WRANGLE REPORT

GATHERING THE DATA

The first thing I did when I started working on this project was to manually download the file called "twitter-archive-enhanced.csv" provided for us by Udacity and stored it in a dataframe called "twitter_archive". After that, I used the requests library to programmatically download the file "image-predictions.tsv" from Udacity's server and named the data frame "image_prediction". Lastly, I tried to use the Twitter API, but even after debugging, it didn't work. That's why I used the alternative JSON file that was given. I downloaded the tweet_json text file that was provided to us by Udacity and selected the columns I was interested in; id, retweet_count and favorite_count.

I got my 3 data frames; twitter_archive, image_prediction, tweet_json and I moved on to the assessing phase.

ASSESSING THE DATA

The three tables were found to have some quality and tidiness issues.

QUALITY ISSUES

twitter_archive table

1. Incorrect datatypes were found in the following columns; tweet_id, in_reply_to_status_id, in_reply_to_user_id, timestamp, retweeted_status_id, source, retweeted_status_user_id, retweeted_status_timestamp
2. typos in dog names like "a," "an," and "really" are not valid dog names.
3. some entries in some columns such as 'name' had a lot of missing values set to null
4. The source column had entries in html format which is non readable to humans
5. The tweet rating_numerator and denominator values consists of some high values or decimal values that are considered to be inaccurate

image prediction table

6. some columns are do not have proper descriptive names for clearer understanding, e.g., p1, p2
7. The jpg_url column has some duplicates, and these duplicates has different IDs.
8. Some entries have p1_dog, p2_dog, p3_dog set to false. These are not dogs

tweet_json table

9. incorrect datatype; tweet_id

TIDINESS ISSUES

1. The columns, doggo, floofer, pupper and puppo in the tweet archive table which are all dog types are in separate columns.
2. Tweet_json dataframe should be merged with twitter_archive dataframe
3. All the tables should be combined in one dataframe

CLEANING THE DATA

I first made copies of the data frames.

1. Those columns with a lot of null entries were dropped because they are not needed for analysis.
2. I converted tweet_id from integer to string datatype, and also converted timestamp from object datatype to DateTime data type in the twitter_archive table.
3. I removed HTML from rows in source column
4. I replace all the invalid names (lower case names) with NaN
5. I corrected numerator_ratings with decimals
6. I renamed columns for a better description
7. I dropped duplicated URL in Jpg_url column
8. I Drop rows that have p1_dog, p2_dog, p3_dog values set to false
9. I converted all the tweet_id columns to datatype string, for easy merging
10. I combined the 4 columns; doggo, floofer, pupper and puppo into one column named dog_type
11. I merged the tweet_json data frame with twitter_archive data frame into a data frame named twitter_df
12. I merged the tweet_json data frame with twitter_archive data frame into a data frame named twitter_df

STORING THE DATA

Now the dataset is clean and ready for analysis. I saved the dataframe to twitter_archive_master.csv. Then I started my investigation.