**DigiSyska – Excel**

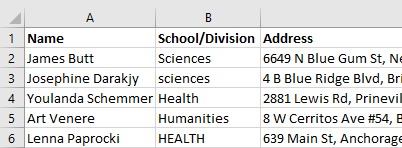
4 September 2019

**Part 1 – Filtering and Deduplicating**

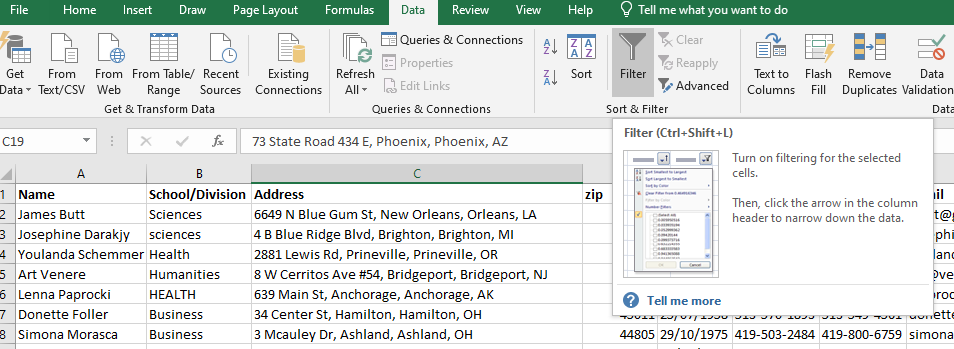
**Filtering**

The *Filter* tool allows you to sort, filter and narrow down your data by hiding parts of it from view. Filtering allows you to qualify and display only the data that interests you. For example, you could filter a list of students based on School/Division.

Step 1: Ensure your data has a header row that identifies the contents of each column.

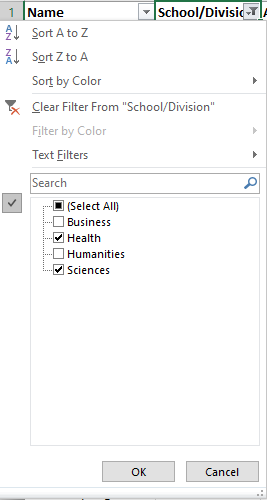


Step 2: Select the *Data* Tab and click Filter.



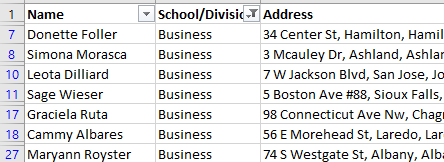
Drop down arrows will appear in the column headers – note there is no need to select the headers beforehand, Excel will identify the header row automatically.

Step 3: Click the dropdown arrow for the column you want to filter. In this example, we will filter using the School/Division column. Uncheck the boxes of the data you don’t want to view. In this case we only want to view students in Science and Health.

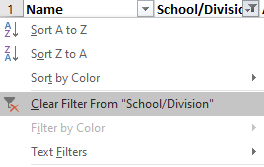


When you have made your selection click ok. To undo this redo Step 3 and choose the *Select All* option.

Note: There are two methods to identify if there is a current filter present. Firstly, the drop down for the column will appear as a funnel. Secondly, the Row count on the left hand side will show missing rows.



To clear all filters from the worksheet, select the drop down with the active filter and then click *Clear Filter From “XXX*”



**p**

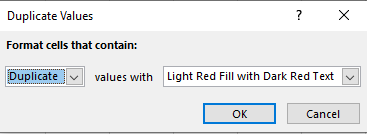
**Deduplicating**

Excel allows you to highlight duplicate data within cells. In this example, we wish to identify if there are duplicate Student entries in our data.

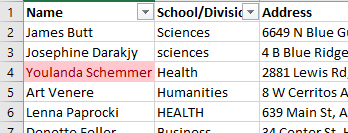
Step 1: Ensuring there are no filters present, Select the range of cells you wish to check for duplicate data. In this case we select the entire *Name* column by clicking above the header row.

Step 2: From the *Home* tab select *Conditional Formatting >Highlight Cell Rules > Duplicate Values.*

From the pop up box select OK.



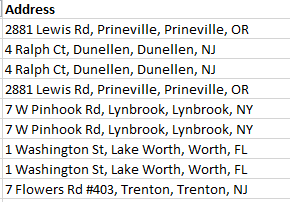
As you scroll through your list you will find duplicates show in Red. You can investigate these and delete where necessary.



Tasks1: You will notice that the duplicates aren’t listed together. How would you group them together easily to see them? Note that there are multiple ways to do this.

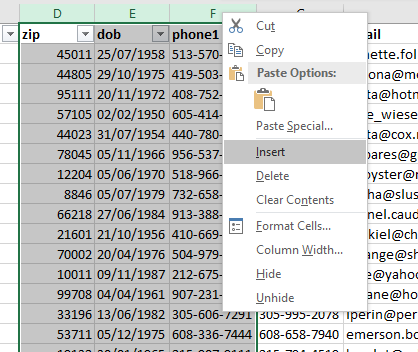
**Part 2 Text to Columns**

The *Text to Columns* tool allows you to split one cell into a number of columns based on a chosen delimiter. In this case we want to separate the address into the multiple columns to fix an error with the data. The third value in the Address information is incorrect and needs removed. As the data is constantly formed with commas we are able to delimit using comma.



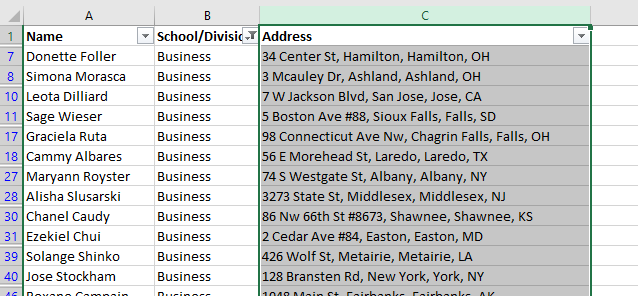
Step 1: To split the data into columns we need to insert the correct number of columns to the right. We can see in our data we have 4 comma delimited pieces of data (Address, City, incorrect Information, State), so we will need three columns to the left (the current plus 3).

Select the three columns (above the header) to the right of the address and click *Insert.*

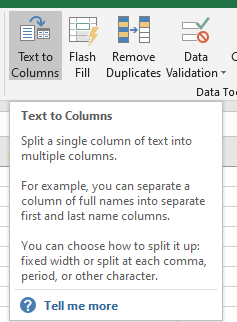


Note: If you don’t have enough empty columns to the right, when you perform the *Text to Columns* operation it will overwrite the data in the adjacent columns.

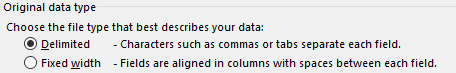
Step 2: Select the column you want to split by clicking above the header. In our case this is the address column.



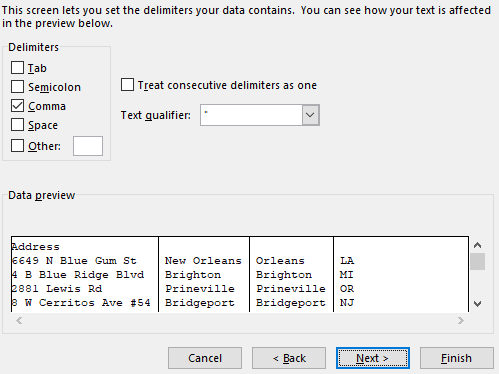
Step 3: From the *Data* tab choose the *Text to Columns* button (ensure the data you want to delimit is selected)



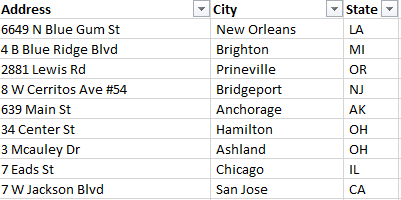
Step 4: Ensure the *Delimited* radio button is checked and click *Next*



Step 5: Ensure the *Comma* tick box is selected. You can preview the operation in the window below. It is a good idea to use the slider to look at the data to see that it has made the changes as you would expect.



Step 6: Click *Finish*. The data should be slip into the additional columns and you can delete the erroneous information. Remember to give your new columns a title.

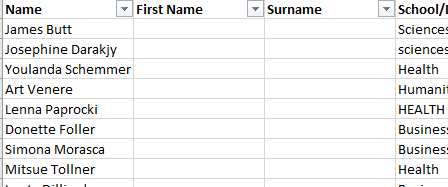


**Part 3 Flash fill**

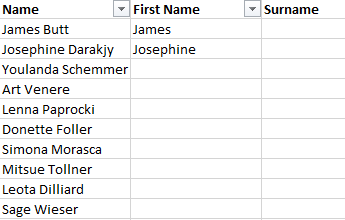
Flash fill is Excel’s attempt at machine learning and artificial intelligence. It tries to mimic repetitive tasks so that you can focus on what you want rather than attempting complicated functions. In our use case today, we want to use tidy data practices and have our name split into separate fields. This is the most simple use case for flash fill, but we will try some additional tasks once we have mastered this.

For this section please ensure all filters are reset.

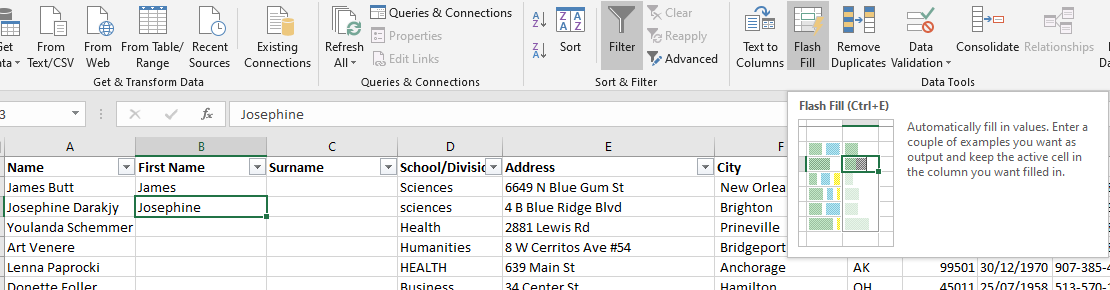
Step 1: Firstly we need to have two blank columns beside name, 1 for First Name, and 1 for last name. Add the columns and give them a title.



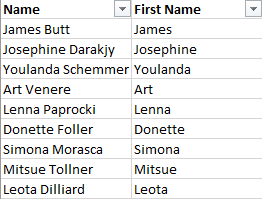
Step 2: This is where we train Excel to what we want. In the *First Name* column type the first name of the student. As this is a simple task Excel can work it out easily, but I usually like to do two so it has a full understanding of what is trying to be achieved.



Step 3: Select the last name you entered, and from the *Data* tab click the *Flash Fill*  button.



Excel will fill the remaining cells.



Task 1: Repeat with Surname

Task 2: Create a new field to create city and state in the following format: City, STATE

Task 3: Fix the capitalisation of the School/Division Field

Task 4: Create a new phone field, and remove the – from phone1

Task 5: Create a new university email address. We will assume the format is first.last@student.otago.ac.nz

Note: Remember to delete any old columns, add headings to new columns, and clean up and working data. It’s all about tidy data practice!