving problems using data. It ensures that data analysis is contextually relevant and accurate. Having domain knowledge can help you identify the right problems to solve and the benefits in solving them. Domain expertise means that you understand the existing systems and processes.

The two graphs showing how penguin beak width and length relate gives us an idea of why domain knowledge is important. If we plotted the bill lengths and widths without knowing about the different types of penguin, it may be unclear to us why there appears to be groupings. Knowing about the different types will make it easier to work out why there might be clustering, leading us to create the graph on the right.

2d - Explain data generated from AI in terms of quality, trust and bias.

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.
- **Bias and Fairness**: Outputs can be biased due to bias in data that the model was trained on.
- **Ethical Considerations**: Sometimes outputs can reinforce harmful stereotypes or messages.

3 Outcome 3 - Gather data to investigate a problem.

3a - Describe Methods to Gather Data

There are various methods to gather data, each suited to different contexts and objectives:

- **Manual Data Capture:** Involves collecting data directly through face-to-face interviews, telephone surveys, or postal questionnaires. It's useful for capturing detailed responses and insights.
- **Online Data Gathering Tools:** Platforms like Google Forms, Microsoft Forms, SurveyMonkey, and Typeform allow for easy distribution and collection of surveys over the internet. They offer scalability and quick analysis.

When planning data collection, consider the study design:

- **Cross-sectional:** Collecting data at a single point in time to understand the current state of a population.
- **Longitudinal:** Gathering data over an extended period to observe changes and trends.
- **Retrospective:** Analyzing existing data from past records to address current research questions.

3b - Describe Best Practices in Survey Design

Designing effective surveys is crucial for collecting reliable and meaningful data. Here are some best practices:

- **Define Clear Objectives:** Clearly outline the purpose of the survey and what you aim to achieve with the collected data.
- **Keep It Concise:** Limit the number of questions to keep the survey short and focused, increasing the likelihood of completion.
- **Use Simple Language:** Write questions in clear and straightforward language to ensure respondents understand what is being asked.
- **Avoid Leading Questions:** Ensure questions are neutral and unbiased to avoid influencing respondents' answers.
- **Use a Mix of Question Types:** Incorporate multiple-choice, Likert scale, and open-ended questions to gather diverse data.
- **Pilot Test the Survey:** Conduct a trial run with a small group to identify any confusing questions or technical issues.
- **Ensure Anonymity and Confidentiality:** Assure respondents that their answers will be kept private to encourage honest and accurate responses.
- **Provide Clear Instructions:** Include detailed instructions on how to complete the survey to avoid misunderstandings.

3c - Describe ways to minimise bias when gathering data.

When gathering data using a survey, you must be aware of potential biases that may arise, and how to minimise them.

Some types of bias that may arise are:

- **Sampling bias**: This type of bias occurs when some individuals of a population have a higher probability of being selected. In surveys, a sample bias might occur when creating a survey on a top musician; it may be more likely that fans of that musician answer the survey, resulting in a biased set of data. To prevent against this type of bias, ensure that the survey is completed by an accurate representation of the population.
- **Non-response bias**: This bias refers to when some subjects don't take part in the study. For example, in a company, employees who are the most busy may fail to answer a survey. One way of preventing this bias is by making the survey compulsory.
- **Response bias**: This type of bias refers to when participants answer questions wrongly or inaccurately. To help to prevent this type of bias, use unambiguous and neutral language, keep surveys consise to prevent response fatigue,
- **Order bias**: This refers to bias caused by ordering of questions in a survey. To prevent this, questions could be given in a new randomised order to each participant.

3d - Design and carry out a survey to investigate a problem, ensuring bias is minimised.

When carrying out a survey, we can use the PPDAC cycle. Below are the steps involved.

Step 1: Identify the *Problem*

Before you design your survey, decide on a problem or question you want to investigate. This could be something relevant to your school, community, or daily life. Some examples include:

- What are the most common methods of transport to school?
- How much exercise do students get each week?
- What types of food do students eat at break time?

Step 2: *Plan* Your Survey

Once you have a clear problem to investigate, plan how you will collect the data. Consider:

- What information do you need? Think about what questions will give you useful answers.
- Who will you ask? Will you survey your classmates, students from different year groups, or your whole school?
- How will you collect the responses? You can use online tools like Microsoft Forms or Google Forms, or you can collect data manually with paper surveys or tally charts.

- Which questions should you ask? Your survey questions should be clear, concise, and unbiased. Use a mix of:
  - Multiple-choice questions (e.g., "How do you usually travel to school? - Walking, Cycling, Bus, Car, Other")
  - Scale or rating questions (e.g., "On a scale of 1-5, how much do you enjoy school lunches?")
  - Short-answer questions (e.g., "What improvements would you like to see in school lunches?")

Step 3: Collect *Data*

- If using an online survey tool such as Microsoft Forms or Google Forms, share the link with your target audience. These tools are useful as they automatically generate visualisations based on the data collected!
- If collecting data manually, make sure to record responses accurately.
- Aim to collect enough responses to make your findings meaningful.

Step 4: *Analyse* Data

Once you have collected your responses, review the results. You can view the built-in graphs and charts created in Microsoft Forms or Google Forms.

Step 5: Draw *Conclusions*

- Look for trends in your data. For example, if most students travel to school by bus, what does that tell you about transport options?
- Think about what your data means and how it might be used to make improvements or inform decisions.

Present Your Findings

After collecting your results, you should present your findings to others, such as peers in your class. Some ways you could choose to present your findings are:

- Create a presentation (PowerPoint, Google Slides)
- Write a short report summarising key findings
- Design a poster with key statistics and graphs
- Make a video or audio recording explaining your findings

4 Outcome 4 - Interpret complex data.

4a - Interpret complex data visualisations to interpret 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.

4b - Evaluate data visualisations in terms of quality, trust and bias.

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 14)

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

Axes not starting at zero

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

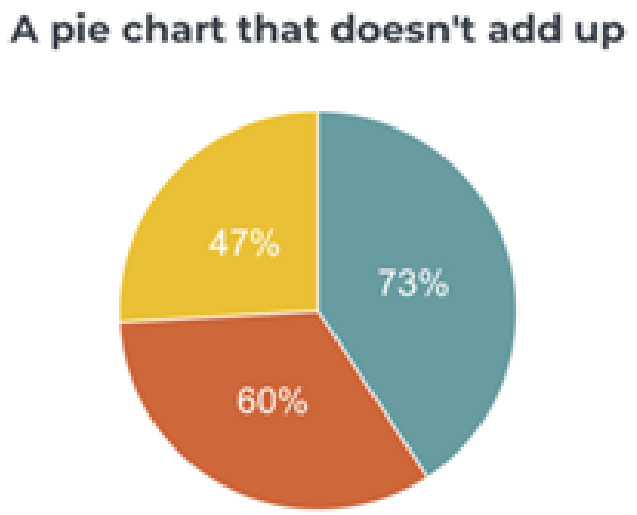

Figure 14: A pie chart that doesn't add up.

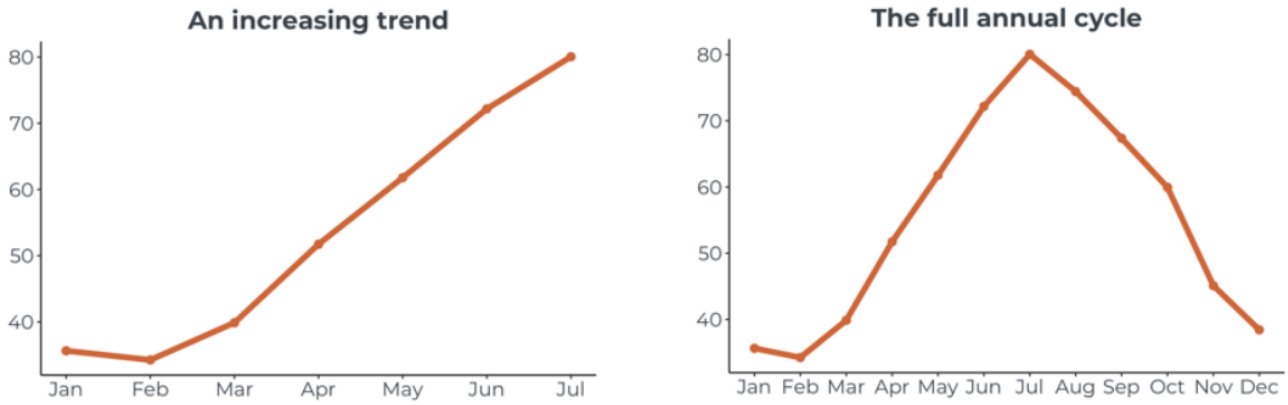

Figure 15: Two line charts. The one on the left has less datapoints than the one on the right.

By choosing only data that fits the creator's objective the reader will not see the full picture. In Figure 15 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 16, all the small categories have been merged together.

4c - Evaluate data generated from AI in terms of quality, trust and bias.

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.

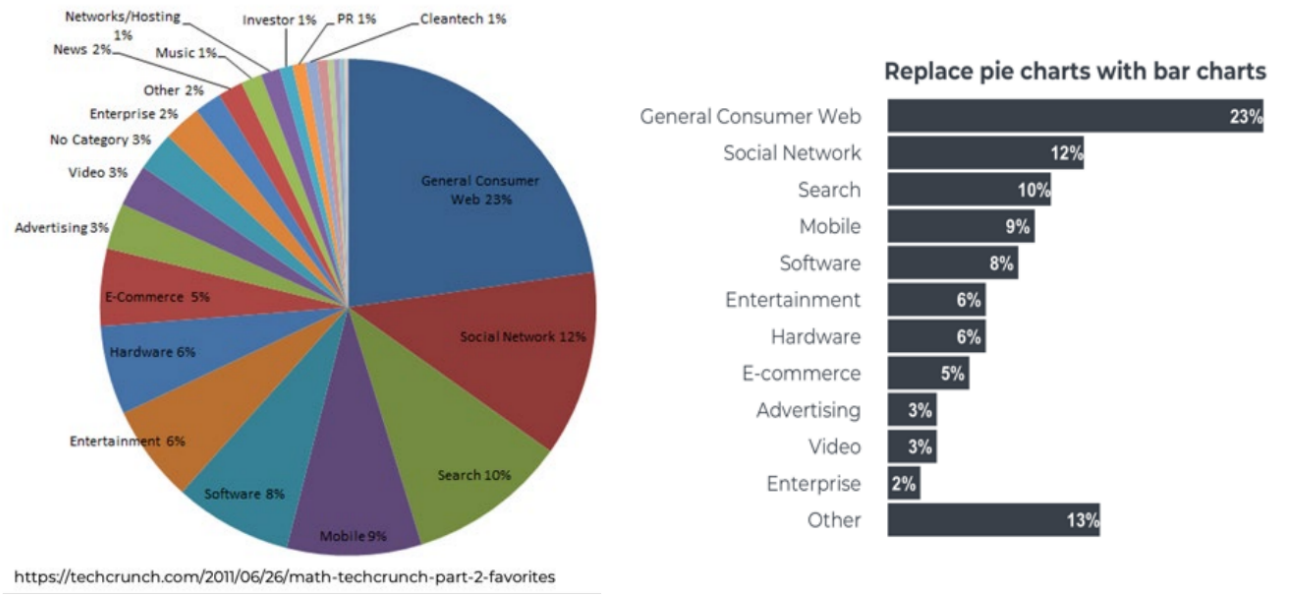

Figure 16: On the left, a pie chart with too many segments. On the right this has been replaced with a bar chart.

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.

- **Bias and Fairness**: Outputs can be biased due to bias in data that the model was trained on.
- **Ethical Considerations**: Sometimes outputs can reinforce harmful stereotypes or messages.

4d - Draw conclusions from data to investigate a problem.

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?”).

4e - Make recommendations based on conclusions and communicate findings.

After analysing data and drawing conclusions, it's essential to make informed recommendations and effectively communicate your findings:

- **Develop Actionable Recommendations:** Based on your conclusions, propose clear and feasible actions that address the identified issues or opportunities.
- **Prioritise Recommendations:** Rank actions by their potential impact and feasibility to guide decision-makers on where to start.
- **Tailor Communication to the Audience:** Present findings and recommendations in a manner suited to the audience, whether they are peers, management, or external stakeholders.
- **Use Visual Aids:** Enhance understanding by incorporating charts, graphs, and tables that illustrate key points and trends.
- **Be Clear and Concise:** Summarise findings and recommendations in a straightforward manner, avoiding technical jargon where possible.
- **Provide Context:** Explain the significance of the findings and how they impact the organisation or issue at hand.
- **Encourage Feedback:** Invite questions and discussions to ensure understanding and gather additional insights.

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

Example 1: Figure 17. 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 18. 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.

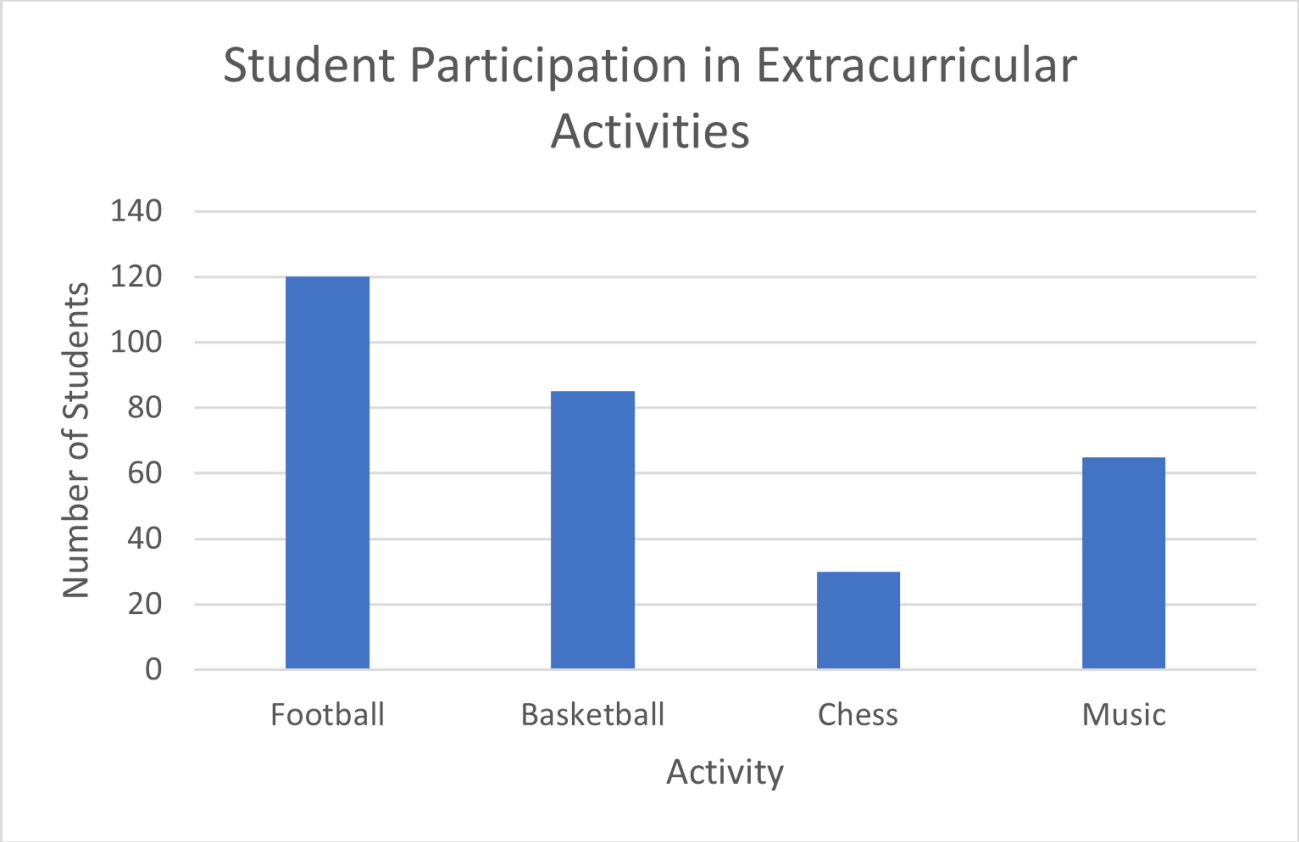

Figure 17: A bar chart showing how many students partake in different extracurricular activities.

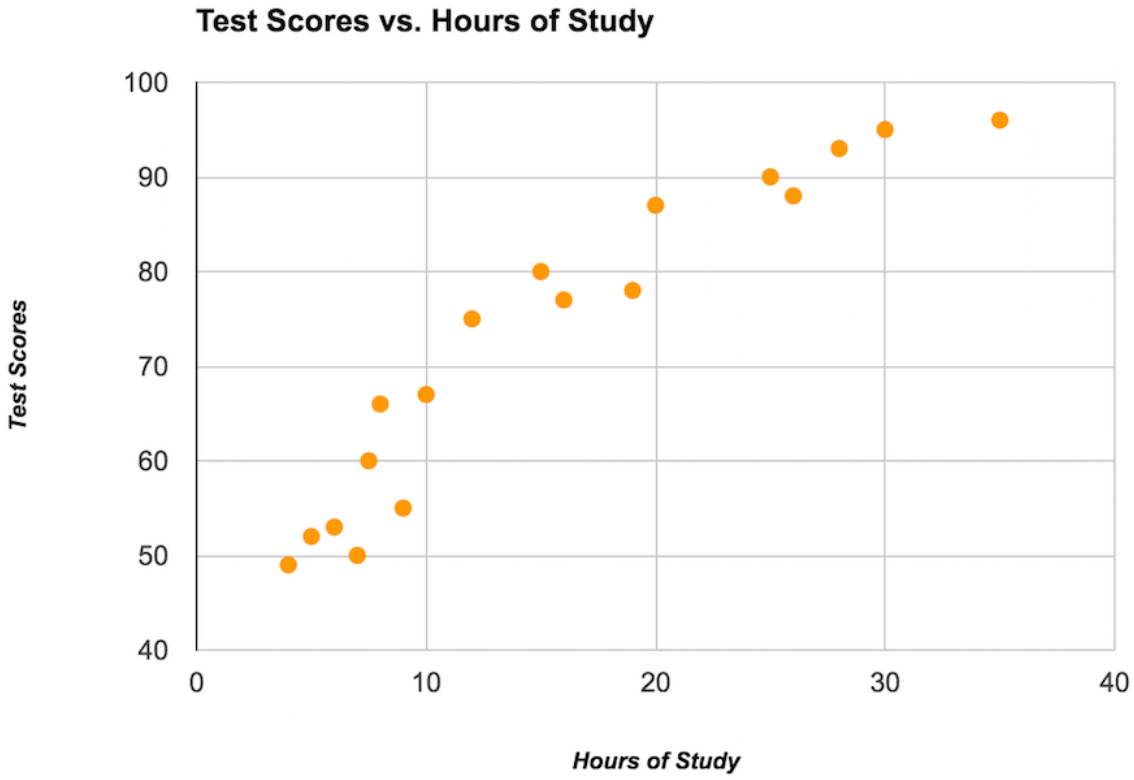

Figure 18: Graph of time studied against exam performance.