Wrangle Report

Data Gathering:

Three datasets were used in this project:

1. The "twitter-archive-enhanced.csv" dataset was provided by Udacity academy.
2. The "image-predictions.tsv" dataset was provided by Udacity academy.
3. Since I was not able to use Tweety, Udacity academy provided the "tweet-json.txt" dataset.

Assessing Data:

After visualizing the data, I identify some quality and tidiness issues.

Quality issues:

1. Converting (retweet_count , favorite_count) in count_df table from "object" to "int".

2. The (timestamp) column in twitter_arch table should be converted to "dateTime" instead of "object".

3. The (rating_denominator) values in table twitter_arch should be all converted to 10.

4. The (rating_numerator) should be greater than 10. any number less than or equal to 10 will be converted to 11.

5. Remove rows that are considered as "retweet" or "reply" since they are not real tweets.

6. Drop (in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp) in twitter_arch table sense there are a lot of missing values.

7. Delete unnecessary information from "source" column in twitter_arch table (for example: converting from'<a href="http://vine.co" rel="nofollow">Vine - Make a Scene</a>' ---> "Vine - Make a Scene").

8. Fixing some dogs name (for example: "None" , "a" , "O").

9. Drop "jpg_url" column in image_prediction table.

10. Convert (tweet_id) in image_prediction and twitter_arch tables from "int" to "str"/"object".

Tidiness issues:

1. Converting "doggo" , "floofer" , "pupper" , and "puppo" columns from twitter_arch table into one column.
2. Adding the three datasets together using tweet_id.

Cleaning Data:

The first step in cleaning the data was making copies from each dataset. Next, I solved each quality and tidiness issues that I have found in the datasets by using the define-code-test steps. After solving the issues, I merged the datasets into one dataset using tweet id key. Lastly, I saved the data with the name "twitter_archive_master.csv".

Analyzing and Visualizing Data:

In this step, I used the "twitter_archive_master.csv" to visualize and analyze some important information using graphs such as bar graph and scatter plot.