Wrangle Report

Jessica Vargas

For this project, I was tasked to wrangle and analyze WeRateDogs twitter data. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

Gathering data

The first task was to gather each of the three pieces of data: WeRatedogs Twitter archive, tweet image predictions, and each tweet's retweet count and favorite count. The first piece was manually downloaded while the tweet image prediction file was downloaded programmatically using the Requests library. Lastly, I retrieved retweet and favorite count data by querying the Twitter API for each tweet's JSON data using the Tweepy library. I then stored the JSON data into a text file tweet_json.txt.

Assessing & Cleaning

There were several quality issues I tackled in these 3 datasets. Before making any changes to each dataframe, I made copies to keep the original data unmodified. I made most of my modifications to the Twitter Archive dataset. I examined the data for null values and duplicates, cleaned the data and merged them together. Here were the issues I addressed:

Quality issues

Twitter Archive Dataset Changes

1. **Added ratings variable**
   I wanted to add a normalized ratings variable because there were several instances where the numerator was greater than the denominator. There are also many cases where the denominator had a value of something other than 10. There is no standard value, so I added a the ratings column which is essentially: (numerator/denominator)*100.
2. **Changed timestamp variable to datetime data type from object datatype**
3. **Dropped rows with invalid dog names**

   Since I did plan to use the name variable in my analysis, I dropped rows that contained invalid dog names. Most of these values were articles and words like 'None' instead of names.
4. **Removed retweeted ratings since we only wanted original tweets**
5. **Made corrections to some denominators/numerator ratings**

   I looked at Tweets that had denominator greater than 10 to validate and it turned out a lot of them were not valid values. These tweet text contained entries such as 7/11 (the store) or 4/20 (the date). I manually updated these values with the correct ratings after looking at the text.
6. **Made names all uppercase to make sure there were no duplicates (Susan v SUSAN)**
7. **Modified dog type values to 0/1s**

   I just replaced dog type values with 1s and the None value with 0 to be a little cleaner and more dummy data like.
8. **Updated json tweet column name to match other datasets**

   Changed tweet column id to tweet_id to match the other datasets before merging the 3 datasets together.
9. **Changed source value text to shortened value for easier reading**

The source variable from the Twitter Archive dataset contains 4 possible values. They are different links from different platforms. I replaced the links with iPhone, Web Client, TweetDeck and Vine.

Tidiness

1. **Removed extra columns**
   Dropped columns that I did not want to use in analysis
2. **Merged 3 different datasets into one single dataset**