**Wrangle Report**

The dataset wrangled in the project is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

The WeRateDogs Twitter project goals included:

1. Wrangling the twitter data through the following processes:
    - Gathering Data
    - Assessing Data
    - Cleaning Data
2. Storing, analyzing, and visualizing your wrangled data
3. Reporting on the data wrangling efforts and data analyses and visualizations

**Gathering Data**
My wrangling efforts for the WeRateDogs Twitter project included gathering data from the following sources:

- The WeRateDogs Twitter archive. The twitter_archive_enhanced.csv file was provided to Udacity students. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.
- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity students.
- Twitter API and Python's Tweepy library to gather each tweet's retweet count and favorite ("like") count at minimum, and any additional data I find interesting.

**Assessing Data**
Once the data was gathered, I began to assess the data on both quality and tidiness issues.

**Quality Issues**

Completeness, Validity, Accuracy, Consistency – content issues

**twitter_archive dataset**

- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id should be integers/strings instead of float.
- retweeted_status_timestamp, timestamp should be datetime instead of object (string).
- The numerator and denominator columns have invalid values.
- In several columns null objects are non-null (None to NaN).
- Name column have invalid names i.e 'None', 'a', 'an' and less than 3 characters.
- We only want original ratings (no retweets) that have images.

- We may want to change this columns type (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id and tweet_id) to string because We don't want any operations on them.
- Sources difficult to read.

**image_predictions dataset**

- Missing values from images dataset (2075 rows instead of 2356)
- Some tweet_ids have the same jpg_url
- Some tweets are have 2 different tweet_id one redirect to the other (Dataset contains retweets)

**tweet_data dataset**

- This tweet_id (666020888022790149) duplicated 8 times

**Tidiness Issues**

- No need to all the informations in images dataset, (tweet_id and jpg_url what matters)
- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo
- Join 'tweet_info' and 'image_predictions' to 'twitter_archive'

**Cleaning Data**
After the assessment, I cleaned the data through the following means:

**Define, Code and Test**

1. Add tweet_info and image_predictions to twitter_archive table.
2. Melt the 'doggo', 'floofer', 'pupper' and 'puppo' columns into one column 'dog_stage'.
3. Clean rows and columns that we will not need
4. Get rid of image prediction columns
5. Fix rating numerator and denominators that are not actually ratings
6. Fix rating numerator that have decimals.
7. Get Dogs gender column from text column
8. Convert the null values to None type
9. Change Datatypes