The analysis process

Learning intentions

We will be learning what is involved in the analysis process, specifically,

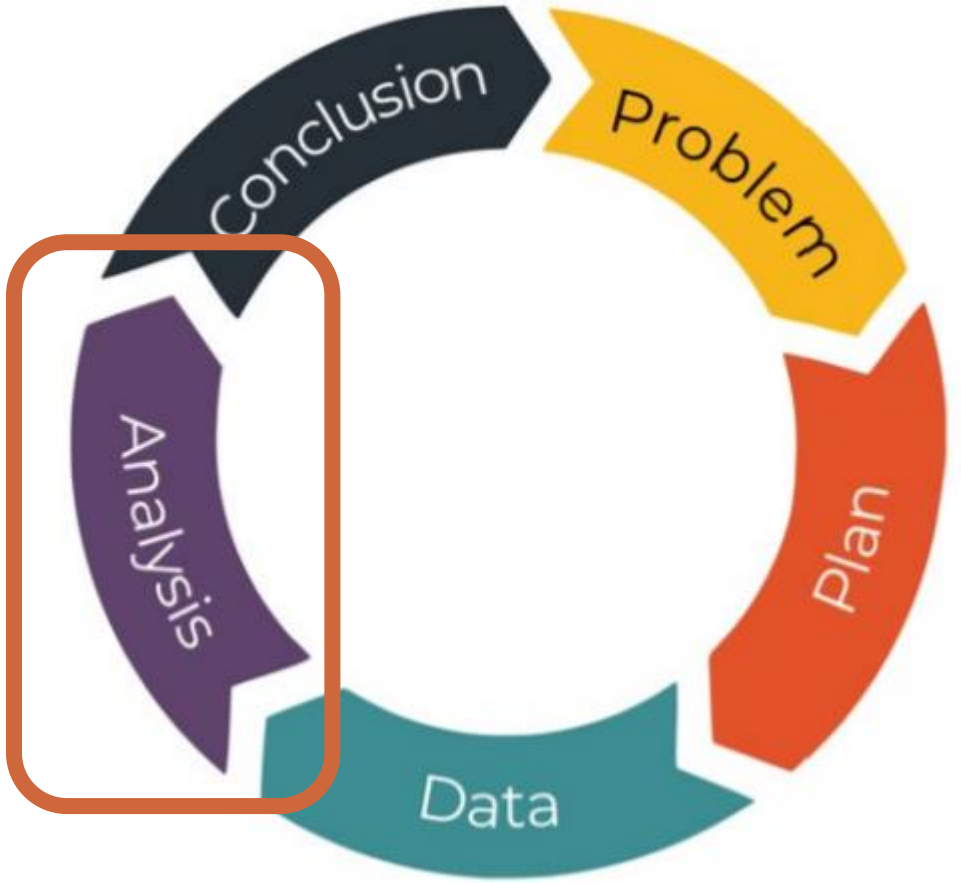
- what we **mean by analysis**
- a structured way of performing analysis (**the analysis steps**)
- how to understand data through **visual inspection**

Background

Data science is about trying to solve a problem.

Making sure that you are analysing the data in the most efficient and accurate way possible to solve the problem is very important.

In this lesson, we will look at what is involved in **analysing data** and **steps you can follow** to complete your analysis.



Definition

Data analysis

The process of examining data to identify patterns and draw conclusions that will inform decision-making

What do we mean by analysis?

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

This could involve any of the following processes:

- Reformatting
- Aggregating
- Cleansing
- Summarising
- Modelling
- Visualising
- Interpreting

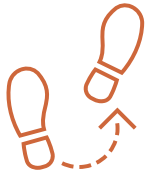
The analysis steps

It can be helpful to breakdown the analysis process into these 5 steps,

Why use the analysis steps?

Some benefits of following the analysis steps are:

Minimise mistakes



Easier to **reproduce/share work** with others if you use a structured approach

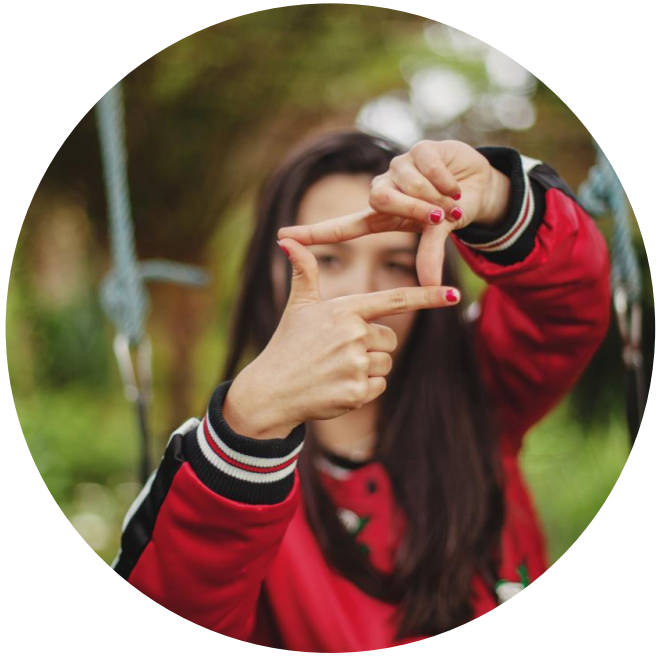
Maximise the confidence of the conclusions or insights extracted from the data

The analysis steps

Data understanding → Data tidying and cleansing → Data manipulation → Identifying patterns → Extracting insights

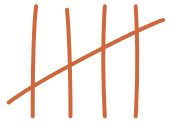
You are looking to

- get a **feel** for the data
- understand its **size** and **shape**
- identify any **obvious issues** with the data that may need to be addressed.



Why data understanding is important?

Makes sure the calculations you are planning to use are suitable for different **data types**



Checks that the data has the **number of rows and columns** you are expecting

Helps you **identify any missing values**

Show me….data understanding

In this dataset we can see there are mixture of **quantitative** and **qualitative** data items that would need to be handled differently.

release_yea	rating	duratio	listed_i	description
2020	PG-13	90 min	Document	As her father nears the end of his life, filmmaker Kirst
2021	PG-13	104 min	Comedies,	A woman adjusting to life after a loss contends with a
2021	TV-MA	116 min	Dramas, In	A three-person crew on a mission to Mars faces an im
2021	TV-Y	2 Seasons	Kids' TV	Young koala caretaker Izzy Bee and her family rescue
2013	TV-Y	76 min	Children &	For Rohan and his magical pal, Keymon, a trip to visit
2018	TV-Y7	80 min	Children &	A time machine sends Motu and Patlu back to the din
2017	TV-Y	81 min	Children &	While returning a goldfish and an octopus from an aq
2019	TV-Y	84 min	Children &	For Motu, facing off against three children becomes a
2019	TV-Y7	2 Seasons	Kids' TV	An intergalactic device transforms toy cars into robot
2020	TV-Y7	87 min	Children &	Kid magician Rudra sets out to save Earth from the de

This is a section of a dataset showing the programs available on Netflix https://www.kaggle.com/shivamb/netflix-shows

Show me….data understanding

Some of columns would need to be manipulated before any analysis could be performed.

release_yea	rating	duratio	listed_i	description
2020	PG-13	90 min	Document	As her father nears the end of his life, filmmaker Kirst
2021	PG-13	104 min	Comedies,	A woman adjusting to life after a loss contends with a
2021	TV-MA	116 min	Dramas, In	A three-person crew on a mission to Mars faces an im
2021	TV-Y	2 Seasons	Kids' TV	Young koala caretaker Izzy Bee and her family rescue
2013	TV-Y	76 min	Children &	For Rohan and his magical pal, Keymon, a trip to visit
2018	TV-Y7	80 min	Children &	A time machine sends Motu and Patlu back to the din
2017	TV-Y	81 min	Children &	While returning a goldfish and an octopus from an aq
2019	TV-Y	84 min	Children &	For Motu, facing off against three children becomes a
2019	TV-Y7	2 Seasons	Kids' TV	An intergalactic device transforms toy cars into robot
2020	TV-Y7	87 min	Children &	Kid magician Rudra sets out to save Earth from the de

rating: categorical variables with fixed sets of valid values

duration: mixture of different time periods and would need to be manipulated to perform any analysis on it

description: free text and difficult to analyse

This is a section of a dataset showing the programs available on Netflix https://www.kaggle.com/shivamb/netflix-shows

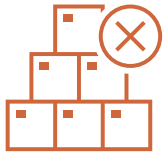
The analysis steps

Data understanding | **Data tidying and cleansing** | Data manipulation | Identifying patterns | Extracting insights

Tidy and clean the data so it's fit for purpose, such as,

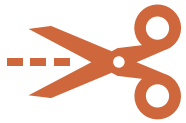
- **Reformatting** data
- Reviewing **missing values** and outliers
- Removing **duplicates**

Why data tidying and cleansing is important?



Keeping any duplicate rows could result in **incorrect answers to calculations**

Reformatting data can make it **easier to view** the data



If you have any **missing values**, you can decide how you are going to handle them before performing any calculations e.g. remove the rows

Show me….data tidying and cleansing

The **date_of_birth** is difficult to understand in this dataset…..

name	date_of_birth
Taylor Swift	32,855
Prince	21,343
Britney Spears	29,922
Madonna	21,413
Lewis Capaldi	35,345
Elvis Presley	12,792

Dataset tidied and cleansed through reformatting

… but now it has been reformatted it is easier to use.

name	date_of_birth
Taylor Swift	13 December 1989
Prince	07 June 1958
Britney Spears	02 December 1981
Madonna	16 August 1958
Lewis Capaldi	07 October 1996
Elvis Presley	08 January 1935

Reminder: dates are stored as the number of days or seconds passed the ‘epoch’ date. Reformatting changes the display format.

The analysis steps

Data understanding → Data tidying and cleansing → **Data manipulation** → Identifying patterns → Extracting insights

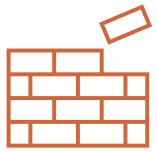
The data manipulation can be done through processes such as,

- **Extracting**, selecting and reordering data
- **Summarising** and grouping data
- Merging or **joining** data

Why data manipulation is important?

Join and merging multiple datasets allows you **bring together data** from different tables into a single dataset

Focusing on only the information you need can **minimise mistakes** and make any **patterns easier to spot**



Creating new variables through extracting or calculations allows you to **understand more about the data**

Your turn…

Can you think of any processes you could use to **manipulate the rows or columns** in a dataset?

For example, you could ,

- **select** (or choose to keep) only the columns you need
- **create a new variable by extracting** data items

Your turn…

Here are some processes you could use to **manipulate the rows or columns** in a dataset.

- Select columns
- Create new variable by calculation
- Subset data items
- Reorder columns
- Reformat columns
- Remove duplicates
- Sort or filter rows
- Creating new variables by extracting
- Creating new variable by combining data items
- Group and summarise

The analysis steps

Data understanding → Data tidying and cleansing → Data manipulation → **Identifying patterns** → Extracting insights

In this stage you look for **relationships** and **patterns** in the data.

This can be done through plotting graphs or using calculations designed to measure relationships.

Why identifying patterns is important?

You can **test ideas for relationships** or patterns in the datasets

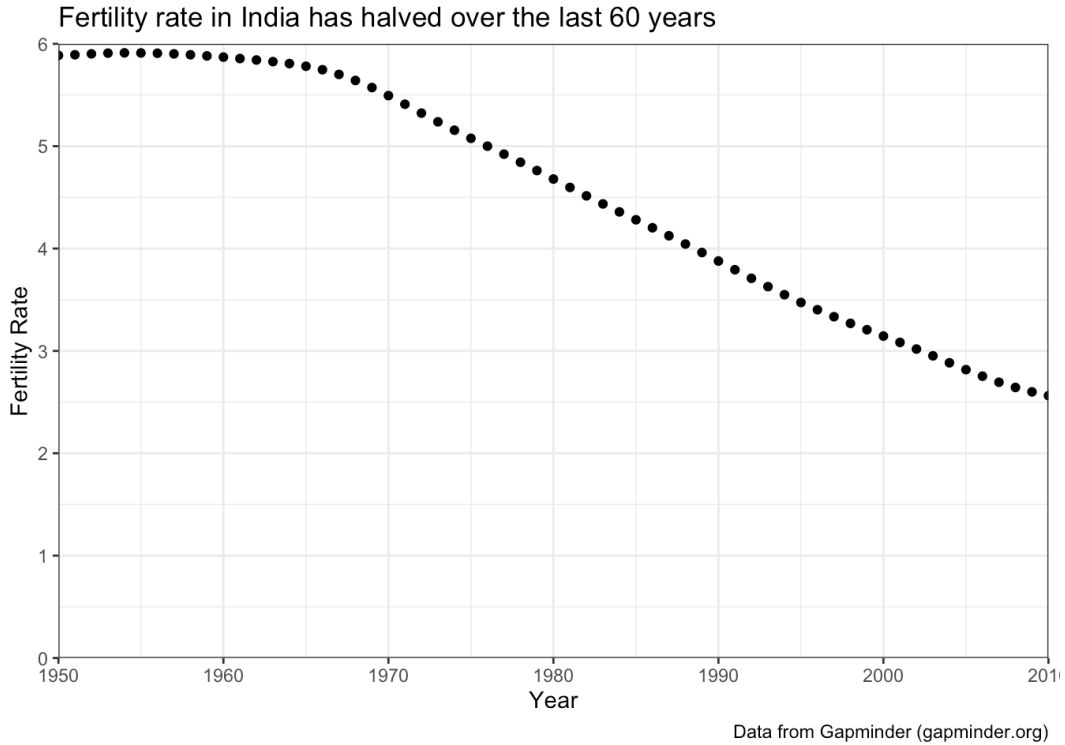
Identify trends in a dataset

Spot unusual data items that might **need more investigation**

Show me….identifying patterns

This graph shows the fertility rate in India over the last 60 years.

This would have been created during the **identifying patterns** of the analysis steps.



The analysis steps

Data understanding → Data tidying and cleansing → Data manipulation → Identifying patterns → **Extracting insights**

At this stage you need to highlight any patterns/relationships that will allow someone to **take action to solve the problem** you are investigating.

All patterns/relationships should be double checked to see if insights are true in other datasets or time frames. Also that no mistakes have been made during any of the analysis stages.

Why extracting insights is important?

Turns **data into actions** that can solve problems

Allows you **check your conclusions in different datasets**/time frames so you are confident in the results

Sense **check your results against how you expect** them to look

Your turn…

Imagine you are given a dataset that looks at the average temperature of the world by year to analyse.

You are told it is ready for the **identifying patterns** analysis step.

However it contains **duplicate rows** as the **data cleansing and tidying step was missed.**

What impact do you think this could have on the results of the analysis?

Your turn…

Some of the issues could be,

- Calculations give the wrong answers
- Incorrect patterns/relationships identified
- Propose the wrong action to solve the problem.

Next steps

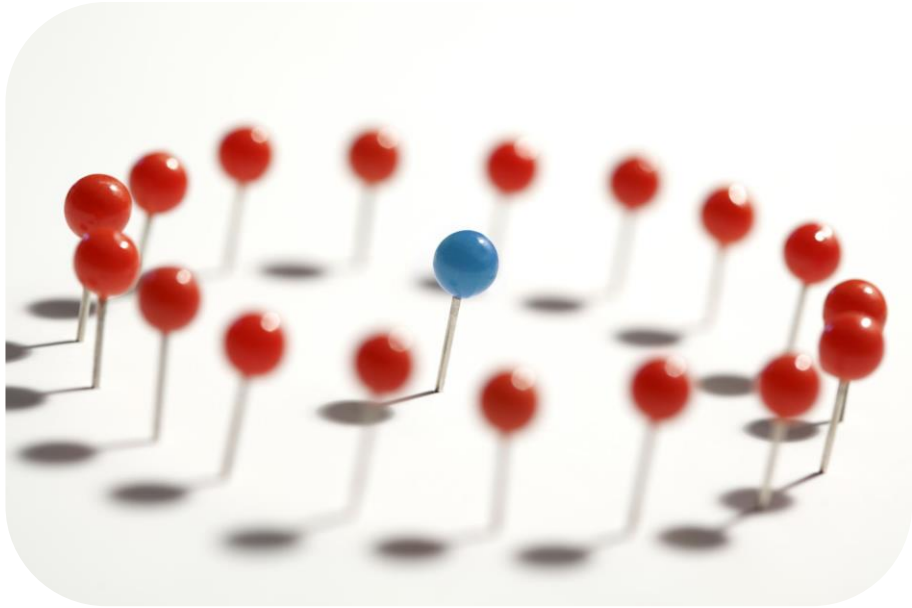
Complete **sections 1 and 2** of the ‘The analysis process’ workbook.

Data understanding

All analysis activities should start with an understanding of the data being analysed.

This involves

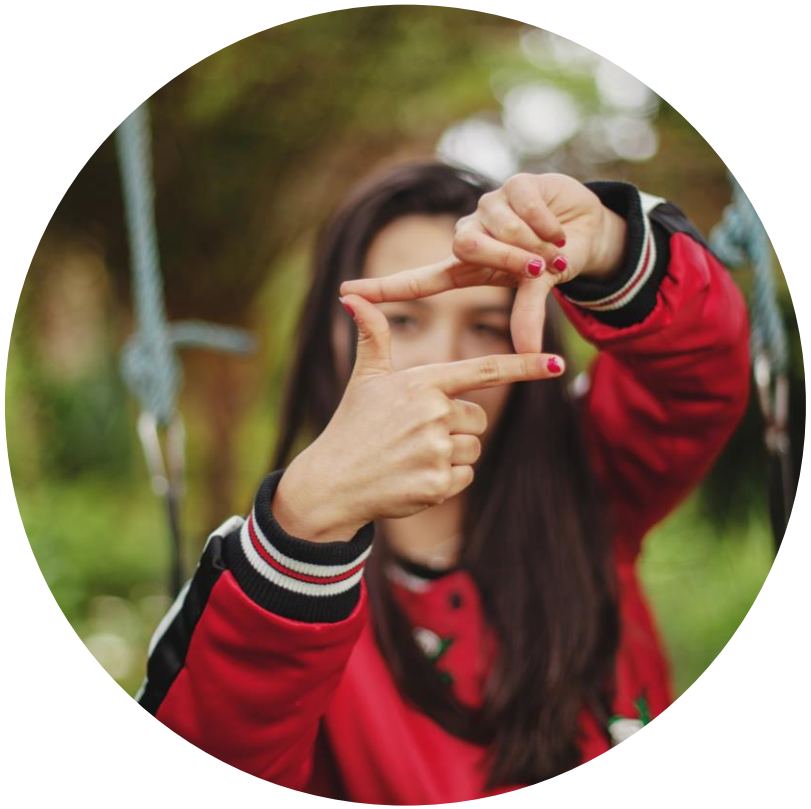
- **Visual inspection** of the dataset
- Reviewing any associated **data dictionaries** for the dataset
- Understanding the **size and shape**
- Identifying any **obvious issues**
- Calculating **summary statistics** (e.g. count, max, mean)



Data understanding through visual inspection

In this lesson we are now going to look at **visual inspection part of data understanding**.

Visual inspection can be done without writing any code or creating new datasets.



Definition

Visual inspection

To get a basic understanding of a dataset by looking at it

Show me…

Through a visual inspection of this dataset we can see,

- Mix of string and numeric data items
- There are missing values (NA) in the **hair_colour** column
- Hair and skin colours can take multiple values.

name	height	mass	hair_colour	skin_colour	gender	homeworld	species
Luke Skywalker	172	77	blond	fair	masculine	Tatooine	Human
C-3PO	167	75	NA	gold	masculine	Tatooine	Droid
R2-D2	96	32	NA	white, blue	masculine	Naboo	Droid
Darth Vader	202	136	none	white	masculine	Tatooine	Human
Leia Organa	150	49	brown	light	feminine	Alderaan	Human
Owen Lars	178	120	brown, grey	light	masculine	Tatooine	Human
Beru Whitesun lars	165	75	brown	light	feminine	Tatooine	Human
R5-D4	97	32	NA	white, red	masculine	Tatooine	Droid
Biggs Darklighter	183	84	black	light	masculine	Tatooine	Human
Obi-Wan Kenobi	182	77	auburn, white	fair	masculine	Stewjon	Human
Anakin Skywalker	188	84	blond	fair	masculine	Tatooine	Human
Wilhuff Tarkin	180	NA	auburn, grey	fair	masculine	Eriadu	Human
Chewbacca	228	112	brown	unknown	masculine	Kashyyyk	Wookiee
Han Solo	180	80	brown	fair	masculine	Corellia	Human
Greedo	173	74	NA	green	masculine	Rodia	Rodian
Jabba Desilijic Tiure	175	1358	NA	green-tan, brown	masculine	Nal Hutta	Hutt
Wedge Antilles	170	77	brown	fair	masculine	Corellia	Human
Jek Tono Porkins	180	110	brown	fair	masculine	Bestine IV	Human
Yoda	66	17	white	green	masculine	NA	Yoda's species
Palpatine	170	75	grey	pale	masculine	Naboo	Human

Sample of information about Star Wars© characters from SWAPI (www.swapi.tech)

Show me…

Through a visual inspection of this dataset we can see,

- Mix of **string** and **numeric** data items
- There are **duplicate rows** that would need to be handled in the data tidying and cleansing step.

name	year_birth	also_known_as
Macbeth	1040	The Red King
William I	1165	The Rough
Malcolm III	1058	Great Chief
David I	1124	The Saint
William I	1165	The Rough
Malcolm III	1058	Great Chief
David I	1124	The Saint

Your turn…

By looking at this dataset, what can you see about the,

- Data types?
- Any missing values?
- Any duplicate rows?

location	temperature	dawn	dusk
Edinburgh	15	03:30	22:50
Paris	21	05:09	22:28
Sydney	17	06:25	17:21
New York	NA	04:54	20:55
Edinburgh	15	03:30	22:50

Your turn…

Data types

- Mix of string, numeric and dates

Any missing values

- There is a missing **temperature** for New York

Any duplicate rows

- The row containing Edinburgh is in the dataset twice

location	temperature	dawn	dusk
Edinburgh	15	03:30	22:50
Paris	21	05:09	22:28
Sydney	17	06:25	17:21
New York	NA	04:54	20:55
Edinburgh	15	03:30	22:50

Next steps

Complete **section 3 and 4** of the ‘The analysis process’ workbook.

Learning checklist

I can *describe* what is meant by analysis

I can *describe/explain* a structured way of performing analysis (the analysis steps)

I can *describe* what it means to visually inspect data as part of the data understanding step

I can *perform* a visual inspection of a simple dataset

How you can use this lesson

You are free to:

- **Share** – copy and redistribute the material in any medium or format
- **Adapt** – remix, transform and build upon the material

Under the following terms:

- **Attribution** — You must give appropriate credit, provide a link to the license, and indicate if changes were made. You may do so in any reasonable manner, but not in any way that suggests the licensor endorses you or your use.
- **NonCommercial** — You may not use the material for commercial purposes.
- **ShareAlike** — If you remix, transform, or build upon the material, you must distribute your contributions under the same license as the original.

Created by Effini in partnership with Data Education in Schools, The Data Lab and Data Skills for Work, with funding from the Scottish Government.

Effini

THE DATA LAB
value from data

Data Skills FOR WORK

Data Education in Schools