Data Wrangling Summary

Data wrangling template link

Gather

- Depending on the source of your data, and what format it's in, the steps i
- High-level gathering process: obtaining data (downloading a file from the page, querying an API, etc.) and importing that data into your programmi Jupyter Notebook).

Assess

- Assess data for:
  - Quality: issues with content. Low quality data is also known as dirty
  - Tidiness: issues with structure that prevent easy analysis. Untidy dat messy data. Tidy data requirements:
    1. Each variable forms a column.
    2. Each observation forms a row.
    3. Each type of observational unit forms a table.
- Types of assessment:
  - Visual assessment: scrolling through the data in your preferred soft Sheets, Excel, a text editor, etc.).
  - Programmatic assessment: using code to view specific portions and (pandas' `head`, `tail`, and `info` methods, for example).

Clean

- Types of cleaning:
  - Manual (not recommended unless the issues are single occurrences
  - Programmatic
- The programmatic data cleaning process:
  1. Define: convert our assessments into defined cleaning tasks. These an instruction list so others (or yourself in the future) can look at you it.
  2. Code: convert those definitions to code and run that code.
  3. Test: test your dataset, visually or with code, to make sure your clea
- Always make copies of the original pieces of data before cleaning!

Reassess and Iterate

- After cleaning, always reassess and iterate on any of the data wrangling s

Store (Optional)

- Store data, in a file or database for example, if you need to use it in the fu