Practise data understanding 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 understanding techniques to **understand an unfamiliar dataset using Excel**, specifically,

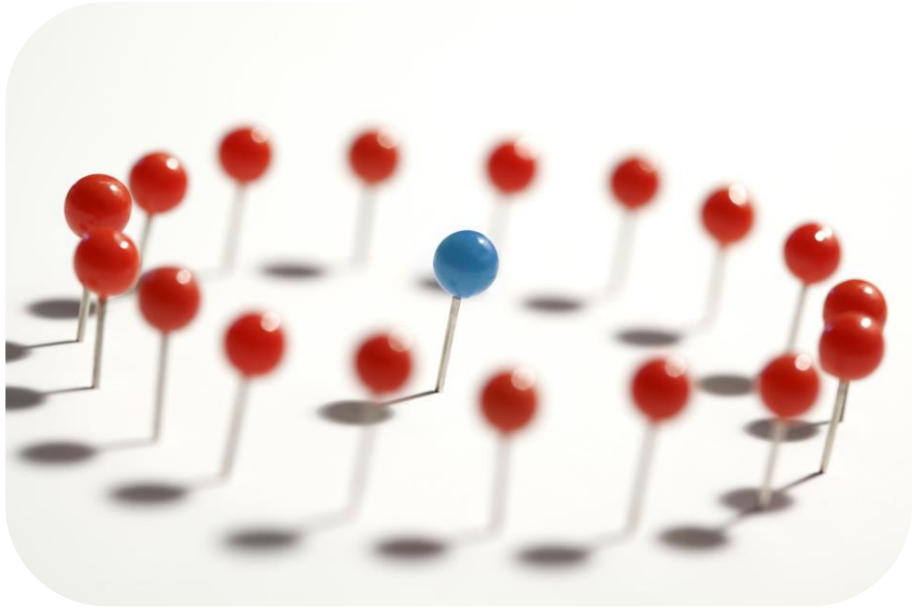
- how to **import** a csv dataset and how to remove the **metadata**
- how to use a **data dictionary** to find out about a dataset
- how to find the **shape**, **size** and **format** of datasets, using Excel
- how to find the **data types** of variables in a dataset, using Excel
- how to **identify outliers** and **missing values** in Excel

Background

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**



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

The data

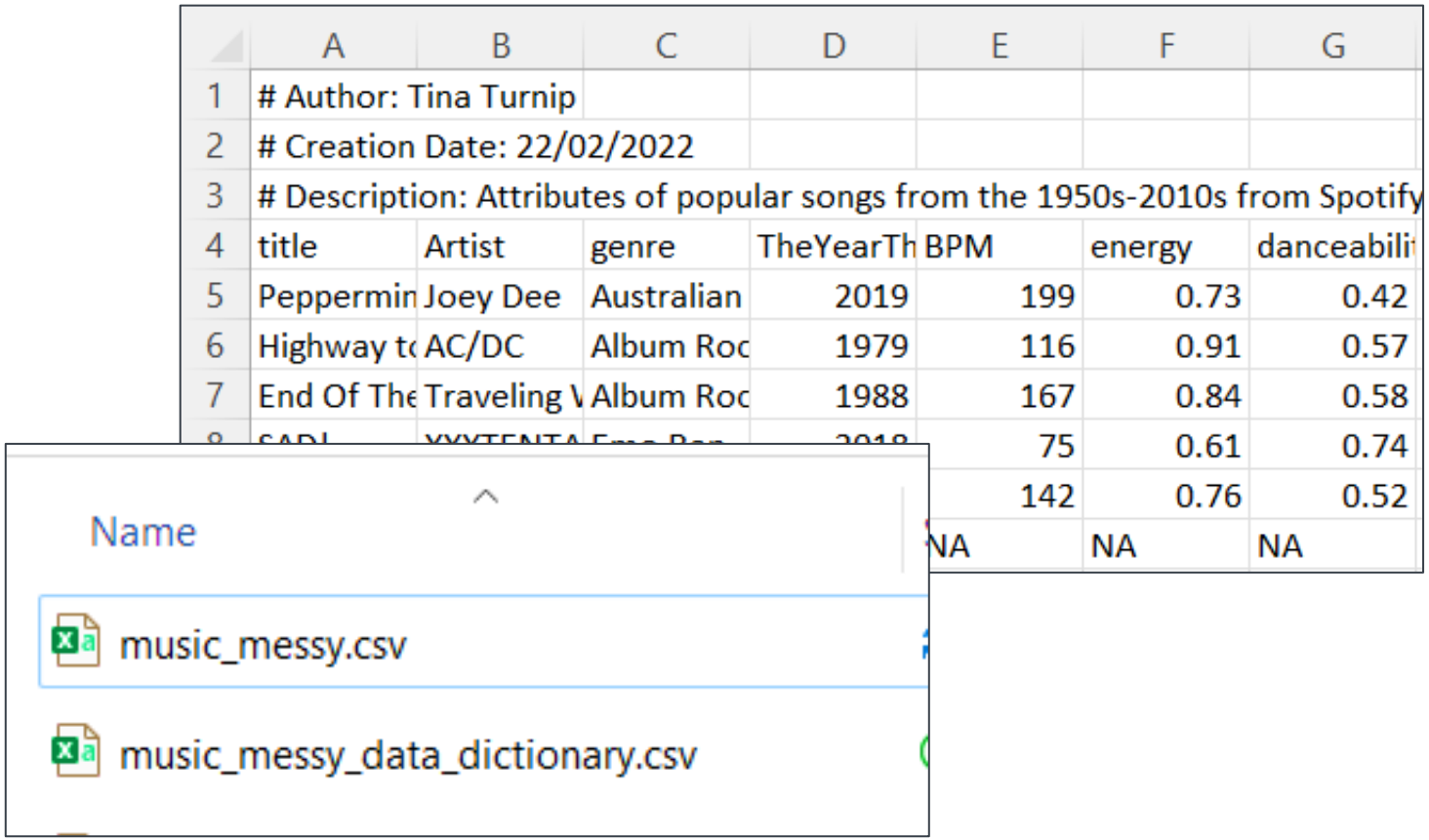
In this lesson you will be **understanding a music dataset** that contains information about popular songs from 1945-2019. The data is originally from Spotify.

	title	Artist	genre	TheYearTheSongWasReleased	BPM	energy	danceability	duration	decade	loudness
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

File format of the music dataset

This music dataset is saved in a **.csv file** and has metadata (top 3 rows).

To complete the analysis of this dataset first we need to **import (or copy) the data** into a standard Excel file.



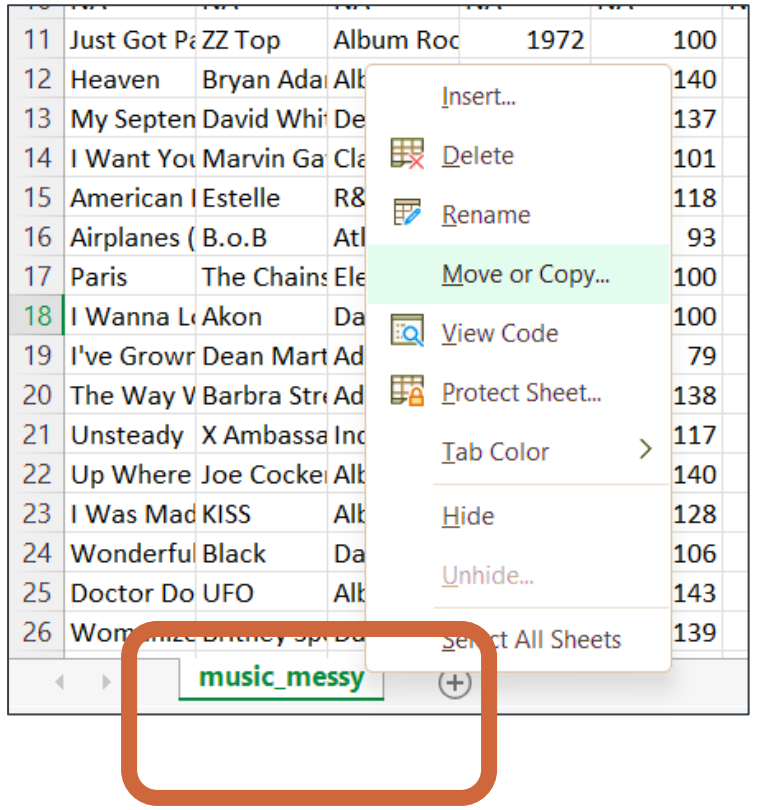
	A	B	C	D	E	F	G
1	# Author: Tina Turnip						
2	# Creation Date: 22/02/2022						
3	# Description: Attributes of popular songs from the 1950s-2010s from Spotify						
4	title	Artist	genre	TheYearTh	BPM	energy	danceabili
5	Peppermin	Joey Dee	Australian	2019	199	0.73	0.42
6	Highway t	AC/DC	Album Roc	1979	116	0.91	0.57
7	End Of The	Traveling \	Album Roc	1988	167	0.84	0.58
					75	0.61	0.74
					142	0.76	0.52
					NA	NA	NA

How to copy/import data into Excel

Once you have your .csv file open you need to copy it into the Excel workbook you will be analysing it in.

Step 1.

Right click on the name of the tab (e.g. music_messy) in the csv file and select **‘Move or Copy…’**



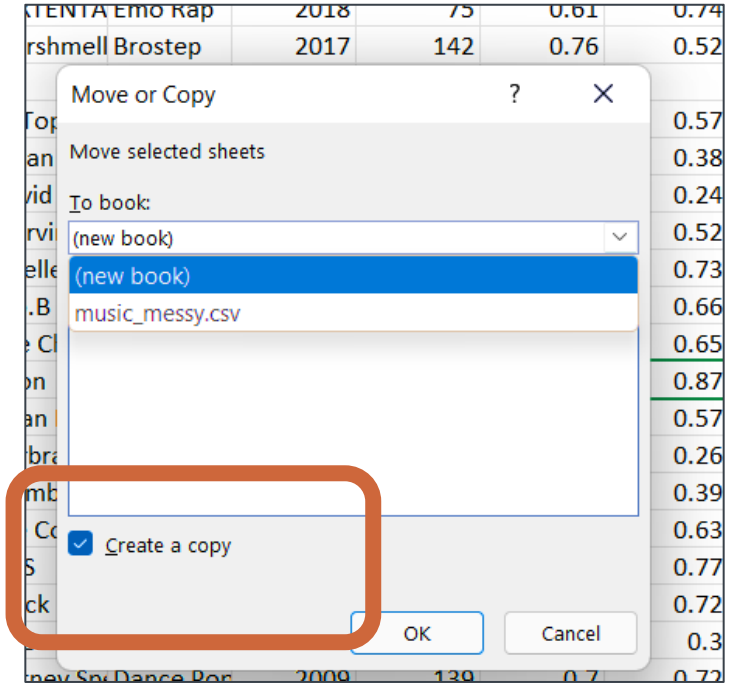
How to copy/import data into Excel

Step 2.

Select the **‘Create a copy’ tick box** and choose the Excel workbook you want to import it into.

In this case it would be a **“(new book)”.**

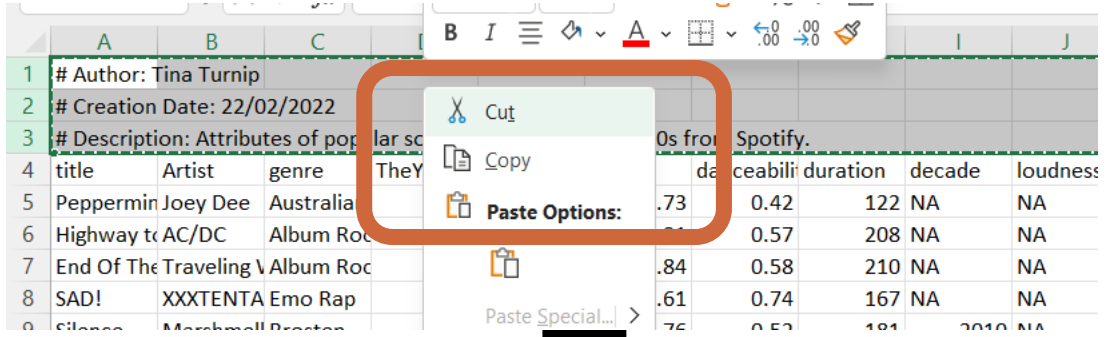
Then press OK.

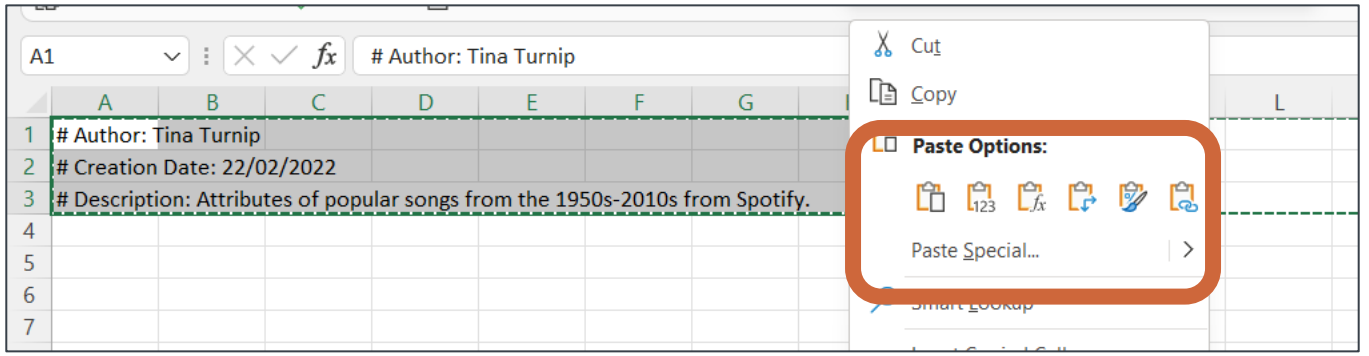


You will now have copy of the dataset in a standard Excel file.

How to remove metadata in Excel

If you have a dataset that has metadata attached to it, it is best practise to **cut and paste it into a different tab** of your workbook.



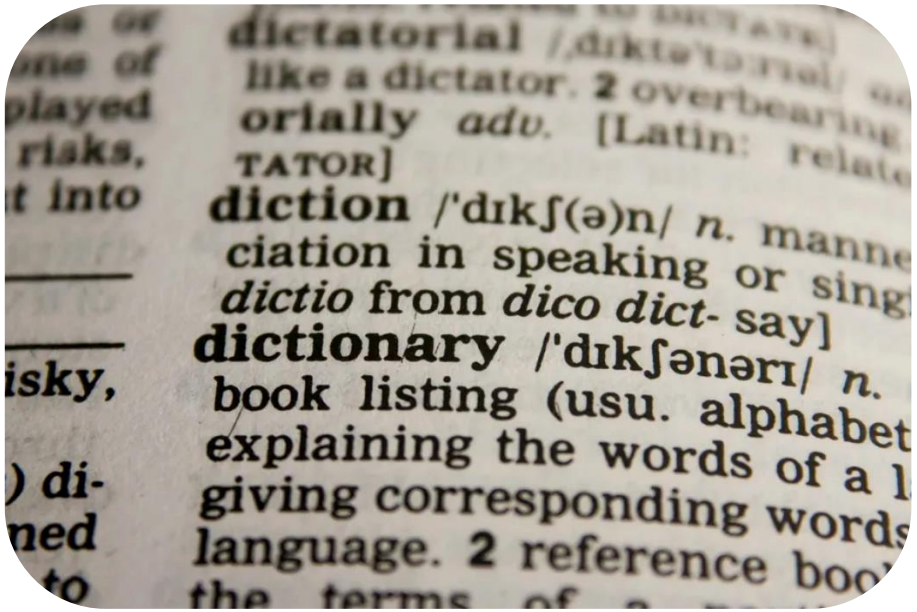


Next steps

Complete **questions 1 to 4**
of **section 1** in the
‘Practise Data Understanding
in Excel’ workbook.

Data dictionary

The data dictionary for this dataset is stored as a csv file and can be copied/imported into your workbook in the same way as the dataset.



The data dictionary contains for following information for each variable,

- **name**
- **data type**
- **definition**

Why use a data dictionary? A data dictionary give you a **good overview of each variable** in the dataset.

Reminder of Excel functions

As well as reviewing associated data dictionaries to support your understanding of the dataset, you can use Excel functions to help you understand the following properties of the dataset,

- Size
- Missing values
- Outlying values

Size of a dataset 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

Maximum and minimum in Excel

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

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

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

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()

Identifying the format of the dataset

What is the format of the dataset: **wide** or **long?**

Why identify the format of the dataset? The analysis you want to perform using the dataset will determine the shape you need your data in. Knowing whether the dataset is wide or long will determine whether you need to **reshape** it or not.

You can identify the format of the dataset through **visual inspection**.

wide

id	x	y	z
1	a	c	e
2	b	d	f

long

id	key	val
1	x	a
2	x	b
1	y	c
2	y	d
1	z	e
2	z	f

Source: Garrick Aden-Buie's Tidy Animated Verbs

Next steps

Complete **questions 1 to 12**
of **section 2** in the
‘Practise Data Understanding
in Excel’ workbook.

Learning checklist

I can *import* a dataset and remove any metadata in Excel

I can *use* a data dictionary to find out about a dataset

I can *find* the shape, size and format of datasets, using Excel

I can *identify* outliers and missing values 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