LESSON HANDOUT

Cleaning data with text manipulation

Why

Text manipulation will allow you to:

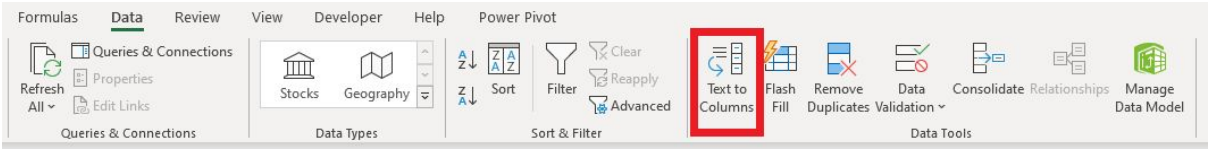
- Standardise classifications across different categorical text values
- Fix inconsistencies with issues such as spelling and incorrect formatting

Splitting by delimiter

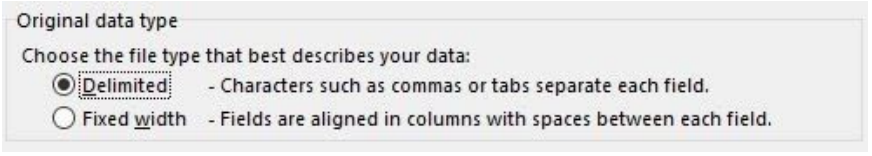
Why: Data is often stored in text files where values are separated by characters to enable import into multiple programs

How:

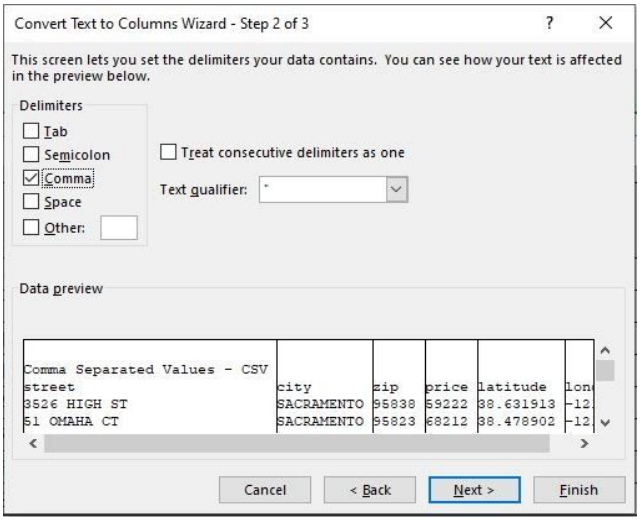
- Highlight cells
- In the ribbon, on the 'Data' tab, select "Text to Columns"



- Select delimited > next



- Choose the delimiter against which to split. The preview pane will show you what the data will look like.



- Choose the destination for the split data and finish.

Text manipulation functions

- **LEN** - Returns the number of characters in a text string (useful in combination with other functions)
    - =LEN(“text” or cell reference)
    - Example: =LEN(“Australia”) will return **9**, as “Australia” has 9 characters
- **TRIM** - Removes all spaces from a text string (except for the single spaces between words) - such as excess spaces before or after a word.
    - =TRIM(text to trim, or cell reference to trim)
    - Example: =TRIM(“ First Name ”) will return “First Name”
- **LEFT** - Returns the number of text characters from the left of the cell
    - =LEFT(Cell, number of characters to return from the left)
    - Example: =LEFT(“Australia”, 3) will return “Aus”
- **RIGHT** - The inverse of LEFT
    - =RIGHT(Cell, number of characters to return from the right)
    - Example: =RIGHT(“dataframe”, 5) will return “frame”
- **MID** - Returns a specified number of text characters from the middle of a cell
    - =MID(Cell, the number at which to start, the number of characters to return)
    - Example: =MID(“169productAUS”, 4 , 7) will return “product”
- **FIND** - Returns the starting number of a character or word within a text cell
    - =FIND(character to find, cell)
    - Example: =FIND("." ,"cental.station") will return **7**, as “**.**” is the 7th character within the string “central.station”.
- **CONCATENATE** - Groups text values together
    - =CONCATENATE(Cell 1 or value 1, Cell 2 or value 2, e.t.c)
    - Example: =CONCATENATE(“cheese”, “ “, “sandwich”) will return “cheese sandwich”

Combining text functions

- **FIND & LEFT/RIGHT/MID**- Use **FIND** to designate the start number for a **LEFT/RIGHT/MID** formula.
    - = MID(Cell, FIND(character, cell), no of characters to return)
        - **Note**: to return characters to the *right* of the character you’ve found using FIND, add (+1). See the example below.
    - Example: both of these formulas will return “rider”: the FIND formula finds the “.”, the MID formula starts at +1 character after and returns 5 characters.
        - =MID("free.rider", FIND("." , "free.rider") +1, 5)
        - =MID("joy.rider", FIND("." , "joy.rider") +1, 5)

- **LEN & MID** - Use **LEN** in combination with the **MID** formula to return the same centre text irrespective of the length of the string.

- =MID(Cell, start number, LEN(Cell))
- Example: both of these formulas will return the model of the phone. The MID specifies to start at the 4th character, then return the LENgth of the string (minus 6 characters - 3 at the beginning and 3 at the end).
    - = MID (“AUSiphone123”, 4, (LEN(“AUSiphone123”)-6))
    - = MID (“GBRmotorolla453”, 4, (LEN(“AUSmotorolla123”)-6))