Advanced data cleansing in Excel (Answers)

Worksheet section	Contents
1	Converting data types
2	Fixing strings
3	Fixing missing and outlying values

Version: 1.0

This lesson has been created by Effini in partnership with Data Education in Schools, The Data Lab and Data Skills for Work, with funding from the Scottish Government.

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

1. Converting data types

Section 1.1 (recall)

1) Can you fill in the gap in this definition of convert.

Convert To change a data item from one [data type] to another.

2) Can you think of any reasons why converting data items is important? State at least one reason.

1. If numerical data is stored as strings you won't be able to perform calculations on them (e.g. average)
2. Phone numbers need to be stored as strings, otherwise the leading zero (0) will not appear.
3. Create date time variable to allow date time calculations to work correctly.

Section 1.2 (define)

3) Can you explain the difference between changing the data type and the display format?

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

Section 1.3 (apply)

4) The variable **height_m** is stored as a string. Using the VALUE() function convert it from a string to an integer.

building_or_statue	height_m	height_m_integer
Scott Monument	61	61
The Glasgow Tower	127	127
Shanghai Tower	632	632
Empire State building	381	381
Nelson's column	51.6	52
The Shard	310	310

5) This dataset contains the launch dates of space probes and the date they flew past Jupiter. You need to work out how long it took them to get to Jupiter.

Enter the formula, **jupiter_flybe** - **launched** into the variable **time_to_jupiter**.

deep_probe_name	launched	jupiter_flybe	time_to_jupiter
Voyager 1	Sept 1977	March 1979	=E40-D40
Voyager 2	August 1977	July 1979	=E41-D41
New Horizons	Jan 2006	Feb 2007	=E42-D42
Pioneer 11	April 1973	Dec 1974	=E43-D43

Why do you think the formula has not worked?

The dates are strings so the formula is also a string.
Note for teachers: This is caused by Excel trying to assign a data type to a cell. In this case it thinks the cells with formula are also strings.

1. Converting data types

Using the DATEVALUE() function convert the dates from strings to datetime values then calculate the **time_to_jupiter_days**.

deep_probe_name	launched	jupiter_flybe	time_to_jupiter_days
Voyager 1	01/09/1977	01/03/1979	546
Voyager 2	01/08/1977	01/07/1979	699
New Horizons	01/01/2006	01/02/2007	396
Pioneer 11	01/04/1973	01/12/1974	609

6) This dataset shows the swimming speed of the fastest fish, you need to calculate the average swim speed.

Using the LEFT() function to extract the number from the **swimming_speed** data items.

Then use the VALUE() to convert the extracted number from a string to an integer.

animal	swimming_speed	speed_extract	speed_integer
Sailfish	110 kph	110	110
Striped marlin	80 kph	80	80
Blue-fish tuna	71 kph	71	71
Blue shark	69 kph	69	69
Swordfish	64 kph	64	64
Average	**#DIV/0!**	**#DIV/0!**	**79**

7) When this dataset has been extracted the phone numbers have been extracted as integers rather than strings.
Using the & function to add a 0 (zero) at the front of each of the phone numbers.

By using the & function to combine data items, Excel will automatically convert them to a string.

Location	PhoneNumber	PhoneNumber2
Gas emergency number	800,111,999	0800111999
Childline	8,001,111	08001111
House of Commons	2,072,194,272	02072194272
Scottish Parliament	1,313,485,000	01313485000
Edinburgh Castle	1,312,259,846	01312259846
Glasgow Science Centre	1,414,205,000	01414205000

1. Converting data types

Section 1.4 (active)

8) This dataset below is looking at events that happened in the 21st Century. Add in 5 more events that happened in the 21st Century to the dataset.

Then using the DATEVALUE() function to convert the data type, then in the date_display_format variable change the display format to make the dates easier for the user to understand.

event_21st_century	date	date_convert	date_display_format
e.g. Facebook is formed	4 Feb 2004	38021	04 Feb 2004
			0
			0
			0
			0
			0

2. Fixing strings

Section 2.1 (Recall)

1) Below are descriptions of functions you can use in Excel to change the case of strings in Excel. Can you fill in the name of the function that match the descriptions?

Name	Description
UPPER	Converts text to uppercase.
LOWER	Converts all uppercase letters in a text string to lowercase.
PROPER	Capitalises the first letter in a text string and any other letters in text that follow any character other than a letter. Converts all other letters to lowercase letters.

2) What is the function that is used in Excel to remove the extra spaces before, after or in the middle of strings?

TRIM()

Section 2.2 (apply)

3) This dataset contains the names of mountains in Scotland. However the strings are in a mix of cases. Using the PROPER() function, change the case of these mountains.

mountain_name	mountain_reformatted
BEN Lomond	Ben Lomond
SCHIEHALLION	Schiehallion
BEN More	Ben More
BEN Nevis	Ben Nevis
ben LAWERS	Ben Lawers

4) Using the PROPER() function, change the names of these people from upper case to proper case.

famous_scots	famous_scots_reformatted
ELSIE INGLIS	Elsie Inglis
VICTORIA DRUMMOND	Victoria Drummond
MARY SOMERVILLE	Mary Somerville
FRANCES RIGHT	Frances Right

2. Fixing strings

Section 2.1 (Recall)

5) These months have been all entered in lower case, however for the next stage of the analysis they need to be in upper case. Using the UPPER() function, change these months.

month	month_reformatted
jan	JAN
feb	FEB
mar	MAR
apr	APR
may	MAY
jun	JUN
jul	JUL
aug	AUG
sep	SEP
oct	OCT
nov	NOV
dec	DEC

6) This dataset contains famous phrases related to data, but they are difficult to read due to the extra spaces. Use the TRIM() function to remove the extra spaces.

famous_phrases	famous_phrases_trimmed
If the statistics are boring, you've got the wrong numbers	If the statistics are boring, you've got the wrong numbers
Without data you're just another person with an opinion	Without data you're just another person with an opinion
Not everything that can be counted counts and not everything that counts can be counted.	Not everything that can be counted counts and not everything that counts can be counted.

7) A survey of the number of trees by type has been conducted in 3 forests. The first dataset contains the results and the second contains a summary of the total number of trees by type.

2. Fixing strings

Section 2.1 (Recall)

forest	tree_type	num_trees
Eden	Scots pine	145
Northwood	Scots pine	14
Eden	Oak	854
Northwood	Oak	66
Westfield	Oak	12
Eden	Scots pine	1
Northwood	Ash	32

tree_type	total_trees
Scots pine	145
Scots pine	15
Oak	866
Oak	66
Ash	32

The UNIQUE() function and SUMIFS() have been used to create the summary dataset. Scots pine and Oak are listed twice. What do you think has caused this?

There are extra spaces in Scots pine and in Oak.

Use the TRIM() function to remove the extra spaces in the tree_type_trim variable below. Notice what happens to the summary dataset below as you enter the TRIM() function.

tree_type	tree_type_trim	num_trees
Scots pine	Scots pine	145
Scots pine	Scots pine	14
Oak	Oak	854
Oak	Oak	66
Oak	Oak	12
Scots pine	Scots pine	1
Ash	Ash	32

2. Fixing strings

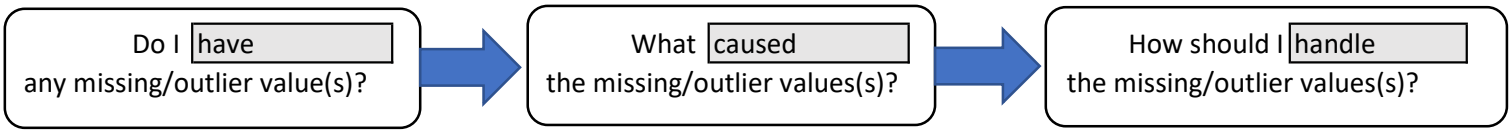
Section 2.1 (Recall)

tree_type_trim	total_trees
Scots pine	145
Oak	866
Ash	32

3. Fixing missing and outlying values

Section 3.1 (Recall)

1) In the lesson we went through a thought process you can follow for reviewing missing/outliers. Can you fill in the gaps in the thought process?



2) If an outlier has been identified as part of the Data understanding stage of the analysis steps, there are 3 options for fixing it in the data cleansing stage. Can you write down the 3 options you have for fixing an outlying value?

You can replace, remove or leave as is.

Section 3.2 (Apply)

3) This dataset contains details of people who have won the Nobel prize.

firstName	surname	born	died
Elinor	Ostrom	08/07/1933	06/12/2012
David	MacMillan	16/03/1968	NULL
William	Ramsay	2/10/1852	23/07/1916
Alexander	Fleming	06/08/1881	11/03/1955

Does the dataset have any missing values?

Yes, David MacMillan has a missing value.

What could have caused the missing value? Fill in the table below (with Yes or No) depending on whether you think it could have caused the missing value.

	Yes/No	Reason
Data entry error	No	
People didn't fill in the question in a survey	No	
Definition has changed	No	
True outlier/missing value	Yes	David is still alive so it's a true missing value
Calculation error	No	

How do you think you could handle this missing value?

Leave as it is, as it's a true missing value.

3. Fixing missing and outlying values

4) This dataset contains the world population split by urban and rural since 1960.

pop_type	year	world_population
urban	2020	43,789,939,440
rural	2020	3,415,804,795
urban	2000	2,868,307,513
rural	2000	3,275,186,310
urban	1980	1,754,201,029
rural	1980	2,703,802,485
urban	1960	1,023,845,517
rural	1960	2,011,104,231

Do you have any missing or outlying values?

Yes, the urban world_population in 2020 is an outlier.

What could have caused the missing/outlier value? Fill in the table below (with Yes or No) depending on whether you think it could have caused the missing/outlier value.

	Yes/No	Reason
Data entry error	Yes	0 has been added to the end of the number.
People didn't fill in the question in a survey	No	
Definition have changed	No	
True outlier/missing value	No	
Calculation error	No	

How do you think you could handle this missing value?

Replace the number with the correct number.

5) This dataset is an extract from a survey completed by customers visiting a café.

ID	What is your first name?	What is your last name?	Which of the following age groups are you in?	Were you happy with the café?
1	Rosie	Love	25-44	Yes
2	Greg	Hill	65 over	Yes
3	Molly	Jones	Under 25	No
4	Isla	Well	45-64	Yes
5	.	.	Prefer not to say	.
6	No	No	45-64	Yes
7	Rachel	Rock	Under 25	Yes
8	Barry	McNeil	45-64	Yes

Do you have any missing or outlying values?

ID 5 has missing values.

What could have caused the missing/outlier value?

People didn't want to fill in their details

3. Fixing missing and outlying values

How do you think you could handle this missing/outlier value?

Remove the row as there is no value to any of the data.

Section 3.3 (Active)

5) This dataset contains the number of pages and the genre of books. Add the details of 3 books you like to the list.

book	num_pages	genre
War and peace	1225	Romance novel
Winnie the Pooh	160	Children's fiction
To kill a mockingbird	281	Southern gothic
Jane Eyre	592	Romance novel
Heart of Midlothian	469	Historial fiction

Do you have any missing or outlying values?

War and peace is an outlier, but based on the books entered by the learners there may be more.

What could have caused the missing/outlier value?

Depends on the data entered by the learner.

How do you think you could handle this missing/outlier value?

Most likely they are true missing/outlier values, but depends on the data entered by the learner.