Data Wrangling Summary

Data wrangling template link

Gather

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

Assess

- Assess data for:
    - Quality: issues with content. Low quality data is also known as dirty data.
    - Tidiness: issues with structure that prevent easy analysis. Untidy data is also known as 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 software application (Google Sheets, Excel, a text editor, etc.).
    - Programmatic assessment: using code to view specific portions and summaries of the data (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:

your work and reproduce 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 cleaning operations worked.

- 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 steps if necessary.

Store (Optional)

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

NEXT