Dataset understanding in Excel

Version: 1.0

Effini

THE DATA LAB
value from data

Data Skills FOR WORK

Data Education in Schools

Learning intentions

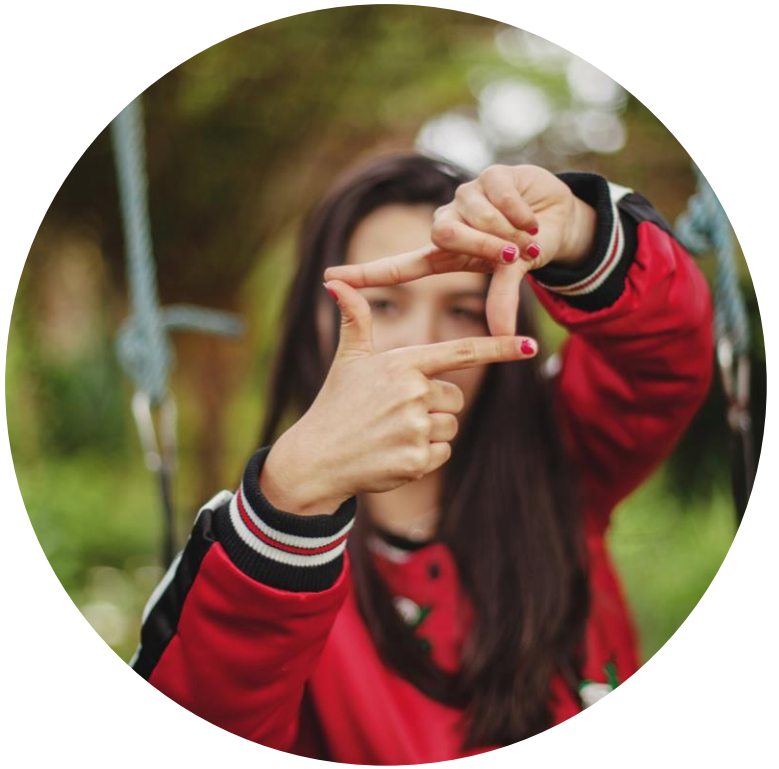
We will be learning about the **data understanding** part of the analysis process, specifically,

- what is **metadata** and the importance of a **data dictionary**
- how to identify the **shape and size** of a dataset in Excel and **data types** of variables
- how to **identify missing values and outliers** in Excel

Background

By making sure that your analysis is conducted in a structured and efficient way it will help you reduce chances of making a mistake and maximise confidence in your results.

In this lesson we will look in detail at the **data understanding** stage of the analysis steps.

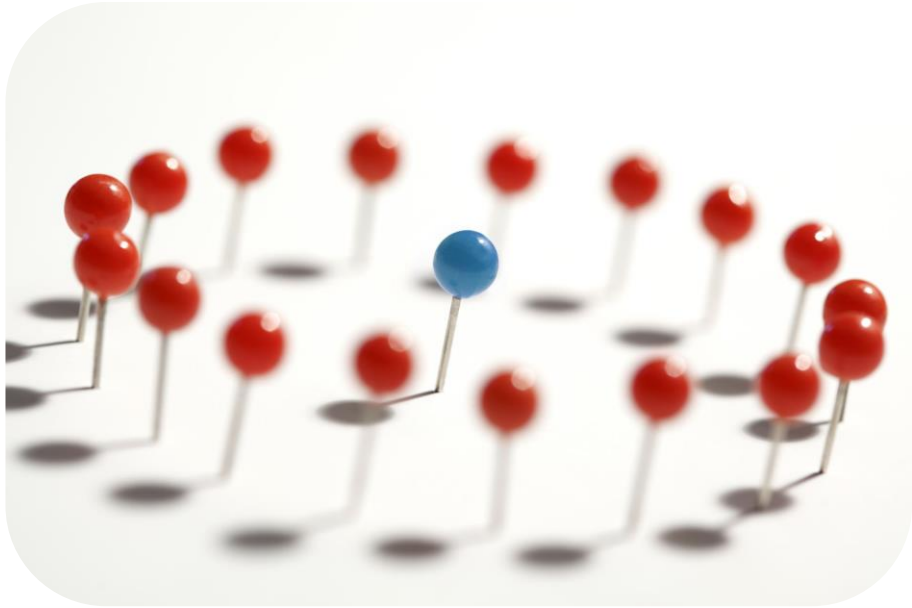


What is dataset understanding?

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

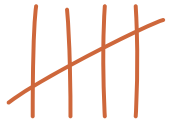
This involves

- **Visual inspection** of the dataset
- Reviewing any associated **data dictionaries** for the dataset
- Understanding the **size, shape** and **data types**
- Identifying any **missing or outlying values**



Why data understanding is important?

Helps you get a sense of whether the **dataset contains data that will help** you answer your question



Checks that the **data looks similar to what you are expecting** e.g. roughly the right number of rows and columns

Helps you **identify any outlying or missing values**

Definition

Metadata

Data about the data

Metadata

Metadata provides information about the data, it can describe,

- content
- structure
- owners and management of the dataset
- when the data might be updated

Show me…

The **metadata** for a photograph would be,

- Date and time of when the photo was taken
- Details of the camera settings
- Geotagging (where the photo was taken)

Show me…

This is an example of the metadata that is provided for datasets on Kaggle (www.kaggle.com). It provides information on the owner, when it was created and how you can use it.

Metadata

Feedback

Usage Information	License	CC0: Public Domain
	Visibility	Public
Provenance	Sources	https://worldhappiness.report/ed/2021/
	Collection methodology	Collected mortality data across the globe from the World Happiness Report 2021 given in the above citation.
Maintainers	Dataset owner	John Harshith
Updates	Expected update frequency	Never
	Last updated	2021-11-29
	Date created	2021-11-24
	Current version	Version 2

Show me…

This is an example of the metadata from a report that has been exported to a csv file.

Rows 1 to 7 contain the **metadata** (the data about the data).

Rows 9 onwards contain the dataset.

	A	B	C
1	ID	5412584	
2	Date/Time Generated	11/11/2021 15:45	
3	Account	145DCS	
4	Account owner	Unknown	
5	Period	2020-01-01 to 2020-12-31	
6	Report Type	Subset	
7	Report Name	Annual product report	
8			
9	Product	ProductID	Stock
10	Black chair	14514	4
11	Blue table	14444	45
12	Blue table extended	75475	1
13	White chair large	14641	87
14	White chair small	44457	3

Removing metadata best practice

Before analysing a dataset that has metadata at the top of the file, it is best practice to remove the rows.

The rows containing the metadata should be saved separately to the dataset itself.

Why is metadata important?

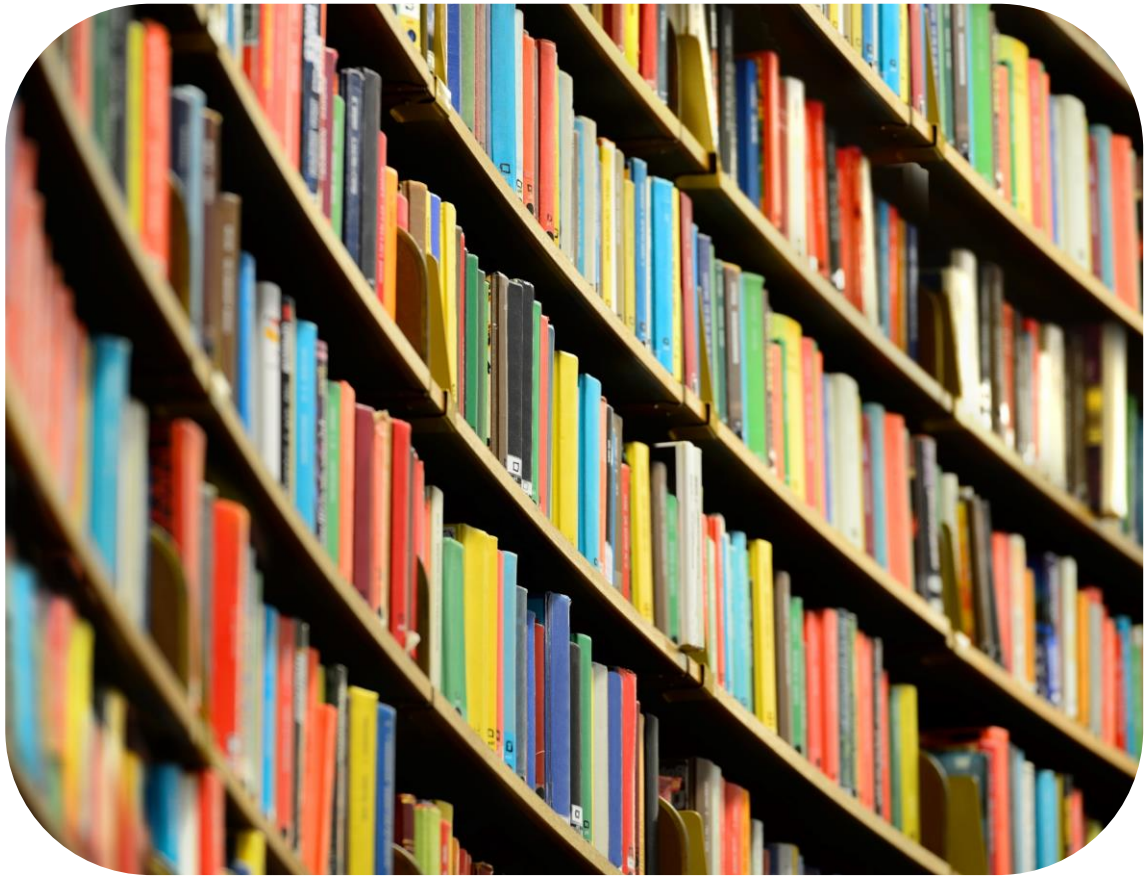
Makes it **easier to find** relevant information and helps make sure the data is being **used correctly**

Helps you to **preserve and re-use** data in the future

Can **prevent incorrect access** and use whilst providing oversight and control over all the data

Your turn…

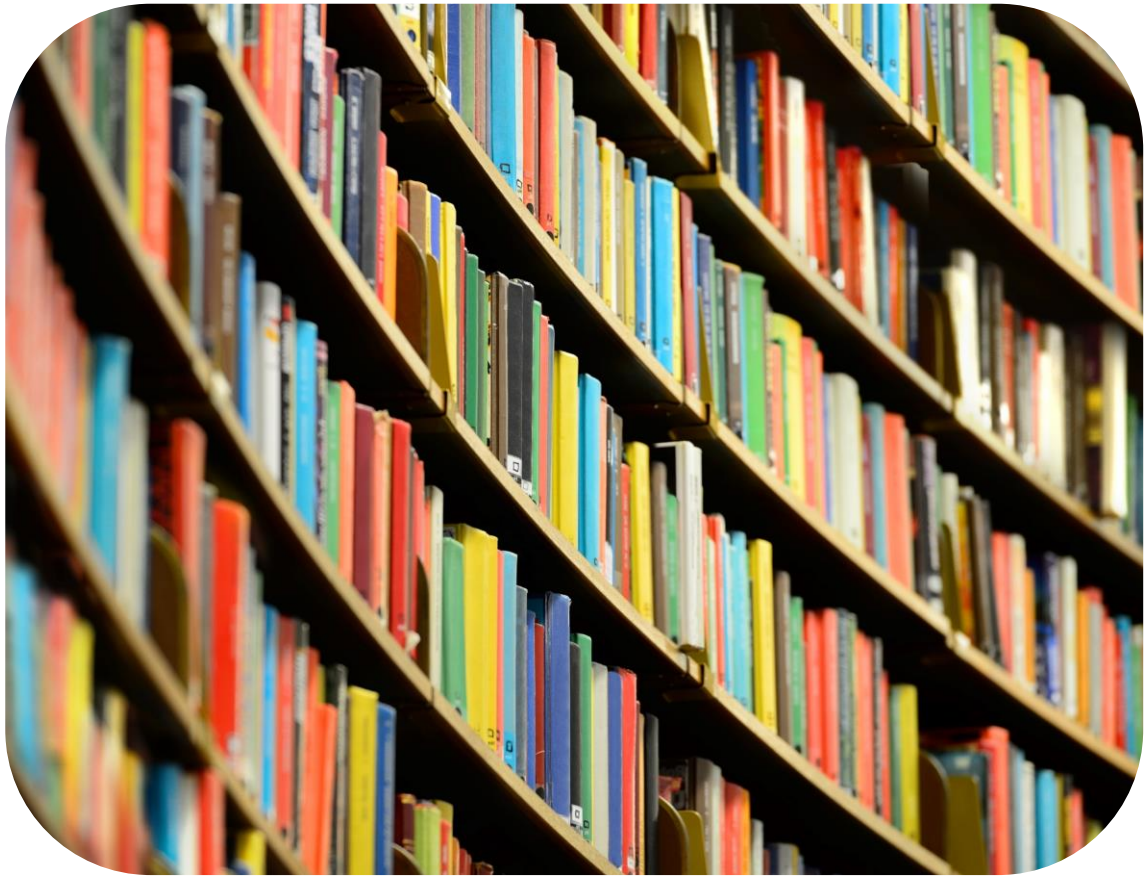
What do you think would be the **metadata** of a book?



Your turn…

The **metadata** would be the data about the book. Some examples are,

- Title
- Author
- Date of publication
- Subject
- ISBN
- Dimensions
- Number of pages
- Language



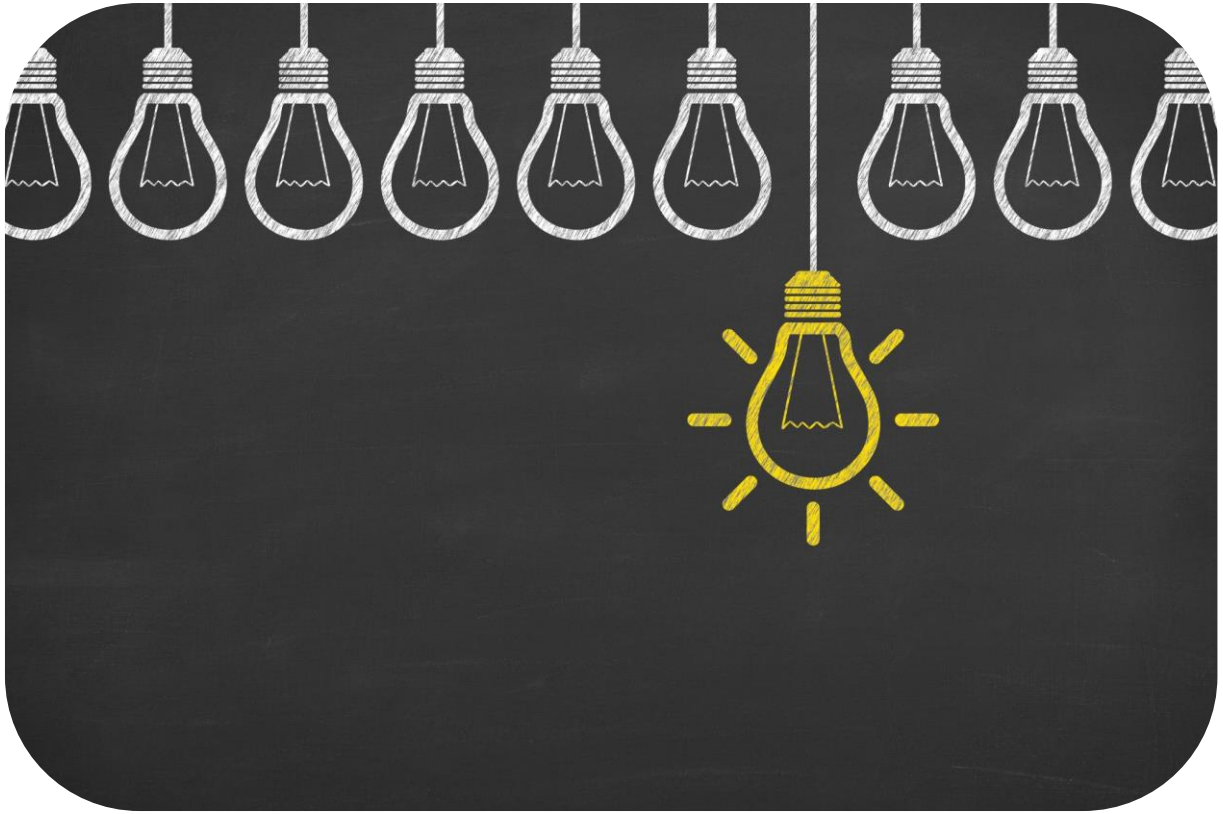
Metadata of a book

In the same way that correctly cataloguing books in a library makes it is easier to find information you are looking for; **metadata makes it easier** to find relevant information in a dataset.

Metadata and data dictionary

Without metadata it would be very difficult to work with any dataset.

A **data dictionary** is one of the most important pieces of metadata.



Definition

Data dictionary

the names, definitions and attributes of the elements in a dataset

Show me…

This dataset contains the historic value of gold from Kaggle (www.kaggle.com)

The associated **data dictionary** describes what is held in each of the columns.

This contains data files of gold historical data (USD).

Year:	Year of observation
AvgClosePrice:	The average close price in the year
YearOpen:	Opening price in the year
YearHigh:	Highest price in the year
YearLow:	Lowest price in the year
YearClose:	Closing price in the year
Annual%Change:	Percent change of the previous and current year price

Year	AvgClose Price	YearOpen	YearHigh	YearLow	YearClose	Annual % Change
2021	1799.1	1946.6	1954.4	1678	1783.9	-0.0587
2020	1773.73	1520.55	2058.4	1472.35	1895.1	0.2443
2019	1393.34	1287.2	1542.6	1270.05	1523	0.1883
2018	1268.93	1312.8	1360.25	1176.7	1281.65	-0.0115
2017	1260.39	1162	1351.2	1162	1296.5	0.1257
2016	1251.92	1075.2	1372.6	1073.6	1151.7	0.0863
2015	1158.86	1184.25	1298	1049.6	1060.2	-0.1159
2014	1266.06	1219.75	1379	1144.5	1199.25	-0.0019
2013	1409.51	1681.5	1692.5	1192.75	1201.5	-0.2779
2012	1668.86	1590	1790	1537.5	1664	0.0568
2011	1573.16	1405.5	1896.5	1316	1574.5	0.1165
2010	1226.66	1113	1426	1052.25	1410.25	0.2774
2009	973.66	869.75	1218.25	813	1104	0.2763
2008	872.37	840.75	1023.5	692.5	865	0.0341
2007	696.43	640.75	841.75	608.3	836.5	0.3159
2006	604.34	520.75	725.75	520.75	635.7	0.2392
2005	444.99	426.8	537.5	411.5	513	0.1712
2004	409.53	415.2	455.75	373.5	438	0.0497
2003	363.83	342.2	417.25	319.75	417.25	0.2174
2002	310.08	278.1	348.5	277.8	342.75	0.2396
2001	271.19	272.8	292.85	256.7	276.5	0.0141

Why are data dictionaries important?

Saves time figuring out what the data means

Provides details of **how variables have been created** (e.g. through calculation)

Helps ensures everyone is **using the datasets consistently** with the same definitions

Show me…

Companies create identifiers for each of their customers in their tools and systems.

However, different systems across different departments can often create different **customer identifiers** leading to confusion when attempting to join up systems.

By using data dictionaries, you can **compare the definitions of variables across multiple datasets** to see if you need to manipulate any variables before you join/merge the data.

46272

McA46272

R McAndrew

Rowan McAndrew

Worked example

The dataset below contains the length of rivers.

However, you need a data dictionary to know whether the length is in miles, km etc

river	length
Nile	6693
Amazon	6400
Mississippi	3766
Yangtze	6300
Congo	4700
Niger	4200

Worked example

river	length
Nile	6693
Amazon	6400
Mississippi	3766
Yangtze	6300
Congo	4700
Niger	4200

Data dictionary

river: name of the river
length: distance from the source to the mouth of the **river in kilometres**

Next steps

Complete **questions 1 to 8**
in **section 1** of the
‘Dataset understanding in Excel’ workbook.

Properties of a dataset

Another part of data understanding is to look at these properties of a dataset,

- Size
- Shape
- Data types

Show me…

This wide dataset has 7 columns and 13 rows.

The column names and data types are,

column_names	col_types
name	string
height	integer
mass	integer
hair_colour	string
skin_colour	string
gender	string
homeworld	string

name	height	mass	hair_colour	skin_colour	gender	homeworld
Luke Skywalker	172	77	blond	fair	masculine	Tatooine
C-3PO	167	75	NA	gold	masculine	Tatooine
R2-D2	96	32	NA	white, blue	masculine	Naboo
Darth Vader	202	136	none	white	masculine	Tatooine
Leia Organa	150	49	brown	light	feminine	Alderaan
Owen Lars	178	120	brown, grey	light	masculine	Tatooine
Beru Whitesun lars	165	75	brown	light	feminine	Tatooine
R5-D4	97	32	NA	white, red	masculine	Tatooine
Biggs Darklighter	183	84	black	light	masculine	Tatooine
Obi-Wan Kenobi	182	77	auburn, white	fair	masculine	Stewjon
Anakin Skywalker	188	84	blond	fair	masculine	Tatooine
Yoda	66	17	white	green	masculine	NA
Palpatine	170	75	grey	pale	masculine	Naboo

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

How to calculate size in Excel

The size of a dataset is the number of columns and the number of rows.

In Excel you can use **counta()**, which counts the number of non-empty cells.

Calculation	Formula
Count	=**counta**(A,B,C,D,…)

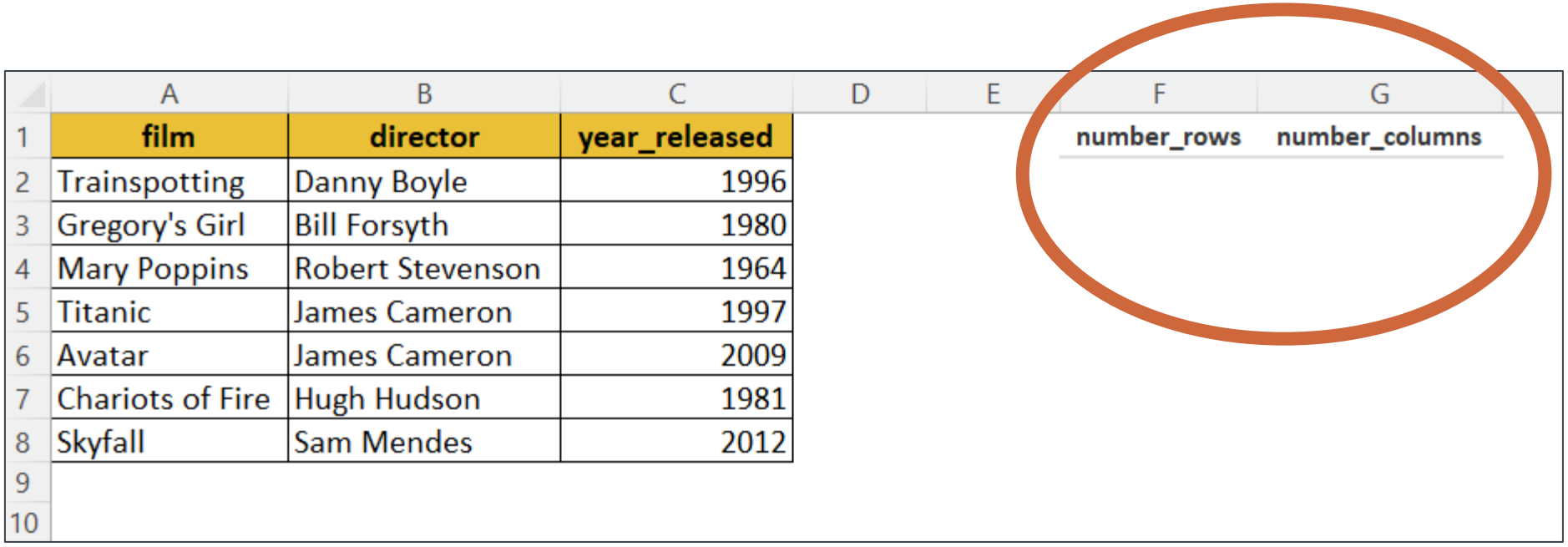
Note: this is different from count() and countif().

For more details of the formulas in Excel, see https://support.microsoft.com/en-us/excel

Set up your dataset

Step 1.

Once you have your dataset open, create a new dataset for the number of columns and number of rows.

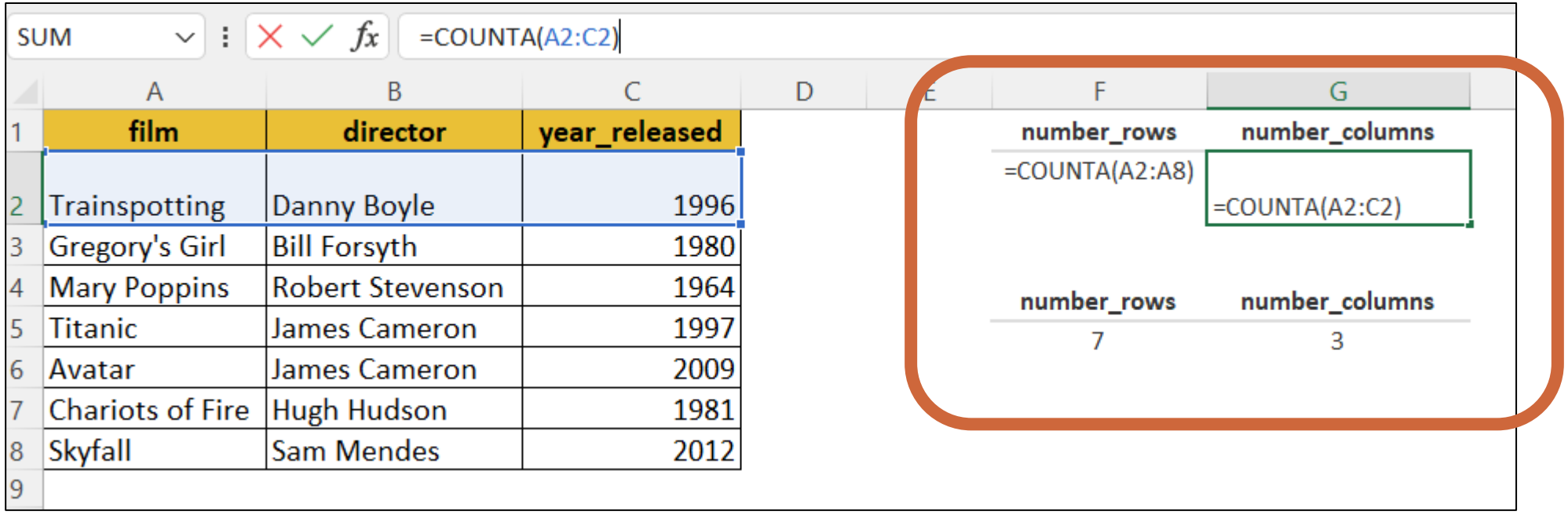


	A	B	C	D	E	F	G
1	film	director	year_released			number_rows	number_columns
2	Trainspotting	Danny Boyle	1996				
3	Gregory's Girl	Bill Forsyth	1980				
4	Mary Poppins	Robert Stevenson	1964				
5	Titanic	James Cameron	1997				
6	Avatar	James Cameron	2009				
7	Chariots of Fire	Hugh Hudson	1981				
8	Skyfall	Sam Mendes	2012				
9							
10							

Using counta()

Step 2.

Using the **counta()** function twice, once to count the number of rows and then again to count the number of columns.



Data types in a dataset

Each column of data should be reviewed to understand the type of data and whether it needs to be reformatted. The table below is a reminder of the different data types.

Data type	Definition
Integer	Number (positive or negative) with no decimal or fractional parts
Floating point	Number that contain a decimal or fractional part
String	A collection of characters combined to create alphanumeric text
Boolean	Can only take two possible values, such as true/false or yes/no
Date and time	The number of days or seconds passed since the 'epoch' date

Show me…

This dataset has 7 rows and 3 columns, calculated using counta().

It has 2 columns containing strings and 1 with integers.

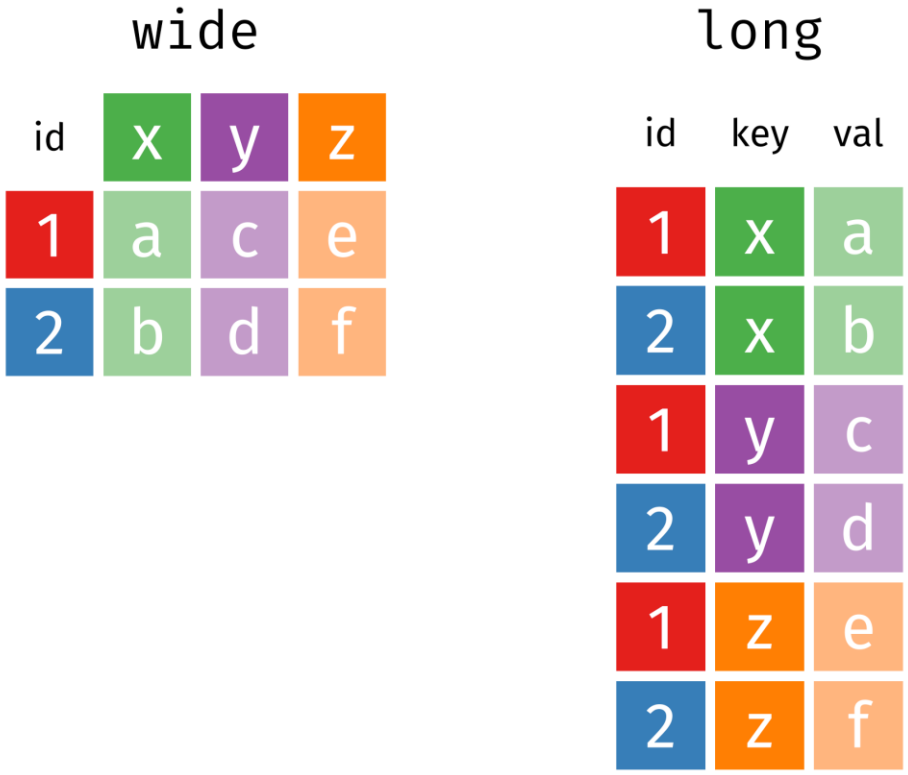
	A	B	C	D	E	F	G
1	**film**	**director**	**year_released**			**number_rows**	**number_columns**
2	Trainspotting	Danny Boyle	1996			7	3
3	Gregory's Girl	Bill Forsyth	1980				
4	Mary Poppins	Robert Stevenson	1964			**column_names**	**col_types**
5	Titanic	James Cameron	1997			film	string
6	Avatar	James Cameron	2009			director	string
7	Chariots of Fire	Hugh Hudson	1981			year_released	integer
8	Skyfall	Sam Mendes	2012				
9							

Format of a dataset

The look of a dataset is often determined by how the data items are collected.

However, it may not be the best form for the analysis you want to perform on it.

By identifying the shape of the dataset during the data understanding stage will help you **decide if you need to reshape** before you manipulate it.



Source: Garrick Aden-Buie's (@grrrck) Tidy Animated Verbs

Next steps

Complete **questions 1 to 6**
in **section 2** of the
‘Dataset understanding in Excel’ workbook.

Identifying outliers and missing values

As part of data understanding, you need to identify any outliers and missing values.

Once they have been identified, they will be reviewed and possibly removed/replaced as part of the data tidying and cleansing stage of the analysis steps.

Definition

Outlier

an observation (or set of observations) which appear to be inconsistent with the set of data

Quantitative outliers in Excel

Reminder: Quantitative
Measures of values or counts and expressed as numbers

When looking at quantitative data, you can identify any outliers by looking for the minimum and maximum values.

planet	planet_type	size_km	
Mercury	terrestrial	2,440	Minimum
Venus	terrestrial	6,052	
Earth	terrestrial	6,371	
Mars	terrestrial	3,390	
Jupiter	gas giant	69,911	Maximum
Saturn	gas giant	58,232	
Uranus	ice giant	25,362	
Neptune	ice giant	24,622	

Quantitative outliers in Excel

Minimum and maximum may help you identify the outliers but doesn't necessarily mean the min and max will *be* outliers.

Maximum and minimum in Excel

In Excel you can use the =max() and =min() functions to find the outliers.

Calculation	Formula
Maximum	=max(A,B,C,D,…)
Minimum	=min(A,B,C,D,…)

Show me…

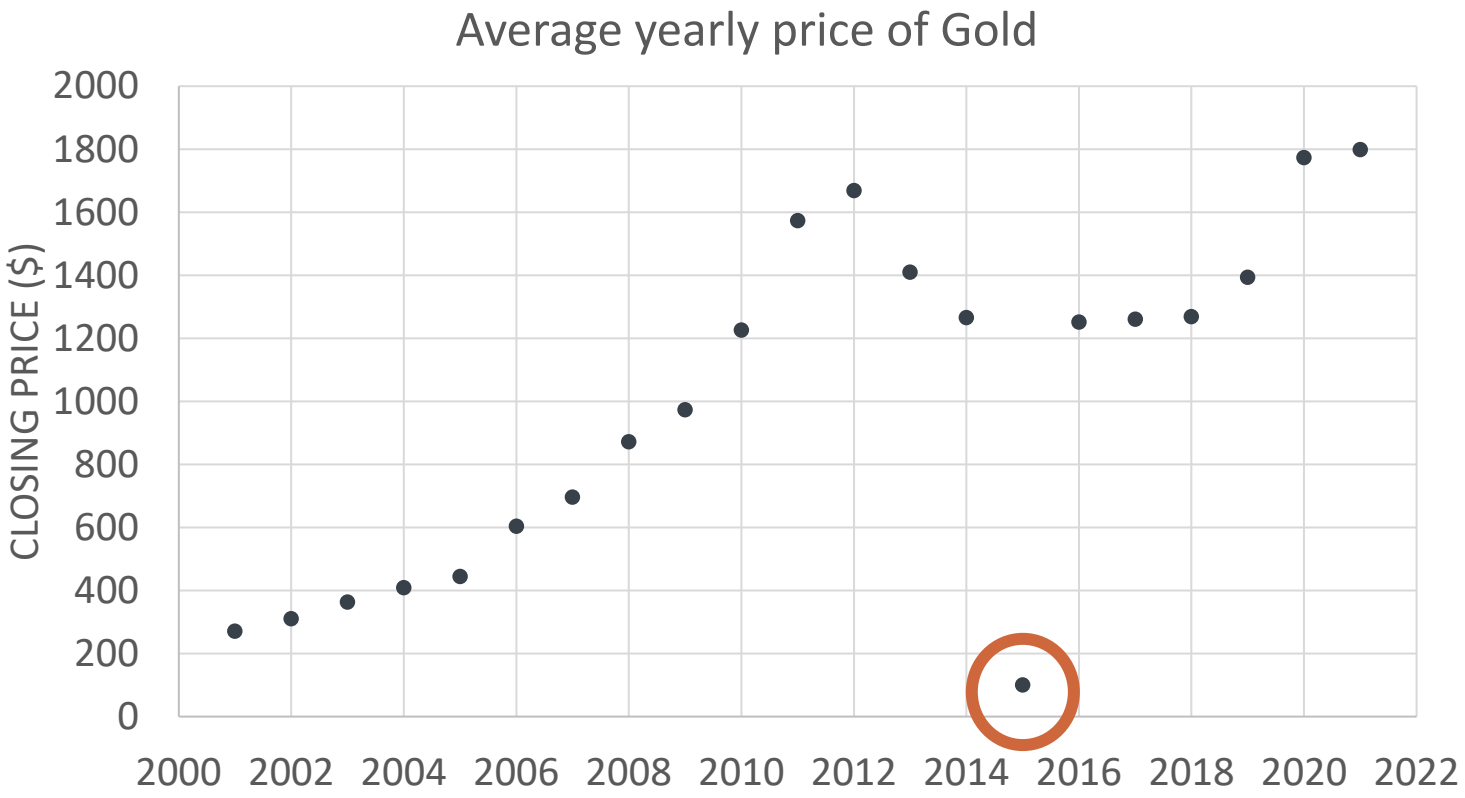
The dataset below contains the boiling point of elements in the periodic table. By using the max and min functions you can see that the boiling point ranges from around -300°C to +6,000°C.

	A	B	C	D	E	F	G	H
1	element_nam	boiling_point_celsiu		number_rows	number_columns		max_boiling_point	min_boiling_point
2	Hydrogen	-252.87		118	2		6,073	-269
3	Helium	-268.93					=MAX(B2:B119)	=MIN(B2:B119)
4	Lithium	1286.85						
5	Beryllium	2468.85						
6	Boron	3926.85						
7	Carbon	4026.85						
8	Nitrogen	-195.79						
9	Oxygen	-182.95						
10	Fluorine	-188.12						
11	Neon	-246.08						

Source: https://en.wikipedia.org/wiki/List_of_chemical_elements, list correct as of Jan 2022

Show me…

This graph shows the average price of Gold by year. The value in **year 2015 is an outlier** and would need investigating.

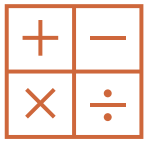


https://www.kaggle.com/nward7/gold-historical-datasets

Why is identifying outliers important?

May indicate **errors in the dataset**, e.g. information has been entered incorrectly, instrument has incorrectly measured a value

Highlight data items that need investigating, e.g. **why is it an outlier**?



Might impact on your **statistical calculations** e.g. mean average

Next steps

Complete **questions 1 to 6**
in **section 3** of the
‘Dataset understanding in Excel’ workbook.

Missing values in Excel

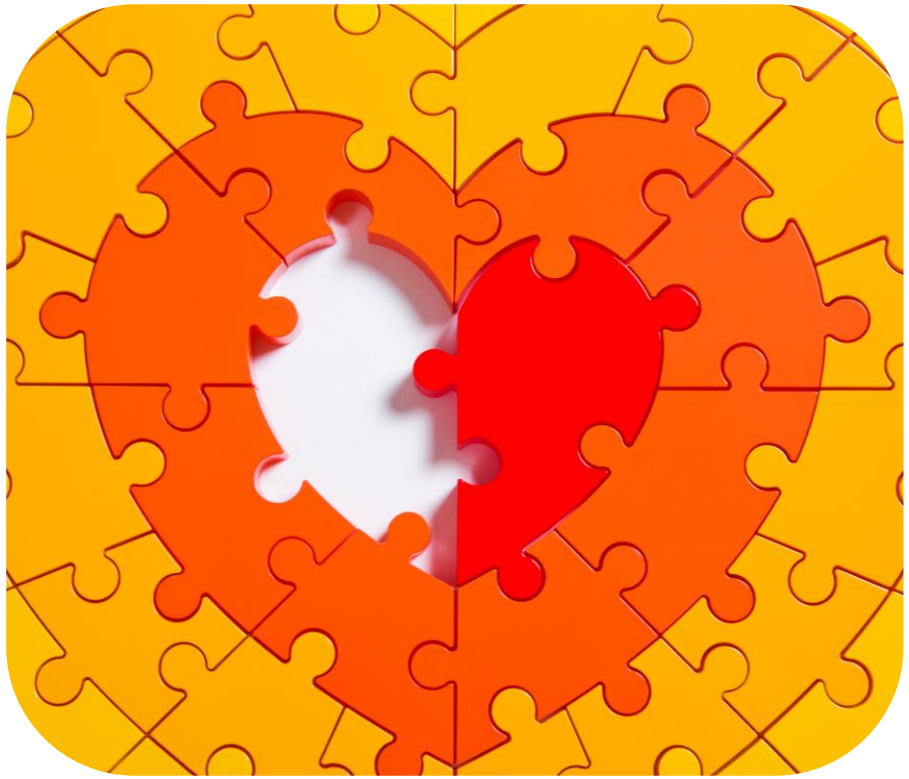
Datasets can often **contain missing or blank values.**

Identifying them during the data understanding stage will save you time later on.

How missing values might look

Missing values can be shown in the dataset as,

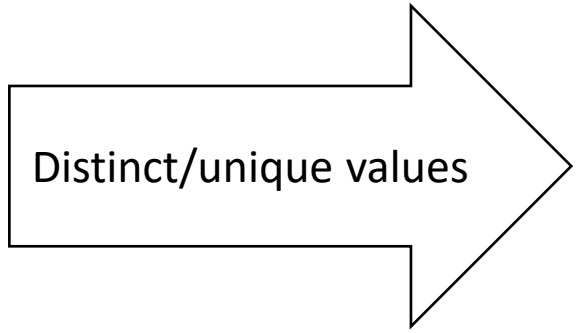
- Empty/blank cells
- Errors e.g. #DIV/0!, #N/A
- NULL
- NA
- NaN
- TBC or missing
- Unknown
- 0



Missing values in Excel

A way to find if a variable contains missing values is to look at all unique or distinct data items contained in the column.

film	year_released
Encanto	2021
No time to die	2021
Parasite	2020
The Favourite	2019
All Is True	2019
Green Book	Unknown



year_released
2021
2020
2019
Unknown

Distinct values in Excel

In Excel, you can use the function =unique() to output all the distinct values in a set of data items.

Calculation	Formula
What distinct values are there in a set of values?	=unique()

Worked example

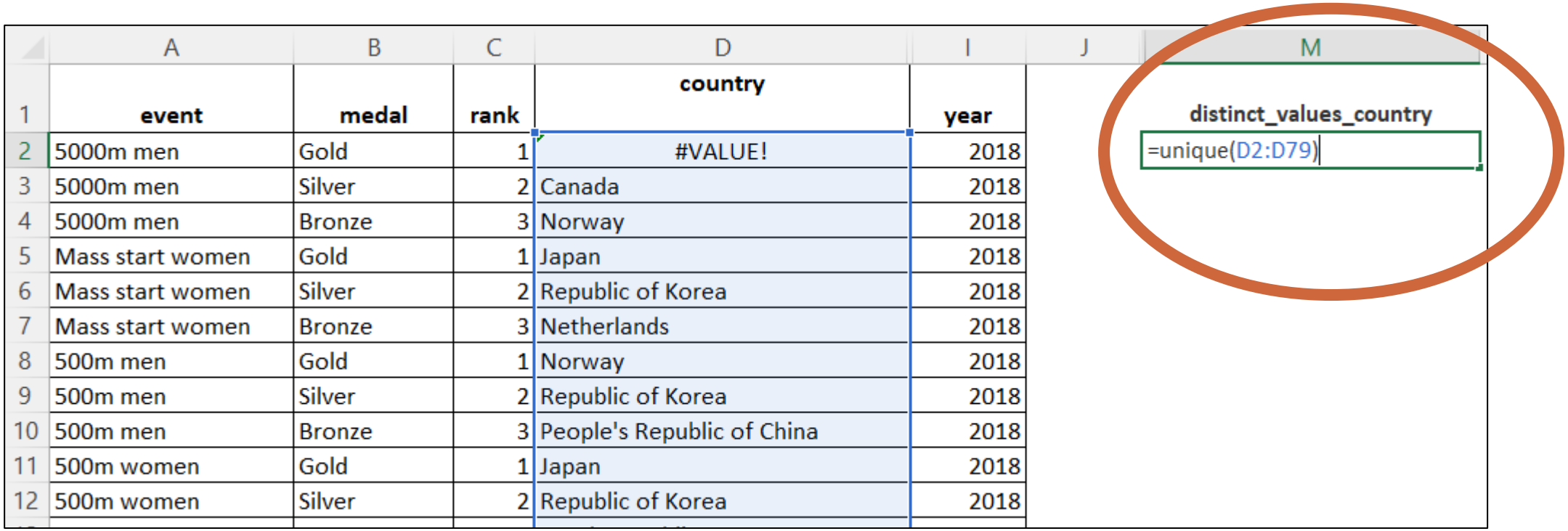
We have a dataset the contains Olympic speedskating data. As part of the data understanding we are going to see what countries are in the dataset.

A	B	C	D
event	**medal**	**rank**	**country**
5000m women	Gold	1	Czech Republic
5000m women	Silver	2	Netherlands
5000m women	Bronze	3	Netherlands
1000m men	Gold	1	Netherlands
1000m men	Silver	2	Canada
1000m men	Bronze	3	Netherlands
3000m women	Gold	1	Netherlands
3000m women	Silver	2	Czech Republic
3000m women	Bronze	3	Russian Federation
1500m men	Gold	1	Poland

Using unique() formula

Step 1.

Type in the heading for the calculation you are about to create, in this case the distinct countries.
Next type in the formula **=unique(*range*)**



	A	B	C	D	I	J	M
1	event	medal	rank	country	year		distinct_values_country
2	5000m men	Gold	1	#VALUE!	2018		=unique(D2:D79)
3	5000m men	Silver	2	Canada	2018		
4	5000m men	Bronze	3	Norway	2018		
5	Mass start women	Gold	1	Japan	2018		
6	Mass start women	Silver	2	Republic of Korea	2018		
7	Mass start women	Bronze	3	Netherlands	2018		
8	500m men	Gold	1	Norway	2018		
9	500m men	Silver	2	Republic of Korea	2018		
10	500m men	Bronze	3	People's Republic of China	2018		
11	500m women	Gold	1	Japan	2018		
12	500m women	Silver	2	Republic of Korea	2018		

This example uses a dataset from Kaggle on Speedskating at the winter Olympics
(www.kaggle.com/niekvanderzwaag/speedskating-at-the-winter-olympics?select=speedskating.csv)

Review the list

Step 2.

When you press enter on the formula, the list of distinct values will appear.

	A	B	C	D	I	J	M
1	event	medal	rank	country	year		distinct_values_country
2	5000m men	Gold	1	#VALUE!	2018		#VALUE!
3	5000m men	Silver	2	Canada	2018		Canada
4	5000m men	Bronze	3	Norway	2018		Norway
5	Mass start women	Gold	1	Japan	2018		Japan
6	Mass start women	Silver	2	Republic of Korea	2018		Republic of Korea
7	Mass start women	Bronze	3	Netherlands	2018		Netherlands
8	500m men	Gold	1	Norway	2018		People's Republic of China
9	500m men	Silver	2	Republic of Korea	2018		Czech Republic
10	500m men	Bronze	3	People's Republic of China	2018		Italy
11	500m women	Gold	1	Japan	2018		United States of America
12	500m women	Silver	2	Republic of Korea	2018		Olympic Athletes from Russia
13	500m women	Bronze	3	Czech Republic	2018		Belgium
14	Team pursuit men	Gold	1	Norway	2018		Poland
15	Team pursuit men	Silver	2	Republic of Korea	2018		Russian Federation

You can review the list of distinct values and spot missing values

Next steps

Complete **questions 1 to 6**
in **section 4** of the
‘Dataset understanding in Excel’ workbook

Then complete the **extension question**
in **section 5** of the workbook

Learning checklist

I can *describe* what metadata is and how it can be used.

I can *describe* what is a data dictionary and how it can be used.

I can *describe* the size and format of datasets in Excel.

I can *identify* missing values and outliers of a dataset using min(), max() and unique() functions in Excel.

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

Alternative format

If you require this document in an alternative format, such as large print or a coloured background, please contact

hello@effini.com

or

4th Floor, The Bayes Centre
47 Potterrow
Edinburgh
EH8 9BT

Effini

THE DATA LAB
value from data

Data Skills FOR WORK

Data Education in Schools