Data Preparation

Data Preparation Step 1: Find the dirty data

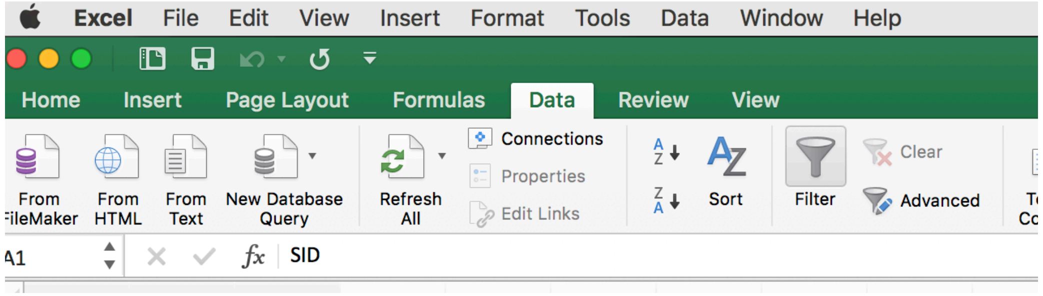
- **If you are editing a .csv file, open it in Excel and save as an Excel workbook first**

- Some things to look out for:
  - Are there missing (empty) rows or values?
  - Are there text values where numerical values should be (or vice versa)?
  - Are there invalid values? If we know a certain value should fall within a range, we can check for invalid values. Eg: If the data contains room temperature in Melbourne, logically, we know the values should be within 0 – 30 C (assuming there was no fire!).
  - Are there NULL values recorded?
  - Are there duplicate values?
  - Do the Column names make sense?
  - Does the data match the column label?

Data Preparation Step 1: Find the dirty data

- **Many ways to do this – You are free to select methods/tools that work for you.**

- You can use Excel’s built in Sort & Filter to find certain dirty fields.



- Hint: If you **sort** a column ascending, the lowest values will be on top & the highest at the bottom. This can easily show you if you have values out of range, if you have empty or NULL values.
- Here is a link with some useful Excel functions: https://multimedia.journalism.berkeley.edu/tutorials/cleaning-data/
- You can also check the data manually – or if you know how to code, write a script to check the data, or use any other tool you like.

Data Preparation Step 2: Clean the Dirty Data

You can :

- Remove the rows containing dirty data
- Replace the dirty data with default values
- Replace the dirty data with calculated values (if possible)
- Rename column names to make more sense …
- …. (you can think of many more ways depending on the context)

More info…

- https://medium.com/towards-data-science/data-cleaning-101-948d22a92e4
- https://www.ringlead.com/flush-bad-leads-improve-data-quality/#.WW7EdhN94k8