Data Science

Learners' Guide to the National Progression Award

Download this guide and other data science resources from **teachdata.science**

This guide has been developed thanks to the kind support from:

CONTENTS

ABOUT THIS GUIDE

This **Learners Guide** has been produced with an accompanying **Educators Guide**. Both guides have been created to support the core units of the National Progression Award (NPA) in Data Science at SCQF Levels 4, 5 and 6.

The **Learners Guide** is a summary document covering the core concepts that learners will need to know in order to learn about Data Science and undertake the assessments. It can be used by educators to introduce each topic, or as a summary or revision guide prior to assessments.

The **Educators Guide** covers information for teachers and lecturers when they are first considering selecting the NPA in Data Science, advice for planning for delivery of the course, and information they may find useful when delivering the two core units of the NPA.

These two guides have also been produced alongside an online guide developed by the University of Edinburgh that details all of the concepts covered in the Learners Guide in greater depth. The University are also working with schools and colleges to develop and trial teaching materials. You can find more information about this in the 'Support and Resources' section of the Educators Guide.

There are many exciting and engaging contexts for learning about data science, and many different tools that can be used to gain the practical skills involved in the course. The Learners Guide will not include practical tasks for particular tools. Instead, this guide will give advice on the best tools that can be used in teaching and learning, based on the experience of educators, the level and interests of learners, and any technological constraints within your school or college.

About the Authors

Kate Farrell is an experienced Computing Science teacher. She works for the Data Education in Schools project at the University of Edinburgh's Moray House School of Education and Sport. She was the Lead Developer for the NPA in Data Science at SCQF Levels 4, 5 and 6, and wrote the SOLAR unit assessments for the core units.

Dr Jo Watts is an experienced Data Scientist and founder of Effini, a data science company. She wrote the NPA Data Science core units, the Data Science project units, and the SOLAR unit assessments for the core and project units. She was also lead developer for the PDA in Data Science at SCQF Levels 7, 8 and 9.

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. The project is developing and adapting resources and are keen to support Centres to work together in partnership. 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.

Teaching materials will feature local datasets and case studies from industry. The Data Education in Schools team are actively working with a range of industry sectors, academics and researchers across the University, and with schools and colleges to trial resources.

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

Visit **dataed.in/NPADS** for more information about the qualification.

The following key at the right-hand side has been used to identify the material suitable to each level:

L4-6: Required knowledge for all levels

L5/6: Required knowledge for level 5

L6: Required knowledge for level 6

WHAT IS DATA?

WHERE DOES DATA COME FROM?

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 is being created all the time, from phones, sensors all around us and the internet. Data is everywhere, but it is what it is used for that makes it special.

Systems

Lots of data is created by software systems. Any interaction with an organisation will result in data about who, what and when something happened. This could be booking, manufacturing, retail purchasing systems or even banks.

The internet

The internet is a special system, not only creating data as individuals browse it, such as "likes" and "shares", but also providing an access point to download datasets either directly or through APIs.

Devices and sensors

Devices, such as mobile phones or fitness trackers, or sensors such as temperature, light, sound or pressure are constantly monitoring their environment and creating data all the time.

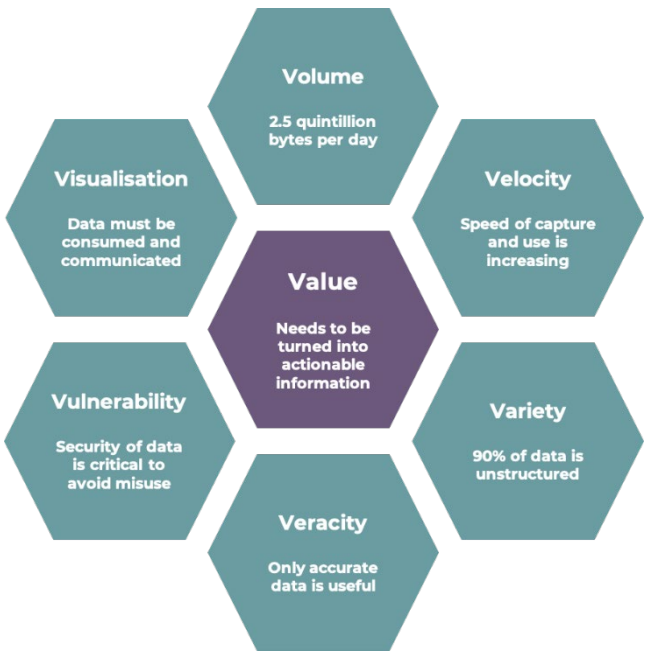
Manually captured

Data can be manually captured through surveys. This can be done either using an online tool or by filling in paper forms, which are often then digitised to enable analysis.

WHY IS DATA IMPORTANT?

Data on its own is not valuable. Data is raw un-organised facts that needs 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**.

The main ways to create value from data are:



DATA CLASSIFICATIONS

OPEN DATA

Open Data: data that can be freely used, shared and built-on by anyone, anywhere, for any purpose

The principles behind open data make it very powerful:

Availability and access: everyone can access the data
Re-use and redistribution: everyone can share and reuse the data
Universal participation: anyone can use the data

Types of open data

Cultural: gallery, museum, library and archive collections
Science: outputs of scientific research
Finance: government accounts and financial markets
Statistics: outputs from government statistical offices
Weather: all types of weather information
Environment: examples include river quality and air pollution levels
Charity: progress towards development goals

Benefits of open data

- Improved public services
- Transparency and democratic control
- Ability to spot global patterns
- Measurement of government policies
- Increased government effectiveness
- Facilitates innovation

PUBLIC AND PRIVATE DATA

Public Data: data that anyone can access, but there may be restrictions in place about its use and sharing

Much data is not formally classified as open data, but can still be freely used, perhaps by registering to access.

Private Data: data that a limited set of people can access and use

Most data is private. This may be to protect privacy, maintain commercial competitive advantage or comply with legislation.

Examples of data classifications

Open Data	Public Data	Private Data
Just Eat Cycles **(dataed.in/jeat)**	IMDB datasets (**dataed.in/imdb**)	Contact details
UK Open Data (**data.gov.uk**)	FiveThirtyEight (**dataed.in/538**)	Medical records
NHS Scotland (**dataed.in/nhs**)	Amazon reviews (**dataed.in/aws**)	Banking records
Unicef **(dataed.in/uni)**	Twitter public (**dataed.in/twit**)	School attendance

INTERPRETING DATA

THE IMPORTANCE OF GRAPHS

What is a graph?

A graph is a visual representation of data. Graphs are easier to interpret than tables of numbers and they allow the reader to understand relationships between different items to spot patterns and trends.

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

History of graphics

A Scot, William Playfair (1759-1823) is considered the father of statistical graphics. He is credited with inventing the line, bar and pie charts.

After Playfair came Florence Nightingale (1920-1910) and her Coxcomb diagrams (**dataed.in/cox**) of mortality causes in the Crimean War.

The 20th century saw the biggest progress in the art and science of data visualisation with thinking from people such as Edward Tufte (**dataed.in/tufte**)and Leland Wilkinson (**dataed.in/gog**) and an explosion in tools capable of manipulating and visualising data.

Where can graphs be found?

In the news or newspapers:

- The Guardian (**dataed.in/guard**)
- The Economist (**dataed.in/econ**)
- The Financial Times (**dataed.in/ftgraph**)
- FiveThirtyEight (**dataed.in/538**)
- NY Times (**dataed.in/nyt**)

Open data sources and reports:

- Gapminder (**dataed.in/gapdata**)
- World Bank (**dataed.in/wb**)
- World Health Organisation (**dataed.in/who**)
- UNICEF (**dataed.in/univiz**)

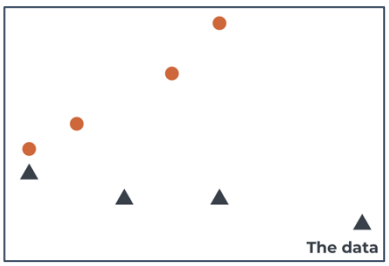
Academic texts and papers

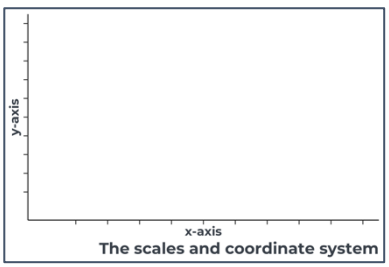
Workplace reports and dashboards

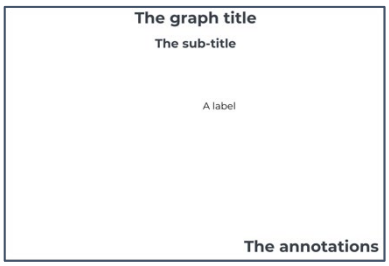
Graphical interpretation of current affairs:

- Graph of the Week (**dataed.in/turner**)
- Spurious Correlations (**dataed.in/spur**)
- The Pudding (**pudding.cool**)
- chartr: Data Storytelling (**chartr.co**)
- Information is Beautiful (**dataed.in/iib**)

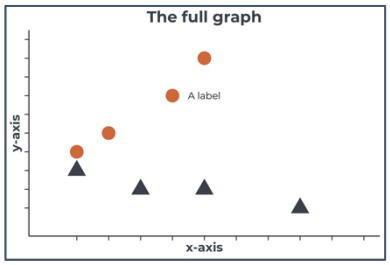
ANATOMY OF A GRAPH







A graph is made up of three separate components: **data**, **scales and coordinates** and **annotations**. Each component can be changed independently, leading to very different resulting graphs.



Data: the data on its own doesn't mean anything. It requires context.
Scales and coordinates: these provide the context for the data
Annotations: these inform the reader to understand what the graph is showing.

Graphical Crimes

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.

Good graphs

- Easy to interpret at a glance
- Use the correct graphics choice for the type of data and insights being communicated
- The colours are accessible for colour-blind readers
- Carefully labelled so the reader does not have to make guesses around what is being shown

Bad graphs

- Can be deliberately misleading
- Are hard to decipher
- Use too many colours
- Use the wrong graphics type
- Contain mistakes
- Are missing explanatory text such as axis labels or titles

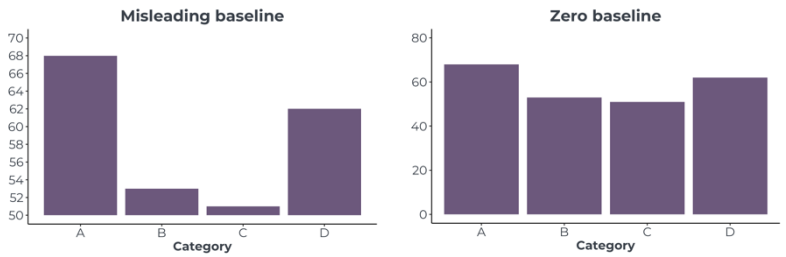
Examples of graphical crimes

Proportions not adding up to 100%

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

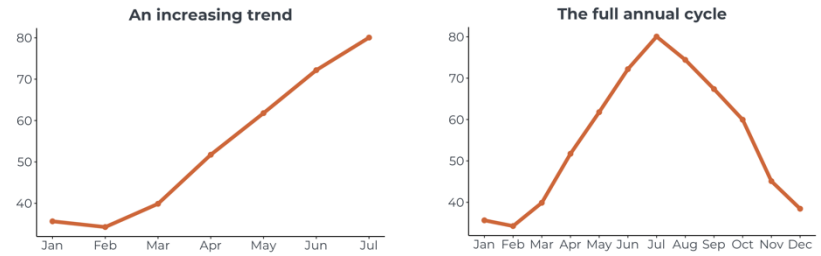
Axes not starting at zero

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

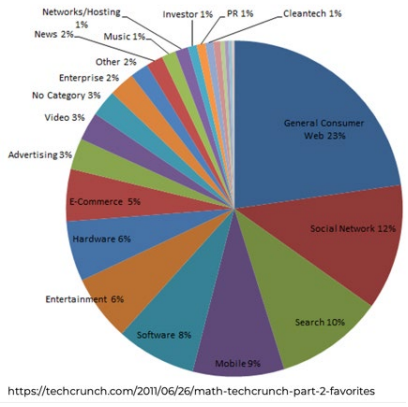
By choosing only data that fits the creator's objective the reader will not see the full picture. In this example only half a year of data is shown to imply a trend that doesn't exist in the second half of the year.

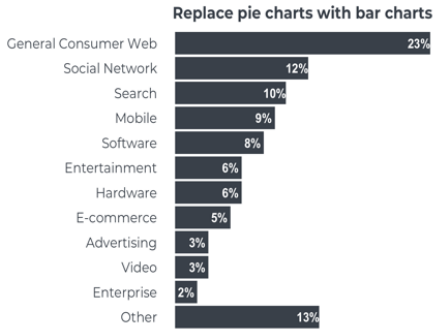


Too many colours and segments

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 this bar chart, all the small categories have been merged together.

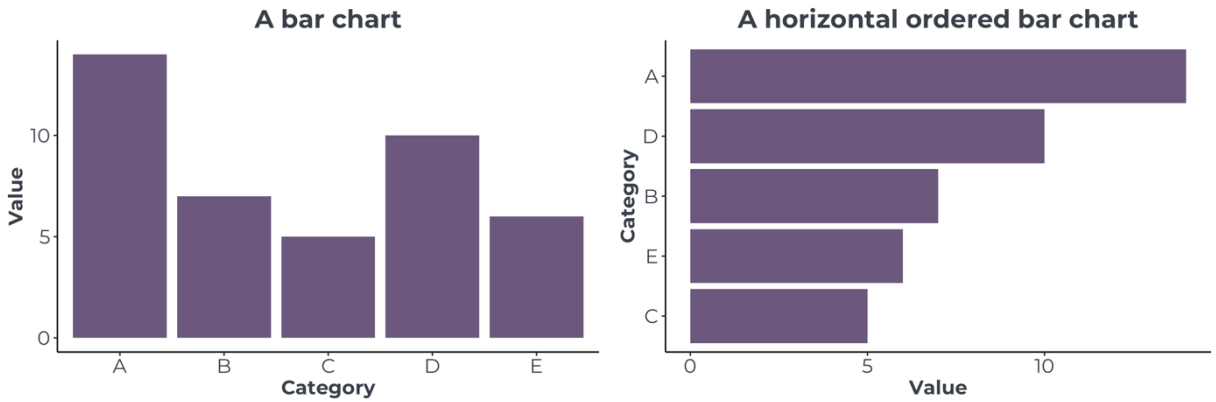




GRAPHICS TYPES

COMMON GRAPHICS TYPES

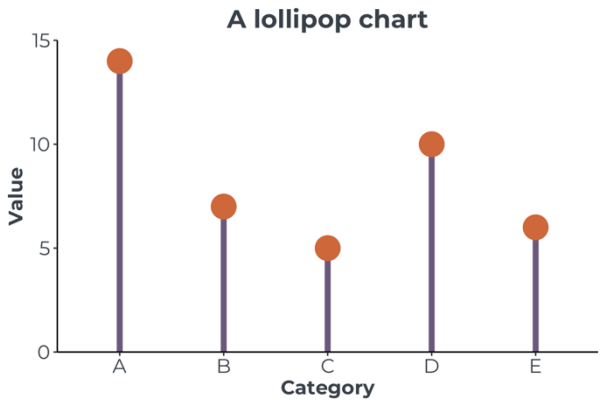
Bar charts



Bar charts 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.

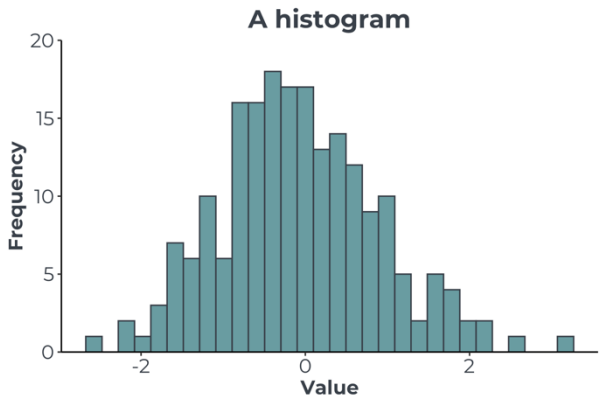
A horizontal bar graph is often a good option when there are many categories, or the category labels are long. It is also possible to reorder the bars, which makes it easier to see the smallest and largest categories. These can also be highlighted by using different colours.

A **lollipop chart** is identical to a bar chart but uses lollipops instead of bars. They are useful either when there are many categories, or the focus is on the actual values.



Histograms

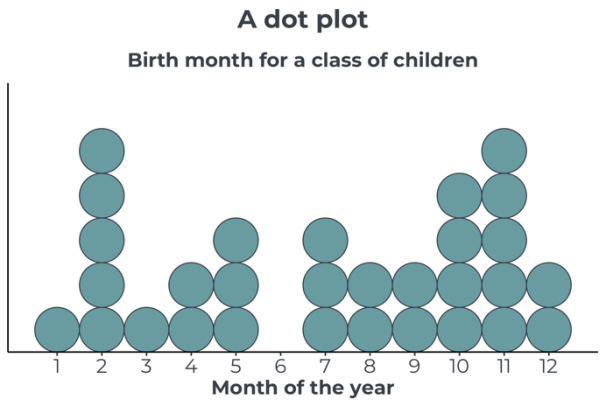
A **histogram** might look very similar to a bar chart, but it is fundamentally different since it is plotting numerical rather than categorical data.



Histograms 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.

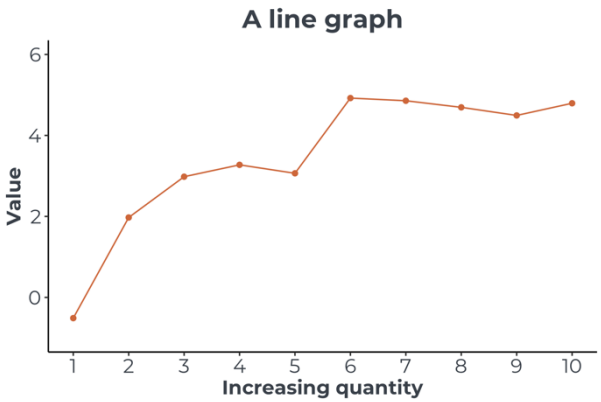
The bars on a histogram should always be displayed as touching, since the variable is continuous. If there are gaps between the bars, then that implies missing data in that range.

A **dot plot** can also be used, where each dot counts 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.

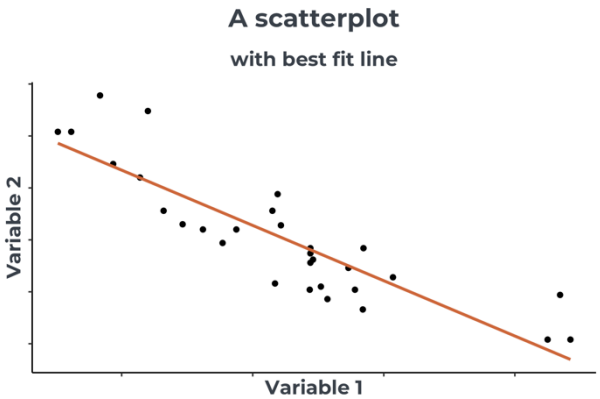


Line graphs

Line graphs 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.

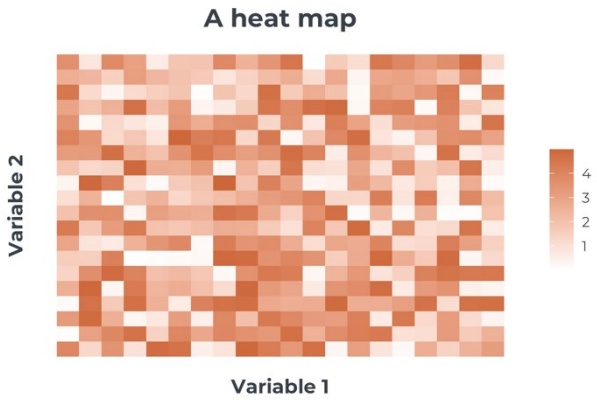


Scatterplots and heat maps



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

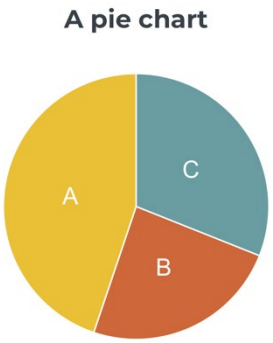
A **heat map** is able to show patterns between three variables. The first two variables are demonstrated spatially, and the third variable utilises a colour scale. Heat maps are best for identifying spatial patterns rather than reading off accurate values.

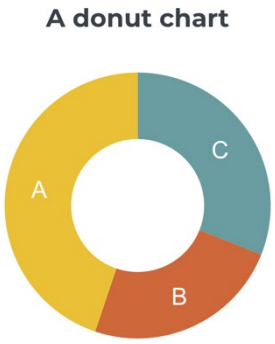


Pie charts

Pie charts 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 is very difficult to pick out slivers or compare relative segment sizes. A bar chart can always be used in place of a pie chart and is much clearer to read.

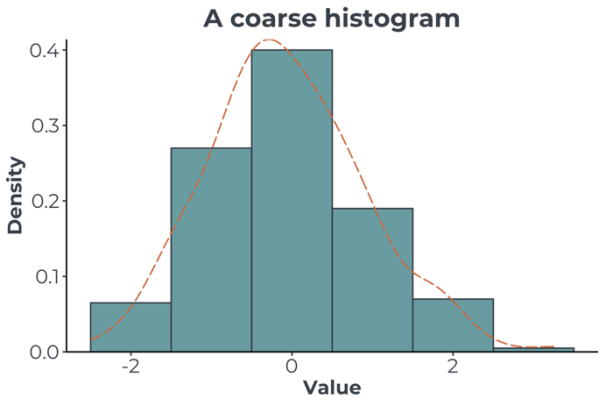
A **donut chart** is a version of a pie chart with the centre removed. They suffer from the same problems as a pie chart and again, bar charts are generally a better choice.





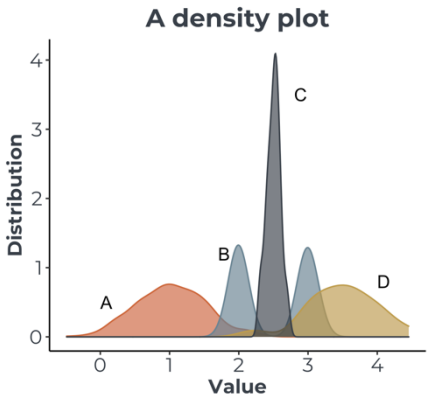
COMPLEX GRAPHICS TYPES

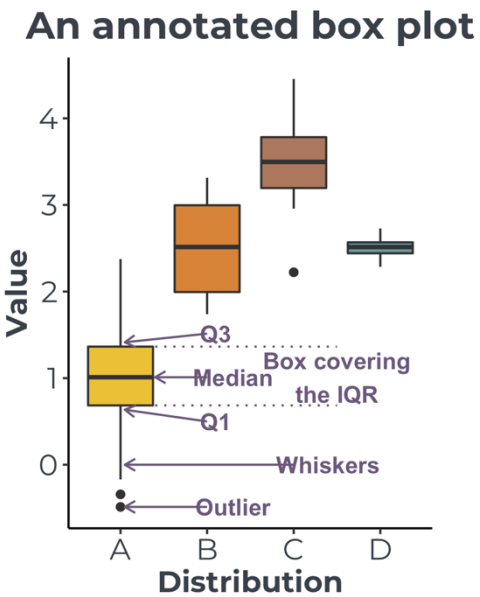
Distributions and densities



Histograms give a sense of the underlying distribution of the data, as demonstrated by this overlaid distribution or density function. The sum of a density function should always add up to 1 or 100% of the measurements.

It is also possible to compare the different shapes of distributions as a **density chart**. In this example transparency is used to make it possible to view the distributions on top of each other, however this can easily lead to a cluttered graph.





Box plots 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.

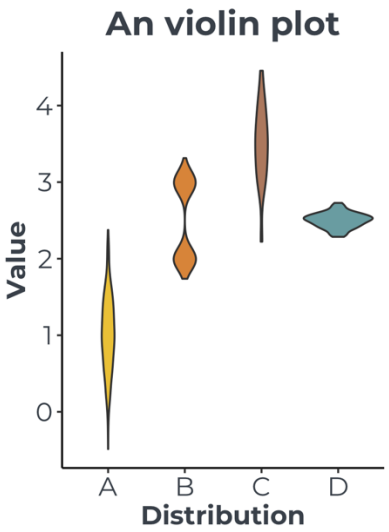
The middle part of the box shows the interquartile range (IQR) from the 25th percentile (Q1) up to the 75th percentile (Q3), with the central line being the median of the data. The lines coming out of the box (the **whiskers**) are set at 1.5*IQR beyond quartiles 1 and 3. Any values outside this range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

Find out more about interquartile range in the **Dispersion** section.

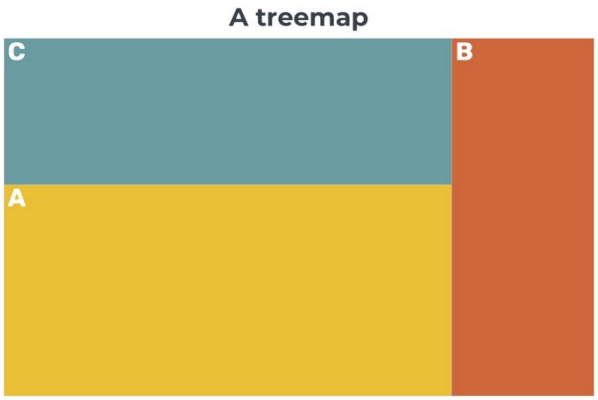
The shape of the box gives a feel for the shape of the distribution. Boxes that are equal widths either side of the median, imply symmetric distributions, whereas boxes with different widths imply skewed distributions. A shorter box implies a distribution with lower spread and a wide box means the data varies more and the distribution has greater spread.

The problem with box plots is that because the information is summarised, detail about the actual shapes of the distributions can be lost. This is where a **violin chart** is useful, which replaces the box with the shape of the distribution.

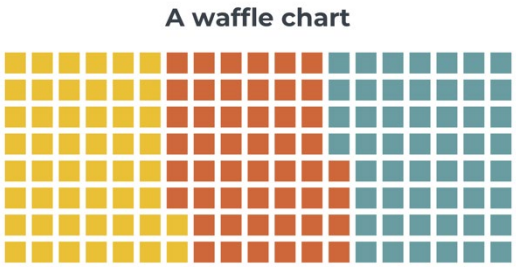
In this example it is now very clear that the shape of distributions A and B are very different.



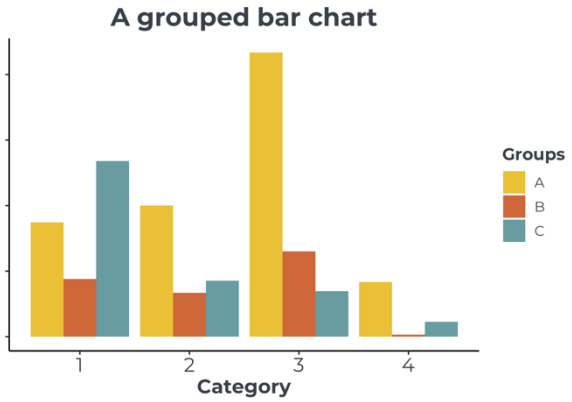
More proportions

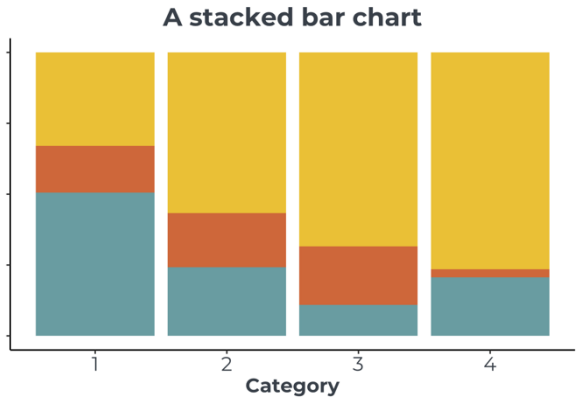


A **treemap** represents the **proportions of a whole**, but area rather than angles are used. They are useful for showing subcategories as well as high-level categories.



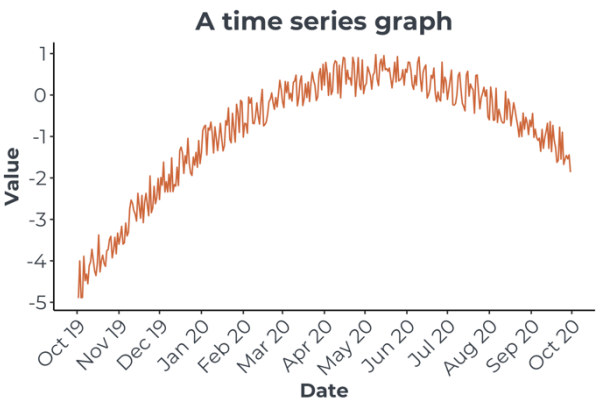
A **waffle or unit chart** shows proportions of a whole, but with each square or icon showing a tally. These are very popular in infographics and provide a clear at-a-glance view of proportions the case of limited numbers of categories.





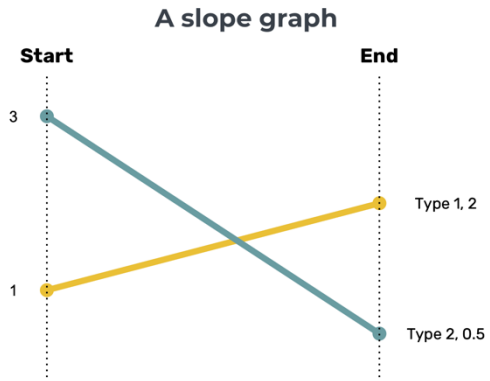
Grouped or stacked bar charts can show proportions of a whole and how they change across a varying quantity such as time. Like an area chart they are able to demonstrate trends in proportions.

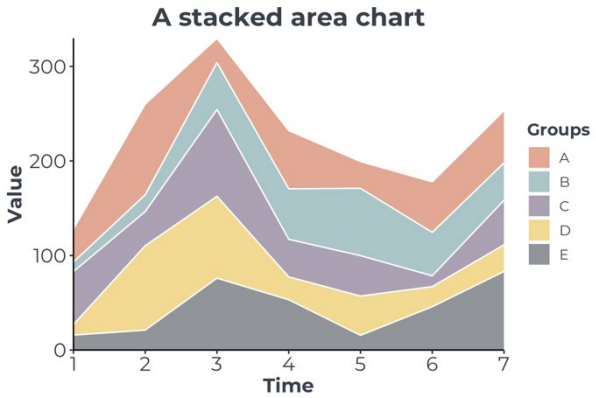
More line graphs



A **time series graph** 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.

A **slope graph** is a special case of a line graph, **showing change** over only two points in the varying quantity, say the start and end points. They are good for demonstrating trends, especially when there are many lines, but lose the detail of what is happening between the two points.

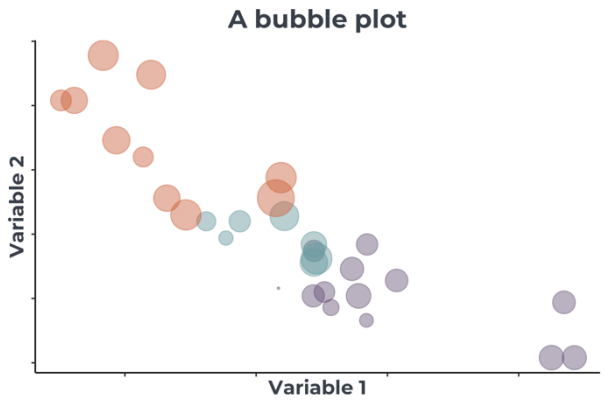




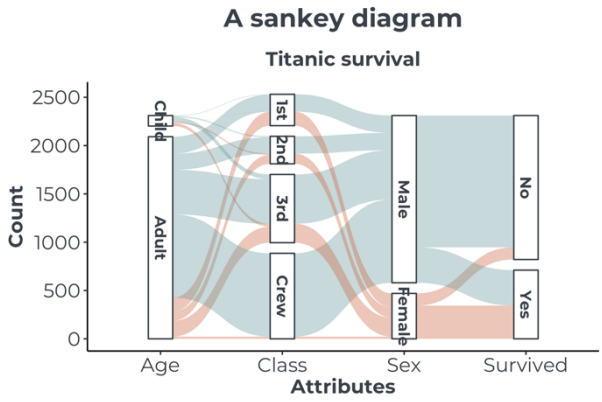
An **area chart or a stacked area chart** highlights **proportions changing** over the varying quantity. As well as raw data volumes, this can also be done as proportions of the whole, which is useful for example in demonstrating the change in survey results over time.

More relationships

A **bubble plot** 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.

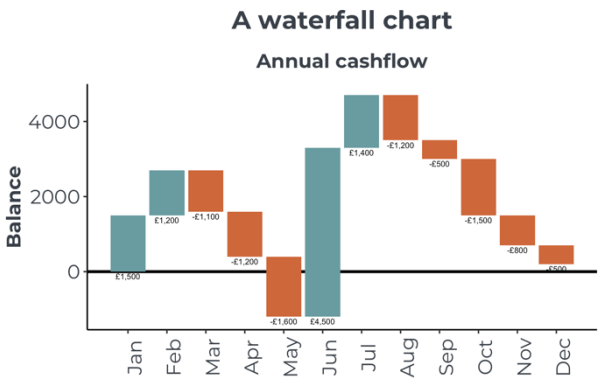


Demonstrating flow and change



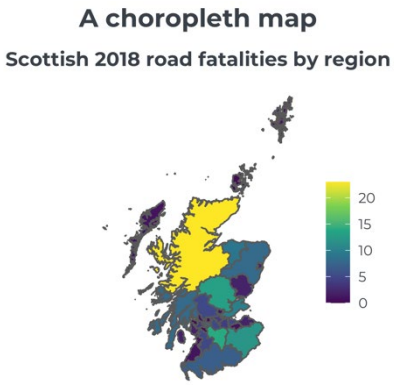
A **sankey chart** or alluvial diagram shows **flow**. This can be either through different states or over time. The width of the bars is proportional to the quantity flowing through it.

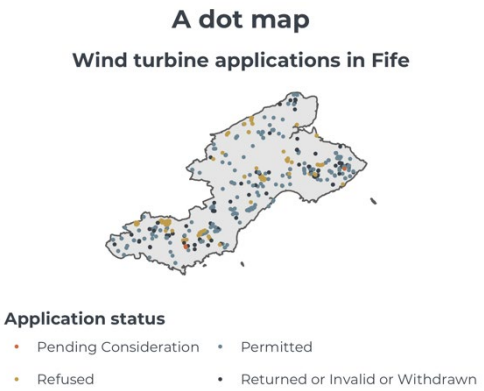
A **waterfall chart** is useful to demonstrate **changes in quantities** from a start point to an end point. The start and end points could be a time period, or often in the case of corporate waterfall charts, departments, or financial categories. The quantities are often monetary, stock levels or headcount. This example shows a fictitious cashflow over a year, with the money out in a different colour to the money in.



Maps

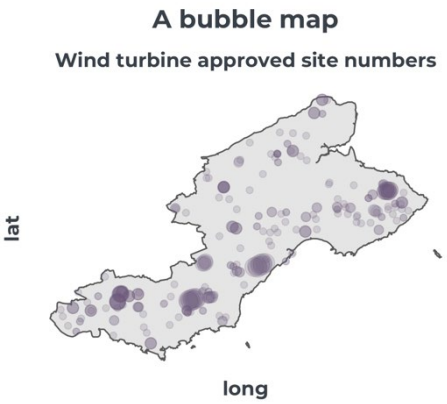
Maps are very useful for demonstrating **spatial patterns** in data. A **choropleth map** uses a colour scale to represent values of a quantity. In the example, it can clearly be seen that the Highland region had the highest number of road fatalities in the year. However, the size of the region is not 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.





This 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.

A **bubble map** 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.



WHAT IS DATA SCIENCE?

Data Science: an area of study that uses maths and statistics combined with computing science to find answers and solve problems in business and society

Analytics: the discovery of patterns and trends from data

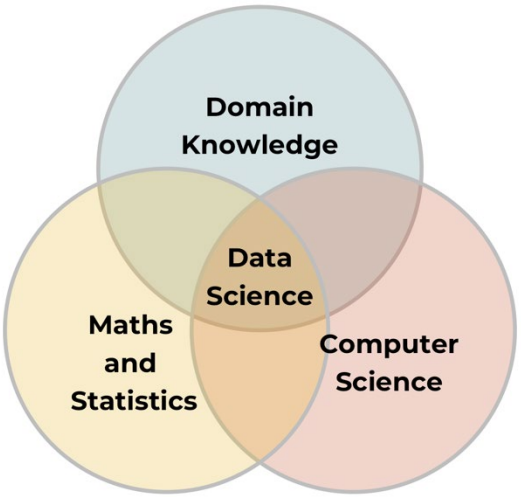
Big Data: a dataset that is too large to be processed by traditional tools and software

Artificial Intelligence (AI): the broad concept of machines being able to carry out tasks in a way that we would consider "smart"

Machine Learning: an application of AI based on giving machines access to data and letting them learn for themselves

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.

Applications of Data Science

Computers do not need to eat or sleep so can be used 24 hours a day to make consistent predictions. Although more consistent than a human, they are not necessarily more accurate. Machine learning models require updated data to constantly improve their predictions as the world around them changes.

Applications of data science and machine learning are far and wide, touching every aspect of our lives:

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

The Data Science Lifecycle

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

1. **Problem:** Identify the question being solved
2. **Plan:** Decide how to answer the question
3. **Data:** Collect, store and secure data
4. **Analysis:** Look for patterns and insights
5. **Conclusion:** Interpret, communicate and implement findings

Problem

Identify the right question to solve

- How much, or how many?
- Which category or group does this belong to?
- Is this weird?
- Which is the best option to choose?

Plan

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

Data

Collect and store the data securely

- Data quality checks
- Data understanding
- Data dictionary creation

Analysis

Preparation, manipulation, visualisation, data modelling and validation of findings

Conclusion

Summarise and communicate

- How does the data answer the original problem statement?
- Are there any aspects of the original question that has not been addressed?
- What are the conclusions?
- What could be done differently next time?
- 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.

Data Science in Business

Projects

There are almost limitless possibilities to what data can be used for, so deciding on the right problem to solve can often be one of the most challenging aspects of the project.

Data projects take time, effort and therefore money, so it is important to pick the right problem to solve.

Data projects can be created to answer an interesting question, to do something useful for the business or to create something valuable. The best problems address all three.

Data Scientists

Data scientists in business are not only able to write code to manipulate data, but they also use maths and statistics to carry out the analysis.

They combine their knowledge of the business they are working in, with a problem-solving mindset and a curiosity with data to ensure they are working on the right problems.

They also need to be good communicators and storytellers to ensure that they can summarise and share the insights they have captured.

Domain Experts

Domain experts support data scientists to understand the business and the customers. They are not necessarily skilled in data, but they understand the business in detail. They can identify the right problems to solve and the benefits in solving them. They understand the existing systems and processes. They may also understand how the data was initially captured and therefore the meaning of any special values or uses. They may be able to highlight any potential biases or ethical challenges.

WORKING WITH DATA

DATA CATEGORIES

Data can take a variety of forms, from numbers stored in databases to voice recordings and images. Each of these contain information, so they are all data. Most data falls into two main categories:

Qualitative

Descriptive, normally using words rather than numbers

- Difficult to analyse
- Limited visualisation options
- Words often captured against different categories or labels enabling quantification of occurrences in a category
- Summary statistics cannot be calculated
- The data can be subjective, containing opinions rather than facts
- Used to answer questions such as:
  - What?
  - Why?

Quantitative

Measures of values or counts and expressed as numbers

Discrete

- Whole numbers
- Obtained by counting

Continuous

- All possible values
- Obtained by measuring

- Easy to analyse
- Can be visually displayed using many different chart types
- Summary statistics can be calculated
- Used to answer questions such as:
  - How many?
  - How often?

SCALES OF MEASUREMENT

There are four scales of measurement in statistics that align to the categories of data. Nominal and ordinal are qualitative, whilst interval and ratio are quantitative.

Nominal

- Named variables
- No explicit order
- Examples include:
  - Names
  - Colours
  - Gender

Ordinal

- Named variables
- Ordered variables
- Differences cannot be measured
- Examples include:
  - Low / medium / high
  - Likert-scale
  - Age bands

Interval

- Named variables
- Ordered variables
- Differences can be measured
- No true zero
- Examples include:
  - Celsius temperatures
  - Time

Ratio

- Named variables
- Ordered variables
- Differences can be measured
- Has a true zero starting point
- Examples include:
  - Height
  - Weight

THE STRUCTURE OF DATA

Structured

- Ordered in rows and columns
- Defined by a data model
- Can be easily queried

Semi-Structured

- Self-describing structure
- Allows for flexibility
- Examples include HTML, JSON and XML

Unstructured

- No pre-defined structure
- 80-90% of data is in this format
- Difficult to analyse

Tidy data: datasets arranged such that each variable is a column and each observation a row

Storing and using data

Data type: how data is stored internally to the computer

Data structure: an organised collection of data types

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.

Examples of data structures:

- **Strings**: a collection of characters combined to create alphanumeric text
- **Array**: a structure of a fixed size which can hold items of the same data type
- **Vector**: a one-dimensional array
- **List**: a dynamically sized structure which can contain different data types
- **Data frame**: a two-dimensional structure designed for holding datasets. Each column can hold different data types, but all must contain the same number of items.

By using data structures, particularly data frames it allows faster and easier processing of properly structured datasets.

Display formats: control how a value is displayed without affecting the underlying precision

Display formats

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

Data type	Display format	Stored value	Displayed value
Floating point	1 decimal place	22.6176470588235	22.6
Floating point	percent	0.4893	48.9%
Date	%d-%m-%y	18496	22-08-20
Date	%B %d, %Y	18496	August 22, 2020
Date	%b %Y	-2000	Jul 1964
Time	%Y-%m-%d %H:%M:%S	1584801002	2020-03-21 14:30:02
Time	%r	1584801002	02:30:02 pm
Time	%c	-1613826000	Mon 11 Nov 11:00:00 1918
Boolean	TRUE/FALSE	1	TRUE
Floating point	£	24.99	£24.99
Character (Unicode)	emoji	\U0001f63b	😻

File formats, data storage and Sharing

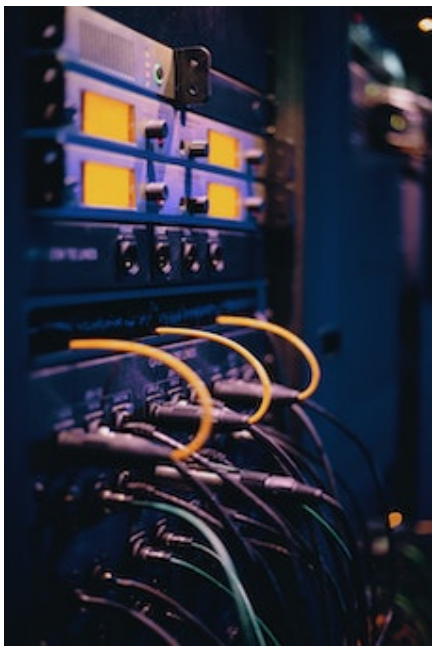
Data is stored in different digital file formats for sharing and transporting. The most common format for tabular data is a **.csv** (comma separated value) file. There are many different ways of storing data, depending on its contents. Examples include database (xml, csv, tab), geospatial shapefiles (shp, dbf), image (png, jpg), audio (mp3, wav), video (mp4, mov).

The chosen format should ensure long-term access and preservation of the data.

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, be they criminals or competitors. It can protect against fines, which can result from the loss of personal data under GDPR, reputational damage or trust in the company. Keeping data safe is everybody's responsibility with human beings often the weakest link. Data security requires protection of the physical locations where data is stored, the devices and accounts that can access it and the data itself when it is either at rest or in transit.

Staying Secure



Physical security

Devices that have data stored on them, or from which data stored in the cloud can be accessed, should be physically protected. This involves not leaving the device unattended, protecting it with a passcode and rebooting or switching it off if is not being used for a period of time.

If data is on a computer's hard drive, that drive should be encrypted to ensure no loss of data if the computer or disk are stolen.

When a computer is no longer required, the hard drive should be securely wiped prior to disposing of the computer, otherwise the data may still be accessible.

Preventing software viruses

Malware (malicious software) is unauthorised software that can end up on a computer and cause damage. There are several different types of malware:

Viruses

These attach themselves to files or software and require human interaction to spread. They can come from clicking on links in emails from unknown sources or downloading infected software.

Worms

These are like viruses but are able to spread without human interaction. The infamous ILOVEYOU (**dataed.in/love**) worm managed to infect 10% of the world's computers in only 10 days.

Ransomware

These can encrypt a user's files and hold data and computers hostage until a ransom is paid. A complete wipe of the computer will fix the problem, but data could be lost if not backed up elsewhere.

Spyware and Trojan horses

These can install themselves onto computers without the user's knowledge. It can then capture information such as passwords by monitoring keystrokes.

Antivirus software prevents malware from being downloaded to a computer and can also remove it if detected. All computers and mobile devices should have a virus-checker installed and scans run regularly. Antivirus software providers often provide a basic version for free.

Data science techniques can be used for detecting viruses and anomalous behaviours.

Other recommended defences

Firewalls

By acting as internet gatekeepers, they can block unwanted traffic from outside. Both routers and computers should have these enabled.

Virtual Private Networks

A VPN prevents data loss in transit by automatically encrypting traffic. They should always be used when accessing public WIFI hotspots.

Software upgrades

Any identified issues with software are regularly fixed by security patches. It is important to keep software up to date, so no vulnerabilities are present.

Keeping Accounts Secure

If the account used to access data is compromised, then all the data associated with that account can be accessed. Accounts are normally secured with a username and password.

Password management best practices

- Use a **strong** password
- Store passwords in a password manager
- Use multi-factor authentication (MFA)
- Use biometrics where possible
- Do not reuse passwords between different accounts
- Change passwords regularly
- Do not use personal information that can be guessed, or words that can be found in a dictionary
- Do not use consecutive letters on a keyboard
- Do not share passwords with other people and if you do, change it immediately

Strong passwords

Strong passwords are difficult to crack
Use a password strength checker (**dataed.in/pstrength**) and password generator (**dataed.in/pgen**) to help create them

123456
passw0rd
qwerty
awesomedog1

9X#u$4Xg9
zHw&kY^S5zR
RedLorryCliffRodeoPolice
niensoffirly

Password managers

Use a password manager to remove the need to ever remember a password again
They can be set up to be available across all devices
Many are available for free or come with browsers:

Lastpass (**lastpass.com**)
Remembear (**remembear.com**)
Dashlane (**dashlane.com**)
Myki (**myki.com**)
LogMeOnce (**logmeonce.com**)

Multi-factor authentication

Where available this provides an extra level of security by requiring multiple factors:

- **Knowledge**: something you know e.g. password
- **Possession**: something you have e.g. phone
- **Inherence**: something you are e.g. biometrics
- **Location**: somewhere you are e.g. a building

Biometrics

A person's physical characteristics can be used to authenticate access. Although convenient the use is highly debated, as if compromised unlike passwords cannot be replaced.

Examples include facial recognition, iris recognition, fingerprint scanners and hand geometry.

Keeping Data Secure

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

Access and control

Backups

To avoid accidental loss of data by deletion or corruption, regular backups should be taken.

Access Limitation

Only users with a valid reason to access the data should be able to. Permissions should be time limited and removed when no longer required.

Testing & Monitoring

Regular ethical hacking tests to identify weaknesses should be carried out. Monitoring of systems access can also identify data breaches.

DATA PRIVACY

Individual privacy: the right to keep one's personal matters and relationships secret.

Information privacy: the right to have some control over how your personal information is collected and used

Data Protection: a legal framework to protect individual privacy by focusing on data privacy

Privacy is a fundamental human right. There are seven types of privacy:

Body | Correspondence | Finance | Identity | Location | Data | Territory

Data protection and data security, although related are not the same. Data protection focuses on the handling, storage, and usage of personal data to maintain a person's right to privacy. Data security focuses on protecting data from unauthorised access. However, if security is compromised and a data breach has occurred, then it is likely that privacy has also been compromised.

GDPR (The General Data Protection Regulation)

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.

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.

Examples of personal data include:

Personal Data

- Names
- Addresses
- Phone numbers
- Identification numbers
- Location data
- Online identifiers
- A combination of identifiers that together can identify an individual
- Pseudonymised data – a key is required to identify the individual

Sensitive Personal Data

- Racial or ethnic origin
- Political opinions
- Religious or philosophical beliefs
- Trade union membership
- Genetic data
- Biometric data, where used for identification
- Health data
- Sexual orientation and activity

Your data rights

Individuals have certain rights under GDPR:

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.

Object

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

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.

Lawful processing

There are six valid reasons for processing data and at least one must be in place to allow personal data to be processed. The lawful basis for processing the data should always be documented in the Privacy Notice

Protecting privacy online

It is important to actively manage your privacy online otherwise more information may be being shared than is 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

This type of information allows companies to build up profiles of individuals and use them for targeting products and services. It is important to regularly review your privacy settings and the information that online companies are storing about you. Most have easy ways to exercise your rights and access this information.

MANUAL DATA CAPTURE

Data can be found in multiple places: systems, the internet, devices and sensors and can also be manually captured. When data is captured manually, it can be obtained in a number of ways. Data can be obtained by **observation**, such as counting traffic, or observing activities. It can also be obtained through **interviews**, although this is quite time consuming if aiming to acquire multiple data points. It can also be obtained by **reviewing individual documents**. The most efficient way to manually capture information is through **surveys** or **questionnaires**.

SURVEYS

Types of survey

- **Cross-sectional:** point in time view of responses from a population
- **Longitudinal:** the same population is followed over a period of time
- **Retrospective:** responses about events in the past

Delivery method

- **Face to face/telephone:** questions asked directly to respondent. High cost.
- **Postal:** paper form filled in remotely. Low response rates. Need to transcribe to digital format
- **Internet:** Web-based survey. Cheap and fast.

Example survey tools

- **Google Forms** (**dataed.in/GForms**)**:** Free. Results stored in a Google sheet
- **Microsoft Forms** (**dataed.in/MSForms**)**:** Free. Results stored in Excel
- **SurveyMonkey.co.uk:** Free to capture. Pay to download results
- **Typeform.com:** Paid for product

Survey question types

Open-ended questions

- Allows respondent to answer in their own words so captures greater detail in responses
- Enables an understanding of **why?**
- Difficult to analyse

Closed-ended questions

- Limited responses capture quantitative information
- Simpler analysis approaches
- Nuanced responses can be lost

Types of closed questions

- **Dichotomous**: two responses, limit the ability to capture neutral responses
- **Multiple-choice**: a closed list of possible responses. Important to include all the possible answers
- **Likert scale**: a 5- or 7-point scale normally from **Strongly Agree** to **Strongly Disagree**. Powerful, but data captured is ordinal, so averages should not be calculated.
- **Rating scales**: 5- or 10-point scale. Similar to the Likert Scale

Personal data in surveys

Does personal data really need to be captured as part of the survey, or could the same outcomes be achieved anonymously? If personal data is needed, a privacy notice is legally required to tell the respondent how their data will be stored and managed. This should be provided before any data is captured to allow the respondent to withdraw from the survey if they wish. Sensitive data should never be requested in a survey.

it is important to tell the respondent:

The personal data that is being captured	The plan for using the data including how long it will be held	Whether the data will be shared with a third-party and if so whom	How respondents can access their responses	How respondents can contact you

Avoiding Bias

Bias can easily creep into even the most carefully designed survey, leading to inaccurate responses, resulting in incorrect insights.

There are different types of bias that should aim to be reduced as much as possible by the design of the survey:

Sampling bias

The individuals surveyed do not match the target population. Some demographic groups will be missed out.

Explicit sampling methods should be used to avoid this.

Non-response bias

Similar to sampling bias, groups of individuals may be either unwilling or unable to respond.

Personal invites and inclusive language can help this.

Response bias

Getting truthful and accurate responses is a challenge. Human tendency is to agree, or to provide the expected "right" answer.

Careful question wording can reduce this.

Order bias

The order of questions can influence responses with answers to earlier questions affecting later responses.

Randomising the question order or grouping by topic can minimise this.

Survey Best practice

A summary of the best tips for survey design:

- **Design**: define a goal for the survey and ensure every question adds value and insight
- **Ordering**: start with the more straightforward questions and keep the personal questions until the end. Structure into logical sections
- **Length**: make it as short and simple as possible
- **Question type**: focus on closed-ended questions
- **Incentives**: carefully word the email invitation to maximise response and encourage feedback
- **Leading questions**: avoid asking leading questions and allow for neutral responses
- **Balance**: keep answer choices balanced
- **Compound questions**: only ask one question at a time
- **Test it out**: pilot the survey on a small population before sending out more widely
- **Language**: use simple, unambiguous language
- **Answer grids**: avoid grids of answers as they are not mobile-friendly and tend to lead to similar responses for all
- **Privacy awareness**: be aware of data protection responsibilities
- **Data quality**: aim to capture high quality data by putting checks in place around allowed input values
- **Feedback**: if possible, communicate back the outcome to make the respondent feel involved

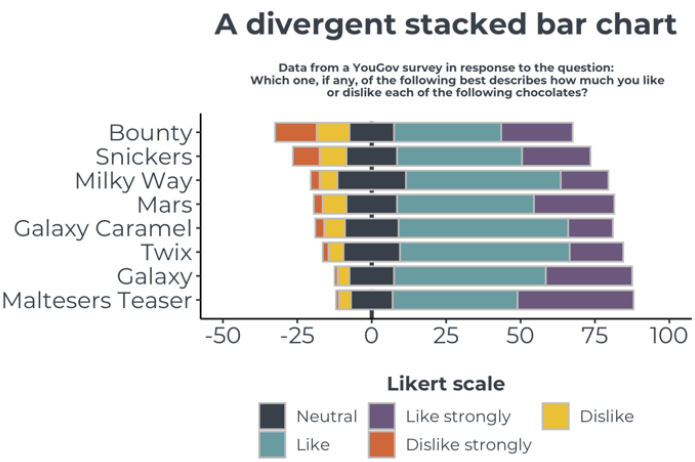
Analysing Surveys

Use the PPDAC framework to analyse surveys:

Problem: Focus on the survey purpose
Plan: Design the questions to test the hypothesis
Data: Collect and store the data securely
Analysis:

1. Frequency by category
2. Cross-tabulate between questions
3. Create derived variables
4. Visualise the information

Conclusion: Draw conclusions and communicate findings



DATA TRANSFORMATION AND MANIPULATION

Data transformation is the process of changing data into something usable. This can involve changing formats, creating new variables, cleaning or summarising data. There are lots of different terms used to describe these activities including: data manipulation, data wrangling, data munging and data preparation, however they all generally mean the same thing.

There are a wide variety of tools that can be used for data transformation, from spreadsheet packages to structured programming languages. These are covered in more detail in **Tools and Languages** section of this document. Most tools will be capable of undertaking a similar set of activities; however, it may be easier to do certain things in one tool than another. The vast majority of a data analyst's time is spent on these activities, so it is important to become adept at manipulating data in whatever tools are available.

Single Table Manipulations

There are a core set of activities that can be carried out on a single dataset. When undertaken in the correct order, they can transform a messy input dataset into clean output dataset, ready for further analysis.

Tables

The main activity that can be carried out on a whole table of data is to **subset** it. This involves taking a copy of just a portion of the data: taking a selection of the rows and columns. This might be required to test out the analysis on a smaller dataset, or to only keep the columns of interest.

In this example only the products bought by the first two customers have been subsetted.

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41
4	orange	3	50
5	melon	1	89

subset

sale	product
1	apple
1	banana
2	pear
2	lemon

Columns

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41

select

sale	product
1	apple
1	banana
2	pear
2	lemon
3	apple

The first thing an analyst might want to do is to **select** the columns of interest. This involves taking a copy of the dataset, but only a subset of the columns. In this example the sales and product columns have been selected along with all the observations or rows. of the data.

The next thing an analyst might choose to do is to **reorder** the columns in the dataset. The reason for this might be to put related columns together to make it easier to work with. The underlying data will remain untouched.

1	2	3	4
sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41

reorder

2	4	1	3
product	price	sale	quantity
apple	41	1	6
banana	27	1	5
pear	46	2	4
lemon	28	2	6
apple	41	3	2

total(p) = quantity x price

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41

create

sale	product	quantity	price	total(p)
1	apple	6	41	246
1	banana	5	27	135
2	pear	4	46	184
2	lemon	6	28	168
3	apple	2	41	82

It is very likely that an analyst will **create new variables** calculated from the existing data in the dataset.

In this example the total price of the sale (in pence) has been calculated from the product price multiplied by the quantity bought. It is important to document the calculations undertaken so that the analysis can be reviewed or reproduced by another team member.

It may also be necessary to **reformat** the values in a column. This may not involve changing the underlying stored value but displaying it differently. In this example the total price has been reformatted to be shown in £s rather than pence. It is common to reformat dates and times, percentages, or the precision of numbers.

sale	product	quantity	price	total(p)
1	apple	6	41	246
1	banana	5	27	135
2	pear	4	46	184
2	lemon	6	28	168
3	apple	2	41	82

reformat

sale	product	quantity	price	total(£)
1	apple	6	41	2.46
1	banana	5	27	1.35
2	pear	4	46	1.84
2	lemon	6	28	1.68
3	apple	2	41	0.82

sale	timestamp
1	2020-09-07 09:23:41 BST
2	2020-09-07 10:16:25 BST
3	2020-09-08 16:02:37 BST

extract

sale	date	time
1	2020-09-07	09:23:41
2	2020-09-07	10:16:25
3	2020-09-08	16:02:37

The last activity that can be carried out on columns is to **extract** information from the data. In this example the timestamp contains both date and time information, which has been split up into date and time separately.

Rows

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41

filter

product equals "apple"

sale	product	quantity	price
1	apple	6	41
3	apple	2	41

It is common to need to **filter** data on a logical criterion, keeping only the rows that satisfy the criteria. In this example only the rows where the product is equal to "apple" have been retained. The logic for filtering data can be as simple or complex as required.

As well as re-ordering columns, rows can also be re-ordered. This is normally called **sorting**. In this example the dataset has been sorted on the total column in descending order, so the largest value is at the top.

sale	product	total	
1	apple	246	**5**
1	banana	135	**2**
2	pear	184	**4**
2	lemon	168	**3**
3	apple	82	**1**

sort

sale	product	total	
1	apple	246	**5**
2	pear	184	**4**
2	lemon	168	**3**
1	banana	135	**2**
3	apple	82	**1**

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
1	banana	5	27
3	apple	2	41

deduplicate

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
3	apple	2	41

It is sometimes necessary to **deduplicate** rows in a dataset. This is only possible if every single value identically matches across all the columns. In this example the sale of bananas to person 1 has been recorded twice, so only a single row remains after deduplication.

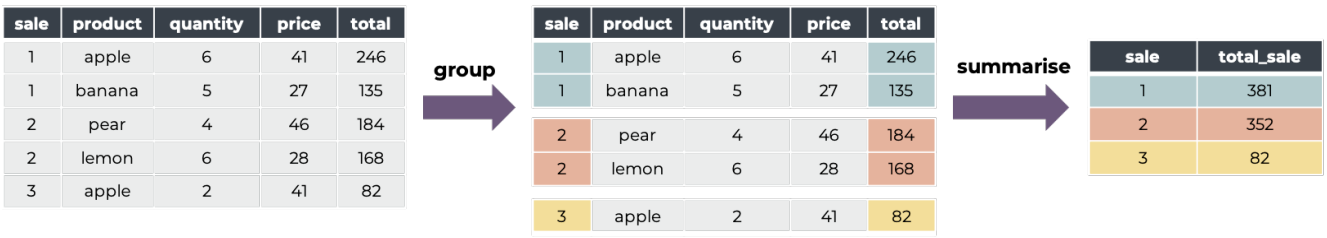
To undertake analysis on datasets it is often necessary to **aggregate** the data.

> **Aggregation:** the summarisation of multiple data points into a single metric

When data is summarised across rows, there are different calculations that can be used. The most common ones are:

- **Counts**: a count of the number of valid data items
- **Totals**: a sum of the values
- **Averages**: a mean, mode or median of the valid values, care should be taken when missing values are present
- **Min/max**: the minimum or maximum of the valid values

The summary does not need to be across the whole dataset but can be within groups. In the example below the dataset is first grouped by the sale and then the total value of each sale is calculated.



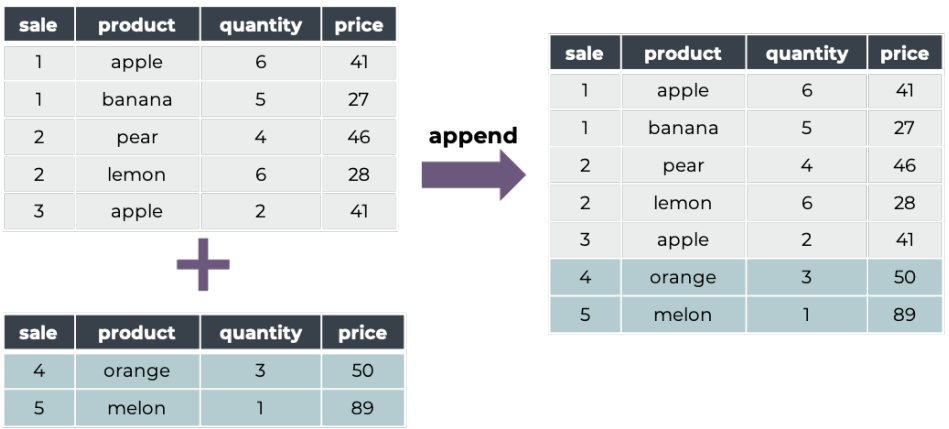
sale	product	quantity	price	total
1	apple	6	41	246
1	banana	5	27	135
2	pear	4	46	184
2	lemon	6	28	168
3	apple	2	41	82

sale	total_sale
1	381
2	352
3	82

Joining

It is often necessary to **join** datasets together. This could be to add additional rows, columns or merge in information from a lookup table.

Appending is to add on to the end of the existing dataset by adding additional rows. Ideally the appended dataset contains identical variables to this original dataset. If this is not the case, then additional columns are added containing missing data.



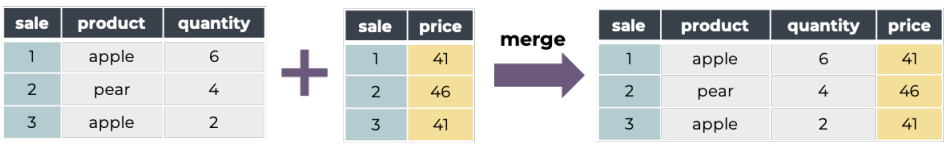
sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41

sale	product	quantity	price
4	orange	3	50
5	melon	1	89

sale	product	quantity	price
1	apple	6	41
1	banana	5	27
2	pear	4	46
2	lemon	6	28
3	apple	2	41
4	orange	3	50
5	melon	1	89

When two datasets are **merged** together, new columns are added, however they need to have at least one column in common. This is called the **key**. This commonly occurs when *reference* or *lookup* data exists in a supporting table.

In this example the *key* is the sale and the price of each sale is being merged into the original table containing products and quantities bought.



sale	product	quantity
1	apple	6
2	pear	4
3	apple	2

sale	price
1	41
2	46
3	41

sale	product	quantity	price
1	apple	6	41
2	pear	4	46
3	apple	2	41

Types of join

There are different ways to join data in a merge, each resulting in different output datasets if not all rows are common to both datasets. Care should be taken to ensure when joining that the output dataset is as expected.

Sales Table

sale	product	quantity
1	apple	6
2	pear	4
5	melon	1

Pricing Table

product	price
apple	41
pear	46
lemon	28

Left Join

sale	product	quantity	price
1	apple	6	41
2	pear	4	46
5	melon	1	

Right Join

sale	product	quantity	price
1	apple	6	41
2	pear	4	46
	lemon		28

Inner Join

sale	product	quantity	price
1	apple	6	41
2	pear	4	46

Full Join

sale	product	quantity	price
1	apple	6	41
2	pear	4	46
5	melon	1	
	lemon		28

In the examples the merge has taken place on the sales variable (*the key*), so this is only included in the output dataset once.

Left join: In this situation we consider the leftmost, or first dataset. It is being joined to the rightmost, or second dataset. For a left join **all** the observations from the leftmost dataset are included and **any** that match with the second dataset. This is the most common type of join.

Right join: This situation is the opposite of the left join. **All** the observations from the rightmost dataset are included, and any that match from the first/leftmost dataset.

Inner join: This situation only looks at information that is contained in **both** datasets being merged. Rows have been lost from both datasets, where there is no overlap. This is a common area for errors to occur, as dataset rows will vanish from the output dataset.

Outer (full) join: An outer or full join ensures that no information is lost, since it merges any data in either the first or second dataset.

Reshaping data

Datasets can also be reshaped. It is common for data not to be captured in an easy structure for analysis. This is often necessary to reshape data to create **tidy** datasets.

weather_station	temp1960	temp1980	temp2000	temp2020
Lerwick	7.5	7.0	7.5	8.0
Armagh	9.4	9.0	9.9	10.7
Eastbourne	10.8	10.6	11.4	12.8

Wide data has a different data variable in each column. This means that the row headers often contain important information. In this example the columns contain the average temperature values for each year observed.

weather_station	year	temperature
Lerwick	1960	7.5
Armagh	1960	9.4
Eastbourne	1960	10.8
Lerwick	1980	7.0
Armagh	1980	9.0
Eastbourne	1980	10.6
Lerwick	2000	7.5
Armagh	2000	9.9
Eastbourne	2000	11.4
Lerwick	2020	8.0
Armagh	2020	10.7
Eastbourne	2020	12.8

Long data has every row containing a single observation belonging to a particular category. That means there will be multiple columns for the categories and a single column for the value.

In this example the temperature value is the observation and the weather station and year are the categories.

The advantage of long data is the ease with which it can be grouped and summarised.

STATISTICS

Statistics can mean two different, but related things:

Field of Statistics: the mathematical science involving the collection, analysis and interpretation of data

and

A statistic: a fact about or summary of data

Data science is often considered to be applied statistics, as it involves the application of statistics to real-world problems. However, data science also involves implementing a solution or acting on the interpretation of data.

Statistics is about uncertainty. It provides the mathematical approaches and tools to deal with uncertainty in a dataset of any size.

Population and Samples

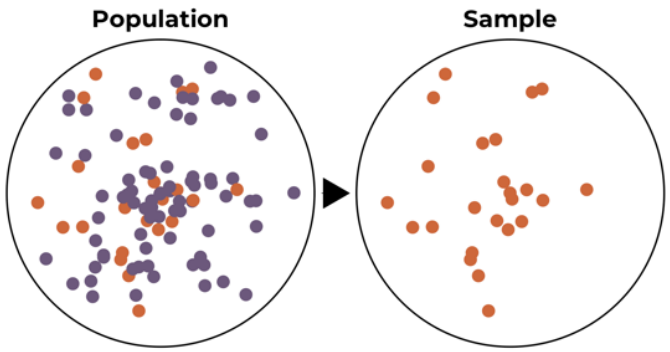
The concept of a population has a special meaning in statistics and can refer to things as well as people.

Population: the total set of observations that can be made

and

Sample: one or more observations drawn from the population. The size of the sample is always less than the size of the population.

Population	Sample
All the people in the UK	All the people in Scotland
All S3 pupils in Scotland	The S3 pupils in a single school
All the items a factory produces in a week	Items extracted from the production line for testing every hour



A statistic is just a number that describes the data. There are two types of numbers that describe data, **statistics** and **parameters**. A parameter is calculated from the full **population** and therefore does not change, whereas a statistic is calculated from a **sample** of the full population and varies depending on the sample chosen. A statistic is used to estimate a parameter. Using a sample to estimate a population parameter is called **inferential statistics**.

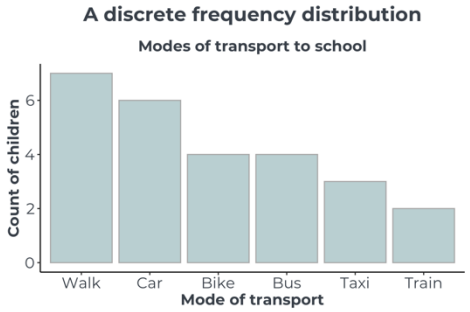
Inferential Statistics: takes data from a sample to make predictions about the population from which the sample was drawn

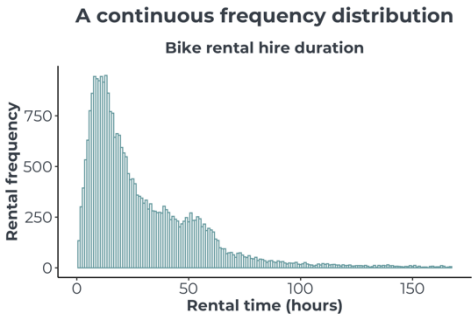
Descriptive Statistics: summarises the data in the sample provided

Summarising Data

To make sense of data it needs to be summarised. The simplest way to summarise it is to **count** it. For discrete data, this involves counting the number of observations in each category. For continuous data, the data first needs to be **binned** into intervals and the number of observations in that bin counted. This is called a **frequency distribution**.

Below are examples of both discrete and continuous frequency distributions:





Centrality

When data needs to be summarised to just a single value, the most common value chosen is the middle value or average value. There are three types of **average** that can be calculated and the relationships between the different midpoints give a feel for the shape of the data.

Mean

The arithmetic mean is the sum of all the values in the dataset variable, divided by the number of dataset points. Care should be taken when missing values are present.

The mean is very sensitive to extreme values or outliers.

Median

The median is the middle value of an ordered dataset variable. If there are an odd number of points the middle value is taken, if an even number, then the mean of the middle two values is used.

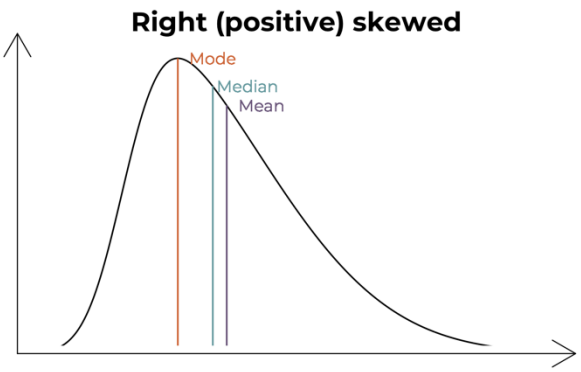
The median is not affected by outliers.

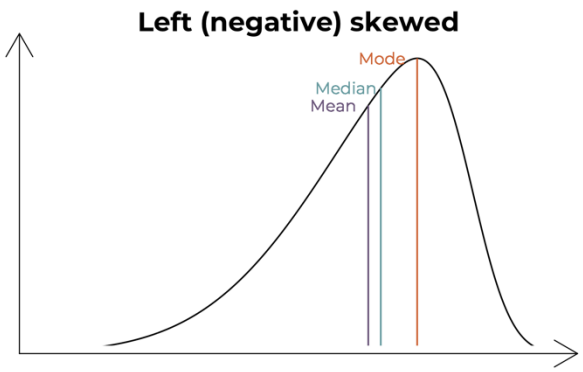
Mode

The mode is the most frequent value of a variable. This can be used with both numeric and non-numeric data.

For continuous frequency distributions it is the maximum value. There can be more than one mode if the distribution has multiple peaks.

Skewness is the amount of asymmetry in the shape of the dataset. If the **mean is roughly equal to the median** then the distribution is **symmetric**, so there is no skewness. If the **mean is greater than both the median and the mode** this implies that there is an extreme high value, as the mean has been pulled up towards the **outlier**.





If the mean is greater than the mode, the distribution is positively skewed.
If the mean is less than the mode, the distribution is negatively skewed.
If the mean is greater than the median, the distribution is positively skewed.
If the mean is less than the median, the distribution is negatively skewed.

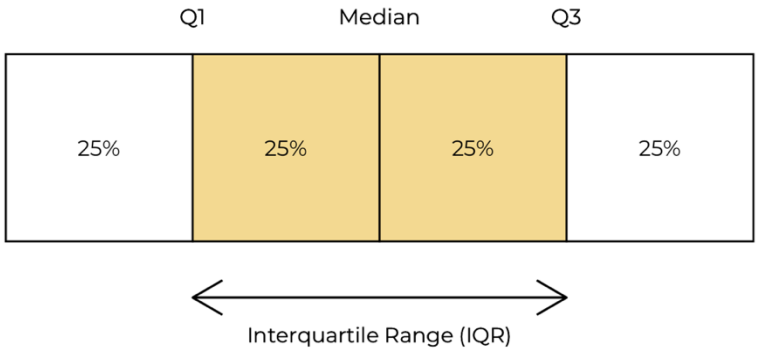
Dispersion

As well has having a feel for the middle of the data, it is also important to understand how it is **spread**, or dispersed, around the middle value. This gives a feel for how well the mean and median can summarise the data.

Range: largest (maximum) value minus the smallest (minimum) value in a dataset variable

Interquartile range (IQR): upper quartile (Q3) minus the lower quartile (Q1) of a dataset variable

In a similar way to the mean, the range is also very affected by outliers. To address this in a similar way to using the median to estimate the middle of the dataset rather than the mean, the interquartile range gives a view of the spread of the middle 50% of the data and is therefore unaffected by extreme values.



Variance

The variance measures the average amount that each point in the data varies from the mean.

The variance is calculated by first calculating the mean. Then subtracting the mean from the value of each data point and squaring those numbers. Finally, the average of the squared differences is calculated. This is calculated slightly differently for populations and samples.

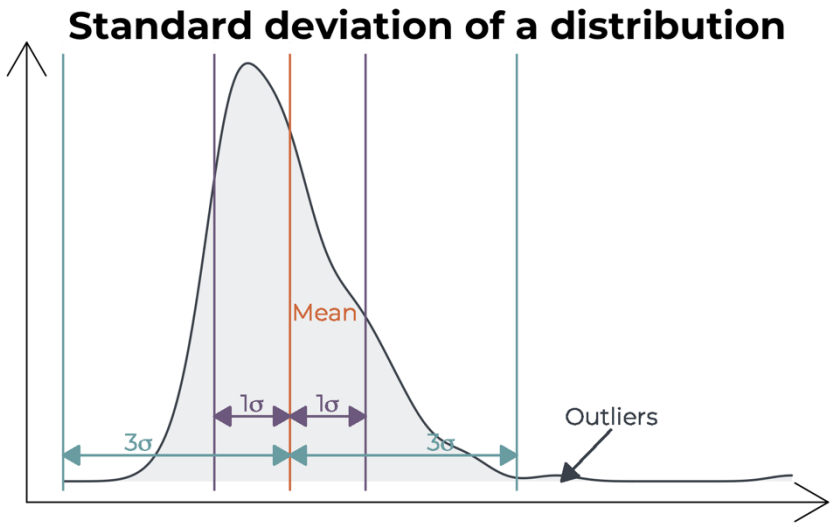
Like the range, the variance is sensitive to outliers.

Standard Deviation (σ)

The standard deviation is the square root of the variance.

The standard deviation is useful as it is in the same units as the mean.

For normally distributed quantities, 68% of all the values will lie within 1 standard distribution either side of the mean and 99.9% of values will lie within 3 standard deviations, making it useful to help define outliers.



An outlier is an unusual data point that differs significantly from other observations. This could either be a real, but extreme value, a measurement error, or highlight a data quality issue.

Suspected outliers are often identified as being either 1.5*IQR or 3σ.

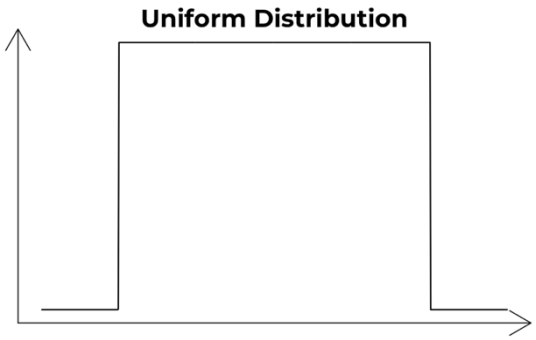
Distributions

Distribution: a function that shows the possible values for a variable and how often they occur

The distribution of data is the shape of the data, gathered from an understanding of its centre and spread. A probability distribution gives the likelihood of obtaining each possible value. Probability distribution functions always add up to 1 since they are the sum of all possible values.

There are a number of common distributions that appear regularly in real-world scenarios, these can be both discrete and continuous.

Discrete distributions

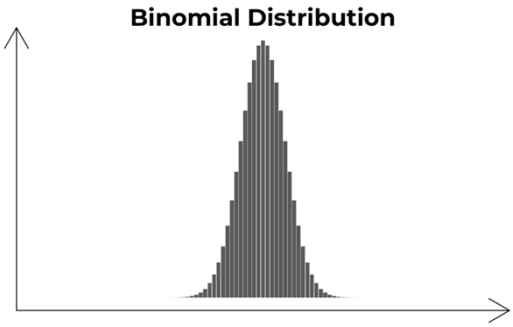


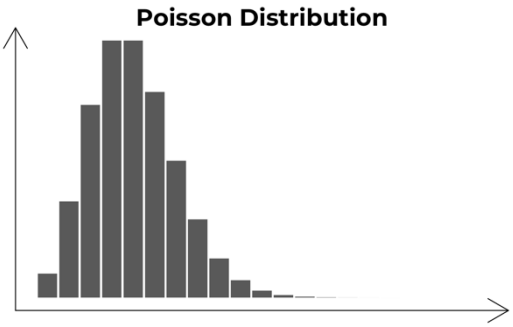
The **uniform distribution** covers activities, such as rolling a die where the outcome is equally likely. For example, the probability of throwing any particular dice value is 1/6.

The **Bernoulli distribution** is similar but represents outcomes that are not equally likely, such as an unfair coin toss.

The **binomial distribution** is the sum of multiple Bernoulli events, such as multiple coin tosses.

The **poisson distribution** predicts the count of events in a specific interval, such as time or distance. An example of this could be the decay from a radioactive source.



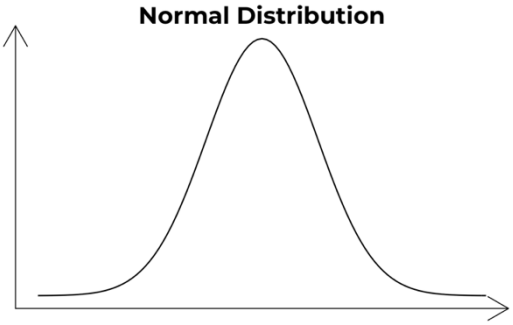


Continuous Distributions

The **normal distribution**, also called the Gaussian distribution or a bell-curve is the most important distribution, as it occurs most commonly in nature. It also has some special properties, as it's mean, mode and median are all equal.

Real-life examples of things that follow normal distributions are:

- Adult heights
- Blood pressure
- IQ
- Test scores



DATA ANALYSIS

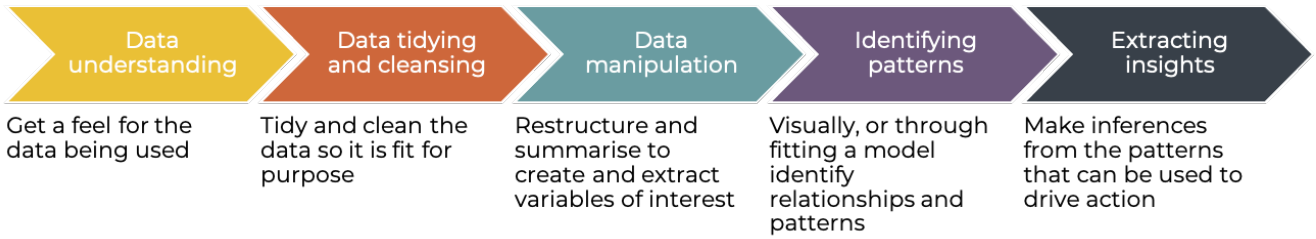
Data Analysis: the process of inspecting, cleansing, transforming and modelling data with the goal of discovering useful information, informing conclusions and supporting decision-making

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

Analysis Steps

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**:



Data Understanding

All analysis activities should start with an understanding of the data being analysed. This allows the analyst to get a feel for the data, understand its size and shape and identify any obvious issues with the data that may need to be addressed.

Visual inspection

By visually inspecting the data as a first step, it will give a feel for the size of the data, the frequency of missing values and the data types and formats.

Size and shape

Information captured at this point should include the number of rows, the number of columns, the column names and the column formats.

Summary statistics

Summary statistics for each variable should be created. This will be different depending on whether the variable is qualitative or quantitative.

sale	product	quantity	price
1	Apple	6	41
2	b@n@na	2	
3	kiwi	3	46
4	lemon	1	28
5			41

Initial observations:
- Mixture of numeric and character data
- Missing values
- Inconsistent format of product

shape	number
Number of rows	5
Number of columns	4

variable name	variable type
sale	integer
product	character
quantity	integer
price	float

var	mean	min	max	missing
sale	3	1	5	0
quantity	3	1	6	1
price	39	28	46	1

var	distinct values	missing
product	4	1

Data dictionaries are often a good place to start to understand the data being analysed. Data dictionaries contain a lot of information about the dataset, not only explanations about the variables, but also formats, valid values and allowed ranges. Data dictionaries can also provide valuable information about the required security classifications and retention periods.

Data Tidying and Cleaning

Tidying data

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

- **Removing metadata** at the top of a file, by either deleting or skipping rows to allow the dataset to be successfully read in.
- **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.
- **Rescaling data** so that numeric values fit within a specific range.

Fixing strings

An example of some string data that might require fixing is: “Star Wars”, “STARWARS”, “star_wars” and “Star wars”. A computer, being case-sensitive will not interpret these as the same.

Options for fixing strings include:

- **Upper case**: THE WHOLE STRING IS IN UPPER CASE
- **Lower case**: the whole string is in lower case
- **Title case**: The First Character of Each Word is Capitalised, Except Common Words
- **Capital case**: The First Character Of Each Word Is Capitalised
- **Sentence case**: Everything is lower case except the first character and I

Common options are upper and lower case, but any can be chosen so long as they are consistent.

Variable naming conventions

A variable name should be descriptive and summarise what is contained in the column. A good variable name makes it easy for someone else to read through code and understand what is going on. Names such as “Var1” and “Var2” should be avoided.

The best practices for naming variables in datasets include:

- **Avoid whitespace in a name:** this is because a name such as “First Name” will always need to be referred to in code using quotes
- **Stick to lower case names:** most computer languages are case sensitive and by sticking to a consistent case ensures that errors are less likely to occur
- **Use a consistent style:** there are a number of common styles to choose from:
    - **Snake case (snake_case):** this puts an underscore _ between each word e.g. first_name
    - **Camel case (camelCase):** this capitalises the first letter of each word, apart from the first one e.g. firstName
    - **Pascal case (PascalCase):** this capitalises the first letter of each word, including the first one e.g. FirstName
    - **Kebab case (kebab-case):** this puts a dash – between each word e.g. first-name

Missing values and outliers

Missing data includes empty cells, some zeroes, blank strings and sometimes system default values. There may be many reasons why data may be missing. A question may not have been asked, or the system may not have captured the response. Without knowing the real reason the data was missing, it is difficult to decide the best approach to deal with missing values.

Similarly, there are a number of causes of outliers which can be both natural and due to system and human error:

- **Data entry errors:** entering a value in grams when the system expects kilograms will lead to an outlier
- **Measurement error:** the instrument has incorrectly measured the value
- **Experimental error:** incorrect data extraction or experimental approach
- **Intentional error:** misreporting of information
- **Data processing errors:** incorrect manipulation of data prior to capture and use
- **Sampling:** incorrect merging of datasets
- **A real outlying data point**

There are four common approaches to deal with both missing data and outliers, and good data analysts will make active decisions around which method to use and why it is suitable for the specific situation:

Remove the record	Treat missing records separately	Impute a suitable value e.g. average, median, best fit	Do nothing

Whichever approach is chosen, the decision should be documented and justified.

Duplicates

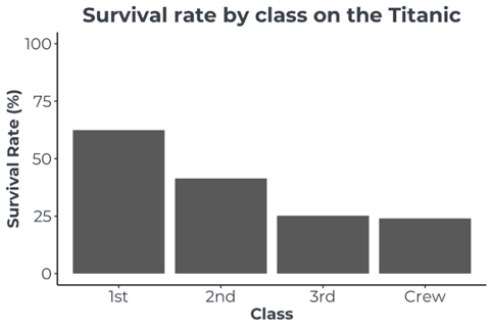
Duplicates can be exact copies of observations, that can be tidied up by deleting the additional copies. However, duplicates can also be similar records, caused for example by an individual responding multiple times. In the case of partial duplicates, it is common for one version to be active and regularly updated and the remaining copies to remain unused. In these cases, care should be taken when deduplicating to ensure the master active record is not deleted. The records should ideally be merged into a single active record prior to removal of the copies.

Manipulation

In addition to the common data manipulations and joins covered in the **Data Manipulation** section and the graphical display of data covered in the **Visualising Data** section it may also be useful to **cross tabulate** between variables to uncover hidden relationships. It is normally carried out on categorical data. This is the type of functionality that Excel's pivot table provides.

Here is an example of the Titanic survival dataset (**dataed.in/titanic**), where the outcome of all 2201 passengers has already been cross-tabulated by Class, Sex and Age:

		Survived		
		No	**Yes**	**Total**
Class	**1st**	122	203	325
	2nd	167	118	285
	3rd	528	178	706
	Crew	673	212	885
	Total	1490	711	2201

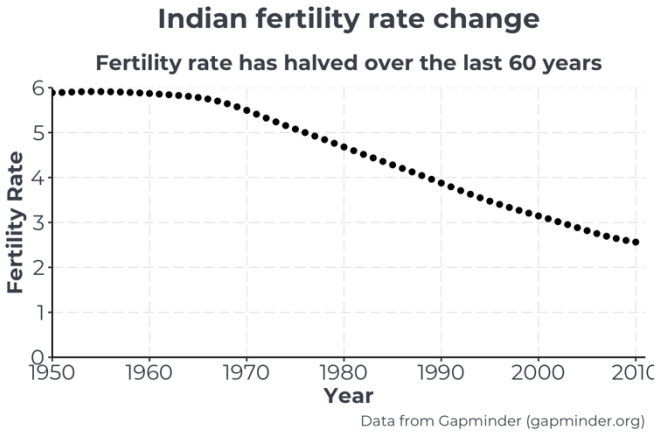


From this, by first cross tabulating the groups, it is now possible to calculated and plot the survival rate by Class.

Identifying Patterns

Once data has been manipulated and visualised a pattern may be immediately obvious. This could be increasing, or decreasing, or cyclical. The pattern could be over time, or by age band, or another independent analysis variable.

For example, this graph from the Gapminder (**dataed.in/gapdata**) dataset shows a very clear decreasing pattern in fertility rate in India over the last half century.

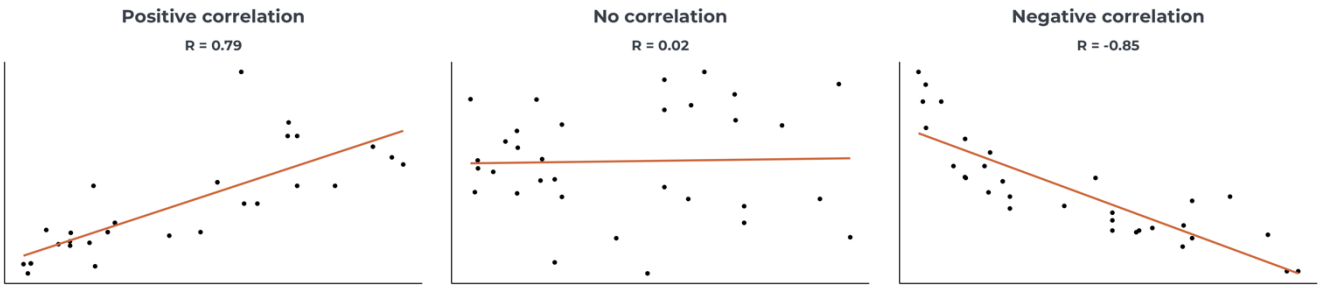


Correlations

Correlation: the relationship between two variables

Sometimes the patterns are less obvious, and it becomes necessary to statistically test for the presence of a relationship. It is then useful to calculate a correlation coefficient.

There are a number of different correlation coefficients, but the most commonly used is Pearson's (**dataed.in/corr**) . This is a number that summarises both the strength and direction of the relationship.



The correlation coefficient (R) takes a value between -1 and 1, with -1 being a perfect negative correlation and +1 being a perfect positive correlation. A value of around 0 implies no correlation or relationship between the variables.

The assumption here is that the relationships are linear, they form a straight line. Relationships can also be non-linear, and it may be necessary to fit a differently shaped function to the data to test for a relationship.

It is important to remember that "correlation does not imply causation". What this means is the change in one variable does not necessarily cause the other variable to change. There may often be a hidden **confounding factor**, which is a variable that related to both the variables being studied, which when changes, causes a change in both the analysis variables.

Extract Insights

Having identified a pattern in the data, this needs to be turned into an **actionable insight**. This ensures that the knowledge of the data showing a pattern provides information that can be used to solve a problem.

Before sharing this knowledge, the insight will need to be tested to ensure that it holds true for further data sources, or over a different time period. It is also important to review the analysis process carefully at this stage to ensure no mistakes have been made in the manipulation. Finally, it is necessary to ensure that the pattern isn't just a consequence of an **underlying bias** in the original data. Only at this point can the insight be used to make a decision, drive a new process, or prompt a different action.

VISUALISATION AND STORYTELLING

Data visualisation: the graphical representation of data

Data is visualised to allow the human brain to spot patterns and identify insights easily and quickly. Data visualisation is both an art and a science. It involves both the manipulation of the data into the correct shape for display alongside an understanding of the use of colour, how the brain consumes visual information and effective design choices. This should then all come together into something visually pleasing that tells a story. This requires both patience and practice.

CHOOSING THE RIGHT GRAPH

Before setting out to create a visualisation it is important to plan. This involves making sure the output addresses the question being asked and presents back the information clearly.

Not all chart types are suitable for all types of data, or the questions being addressed. There are six main types, that can address different questions:

Comparison
Comparing and sorting data points, e.g. bar chart

Composition
Part-to-whole comparisons, e.g. pie chart

Distribution
Comparison of data points along an axis, e.g. histogram

Relationship
Patterns between two or more variables, e.g. line

Evolution
Patterns over time, e.g. scatterplot

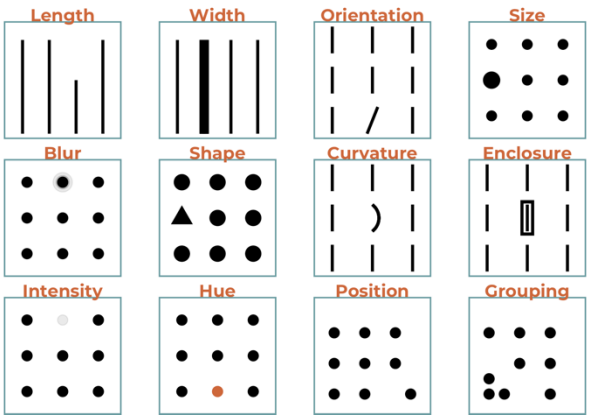
Profiling
Pattern comparison, e.g. grouped bar chart

There are many useful tools which can be used to narrow down the choice of chart to use. A summary of them is given here: **dataed.in/cool**. Some posters and printouts that can be downloaded for reference are:

- data-to-viz (**dataed.in/dtv**)
- Chart.Guide (**dataed.in/cg**)
- Qualitative chart chooser (**dataed.in/cc3**)
- FT Visual Vocabulary (**dataed.in/ftvv**)

VISUAL PERCEPTION

Our brains are particularly good at spotting a number of specific patterns. This means that we process them subconsciously and patterns just “pop-out”. This is called **pre-attentive processing**. By designing graphs that make use of these visual features it ensures that they can be processed at a glance.



The first eight are all ones that make use of **form** or how shapes appear. The next two use **colour**, and finally the last two use the **position** of the marks relative to each other. There is actually a final visual feature that can be processed subconsciously, this is **motion**, but this is not often used for graphics as it can be distracting to have a moving mark on a graph.

Different graph types **encode** data using combinations of these different visual features. For example, a scatterplot uses position, a bubble plot uses size, heatmaps use intensity and a bar chart uses length. Often multiple visual features can be combined together, for example when a highlight colour, shape or enclosure is used to bring attention to a specific data point.

Graphical Best Practices

Use of colour

The different ways to use colour effectively in a data visualisation are:

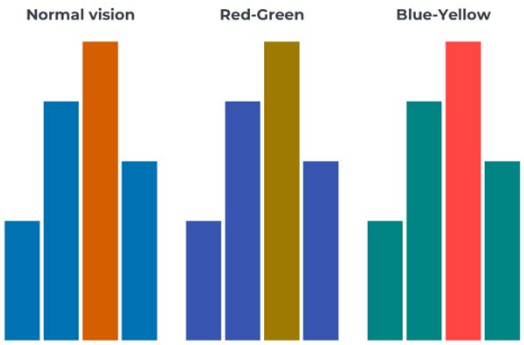
Sequential	Diverging	Categorical	Highlight	Alerting
Single colour scale from light to dark e.g. heatmaps	Two colours with a neutral midpoint e.g. positive / negative value	A different colour per category e.g. line charts, or stacked bar charts	A highlight colour against a neutral colour to bring attention to a specific point	Use of a colour code to show whether on target or not.

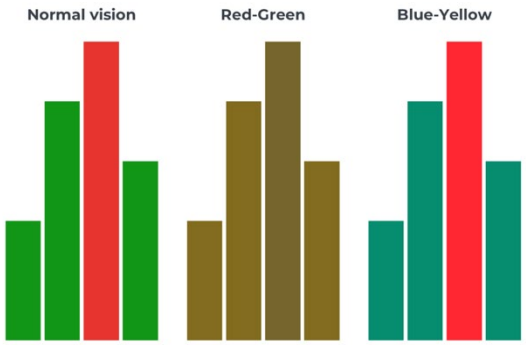
It is best to aim to use no more than three colours on a single graph, otherwise the brain needs to revert to **attentive** processing and the patterns cannot be spotted quickly.

Colour-blindness

Care should be taken with colour on a graph. It can be very effective if used correctly. However, not everyone's perception of colour is identical, so consideration should be given to the 8% of males and 0.5% of females who are colour-blind. Colour-blindness is the decreased ability to differences in particular colours. The majority of people with colour-blindness are red-green colour-blind. The blue-yellow form is much rarer.

To the right is an example of how a red and green highlighted bar chart would look for different types of colour-blindness:





To the left a colour-blind friendly palette has been used. These often use combinations of blue and orange and can ensure accessibility for all types of colour-blindness. To test out how a chart would look for a person with colour-blindness, images can be uploaded to tools such as Coblis (**dataed.in/coblis**).

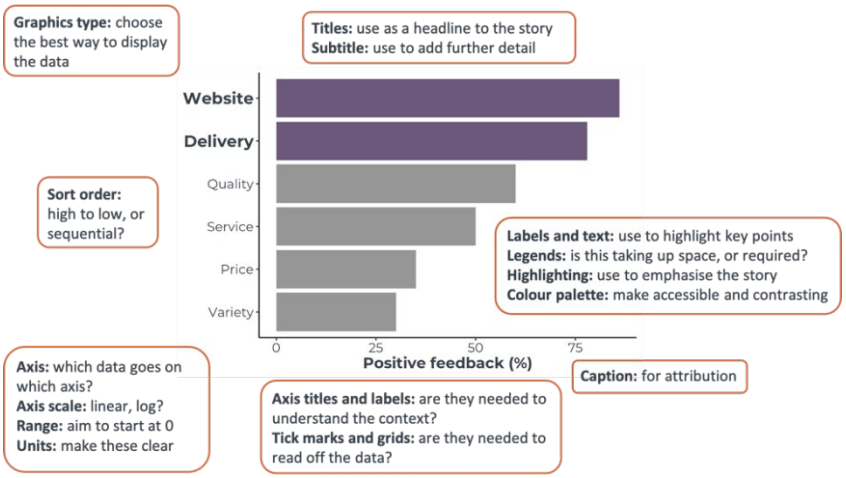
Graphical design considerations

There are many little design decisions to be made when creating a data visualisation.

Here is just a subset of the common design decisions that will be required:

Start by designing on paper first, before any data manipulation, since different graphics types will require differently shaped data.

As a minimum, all graphs should have a title and are labelled sufficiently for the audience to understand them without need for further information.



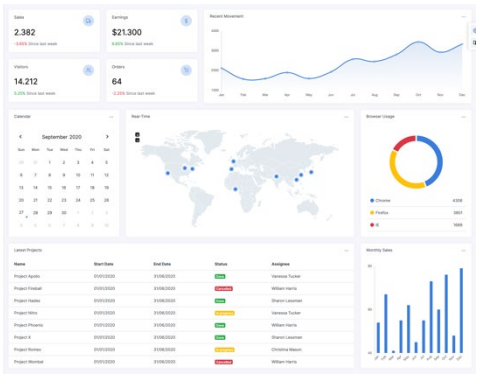
Dashboards

Dashboard: A **visual display** of the **most important information** needed to **achieve one or more objectives** that has been consolidated on a **single computer screen** so that it can be **monitored at a glance**.

Stephen Few – March 2004

A dashboard is a special type of data visualisation that combines graphical information and metrics from various sources together. Similarly, to the car dashboard, where it derives its name, a data dashboard should allow the user to understand what is happening from just a quick glance. This means the information should all fit within a single eye span, or a single computer screen.

Types of dashboard



Dashboards are designed for a specific purpose and will contain the right information for that purpose. There are generally three different types of dashboard in use:

- **Strategic**: this is normally used by the CEO and board of a company. It will contain the key metrics monitored by the business and show progress against these.
- **Operational**: this is generally focused on a specific area or process within a business. Sometimes they may display real-time information on a screen, making information freely available to all. Within a call centre this could demonstrate call waiting times, whereas within a factory, this might show orders or stock levels.
- **Analytical**: this type of dashboard is often interactive, with some filtering capability to dig into the data without the need for a data analyst. They allow domain experts to understand **why** something is happening.

Metrics and KPIs

Dashboards contain a mixture of visual graphics and raw numbers. These raw numbers are referred to as metrics.

Metric: a quantifiable or measurable value

Key Performance Indicator (KPI):
a strategic metric that supports a business goal

Targets, thresholds and actions

For metrics to be useful on a dashboard they need to be given in context, so the user understands whether the raw number is within acceptable limits. Therefore, metrics must have targets and thresholds attached.

A **target** is the goal that is being aimed for.
A **threshold** defines the normal range of the metric, and often has predefined colours when the metric is outside these limits.
A **predefined action** is what will happen when a metric breaches its thresholds.

Examples of KPIs

KPIs are always aligned to business goals. Every business will measure different things depending on their priorities.

Examples of common KPIs include:

- Total number of customers
- Website visitors in the last month
- Absentee rate
- Average delivery time
- Customer satisfaction rate
- Sales revenue in the last month

Communicating Data and Data Analysis

Communicating data analysis is at least as important as the analysis itself. The purposes for communicating data will include:

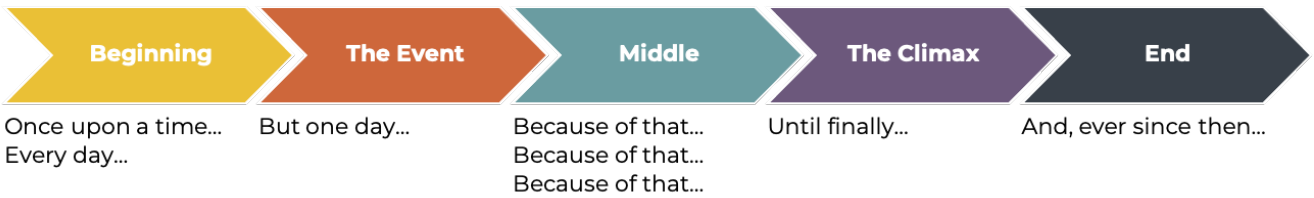
Explain To clearly explain the analysis that was carried	**Clarify** To clarify any assumptions made	**Common Understanding** To ensure a common understanding of what the data is showing
Simplify To simplify what is being shown	**Focus** To focus the audience on the most important areas	**Drive action** To persuade the audience to act on the insights, leading to a positive action

The purpose of communicating data should be kept in mind throughout the analysis and visualisation processes. If the analysis isn't communicated clearly to someone who can act on the insight, then it is likely that no action with be taken and the value of doing the analysis will have been lost.

Storytelling with data

The term **storytelling** is often used in conjunction with communicating data. The reason for this is that data visualisation is not really about data and numbers, but it's about data being used to tell a story. The reason stories are used is that they are engaging and keep the audience's attention. A story is a structured form of communication, also helps with the audience's understanding, building an emotional connection that drives action.

Stories all follow a similar structure, which is summarised in The Story Spine (**dataed.in/story**) by Kenn Adams.



For a data story the structure is similar, but the functions are slightly different. The **beginning** should include the problem and the context behind it. The **event** is the analysis undertaken, the assumptions and the data used. The **middle** is the relationships between the variables and the interpretations of those relationships. The **climax** is the implications of these results. Finally, the **end** includes the recommendations and next steps.

When telling the story, it's important to use compelling graphics that have been designed to communicate the insight clearly. Make sure that the language used is the language of the audience, not analytics or data. Focus on what is most important to the audience to keep them engaged.

Data Communication Checklist

- **Problem**: tell the story of the problem, not the story of the analysis exercise.
- **Language**: don't baffle the audience with technical terms. Tell the story in simple, clear language.
- **Visualisations**: only include visualisations that add to the story. Visualisations that were created to support data understanding may not be important to the story.
- **Metrics**: only present metrics in context, so that the audience is able to understand whether what they show is good or bad.
- **Problem**: revisit the problem to ensure that all aspects of it have been fully addressed and the next steps are clear.

DATA QUALITY AND MANANAGEMENT

Data Quality

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.

Impacts of poor data quality

- Analysis rework
- Organisational inefficiencies
- Customer dissatisfaction
- Opportunity cost of missed sales
- Reputational costs from loss of trust
- Compliance costs/fines from incorrect reporting

Quality data: the degree to which data is fit for its intended purpose

Benefits of high-quality data

- Improved customer experience
- Reduced risk
- Competitive advantage from accurate insights
- Increased revenue
- Higher staff productivity

Measuring Data Quality

Quality data does not mean perfect data. High quality data is data that is good enough for the activity it is being used for.

There are many ways data can be incorrect. It could be missing, out of date, duplicated, in the wrong format, doesn't match the same information elsewhere, or just plain wrong. The different ways that data can be incorrect are organised into six different areas called the **Dimensions of Data Quality**. It is important to check quality against as many of the different dimensions as necessary when reporting data quality.

- **Completeness**: how non-blank or populated the data is
- **Uniqueness**: data is not recorded more than once - identifies duplicates
- **Timeliness**: how up to date the data is
- **Validity**: that data is in the correct format, type and range
- **Accuracy**: how data represents the real-world
- **Consistency**: data matches if two copies of the same information are compared

Data Classification

There are special classifications of data, which mean they should be managed differently:

Metadata

This is data about data. A data dictionary is one of the most important pieces of metadata. Without metadata it would be very difficult to find and work with any dataset.

Reference Data

This is data that is referenced, or used by other data sources, such as a lookup table, or a list of possible values. By tightly controlling the allowed values, it can reduce data quality issues

Master Data

This is data that is also referenced by other datasets, such as customer numbers, or email addresses. Since this data is used to link all the other dataset together it needs to be as accurate and up to date as possible.

Caring for Data

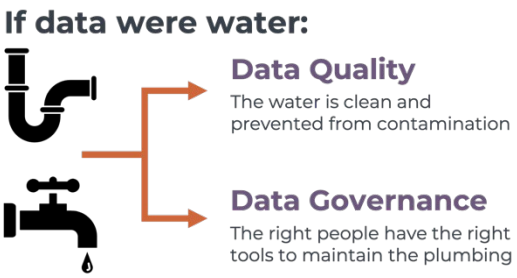
Taking care of data is a difficult task, even when data quality rules are put in place. This is where data management comes in. It involves the care of data throughout its lifecycle, to ensure it can be treated as an asset, like any item of value.

Data Management: the activities involved in treating data as an asset through its entire lifecycle from creation to deletion

Data Governance: the practices and processes which ensure the formal management of data as an asset

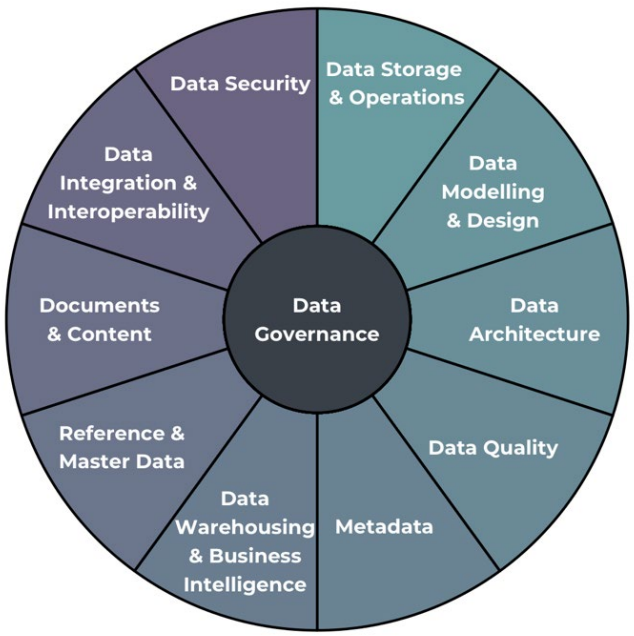
Data management and governance aren't quite the same. Data governance is the glue that holds all the data management processes together. An analogy is given below: The benefits of treating data as an asset include:

- **Increased data security**: access is limited only to those that need it
- **Reduced data loss**: from system failures
- **Improved productivity**: enables quick access to the correct information
- **Increased cost efficiencies**: from avoiding the need to manually fix information
- **Ensures accurate decisions**: accurate analysis leads to better insights and decisions



Data management areas

There are many different activities that fall under data management. DAMA (the International Association for Data Management) break it down into the following areas:



DAMA-DMBOK2 Data Management Framework

- **Data architecture**: the standards for how data is collected, stored and used.
- **Data modelling and design**: a map of how the data relates to the real-world
- **Data storage and operations**: the lifecycle of storage including backing up, archiving and deleting.
- **Data security**: protecting the data from unauthorised access or loss
- **Data integration and interoperability**: how data moves between different systems
- **Documents and content**: caring for data in an unstructured form
- **Reference and master data**: looking after some of the most critical datasets
- **Data warehouse and business intelligence**: managing the data used for reporting
- **Metadata**: understanding about what data is available
- **Data quality**: defining rules to check for and then fix issues with quality
- **Data governance**: the policies, processes and responsibilities that encompass treating data as an asset.

ETHICS AND BIAS

Ethics

The incorrect use of data has the power to cause much harm to individuals and organisations. As the amount of data being created, collected, shared and acted upon grows daily, existing regulations and governance frameworks are not able to keep up with the new risks that are appearing. This could be through the unethical or illegal use of insights which amplify existing social and economic biases, or by using data for purposes that it was not originally collected for without the individual's consent or knowledge. Data ethics covers all of these areas.

Data Ethics: a branch of ethics that evaluates data practices with the potential to adversely impact on people and society – in data collection, sharing and use

Open Data Institute (**theodi.org**)

Data Ethics: the responsible and sustainable use of data. It is about doing the right thing for people and society. Data processes should be designed as sustainable solutions benefitting first and foremost humans

(**DataEthics.eu**)

Ethical risks

The types of problems and risks brought about by the unethical use of data and technology are wide and varied. When assessing risks, it is important to consider risks across all the possible categories.

Truth

Could be used to undermine the truth, spread disinformation or propaganda.

Health

Could affect the mental, physical or social health of users.

e.g. addiction or attention maximisation

Equality

Could amplify existing economic inequalities. Limits access to employment, knowledge or services.

Fairness

Could amplify or reinforce existing biases from models built on datasets containing historical bias

Human rights

Could infringe the human rights of citizens. Through surveillance or targeting of individuals.

Privacy

Personal data could be disclosed without individual's knowledge and control.

Trust and transparency

Users may not be made aware how their data is being used. Trust could become eroded.

Criminality

Could be used for illegal activities, even if it were not originally designed as such.

Ethical Frameworks

Ethical frameworks are a structured approach to ensuring data is used responsibly and all risks are considered:

- **Ethical OS** (**ethicalos.org**): a structured approach to identifying risks
- **MI Garage** - Ethics Framework (**dataed.in/mig**): a comprehensive list of questions for developing a product or service
- **The ODI Data Ethics Canvas** (**dataed.in/dec**): a tool to use to identify issues in a project
- **Deon** (**dataed.in/deon**): a checklist for data scientists to use within their python projects

Examples of ethical challenges

There are many examples where the initial objective was noble, such as to save lives. However, retrospectively additional consideration at the start would have brought the ethical risks to light more quickly. Examples include:

- Facial recognition for policing (**dataed.in/fac**)
- Algorithms for exams (**dataed.in/exam**)
- Samaritans twitter bot (**dataed.in/sam**)
- Homeless database (**dataed.in/home**)
- Medical chatbots (**dataed.in/chat**)

Bias

When bias is discussed it can mean three related, but slightly different things:

Statistical bias: the estimated value of the results is systematically different from the true underlying parameter

Data or sampling bias: the available data is not representative of the population or phenomenon of study

Algorithmic bias: a computer system reflects the implicit values of the humans who created it

When data is biased it leads to an ethical challenge, in particular fairness and equality risks. Data scientists need to be aware of the potential for bias, stay vigilant and do everything they can to minimise it in their models and analysis.

Causes of Bias

Sample bias

The dataset used for model development does not represent the population the model will be used on.

Exclusion bias

Predictive variables are removed prior to the modelling process.

Measurement bias

Arises from a systematic issue with the collection of the data, often through calibration or or faulty detectors.

Confirmation bias

Arises from cherry-picking data or variables that confirm a human's existing beliefs and expectations.

Stereotype bias

Cultural or gender stereotypes may lead to an uneven distribution of modelling data.

Survivorship bias

Arises from concentrating on the successful output of a process and ignoring those that don't survive.

Simpson's paradox

A statistical phenomenon that leads to incorrect inferences when trends for subgroups reverse the overall norm.

Correlation bias

Arises from inferring a correlation where it doesn't exist due to confounding variables that have not been accounted for.

As can be seen from the variety of bias causes, they can sneak into an analysis at every stage. It is important to question all output results and if it doesn't make intuitive sense then to carry out additional validation on an additional dataset or look for additional variables that might be missing or hidden.

Impacts of Bias

There are many examples of bias in machine learning algorithms that have led to a negative outcome for those for which the biased algorithms have been used upon. The impacts vary from perpetuating biases in recruitment processes, to an increased chance of being arrested, to requiring unnecessary surgery, to missing out on a place at university, as was the case in the UK, when an algorithm was used to predict exam grades.

Mitigating Bias

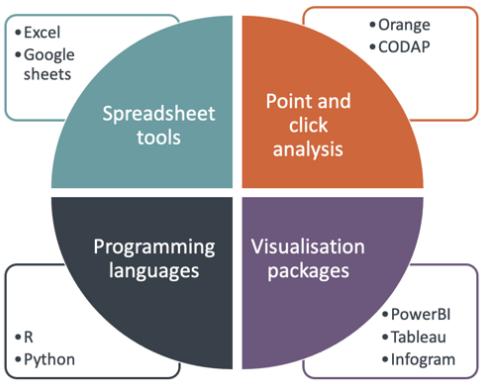
The best way to avoid bias in data science is to have a diverse team who each bring a different perspective to the problem:

- Have multiple people with differing backgrounds involved in the project
- Review interpretations and conclusions with team members to identify gaps in reasoning
- Review the findings with subject matter experts who understand the data and how it was captured
- Include multiple data sources from as wide a range of providers as possible
- Look for simple interpretations to the findings, such as data quality issues and rule these out before moving on to more complex interpretations

Transparency in the model development approach and dataset used, alongside extensive testing and validation will also mitigate bias.
Finally, the ability to put a "human in the loop" for critical decisions should always be available.

TOOLS AND LANGUAGES

There are a number of different tools and languages that can be used for data analysis and visualisation. The choice of tool to use will depend on a number of different factors such as access to the internet, requirements to keep data private, the need for version control, time available to learn to use the tool and the programming experience of the user. Some tools can be used for analysis and visualisation, whilst others specialise in visualisation only.

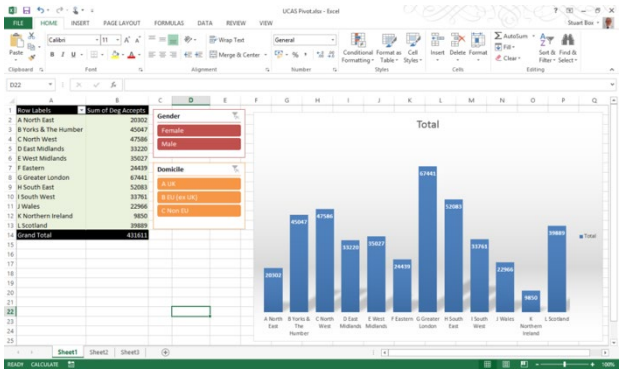


GRAPHICAL TOOL OPTIONS

Spreadsheets

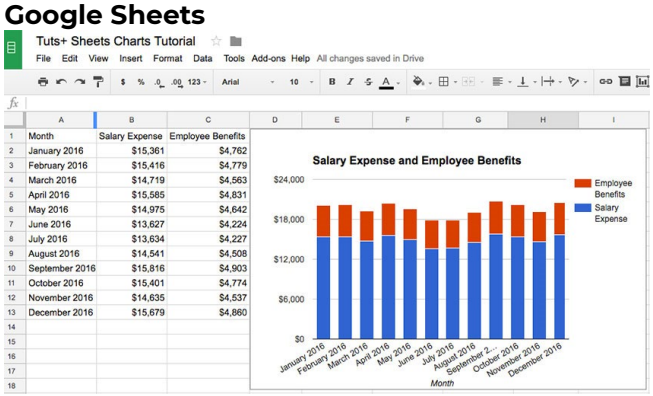
Spreadsheet tools are suitable for quick ad-hoc data analysis and visualisation. The two most popular tools are Microsoft's **Excel** and **Google Sheets**. Both tools have scripting languages which can be used to automate repetitive tasks. Challenges with all spreadsheet tools include auditability,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

Excel



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

Google Sheets



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

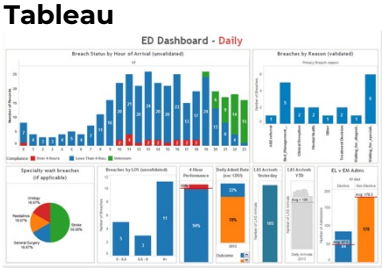
Commercial Visualisation Packages

Although these tools are commercial, they all offer educational licences. They all focus on visualisation, rather than analysis, requiring the manipulation and reshaping of datasets to be carried out externally to the tool. Visualisations can be easily combined into interactive dashboards and most can easily connect with external data sources.

Power BI

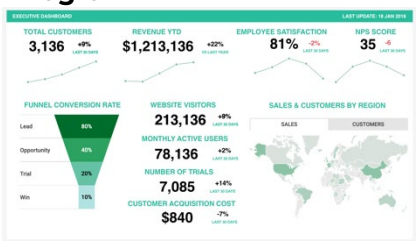
- **powerbi.microsoft.com**
- Part of the office suite
- Easily links to excel for data preparation

Tableau



- **tableau.com**
- Good array of training materials available

Infogram

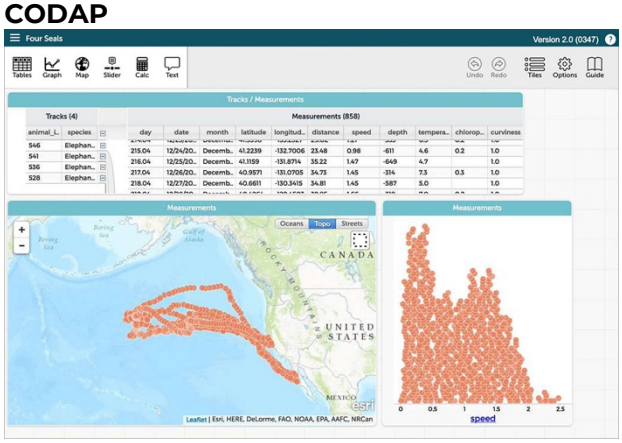


- **infogram.com**
- Basic version is free but with limited functionality

Point and Click

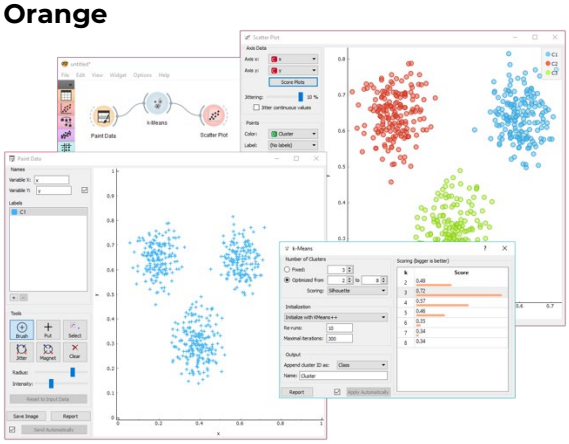
These tools allow analysis and visualisation to be carried out interactively. The examples below are open source and free to use. There are commercial options with educational licences.

CODAP



- **codap.concord.org**
- Designed for educational use online
- Intuitive data exploration
- Datasets are not secure
- No version control

Orange

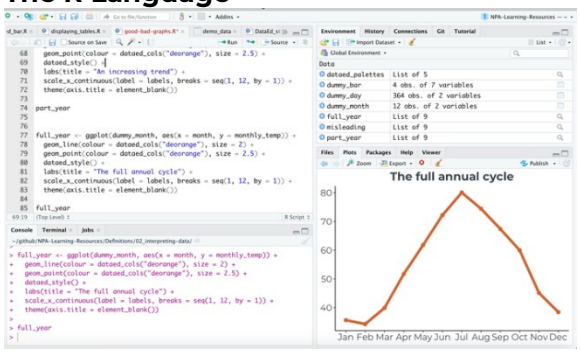


- **orange.biolab.si**
- Must be downloaded locally
- Takes time to become proficient
- Enables advanced analysis such as predictive modelling

Programming languages

The advantages of using a programming language for analysis and visualisation are both the flexibility and reproducibility. The main open source languages that data scientists use are **R** and **Python**, however there are also commercial languages available too. R and Python both have special data types designed for manipulating tidy tabular data. They all have multiple packages for creating visualisations and more complex dashboards that can be implemented in web applications. Both languages have IDEs (interactive development environments) for easy development and integration with source control tools such as **git** (**github.com**). Alternatives to using an IDE include interactive notebooks, which can be used in either language. Popular notebooks include Jupyter (**jupyter.org**) or web-based Google Colab (**dataed.in/colab**)

The R Language



- Language designed for statistical analysis and visualisation
- Has the RStudio development environment
- Popular packages:
  - **Tidyverse** for data manipulation
  - **ggplot** for visualisations
  - **RShiny** for interactive web-based applications

Python

- A full object-oriented programming language
- A recommended IDE to use is Spyder, available as part of the Anaconda suite
- Popular packages for data science:
  - **Pandas** for data manipulation
  - **Numpy** for data functions
  - **Matplotlib or Seaborn** for visualisation
  - **Plotly** for interactive plotting

This guide is available in other formats.
Download from **teachdata.science**.

Data Education in Schools

SQA