Data Cheat Sheet: Cleaning Tricks

Freeing PDFS:

- Tabula: Open source application for turning PDF tables into CSVs
- Cometdocs: Web-based tool for turning PDFs into CSVs (not secure)
- Acrobat: OCR PDFs, if you can't search inside them.
- Preview or PDFsplit: Split or delete pages from PDFs easily (not secure)

Some Useful Formulas for Date Manipulation:

- =day(cell-reference) Returns the day of the month for a given date (cell-reference)
- =month(cell-reference) Returns the month for a given date
- =year(cell-reference) Returns the year for a given date
- =text(cell-reference,"dddd") Returns the day of the week for a given date
- Also useful: You can subtract dates and times just like you do numbers. Example A2-A3 where A2 and A3 are dates would tell you the difference in days between them.

Useful Keyboard Shortcut: To fill a formula all the way down on a Mac, even if there are blank cells present:

- Select the cell with the formula you want to fill down.
- Press shift + command + down arrow
    - This will highlight to the end of the column.
- Press command + d
    - This will fill the formula into the highlighted cells.
- To directly edit a cell: fn + f2

Code Books, Data Dictionaries and Meta Data: Supplementary material that is necessary to understand the data file. For example, this text file may explain that 1 means female and 2 means male, or include information about how the data was collected.

Data Cleaning Tricks:

- **Transpose**: One option under the paste special menu. Transpose flips your data so that the rows become the columns and the columns become the rows.
    - Popular with some government data sources, like BLS.
- **Separating data:** Data -> split text to columns. Good for separating data that follows a clear pattern like first and last name or URLs

- **Wildcards:** Can be used in filters or formulas to work around misspelled or messy data
    - * represents an unknown number of characters
    - ? represents a single character
    - ~ escapes either of the other wildcard characters
    - If you need even more options, try regex
- **=EXACT(value1,value2)** Returns true if two text values are equal and false if they aren't. Useful to check yourself when you are copy-pasting.