Joshua Demeo
171220

Wrangle Report

I began my wrangling efforts by initially going through the WeRateDogs tweeter archive that was provided. I initially checked for missing values, any misrepresented values, columns of data that I found unhelpful for future analysis, and any columns that did not abide to a tidy format. The issues were addressed from what I thought would reduce the size of the data frame greatly before narrowing down singletons throughout the dataset. I began by removing observations related to reply tweets and retweets. I then proceeded to remove any observations that did not contain an expanded_url (tweet's web link) field. I proceeded to removing columns that I did not think would be useful for future analysis or were to be removed regardless such as retweet related data, which was requested to be removed for the project. Given columns (doggo, floofer, pupper, puppo) represented a value rather than a category, these four columns were melted into one column "dog_stage" taking into account any tweets that had multiple occurrences of the four dog stage values. Those were categorized as multiple.

Once the data set was reshaped, individual values in particular columns were adjusted. The dog names associated with each tweet had many values that were not proper names (a, an, the, etc.) and it seemed as though they were all lower case values. A regex pattern corresponding to a name with the first letter being lowercase was used to create a list of those improper name values in the dataset, subsequently identified and converted to "None" in the data set. A number of observations had improper rating denominators and numerators and their proper values were determined by individual inspection of each tweet's text given if the value seemed outside of the most common values for a rating. Once this was cleaned, an additional column of data was created corresponding to a rating score (rating numerator / rating denominator) which would be used later for visualizations an analysis.

Satisfied with the state of the WeRateDogs dataset, the dog classification dataset was downloaded. Data quality issues and tidiness were not apparent upon analysis of the dataset.

Finally, the last of the three datasets was obtained by using the Twitter API to obtain a list of status objects corresponding to the tweet ids found in the WeRateDogs dataset. Each status was then converted to a JSON object before being converted to a string. Once this file was created it was read back in with each JSON line added to a list and then converted to a Pandas data frame. From this point, it seemed like the file had only a few columns of interest ("id", "retweet_count", favorite_count") and was thus converted to its own data frame.

The three datasets were then merged based on matching tweet ids via an inner join. This removed any observations that were not common across all the datasets. Now that all necessary cleaning had occurred, this data frame was converted to a master CSV file, which was used to make visualizations and inferences on WeRateDogs tweets.