

# Advanced data cleansing in Excel



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.

© 2022. This work is licensed under a [CC BY-NC-SA 4.0 license](#).



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.

**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  to another.

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

## Section 1.2 (define)

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

## 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	
The Glasgow Tower	127	
Shanghai Tower	632	
Empire State building	381	
Nelson's column	51.6	
The Shard	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	
Voyager 2	August 1977	July 1979	
New Horizons	Jan 2006	Feb 2007	
Pioneer 11	April 1973	Dec 1974	

Why do you think the formula has not worked?

## 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			
Voyager 2			
New Horizons			
Pioneer 11			

- 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		
Striped marlin	80 kph		
Blue-fish tuna	71 kph		
Blue shark	69 kph		
Swordfish	64 kph		
Average	#DIV/0!	#DIV/0!	#DIV/0!

- 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	
House of Commons	2,072,194,272	
Scottish Parliament	1,313,485,000	
Edinburgh Castle	1,312,259,846	
Glasgow Science Centre	1,414,205,000	

# 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?

<u>Name</u>	<u>Description</u>
<input type="text"/>	Converts text to uppercase.
<input type="text"/>	Converts all uppercase letters in a text string to lowercase.
<input type="text"/>	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?

### 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	
SCHIEHALLION	
BEN More	
BEN Nevis	
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	
VICTORIA DRUMMOND	
MARY SOMERVILLE	
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	
feb	
mar	
apr	
may	
jun	
jul	
aug	
sep	
oct	
nov	
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	
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.	

## 2. Fixing strings

Section 2.1 (Recall)

- 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.

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?



## 2. Fixing strings

Section 2.1 (Recall)

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.

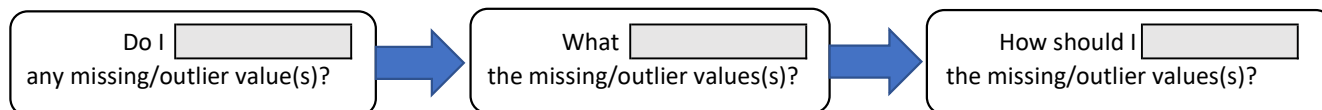
tree_type	tree_type_trim	num_trees
Scots pine		145
Scots pine		14
Oak		854
Oak		66
Oak		12
Scots pine		1
Ash		32

tree_type_trim	total_trees
0	0
	0
	0

### 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?

#### 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?

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
Data entry error	
People didn't fill in the question in a survey	
Definition has changed	
True outlier/missing value	
Calculation error	

How do you think you could handle this missing value?

- 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

3. Fixing missing and outlying values

Do you have any missing or outlying values?

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
Data entry error	
People didn't fill in the question in a survey	
Definition have changed	
True outlier/missing value	
Calculation error	

How do you think you could handle this missing value?

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?

What could have caused the missing/outlier value?

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

Section 3.3 (Active)

6) 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

### 3. Fixing missing and outlying values

---

Do you have any missing or outlying values?

What could have caused the missing/outlier value?

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