WRANGLE REPORT

The wrangling process of the WeRateDogs Analysis involves three steps:

1. Data Gathering
2. Assessing Data
3. Cleaning Data

1. **Data Gathering:** In this step three pieces of data were gathered from three different sources. **The first piece of data is the WeRateDogs Twitter Archive** which was made available to be downloaded manually as a CSV file *(twitter_archive_enhanced.csv).* After downloading the CSV file, it was read into a pandas dataframe.

The second piece of data is the Tweet Image Predictions *(image_predictions.tsv)* which is a file containing the tweets' image predictions according to a neural network. The TSV file was downloaded programmatically from Udacity's server using the Requests library.

The third piece of data is Additional data (specifically, retweet count and favorite count) from Twitter API. The tweet IDs in *twitter_archive_enhanced.csv* were used to query the Twitter API for each tweet's JSON data using Python's Tweepy library. Each tweet's entire JSON object was then written line by line into a .txt file *(tweet_json.txt).* Once this was completed, the *tweet_json.txt* file was read into a pandas dataframe which consists of each tweet's ID, favorite count, and retweet count.

2. **Assessing Data**: The already gathered data were assessed both visually and programmatically in order to detect and document any quality and tidiness issues that they might have.

The following issues were detected from assessing all three data:

<u>Quality</u>

- Some instances are retweets and not original tweets
- Incorrect datatypes
- Nulls represented as "None" in name, doggo, floofer, pupper and puppo columns
- Multiple dog stages for certain tweets
- Inaccurate names in the name column such as 'a', 'an', 'None', etc, and most of these wrong names begin with a lower case letter
- Incorrectly extracted rating_numerator and rating_denominator
- Multiple image predictions per tweet
- Inaccurate image predictions such as paper_towel, orange, bagel, etc.
- The mode of representing the predictions is not valid (Ceratin predictions begin with capital letter while some do not & underscore (_) used in place of space)

Tidiness

- One variable in four columns (dog_stage) in tweets_archive
- favorite_count and retweet_count should be part of tweets_archive

3. Cleaning Data: After the quality and tidiness issues present in the three data have been detected and documented, the issues were cleaned using the define-code-test framework and clearly documented (missing data could not be retrieved).

Tidiness issues identified were cleaned according to the rules of tidy data.

Quality issues were also cleaned to ensure completeness, validity, accuracy and consistency.

After all the data had been cleaned, they were merged into a master dataframe and stored in a CSV file named *twitter_archive_master.csv.*