Goal: Wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations

Sources:

Three different data sources were used to generate the master dataset:

1. Enhanced Twitter Archive csv file for basic Tweet information (Tweet ID, timestamp, and data extracted from the Tweet content)
2. Twitter's API for additional information (e.g. retweet and favorite counts)
3. Image Prediction File produced from a neural network to classify the Tweet's dog breed

Gathering:

The Enhanced Twitter Archive file was manually uploaded to the workspace and loaded to Jupyter via pandas read_csv() function. The Image Prediction File was programmatically pulled from an url. The API data gathering consist of requesting API access, creating a JSON file on the Twitter data found for Tweets listed in the Archive file, and loading the JSON data into a dataframe in Jupyter to assess the data.

Assessing:

Through a combination of programmatic and visual assessment, more than 10 issues were detected. All 3 sources include more than just original tweets, which is the focus of this project. Additionally, some sources are missing original tweets that are in others. Other glaring issues are erroneous data type for timestamps and twitter ID from the Archive data as well as nondescriptive column headers in the Image Prediction data. Further investigation of the Archive data reveals inaccurate ratings and invalid dog names that need to be fixed. Although the different dog stage columns should be consolidated to a single variable column, the data is not accurately populated and cannot be pulled from the text. Per the predictions, there should be around 70% dog tweets yet only 17% of the records are populated for dog stage. Therefore, there is limited value to clean up these columns. However, in terms of tidy data, rating information are in 2 columns and can be recalculated for a single column. Lastly, all the essential columns for analysis can be merged into one table as they all naturally relate to the tweet data.

Cleaning:

Archive data will be the base table for the analysis. First thing is to remove all tweets that are not original tweets. Then, all the nonessential columns should be dropped prior to merging the Archive table with essential columns from the other 2 tables. This inner join ensures that all essential columns are in one location and tweets that are not found in all 3 sources are dropped. Then, quick fixes are done to correct the erroneous data types and rename nondescriptive column headers. Afterwards, I extracted and replaced the names that I found from the Tweet text for some of the records with invalid names.

For the remaining records where names could not be found, I replaced the name to ‘None’ for consistency with other records without names. The final quality issue on the ratings required visual assessment on the 22 records that were potentially inaccurate. After identifying and correcting the rating, I created a new rating column to allow for easier analysis.