Create a 300-600 word written report called wrangle_report.pdf or wrangle_report.html that briefly describes your wrangling efforts. This is to be framed as an internal document.

My wrangling efforts consisted in gathering, assessing, and cleaning the data. Once I performed the preparation, I analyzed the information and found 3 insights from the WeRateDogs twitter trends.

Gathering

We have the following three sets of data:

- Twitter JSON TXT
    - Obtained from the Twitter API, I used the text provided in the classroom due to the amount of computation required to obtain the data live from Twitter. I got the tweet ID, favorite count, and retweet count
- Twitter Archive
    - Obtained from the WeRateDogs database, I used the rating numerator and rating denominator for the analysis
- Twitter Prediction
    - Obtained from the Udacity classroom's tweet predictions. I used the P1 to obtain the type of dog for each tweet ID

Assessing

Our next step will be assessing the data based on the following criteria:

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

I will be evaluating the data in two ways:

- 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).

My goal was to detect and document at least eight (8) quality issues and two (2) tidiness issues

These were some of the issues I found:

Quality Issues

- Rating denominator has values above and below 10, these need to be standardized to 10
- There are some row values with "RT" and a retweet_user_id and retweet_status. We don't want this for our analysis
- There some null values in the "name" column, these seem to not represent dogs

- "Flufffer" doesn't seem to be a dog category, the correct name is "floof"
- Some rating numerators are miscalculated (e.g. 9.75 as 75/10 instead of 9.75/10)
- Some columns could be reduced to ease the analysis (e.g. in_reply_to_status_id / user_id, retweeted columns, source)
- There are some expended_urls with vines instead of tweets, we don't want these for the analysis
- Some dog names only have one letter (e.g. "a")
- Rename ID columns to match across all tables
- The twitter prediction dataset has some values that aren't dogs, for the final dataset we want only dogs

Tidyness Issues

- Dog nicknames are not tidy, there are four columns for what could be a single column with the respective nickname
- Ideally, we want to have one table with the rating for each tweet, one with the favorite count and retweet count, and one with the predictions (3 total tables)

Cleaning

Clean each of the issues you documented while assessing. Perform this cleaning in wrangle_act.ipynb as well. The result should be a high quality and tidy master pandas DataFrame (or DataFrames, if appropriate). Again, the issues that satisfy the Project Motivation must be cleaned.

There are different ways to clean data:

- Manual (not recommended unless the issues are one-off occurrences)
- Programmatic

I will choose to clean it programatically, through the following steps:

- 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.
- Code: convert those definitions to code and run that code.
- Test: test your dataset, visually or with code, to make sure your cleaning operations worked.

Storing & Visualizing Data

Store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv. If additional files exist because multiple tables are required for tidiness, name these files appropriately. Additionally, you may store the cleaned data in a SQLite database (which is to be submitted as well if you do).

Analyze and visualize your wrangled data in your wrangle_act.ipynb Jupyter Notebook. At least three (3) insights and one (1) visualization must be produced.

Total list of dataframes:

- twitter_archive_c_rating
- twitter_archive_c_dog_type
- tweet_prediction_c
- tweet_df_c

Some insights I went to validate are:

1. Which dogs have the most amount of favorites or retweets?
2. Which type of dogs have the best ratings?
3. What is the spread in ratings across dogs? Are there major outliers?