# Data Wrangling Report

## 1. Gathering Data

#### The Dataset

The dataset I'll be wrangling is the tweet archive of Twitter user @dog_rates (https://twitter.com/dog_rates), also known as WeRateDogs. This archive/dataset consists of 2356 basic tweet data from November, 2015 to August, 2017. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

##### Gather Twitter archive CSV file

Using the link provided by Udacity, I downloaded the WeRateDogs Twitter archive manually as twitter_archive_enhanced.csv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archiveenhanced/twitter-archive-enhanced.csv) file and imported this file into a dataframe (twitter_enhanced_data).

#### Gather tweet image predictions

I downloaded the tweet image predictions file hosted on Udacity's servers programmatically using Python's Requests library and saved it locally to image_predictions.tsv file. Then, I imported this file into a Python Pandas dataframe (image_prediction).

#### Gathering data from twitter

Using the tweet IDs in the Twitter archive, i am supposed to access the entire data for every tweet from Twitter API and stored every tweet's entire set of JSON data in a file called tweet_json.txt file. Created a dataframe status_df from this JSON including only tweet_id, retweet_count, favorite_count and display_text_range data, But this was not done, because i had issues installing tweepy on my anaconda, so i proceeded with the json_txt file provided by udacity

## Assessing Data

## Quality issues

1. The in_reply_to_status_id,in_reply_to_user_id , source,retweeted_status_user_id ,retweeted_status_timestamp columns have NaN, hence would not be needed for the purpose of this analysis

2. The rows without extended_url columns is not needed for the sake of the analysis

3. There are some values in the rating_numerator that are to considered as outliers

4. There are some values in the rating_denominator that are to consider as outliers

5. The analysis needs only tweets and not retweets, so we text with 'RT @' can be dropped

6. There are names on the names column that are in lower case - we need to get this repalace them with None

7. The data type for retweeted status time is not in the appropriate time data type

8. some of the dog type columns have more than one dog type

### Tidiness issues

1. The name of the Dog is on separate columns, and it should not be
2. The tweet_data table should be part of the enhanced_archive table and at the end combine all three dataframes

## Cleaning Data

As all the quality and tidiness issues were related to twitter_enhanced_data table, I created a copy of only this table and named it archive_data. For each quality/tidiness issue, I performed the programmatic data cleaning process in 3 stages - Define, Code & Test. During the cleaning process.