Wrangle Report – Seth Brown

To wrangle the data, we had to first gather the data. Different files had to be gathered in their own unique way. For the image prediction file, we had to use the given URL and call that URL to a folder and file we made. The other files we were able to read from a CSV file straight in the notebook using pandas.read function.

Once data had been gathered and data frames made, we needed to assess the data frames, text files and CSV files to understand our data and how to clean in. For the file provided (twitter-archive-enhanced.csv) we found many quality issues. Some of these were removing columns that were not needed for analysis to be able to view cleaner data and correcting the datatype of other columns in that file. Another thing to assess is the tidiness of the data frames. Some tidiness issues we found were that we could combine two of the data frames on the tweet id column to be able to analyze across one data frame instead of having to call two.

Cleaning data is the final stage in the wrangling process where we took our findings in the assess the data phase and used code to clean the data frame for better analysis. Although all of these steps are able to be done over and over again as needed. Starting withing tiding the data, we merged the tweet json data frame and the twitter archive data frame on twee id. This was done so we could analyze our twitter archive data using the retweet and favorite count information that was provided from the tweet json file. After we finished tiding, we started the cleaning process. All cleaning was done to the twitter archive data frame due to this data frame now being the data frame that the analysis would be done on. An extended list of the quality and tidiness problems that were corrected are listed below.

Quality

- timestamp is object instead of datetime
- tweet_id is float64 instead of str
- Columns that are not needed for analysis should be removed
- some of the names in name column are none
- Rating denominator of 313 is 0 and numerator is 960
- df (twitter_archive_enhanced) rate numerator and denominator can be one column called rating and dropping the other two
- retweet_status_timestamp is object instead of datetime
- retweet_status_id and user_id are float instead of string
- Drop max reading for rating

Tidiness

- df_2 (twee_json.txt) should be added to df (twitter archive enhanced)
- df (twitter_archive_enhanced) all different stages of dogs should be combined into one column