WRANGLE REPORT

GATHERING THE DATA

The following steps were taken in gathering the data:

1) I manually download the file called "twitter-archive-enhanced.csv" provided by Udacity.
2) The downloaded csv file was stored in a dataframe called "twitter_archive".
3) The request library was used to programmatically download the file "imagepredictions.tsv" from Udacity's server. It was then saved to a Dataframe "image_prediction".
4) The alternative JSON file that was given was used since the Twitter API didn't work.
5) I downloaded the tweet_json text file that was provided and selected the columns I was interested in. these columns are; id, retweet_count and favorite_count.

After following the outlined stpes above, I got my 3 data frames; twitter_archive, image_prediction, tweet_json. I proceeded to assessing phase were the dataframes were accessed.

ASSESSING THE DATA

The three tables had quality and tidiness issues. These issues are explained in details below.

QUALITY ISSUES

Twitter Archive Table

1. There are retweets and replies in the dataset. Since we want original ratings without retweets and replies I will have to delete them.
2. There are a lot of null values in the following columns: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_user_id, retweeted_status_timestamp. These columns will be dropped.
3. Datatype issues: Timestamp and retweeted_status_timestamp columns should be in date and time datatype. Tweet_id, in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be in string (object) datatype and not int or float
4. The source column has rows with html in them which are not human readable
5. There are some invalid names in the name column such as: a, the, and, very. The names are in lower case.

Image Prediction Table

6. Some column names are not descriptive for a clearer understanding. e.g. p1, p2
7. The Jpg_url column have duplicate entries with different ID's
8. Some entries have p1_dog, p2_dog, p3_dog set to false. These are not dogs

All Tables

9. Tweet_id columns for all the tables should be string datatype for easy merging

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

Since the dataset is clean and ready for analysis, I went ahead to save the dataframe to twitter_archive_master.csv. Then I started my investigation.