Project Details

- Data wrangling, which consists of:
  - Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
  - Assessing data
  - Cleaning data
  - Storing, analyzing, and visualizing your wrangled data Reporting on:
    - your data wrangling efforts.
    - your data analyses and visualizations.

1. Gathering

the first step in data wrangling is to gather the data from diffirent sources to work with.

> downloading The WeRateDogs Twitter archive manualy to get twitter_archive_enhanced.csv.
>
> downloading The tweet image predictions programmatically, which contains breed of dog by using the Requests library and the given link and saving file to disk as .tsv file.
>
> Gathering data about each tweet like retweet count and likes Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

2. Assessing

the second step in data wrangling by assessing data in the 3 dataframes visually and programmaticallyby using pandas functions like:df.info(),df.describe(),df.value_counts(),df.head()...etc.

Quality

- Data quality dimensions help guide your thought process while assessing and also cleaning. The four main data quality dimensions are:
- Completeness: do we have all of the records that we should? Do we have missing records or not? Are there specific rows, columns, or cells missing?.
- Validity: we have the records, but they're not valid, i.e., they don't conform to a defined schema. A schema is a defined set of rules for data. These rules can be real-world constraints (e.g. negative height is impossible) and table-specific constraints (e.g. unique key constraints in tables).
- Accuracy: inaccurate data is wrong data that is valid. It adheres to the defined schema, but it is still incorrect. Example: a patient's weight that is 5 lbs too heavy because the scale was faulty.
- Consistency: inconsistent data is both valid and accurate, but there are multiple correct ways of referring to the same thing. Consistency, i.e., a standard format, in columns that represent the same data across tables and/or within tables is desired.

Tidiness

Tidy data is a standard way of mapping the meaning of a dataset to its structure. A dataset is messy or tidy depending on how rows, columns and tables are matched up with observations, variables and types.

- Each variable forms a column.
- Each observation forms a row.
- Each type of observational unit forms a table.

3. Cleaning

Cleaning the third step in data wrangling. It is where you fix the quality and tidiness issues that you identified in the assess step.

The very first thing to do before any cleaning occurs is to make a copy of each piece of data. All of the cleaning operations will be conducted on this copy so we can still view the original dirty and/or messy dataset later. Copying DataFrames in pandas is done using the copy method.

cleaning types used:

- Manual
- Programmatic
- The programmatic data cleaning process:

  1. Define: convert our assessments into defined cleaning tasks. These definitions also serve as an instruction list so others (or yourself in the future) can look at 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.