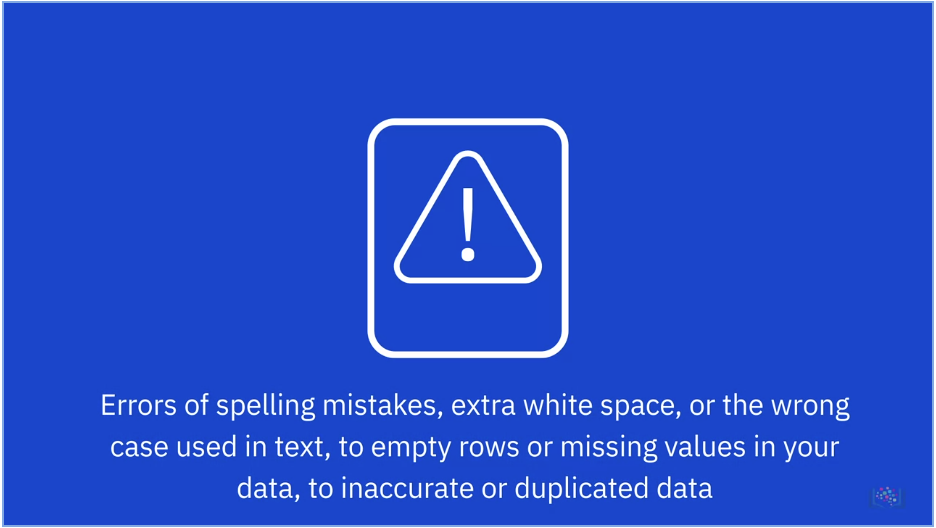
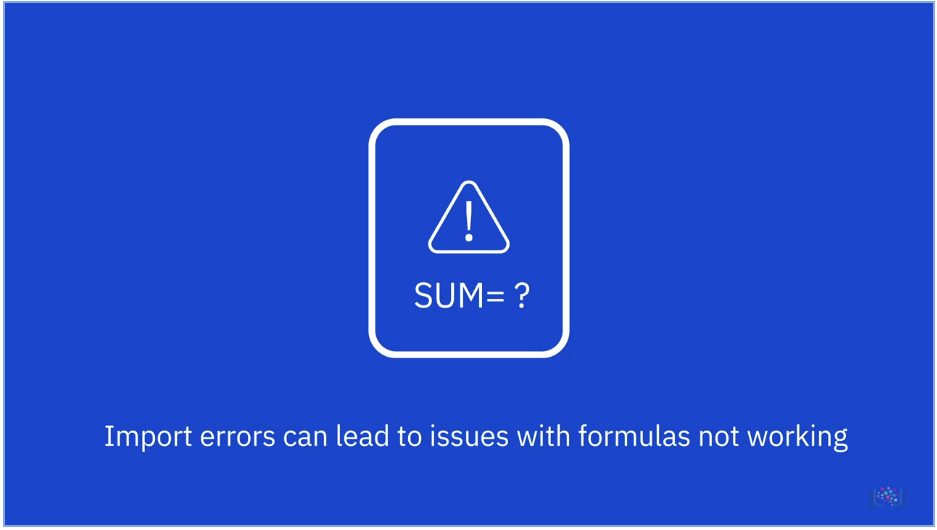
Cleaning Data

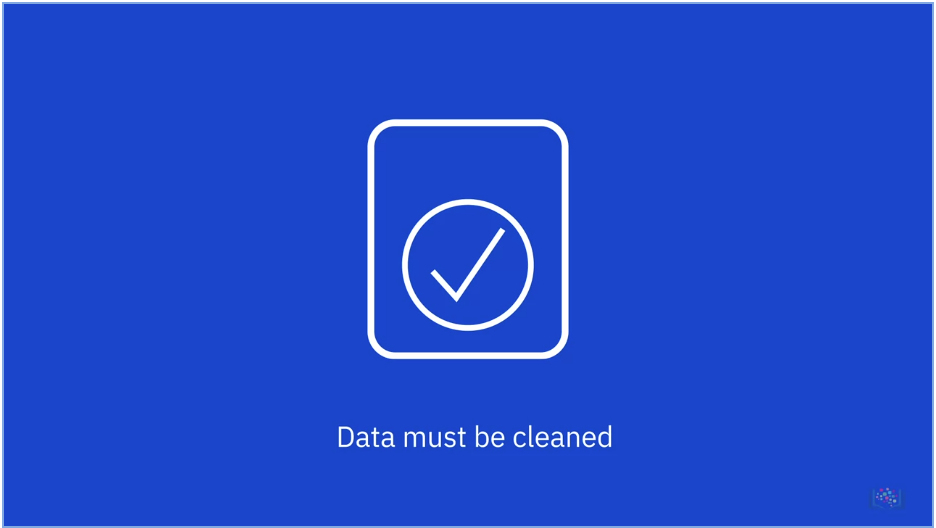
**Removing Duplicated or Inaccurate Data and Empty Rows**











1. Spelling check – Select the column > Review tab > Proofing Group > Spelling

2. To **remove empty rows** in much better way (if we have more than hundreds of empty rows):

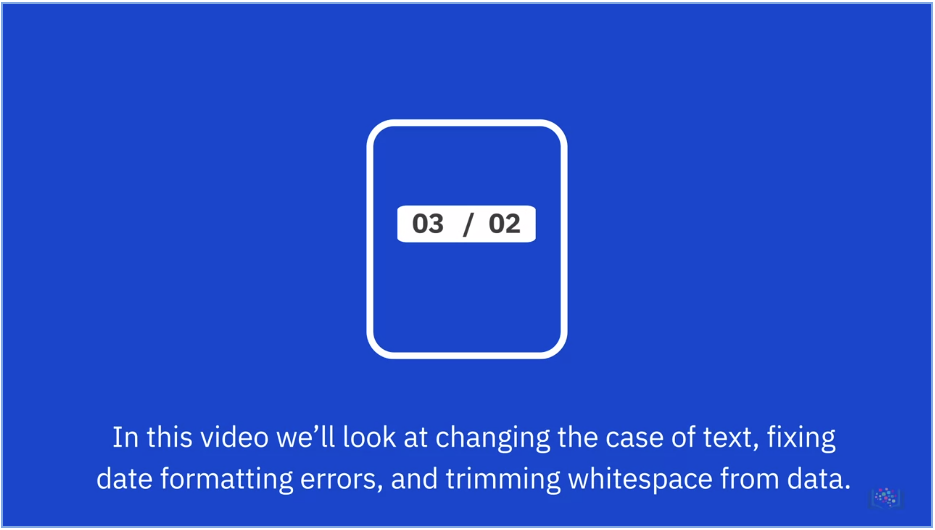
1. Select all of the data (Can use Ctrl + Home to go to the A1 cell) included header by pressing Ctrl + Shift + End keyboard shortcut.
2. Select Filter icon on the Data tab, then we can see each column has a filter icon
3. Choose any one of the columns that doesn’t contain empty cells, click the dropdown list
4. Uncheck ‘Select All’, check the item called ‘Blanks’. This will now show only the empty rows at the top of our sheet
5. We can now select these rows, either using the mouse or going to the first cell in the first empty data row, and then using the Ctrl + Shift + End keyboard shortcut then right-click the selected rows and then click Delete Rows.
6. Finally, on the Data tab, click Clear, then click Filter.

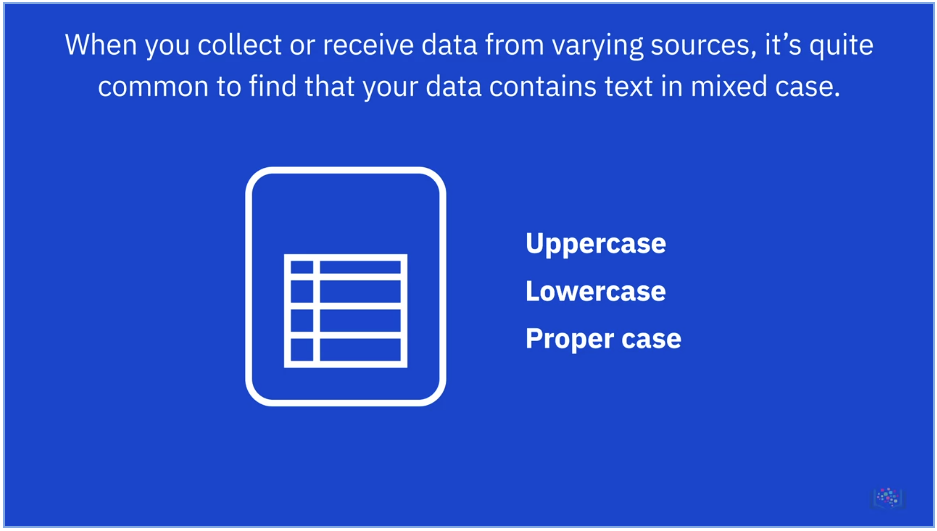
3. The next inconsistency we’ll look for is **duplicated rows of data**. It’s quite common for duplicate data rows to exist in your imported data, caused either by human input error, or an error in the import process. There are two ways of doing this in Excel:

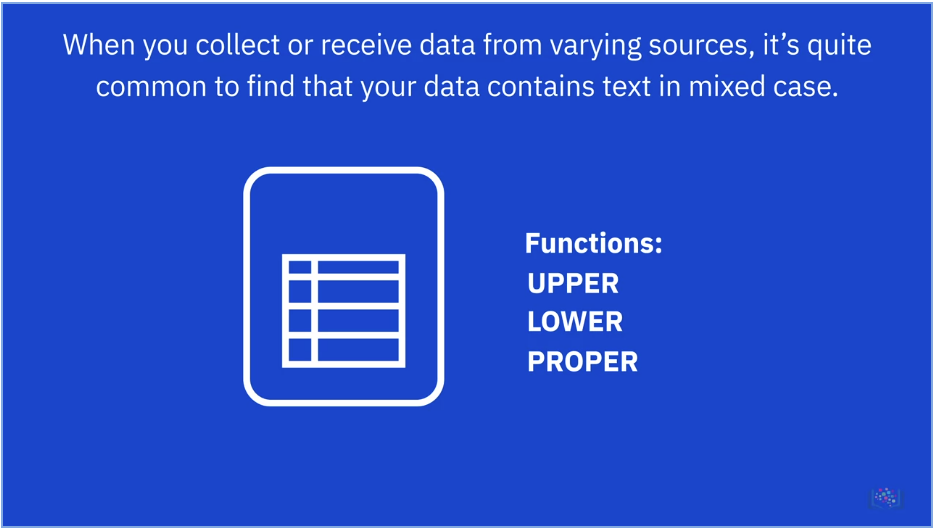
* First method of removing duplicated rows:
* It’s important to select a column of data that you would NOT expect to have duplicate values in. For example, if we consider the Price Unit column, we would expect lots of these values to be repeated, because the unit price of some products is the same, so this is a bad example of a column to use to find duplicates. Instead, we should use the Sales Amount column because it is far less likely that these values will be duplicated in the normal process of things, as they are the total sales for each order.
* So, we select the Sales Amount column > Home tab > Conditional Formatting > Highlight Cells Rules > Duplicate Values. When we click OK, and scroll down the sheet, we can see that only a few values have been identified as being duplicates.
* Second method of removing duplicated rows (Simpler but less secure):
* Press Ctrl + Shift + End to select whole dataset > Data tab > Data Tools Group > Remove Duplicates > Unselect All > check only Sales Amount (same logic as first method)

4. Find and replace – Ctrl + H

**Dealing with Inconsistencies in Data**







Excel doesn’t have a Change Case button like there is in Microsoft Word, so you need to use other methods to perform this data cleaning task. Those methods are functions, namely the UPPER(), LOWER(), and PROPER() functions.

1. Insert a blank column/row beside/below, use the functions: UPPER(), LOWER(),

PROPER() or TRIM().

1. Copy the results and paste on the original column/row with ‘Paste Value option’
2. Delete the row/column with function

For example, change the headers in uppercase to sentence case:

* A1’s row is header, in cell A2, type =PROPER(A1) and press Enter.
* Hover over the bottom-right corner of cell A2, and drag the Fill Handle across to the last column.
* If dragging across is too difficult with the mouse, then select the cells in the row 2 using SHIFT+RIGHT ARROW, then press F2 to put the cursor focus back in cell A2, then hold CTRL while you press Enter.
* Select row 2, then press CTRL+C.
* Select row 1, Right-click and choose Paste Options>Values.
* Select row 2, right-click it and choose Delete Rows.

For whitespaces in the between the text, using Ctrl + H - find 2 empty spaces and replace by 1 empty space.

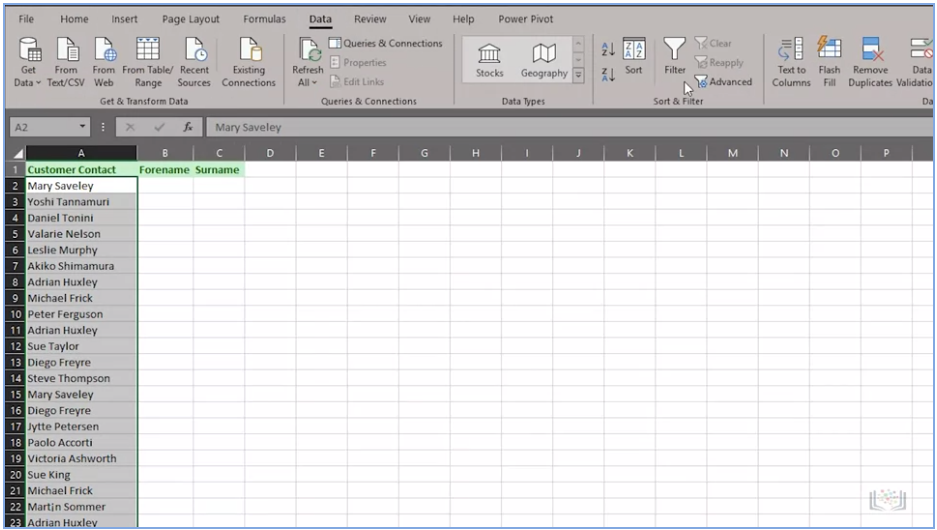
**More Excel Features for Cleaning Data**

*Especially for full name split to first names and last names or vice-versa, there are 3 methods to do this.*

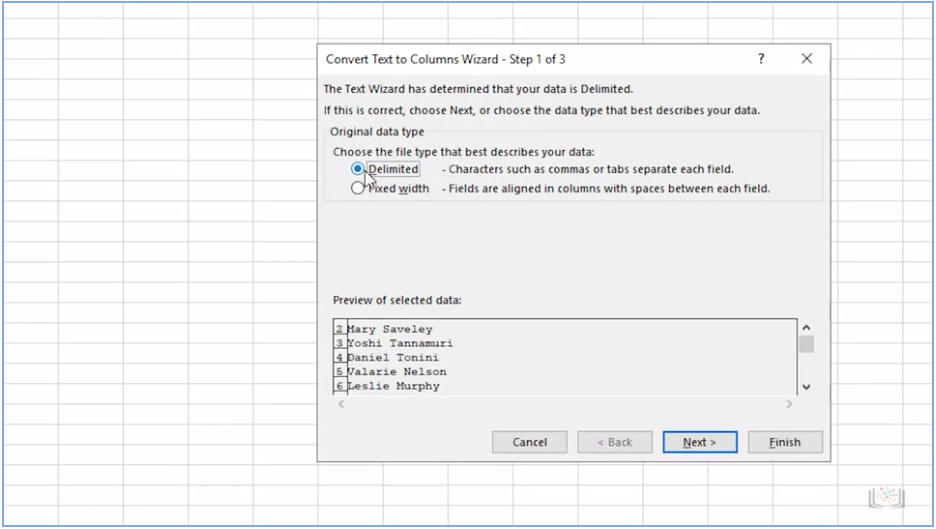


* Method 1: We used Flash Fill briefly earlier as a quick method of entering data that fits a specific pattern, such as the names of months or days of the week, but it can also be useful as a data cleaning tool. It can split a column of full names into two separate columns for the first names and last name, or vice-versa and it can also help to modify the naming convention used in a column of names. Flash Fill can be found on Data tab > Data Tools > Flash Fill or press Ctrl + E

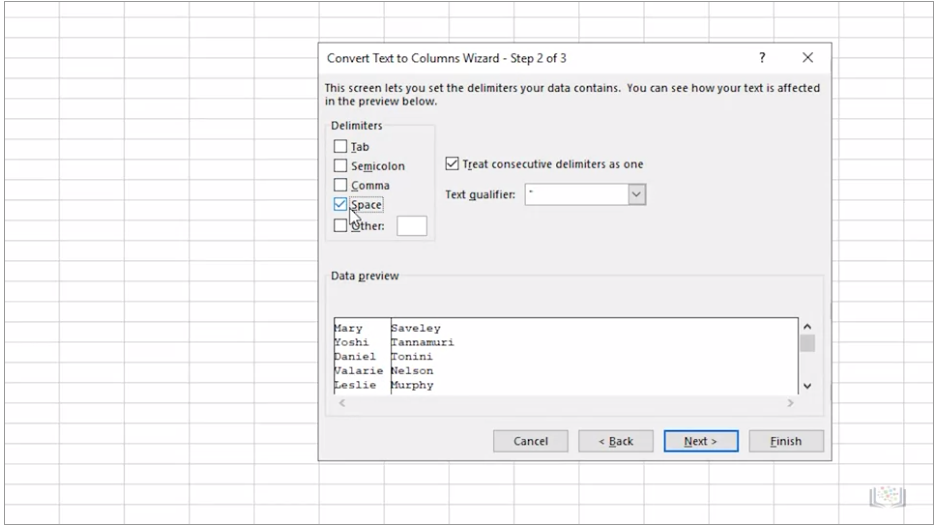
Method 2: There’s also a bit more flexibility with functions, which can be especially useful if you have names that are complex and mixed, such as having hyphenated names or some names with a middle initial, some with two middle initials, and some with no middle initial.



* Full name to be separated into first names and last names.
* Highlight the texts we need to be separated > Data tab > Data Tools > Text to Columns

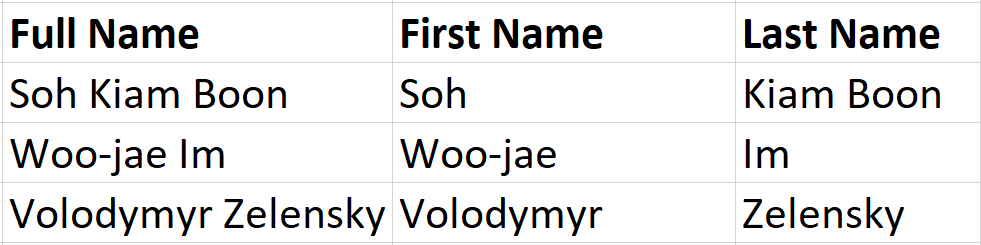


* Convert Text to Columns Wizard will be opened > Ensure Delimited is selected > Next



* Uncheck Tab > Check the Space > Next > Destination (Make sure is absolute reference or locked with $ signs) > Use cursor to select the first cell we need to be converted > Finish

Method 3: We also can achieve the same result using functions.



For first names: =LEFT(A2, FIND(" ", A2, 1) -1)

For last names: =RIGHT(A2, LEN(A2)-FIND(" ", A2, 1))