Wrangle and Analyze Data Project

Gathering Data for this Project

In the initial step, we need to get three pieces of data:

1. WeRateDogs Twitter archive
2. Tweet image predictions
3. Twitter Data using Twitter API

The WeRateDogs dataframe was easy enough to create as the file was given to us and can be created by using read_csv. For tweet image predictions, I used the Requests library and using the provided url, I was able to get the TSV file as needed, and then use the read_csv function to create the dataframe. For the extra twitter data using Twitter API, for privacy reasons I chose to use the provided tweet_json.txt file, then used a for loop to append a dictionary for each row.

After gathering the data, we have 3 data frames that we can work on:

1. **weratedogs_df** = Data frame from WeRateDogs Twitter archive
2. **image_df** = Data frame from Tweet image predictions
3. **api_df** = Data frame from Twitter Data using Twitter API

Assessing Data for this Project

Upon visual and programmatic inspection, I found that following quality issues and tidiness issues with the data frames:

Quality Issue

1. In weratedogs_df, the “name” column has 55 counts of “a” , 7 counts of “an” and 8 counts of “the”. These seem like data entry errors. These should be changed to “None”, or better yet as a NaN.

2. In weratedogs_df, we should remove rows that are retweets. Furthermore, after removing these rows, columns “retweeted_status_id” , “retweeted_status_user_id” and “retweeted_status_timestamp” are unnecessary as project detail states retweets are not considered for the project.

3. In weratedogs_df, the “timestamp” column is a string object. It’s probably a good idea to turn this into a timestamp.

4. In weratedogs_df, the “rating_denominator” column has 23 entries where it does not equal 10. In the project details, it states “These ratings almost always have a denominator of 10”, so other values would be incorrect in this column.

5. In weratedogs_df, the source column is messy due to it containing raw HTML data. We should extract just the text within the HTML tag.

6. In weratedogs_df, consider removing some of the outlier data scores that look suspicious (“1776”, “666”, “420”).

7. In image_df, names in prediction columns (p1, p2, p3) aren’t consistent with capitalization. Lowercase all names.

8. In image_df, column names could be more informative. Change column names to a more understandable title.

Tidiness Issue

1. In weratedogs_df, columns for dog states (e.g. “doggo”) should form one ‘dog_stage’ column, since this is one variable. Drop the four columns after the merge is complete.

2. All three tables should be merged into one dataset. Merge tables based on tweet_id.