Advanced data cleansing 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 **advanced data cleansing** in Excel, specifically,

- how to **convert** between different data types

- how to **fix strings**

- understand the reasons **why there may be missing or outlying values**, and how these reasons affect the ways in which we handle them

Background

In the analysis steps so far, we have looked at,

- Data understanding
- Some of the activities you need to complete during the data tidying and cleansing stage

We are now going to look at the activities involved in **advanced dataset cleansing**.

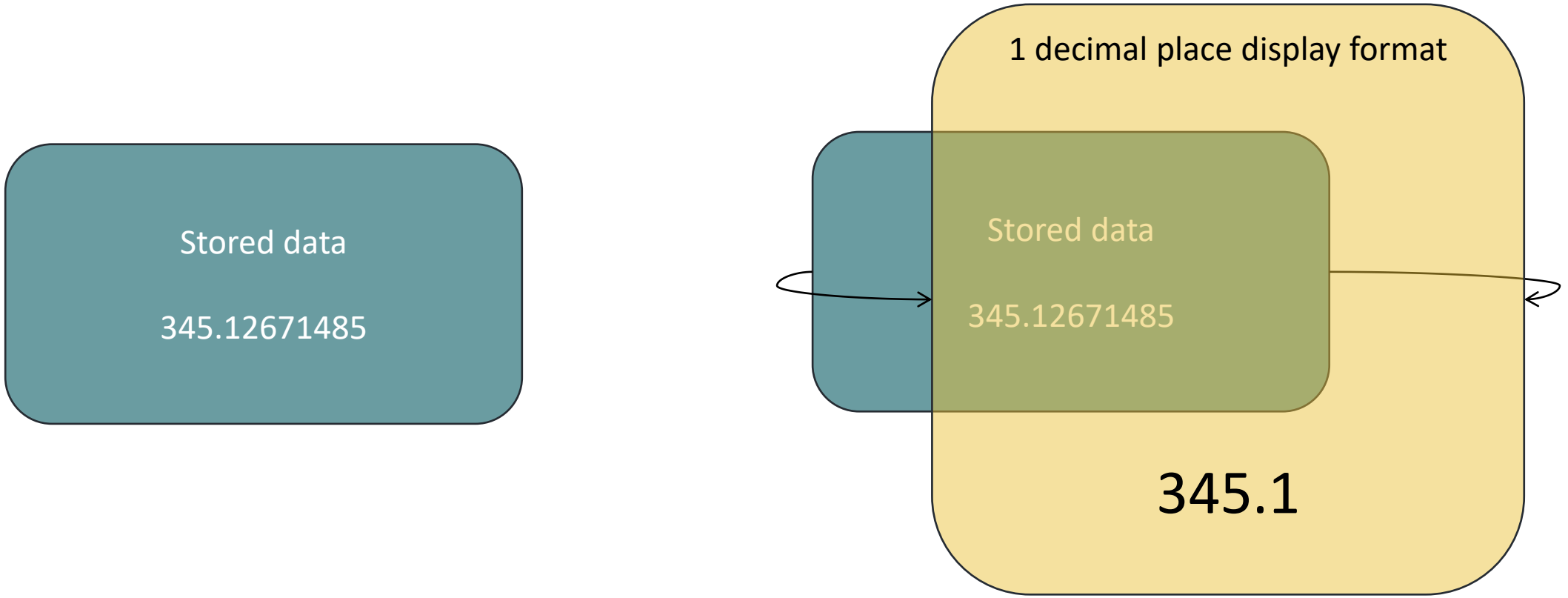
Converting data type

Sometimes a dataset will contain data items that have a data type that will make the next stages of the analysis steps difficult.

Therefore, changing the display format is not enough, the **stored data type needs to be converted**.

Reminder: Display formats vs. data types

The display format only changes how the data appears to the end user. The data stored in the computer stays the same.



Reminder: Data types in a dataset

The table below is a reminder of the different data types you might come across in a dataset.

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

Definition

Convert

To change a data item from one data type to another

Show me…

When data is entered, **Excel tries to assign an appropriate data type** to it. However it is not always suitable.

1/8

1 August

44774

If you have a data item that contains “/” (e.g. a fraction) Excel will think you have entered a date. It will then store the date as the number of days passed since the 'epoch' date.

Show me…

When an address with a flat number is in entered in Excel, (for example 3/1) it might convert the flat number to a date.

Flat 3/Room 1
27 Poppy Lane
The Meadows
Edinburgh

3/1
27 Poppy Lane
The Meadows
Edinburgh

03- Jan
27 Poppy Lane
The Meadows
Edinburgh

Show me…

The first dataset has phone numbers stored as floating point numbers. They need to be converted to strings so that they will start with zero (0).

phone_number
800111999.00
8001111.00
2072194272.00
1313485000.00
1312259846.00
1414205000.00

Convert from floating point to string

phone_number
0800 111 999
0800 1111
020 7219 4272
0131 348 5000
0131 225 9846
0141 420 5000

Why is converting data types important?

If numerical data is stored as strings you won't be able to perform calculations on them (e.g. average)

Phone numbers need to be stored as strings, otherwise the leading zero (0) will not appear.

Create date time variables to allow date time calculations to work correctly.

Converting data best practice

Before converting data items you should **always save a copy of the original** file.

How to convert from string to integer

We are going to look at how to convert variables **from strings to integers.**

We are going to use a dataset that contains the list of the most popular dog names. Although **count** and **rank** look like integers, they are actually stored as strings.

DogName	Count	Rank
BELLA	115	1
LUCY	80	2
SADIE	78	3
MAX	76	4
BUDDY	74	5
BAILEY	64	6
CHARLIE	63	7
DAISY	61	8
JACK	59	9
GINGER	58	10
MOLLY	58	10
LILY	56	12
BEAR	51	13
ROXY	50	14
LUNA	49	15

Source: https://www.kaggle.com/yamqwe/dog-names-over-timee

How to convert from string to integer

Step 1.

Set up a new column that will hold the converted data items.

	A	B	C	D	E	F	G
1	**DogName**	**Count**	**Rank**			**dog_name**	**count_integer**
2	BELLA	115	1			BELLA	
3	LUCY	80	2			LUCY	
4	SADIE	78	3			SADIE	
5	MAX	76	4			MAX	
6	BUDDY	74	5			BUDDY	
7	BAILEY	64	6			BAILEY	
8	CHARLIE	63	7			CHARLIE	
9	DAISY	61	8			DAISY	

Value in Excel

In Excel, you can convert a string to an integer using the VALUE function.

```
= VALUE(text)
```

For more information, please see

https://support.microsoft.com/en-us/office/value-function

How to convert from string to integer

Step 2.

=VALUE(*text)*

Use the **VALUE** function to convert the string data items to integers.

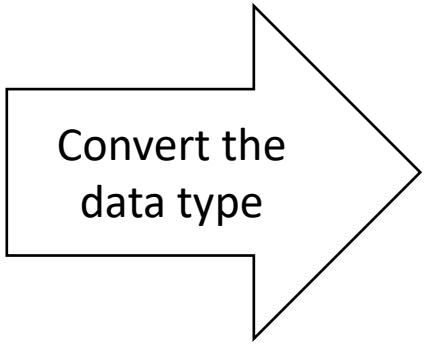
The **Count** variable appears to be the same when you look at it. However Excel will now handle it as a integer rather than a string.

	A	B	C	D	E	F	G
1	**DogName**	**Count**	**Rank**			**dog_name**	**count_integer**
2	BELLA	115	1			BELLA	=VALUE(B2)
3	LUCY	80	2			LUCY	80
4	SADIE	78	3			SADIE	78
5	MAX	76	4			MAX	76
6	BUDDY	74	5			BUDDY	74
7	BAILEY	64	6			BAILEY	64
8	CHARLIE	63	7			CHARLIE	63
9	DAISY	61	8			DAISY	61

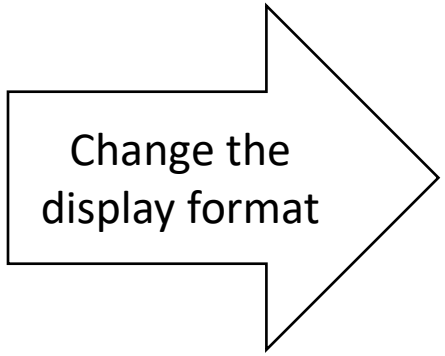
Using VALUE function with currency

The VALUE function will convert strings that look like currency to integers. You can then change the display format to make it look like a currency value.

price_string
£12.50
£9.99
£4.50
£3.45



price_convert
12.50
9.99
4.50
3.45



price_display
£12.50
£9.99
£4.50
£3.45

How to convert from string to datetime

Next, we will look at how to convert variables **from strings to datetime.**

We are going to use a dataset that contains the temperature recorded on Mars. The variable **earth_time_date** is a string, but the next stage of the analysis needs it as a date.

earth_date_time	mars_date_time	max_ground_ temp	min_ground_ temp
2022-01-26 UTC	Month 6 - LS 163°	-3	-71
2022-01-25 UTC	Month 6 - LS 163°	-3	-72
2022-01-24 UTC	Month 6 - LS 162°	-4	-70
2022-01-23 UTC	Month 6 - LS 162°	-6	-70
2022-01-22 UTC	Month 6 - LS 161°	-7	-71
2022-01-21 UTC	Month 6 - LS 161°	-8	-71
2022-01-20 UTC	Month 6 - LS 160°	-4	-72

Source: https://www.kaggle.com/deepcontractor/mars-rover-environmental-monitoring-station

Reminder: Extract in Excel

To convert from strings, you may need to extract part of the variable first.
As a reminder, here are some examples of calculations you can use in Excel to extract data.

Formula	Description	Example	Result
LEFT	Extracts a given number of characters from the **left side**	=LEFT(“apple”,3)	“app”
MID	Extracts a given number of characters from a **defined starting point**	=MID(“abcde”,2,3)	“bcd”
RIGHT	Extracts a given number of characters from the **right side**	=RIGHT(“red”,2)	“ed”

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

How to convert from string to datetime

Step 1.
The variable **earth_date_time** is a string. To convert it to a datetime we need to remove the 'UTC' from each of the data items.

Use the extract function **LEFT** to extract the first part of the string that contains only the date part of the variable.

Although the result looks like a date, it is still stored as a string.

earth_date_time	
2022-01-26 UTC	=LEFT(H2,10)
2022-01-25 UTC	2022-01-25
2022-01-24 UTC	2022-01-24
2022-01-23 UTC	2022-01-23
2022-01-22 UTC	2022-01-22
2022-01-21 UTC	2022-01-21
2022-01-20 UTC	2022-01-20

=LEFT(text,[num_chars])

Date value in Excel

In Excel, you can convert a string so that Excel recognises it as a date.

```
= DATEVALUE(date_text)
```

For more information, please see

https://support.microsoft.com/en-us/office/datevalue-function

How to convert from string to datetime

Step 2.
Once you have extracted the part of the string you need for the data. Use the DATEVALUE function to convert the date from a string to a datetime.

H	I
earth_date_time	**extracted_earth_date_time**
2022-01-26 UTC	=DATEVALUE(LEFT(H2,10))
2022-01-25 UTC	25 January 2022
2022-01-24 UTC	24 January 2022
2022-01-23 UTC	23 January 2022
2022-01-22 UTC	22 January 2022
2022-01-21 UTC	21 January 2022
2022-01-20 UTC	20 January 2022
2022-01-19 UTC	19 January 2022

=**DATEVALUE**(**LEFT**(*text,* [num_chars])

Next steps

Complete **questions 1 to 8**
in **section 1** of the
‘Advanced dataset cleansing in Excel’ workbook.

Fixing strings

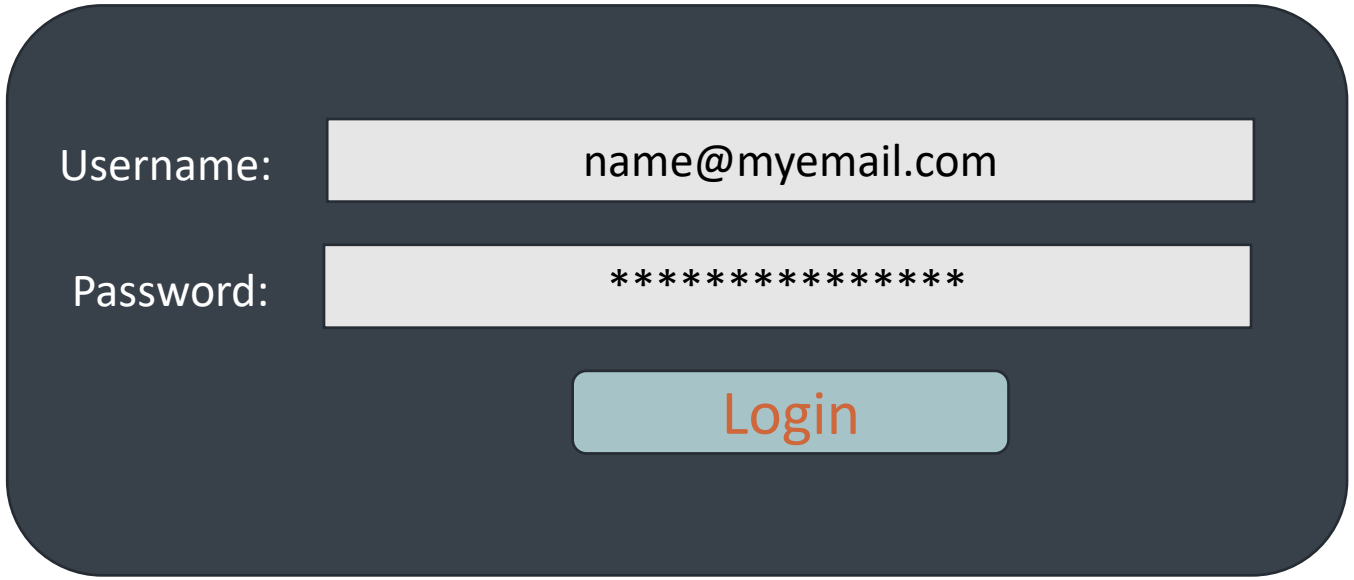
Reminder: String is a collection of characters combined to create alphanumeric text

Two strings may appear the same when compared by a person, but a computer may treat them as being different.

We are going to look at **how to fix strings** so they are consistent and can be compared to each other.

Show me…case-sensitive

Most computer languages are case-sensitive. This means they treat upper and lower case characters differently.



Usernames and password strings are often case-sensitive.

name@myemail.com is not the same as NAME@MYEMAIL.COM

How to change the case of a string

In Excel, there are built in functions that allow you to change the case of a string.

Formula	Description	Example	Result
PROPER	Capitalises the first letter in a string and any other letters in text that follow any character other than a letter. Converts all other letters to lowercase letters.	=PROPER(*"heLLo woRld"*)	Hello World
LOWER	Converts all uppercase letters in a text string to lowercase.	=LOWER(*"heLLo woRld"*)	hello world
UPPER	Converts text to uppercase.	=UPPER(*"heLLo woRld"*)	HELLO WORLD

For more information, please see

https://support.microsoft.com/en-us/office/change-the-case-of-text

Worked example

You need to send out letters to venues around Scotland, however the format of data is all in lower case.

By using **PROPER and UPPER you can convert the data items** into the standard form for writing a letter.

street	town	county	postcode
castlehill	edinburgh	midlothian	eh1 2ng
the helix	falkirk	falkirk	fk2 7zt
castle wynd	stirling	stirlingshire	fk8 1ej
abbey street	melrose	scottish borders	td6 9lg

Change case

street	town	county	postcode
=PROPER(A2)	=PROPER(B2)	=PROPER(C2)	=UPPER(D2)
The Helix	Falkirk	Falkirk	FK2 7ZT
Castle Wynd	Stirling	Stirlingshire	FK8 1EJ
Abbey Street	Melrose	Scottish Borders	TD6 9LG

Show me…extra spaces

Strings can have **extra spaces before, after or in the middle** of the characters. This will make it difficult to complete the next stages of the analysis steps.

film_title
The Imitation Game
A Beautiful Mind
I,Robot
Minority Report
A.I. Artificial Intelligence
Moneyball
21

Your turn…

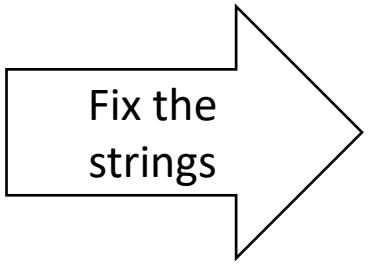
The dataset below has been created by summarising the details of cars sold by a garage. Why do you think there are two rows for the **car_type** “SUV”?

car_type	number_sold
Convertible	12
SUV	10
SUV	15
Minivan	2
Sports car	5
Coupe	8
Sedan	16

Your turn…

There are extra spaces in the beginning of the SUV string. By removing the extra spaces the dataset can be summarised correctly.

car_type	number_sold
Convertible	12
SUV	10
SUV	15
Minivan	2
Sports car	5
Coupe	8
Sedan	16



car_type	number_sold
Convertible	12
SUV	25
Minivan	2
Sports car	5
Coupe	8
Sedan	16

How to remove additional spaces in a string

In Excel you can remove spaces you don’t need by using the trim function.

It removes all spaces from text except for single spaces between words.

= TRIM(*text)*

For more information, please see

https://support.microsoft.com/en-us/office/trim-function

Show me…removing extra spaces

By using the TRIM function you can remove any additional spaces.

film_title
The Imitation Game
A Beautiful Mind
I,Robot
Minority Report
A.I. Artificial Intelligence
Moneyball
21

TRIM

film_title
The Imitation Game
A Beautiful Mind
I,Robot
Minority Report
A.I. Artificial Intelligence
Moneyball
21

Next steps

Complete **questions 1 to 7**
in **section 2** of the
‘Advanced dataset cleansing in Excel’ workbook.

Fixing missing and outlying values - advanced

In dataset understanding, we **identified** missing/outlier values.

In the dataset cleansing, we looked at **how to fix** missing/outlier values by removing, replacing or leaving them.

In this lesson we can going to look some of the **reasons why** we get missing/outlier values and how this helps us **decide how to handle them.**

Missing/outlier value thought process

When looking at missing/outlier values it can useful to think about these questions,

Do I have any missing/outlier value(s)?

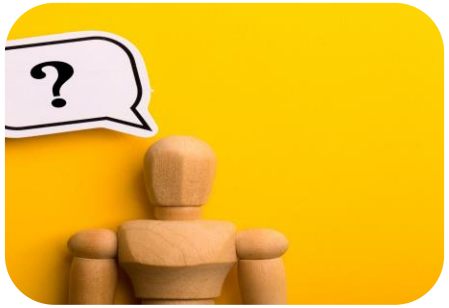
What caused the missing/outlier value(s)?

How should I handle the missing/outlier value(s)?

Why you might have missing/outlier values?

Here are some situations that might cause missing/outlier values in your dataset.

Data entry error



People didn't fill in the question in a survey

Definitions have changed

True outlier/missing value

Calculation error

Show me…data entry error

As part of a health questionnaire, patients have been asked to fill in their height in metres.

However, Lee has filled in their **height** in feet rather than metres. This has caused an **outlier**.

It has been caused by a data entry error.

name	height_m
Rowan	1.68
Sam	1.85
Jamie	1.73
Lee	5.00

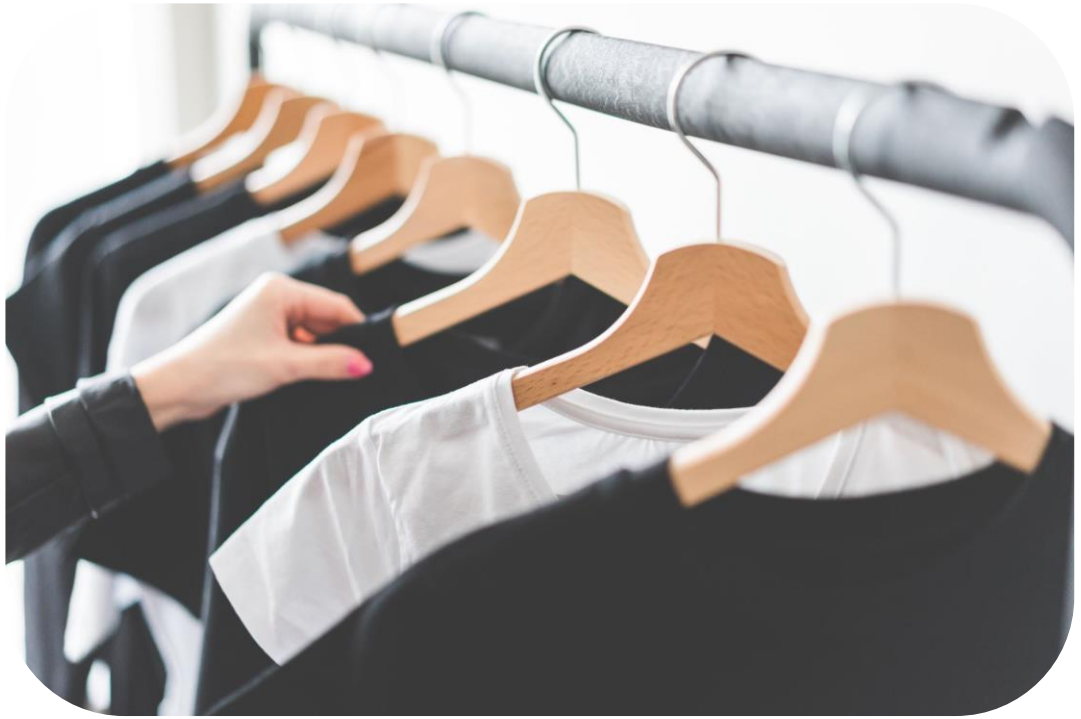
Show me…survey questions

In surveys, people can skip questions (leading to missing values in your dataset) if they do not feel comfortable answering them. Especially if they are not sure how that information will be used.

Thank you for buying a top from us.

Can you tell us, what is your household income?

£0-£15,000
£15,001 - £25,000
£25,001 - £45,000
£45,000+



Show me…definition changed

This dataset shows the maximum temperature by day.

On Monday, the ground temperature was recorded. Whereas, on Tuesday to Sunday the air temperature was recorded.

All the data is correct, but **how it has been measured has changed** which has caused outliers.

Day	MaxTemp
Monday	35
Tuesday	22
Wednesday	18
Thursday	17
Friday	15
Saturday	15
Sunday	13

Show me…true missing value

This dataset records the details of customers phoning a call centre.

It shows the **start time of each call** and whether the **employee hung up the phone** on the customer.

In this case, **a missing value is an useful piece of information** (and should be kept in the dataset) as generally you don't want employees to hang up the phone on customers.

call_start_time	hung_up
10:00	Yes
10:05	
10:45	
11:22	Yes
11:23	

Show me…calculation error

This dataset shows a calculated variable **average_mph**.

average_mph = length_miles/time_hours

journey_id	length_miles	time_hours	average_mph
1	500	10	50
2	75	3	25
3	60	2	30
4	45	5	9
5	0	0	#DIV/0!

In the last row, the calculation is dividing by 0. This results in a **error message causing a missing value**.

Worked example

Do I have a missing/outlier value?

What caused the missing/outlier value?

How should I handle the missing/outlier value?

This dataset is being used to send promotional letters to customers.

There are missing/outlier values in variables **address_3**.

address_1	address_2	address_3
3/1	42 Flower Lane	Dumfries
12 Bluebell Road	Edinburgh	
689 Rose Road	Glasgow	

Worked example

Do I have a missing/outlier value?

What caused the missing/outlier value?

How should I handle the missing/outlier value?

The missing values in **address_3** are blank as they are true missing values as all the address information is in the other variables.

address_1	address_2	address_3
3/1	42 Flower Lane	Dumfries
12 Bluebell Road	Edinburgh	
689 Rose Road	Glasgow	

Worked example

Do I have a missing/outlier value?

What caused the missing/outlier value?

How should I handle the missing/outlier value?

The missing values in **address_3** are true missing values so can be left as there are.

address_1	address_2	address_3
3/4	42 Flower Lane	Dumfries
12 Bluebell Road	Edinburgh	
689 Rose Road	Glasgow	

Your turn…

This dataset has a missing value caused by calculation error. How do you think you should handle it,

a) Leave it
b) Remove it
c) Replace it?

test_1	test_2	average_test
100%	50%	75%
90%	80%	85%
60%	100%	
50%	90%	70%

Data entry error

People didn't fill in the question in a survey

Definitions have changed

True outlier/missing value

Calculation error

Your turn…

c) Replace it

The calculation error can be fixed so the missing value should be replaced.

Test_1	Test_2	average_test
100%	50%	75%
90%	80%	85%
60%	100%	80%
50%	90%	70%

Data entry error

People didn't fill in the question in a survey

Definitions have changed

True outlier/missing value

Calculation error

Next steps

Complete **questions 1 to 6**
in **section 3** of the
‘Advanced dataset cleansing in Excel’ workbook.

Learning checklist

I can *convert* between different data types in Excel.

I can *fix* a string in Excel.

I can *describe* what might cause a missing value and why this is important for fixing them.

I can *describe* what might cause an outlying value and why this is important for fixing them.

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