WRANGLING REPORT

1. Gathering Data

1.1. Data from twitter_archive_enhanced.csv was uploaded and read into the pandas dataframe.

1.2. Data from image_predictions file was downloaded programmatically using this URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. I made a file path to store it, then opened it and finally I read it into a pandas dataframe.

1.3. After reading and understanding how the twitter API code works, I copied and pasted it into my notebook.
Note: I chose not to sign up with twitter.

1.4. I also read the tweet_json.txt file line by line into a pandas dataframe with the following column names; ID, retweet count and favorite count.

2. Assessing Data

2.1. **Visual Assessment** was used to find the following quality and tidiness issues.

2.1.1. some values in twitter_archive_enhanced rating numerator and denominator column are **invalid**

2.1.2. some values in twitter_archive_enhanced are **not equal to 10.**

2.1.3. **Retweeted_status_id**, **retweeted_status_user_id**, **retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id, expanded_urls** and **source** are column names in twitter_archive_enhanced dataframe and we don't really need them.

2.1.4. Strange names like **"a" and "an"** can be seen in the name column of the twitter_archive_enhanced dataframe.

2.1.5. **Doggo**, **floofer**, **pupper**, and **puppo** in twitter_archive_enhanced are in four different columns.

2.1.6. In twitter_archive_enhanced there is a column with name "confg", the name is not descriptive.

2.1.7. **Null** values were represented as **None** in the name column of the twitter_archive_enhanced

2.2. **Programmatic Assessment** was used to find the following quality and tideness issues.

2.2.1. **Tweet_id** in twitter_archive_enhanced should be **string** not **int.**

2.2.2. Timestamp in twitter_archive_enhanced should be **date time** not **string.**

2.2.3. **59** missing values in twitter_archive_enhanced expanded_urls column

2.2.4. **Tweet_id** in image-predictions should be **string** not **int.**

2.2.5. Removing retweeted data and replys, leaving the original data

2.2.6. Image-predictions has missing **IDs**, it had **2075** instead of **2353.**

2.2.7. **ID** column name in tweet_data should be **tweet_id** not **id**

2.2.8. **Tweet_id** in tweet_data should be **string** not **int.**

3. A copy of all datasets was done.

4. Cleaning Data

4.1. Rating_denominator values that are less 10 were replaced to 10.

4.2. Invalid numerator and denominator value were dropped

4.3. **Retweeted_status_id**, **retweeted_status_user_id**, **retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id, expanded_urls** and **source** column names in twitter_archive_enhanced were removed because we don't need them.

4.4. **Tweet_id** of all datasets was converted from int to string.

4.5. **Timestamp** was converted to **datetime** from **string.**

4.6. Using the replace method and the numpy nan method, none was replaced to null.

4.7. Using the rename method confg column was made very descriptive.

4.8. **Doggo**, **floofer**, **pupper**, and **puppo** dog stages were combined into one column by extracting the names from the **text** column. Then dropped the real columns.

4.9. All the three dataframes were combined to one dataframe.

5. Storing Data

5.1. using the ".to_csv" method, the cleaned dataframe was saved as twitter_archive_master.csv