V1.0

Practise data cleansing in Excel

Effini

THE DATA LAB
value from data

Data Skills FOR WORK

Data Education in Schools

Learning intentions

We will be learning how to apply data cleansing techniques to **cleanse a dataset using Excel**, specifically,

- how to **rename variables**
- how to **drop unrequired rows** and **variables**
- how to **drop duplicates**
- how to **handle missing data** and **outliers**

Background

In **Data Cleansing in Excel,** you learnt approaches and techniques for cleaning data.

The purpose of this step in the analysis steps is to **prepare the data for future analysis**.

This is an important step because **datasets are almost never clean**.

Data understanding | Data cleansing | Data manipulation | Identifying patterns | Extracting insights

The data

In this lesson you will be **cleansing a music dataset** that contains information about popular songs from 1945-2019. The data is originally from Spotify.

In the ‘Practise dataset understanding in Excel’ lesson we completed the data understanding of this dataset.

4	title	Artist	genre	TheYearTheSongWasReleased	BPM	energy	danceability	duration	decade	loudness
5	Peppermint Twist Pt.1	Joey Dee	Australian Ta	2019	199	0.73	0.42	122	NA	NA
6	Highway to Hell	AC/DC	Album Rock	1979	116	0.91	0.57	208	NA	NA
7	End Of The Line	Traveling Wilburys	Album Rock	1988	167	0.84	0.58	210	NA	NA
8	SAD!	XXXTENTACION	Emo Rap	2018	75	0.61	0.74	167	NA	NA
9	Silence	Marshmello	Brostep	2017	142	0.76	0.52	181	2010	NA
10	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
11	Just Got Paid	ZZ Top	Album Rock	1972	100	0.69	0.57	267	1970	-4
12	Heaven	Bryan Adams	Album Rock	1984	140	0.59	0.38	243	1980	NA
13	My September Love	David Whitfield	Deep Adult S	2000	137	0.21	0.24	166	1950	NA

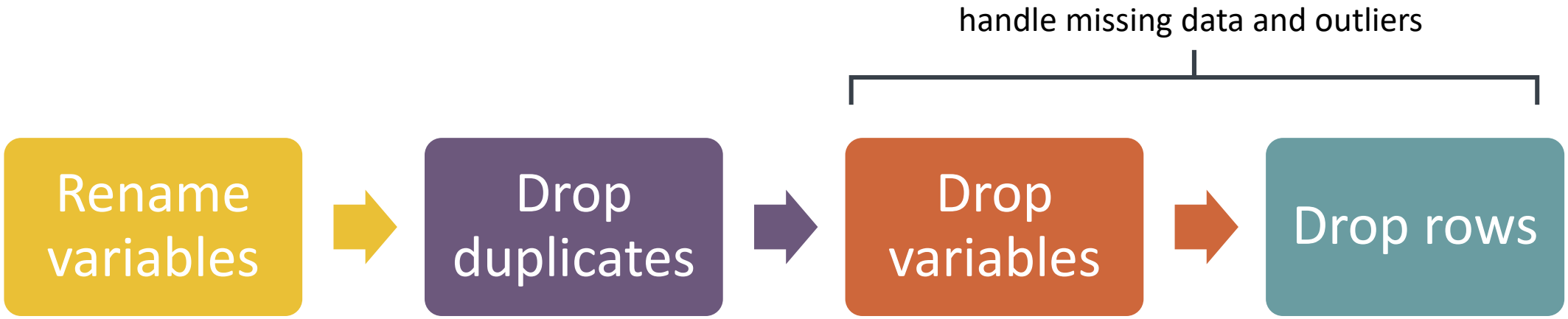
Data understanding

Here is what we discovered:

- there are **647 rows** and **10 variables**
- 3 of the variables are **strings**, 4 are **integers** and 3 are **floats**
- it is in a **wide** format
- the **BPM** variable contains 2 **outliers**: -20 and 2969
- 65% of the values for the **decade** variable are **missing**
- 95% of the values for the **loudness** variable are **missing**
- there are 3 rows in which all the data-items are **missing**

Data cleansing steps

Here are the steps you will take in this lesson,

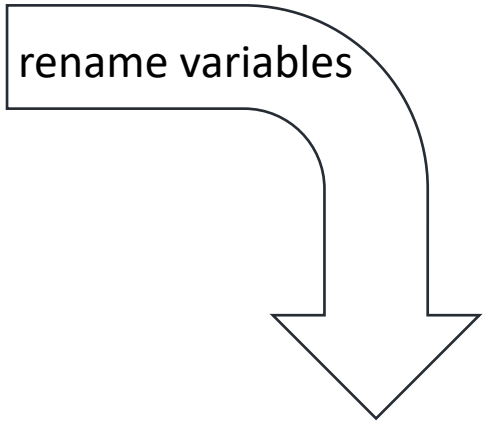


Reminder: Renaming variables

Good variable names:

- are **consistent** (i.e. they use a **naming convention**)
- are **meaningful**
- don't **include characters that might cause issues** (e.g. '?')

?FirstName	LAST-NAME	var3	ret	Response
Rosie	Love	44	22	NaN
Greg	Hill	65	NaN	NaN
Molly	Jones	-2	NaN	NaN
Isla	Well	234	NaN	NaN
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN
Barry	McNeil	22	44	NaN
Barry	McNeil	22	44	NaN



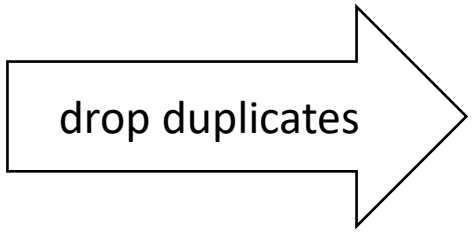
first_name	last_name	age	years_to_retirement	response
Rosie	Love	44	22	NaN
Greg	Hill	65	NaN	NaN
Molly	Jones	-2	NaN	NaN
Isla	Well	234	NaN	NaN
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN
Barry	McNeil	22	44	NaN
Barry	McNeil	22	44	NaN

Why rename variables? Good variable names make it easier to **read** and **understand** the dataset and **avoid issues** with code not working as expected.

Reminder: Dropping duplicate rows

Why drop duplicates? Duplicate rows in your dataset might **change the result of any calculations you perform** (e.g. a count of the number of people who responded to a survey).

	first_name	last_name	age	years_to_retirement
0	Rosie	Love	44	22
1	Greg	Hill	65	1
2	Molly	Jones	-2	68
3	Isla	Well	234	168
6	Barry	McNeil	22	44
7	Barry	McNeil	22	44

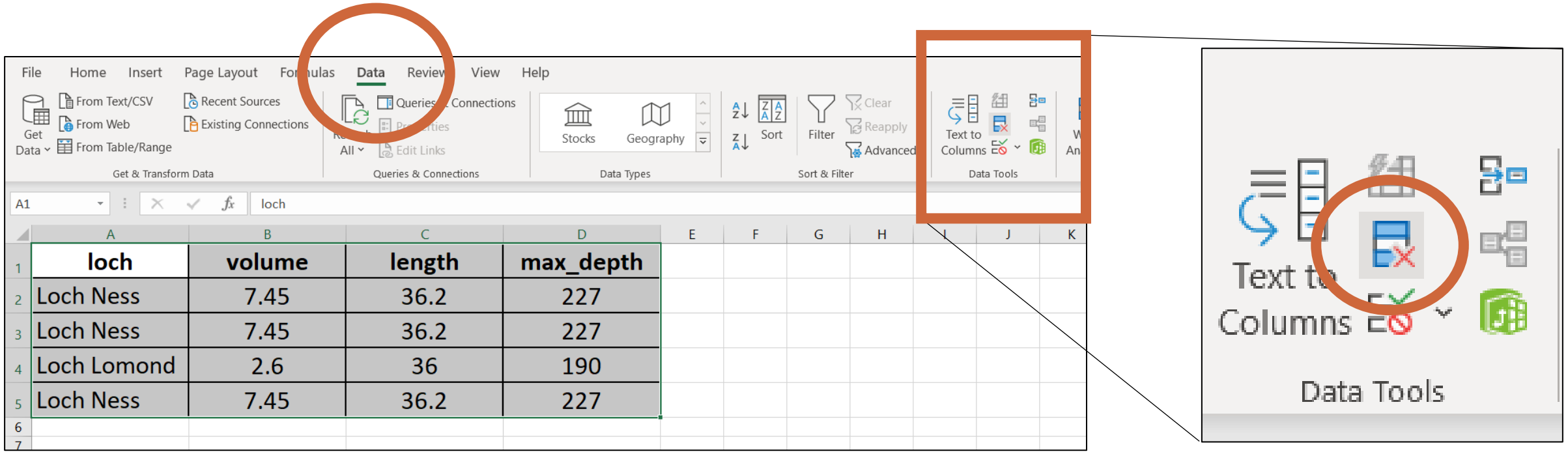


	first_name	last_name	age	years_to_retirement
0	Rosie	Love	44	22
1	Greg	Hill	65	1
2	Molly	Jones	-2	68
3	Isla	Well	234	168
6	Barry	McNeil	22	44

Reminder: remove duplicates in Excel

Highlight the data you would like to remove duplicates from.

On the ‘Data’ ribbon find the option in the ‘Data Tools’ section called ‘**Remove Duplicates**’. It looks like some boxes with a cross next to them.



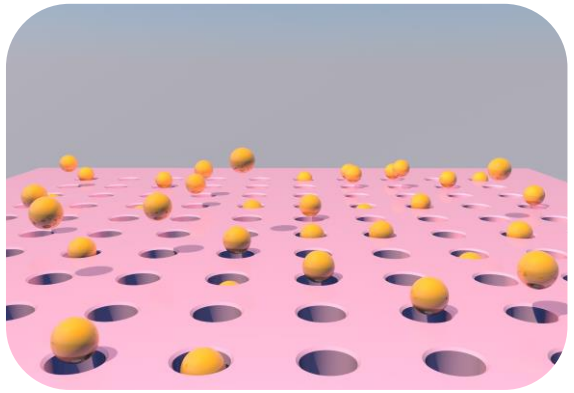
Next steps

Complete **questions 1 & 2**
in **section 1** in the
‘Practise Data Cleansing
in Excel’ workbook.

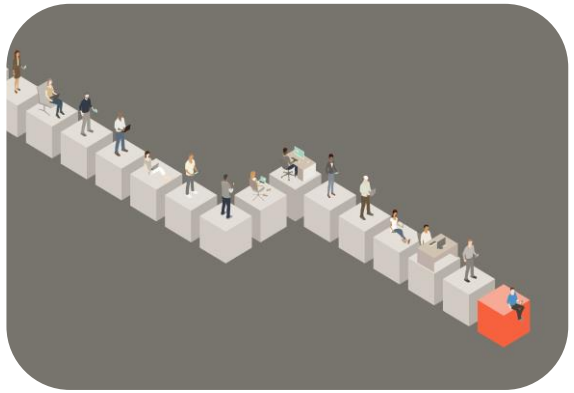
Reminder: Handling missing data or outliers

There are 3 options for handling missing data or outliers.

Remove



Replace



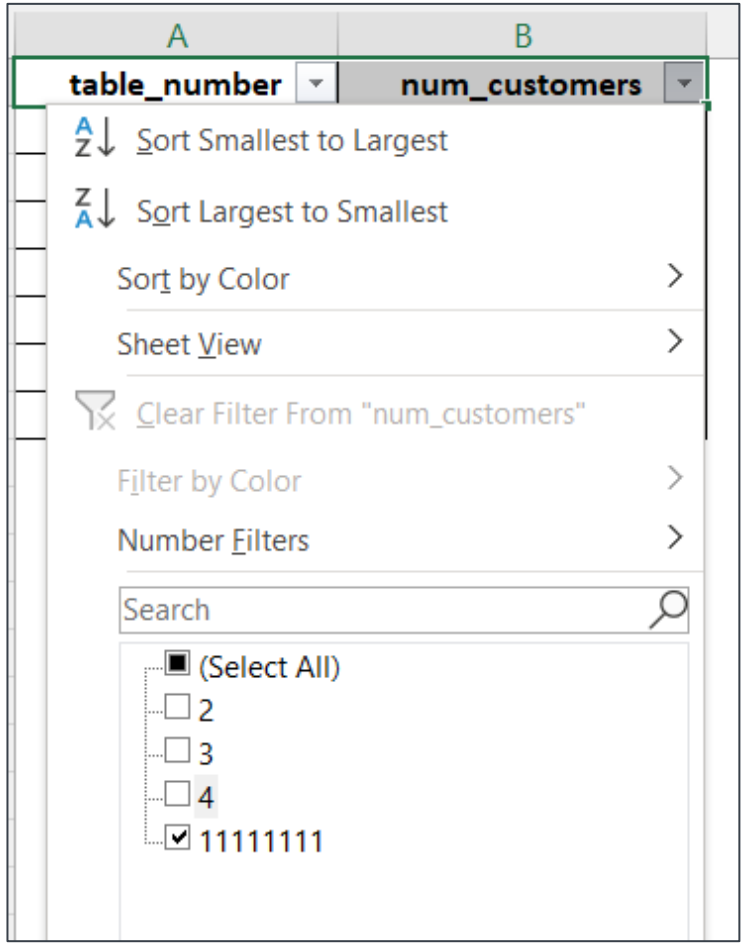
Leave as is

Reminder: Filter to find missing/outlier values

Filter the dataset to show all the values that need to replaced/removed.

Reminder: to turn on the filter press the following

Windows	Ctrl + Shift + L
Mac	Command + Shift +L



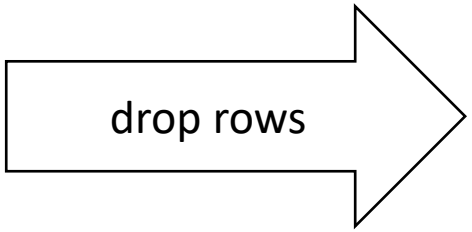
Dropping rows

A dataset may contain **rows** that do **not contain useful information**. For example, all or most of the data-items in the row may be missing.

These rows can be **dropped** without losing any useful information.

Why drop rows? Rows with missing data may change the result of any calculations you perform.

	first_name	last_name	age	years_to_retirement
0	Rosie	Love	44	22
1	Greg	Hill	65	1
2	Molly	Jones	-2	68
3	Isla	Well	234	168
4	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN
6	Barry	McNeil	22	44
7	Barry	McNeil	22	44

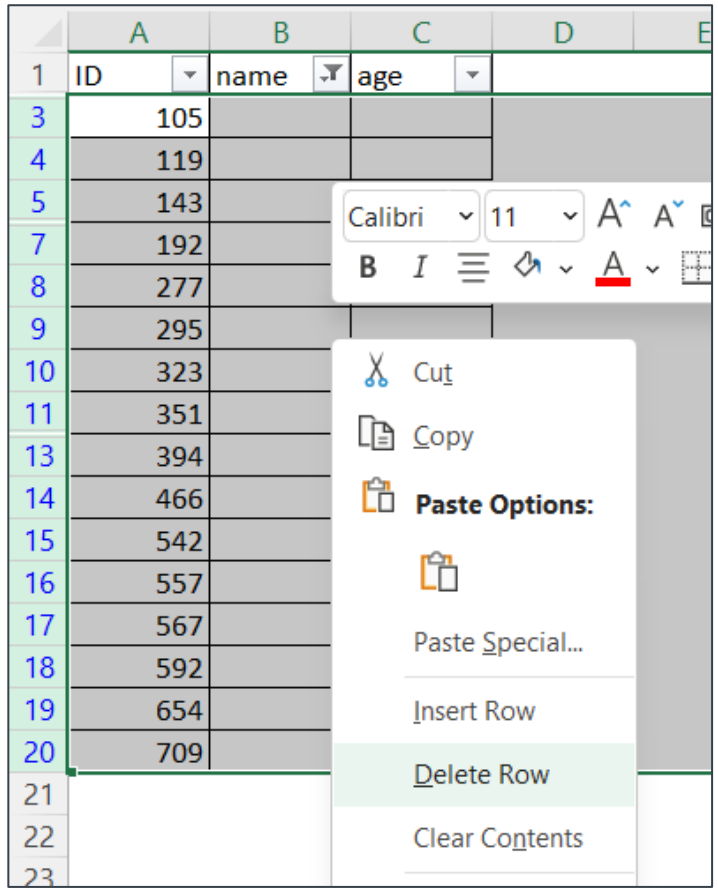


	first_name	last_name	age	years_to_retirement
0	Rosie	Love	44	22
1	Greg	Hill	65	1
2	Molly	Jones	-2	68
3	Isla	Well	234	168
6	Barry	McNeil	22	44
7	Barry	McNeil	22	44

Reminder: Dropping unrequired rows in Excel

If you have **rows** that need to be dropped you can,

1. **Filter** your dataset (*Crtl+Shift+L* to turn on filter) so you can only see the rows that need to dropped
2. **Highlight** all the rows that need to be deleted
3. **Right click** on any cell in the highlighted rows
4. Select **Delete Row**.



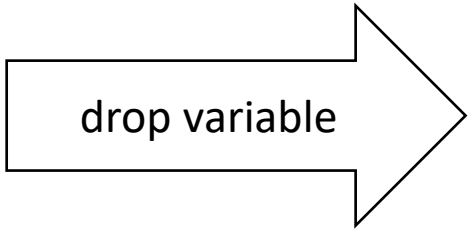
Dropping variables

A dataset may contain variables that are **not needed for the problem you are trying to analyse** or **do not contain useful information**.

These variables can be **dropped** (i.e. removed) without losing any useful information.

Why drop variables? Dropping these variables can make the dataset easier to read and understand.

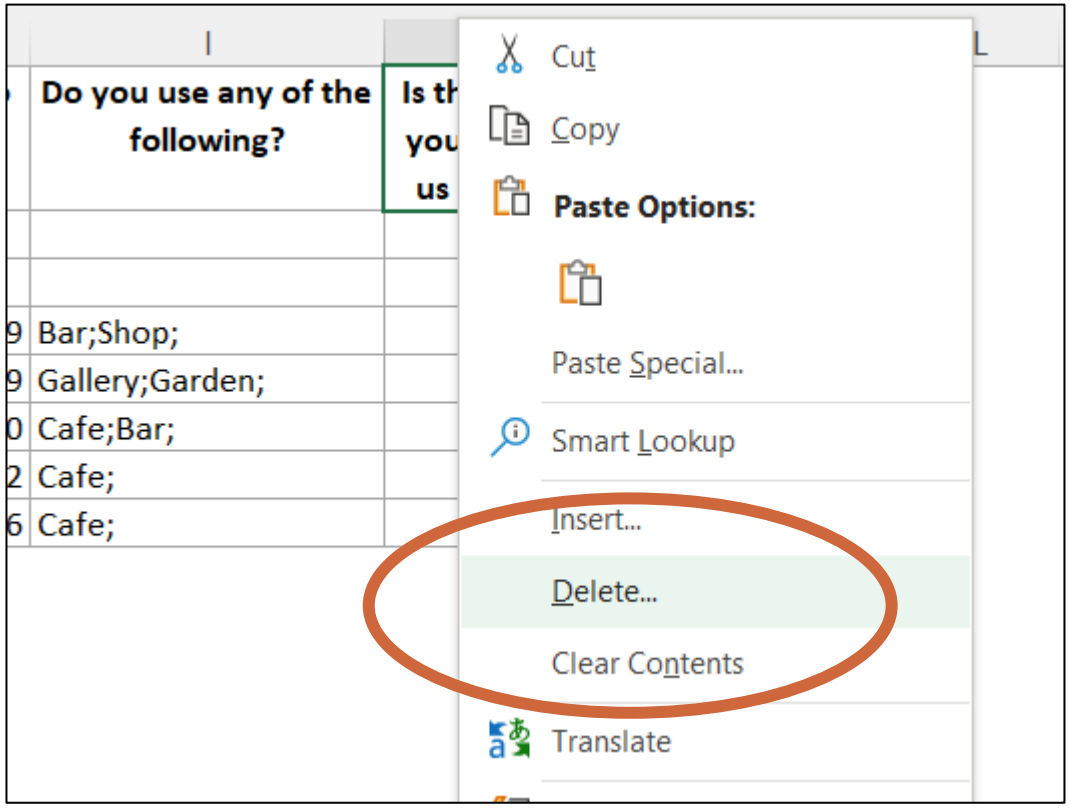
first_name	last_name	age	years_to_ retirement	response
Rosie	Love	44	22	NaN
Greg	Hill	65	1	NaN
Molly	Jones	-2	68	NaN
Isla	Well	234	168	NaN
NaN	NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN	NaN
Barry	McNeil	22	44	NaN
Barry	McNeil	22	44	NaN



first_name	last_name	age	years_to_ retirement
Rosie	Love	44	22
Greg	Hill	65	1
Molly	Jones	-2	68
Isla	Well	234	168
NaN	NaN	NaN	NaN
NaN	NaN	NaN	NaN
Barry	McNeil	22	44
Barry	McNeil	22	44

Reminder: Dropping unrequired columns

Once you have found a column that needs to be deleted, **right click on any cell in that column.** Then click on **Delete.**



Next steps

Complete **questions 1 to 13**
in **section 2** in the
‘Practise Data Cleansing
in Excel’ workbook.

Learning checklist

I can *change* the name of a variable to a chosen naming convention in Excel

I can *remove* rows and variables in Excel

I can *remove* duplicate rows in Excel

I can *remove* rows which contain outliers 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.

© 2022. This work is licensed under a *CC BY-NC-SA 4.0 license*.

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