# WeRateDogs Twitter Archive - WRANGLE REPORT

The WeRateDogs Twitter archive consists of data gathered from the @dog_rates twitter account posted between November 2015 to August 2017. A total of 2356 tweets were extracted from 5000 tweets.

## DATA WRANGLING

The entirety of the process and steps I took in wrangling the 'We rate dogs twitter data' . Four steps guided the wrangling process namely:-

- Gathering Data
- Assessing Data
- Cleaning Data
- Storing Data

## GATHERING DATA

The Data utilised for the wrangling process were gathered from three different sources namely:-

1. The Twitter_archive_enhanced.csv file provided by udacity was downloaded manually, uploaded and read into the pandas dataframe. It was renamed as twitter_data.
2. Using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv, the image_prediction.tsv file was downloaded and read into the jupyter notebook as a CSV file.
3. Using tweet ID, the Twitter API for each tweet's JSON data on the WeRateDogs Twitter archive was queried using Python's Tweepy library and the tweets were stored on a JSON file.

## ASSESSING DATA

The data was assessed to identify quality and tidiness issues and the issues identified in the course of the assessment are hereunder listed

### Quality Issues

- Columns associated with reply and retweets have numerous nan values
- Expanded_urls column have 59 missing urls
- The timestamp column datatype ought to be datetime instead of string
- Inaccurate names like 'O', 'an, 'a' are found on the name column
- Extreme figures are found on the rating_denominator column.
- Extreme figures are found on the rating_numerator column.
- There is need to rename the columns in the predictions dataframe to make them more descriptive
- The columns showing dog breeds p1, p2 and p3 have some name with initial capital while some do not.

### Tidiness Issues

- Dog stages are splitted into four columns rather than one

- There is a need to merge the three dataframes into one for ease of reference and analysis.

**CLEANING DATA**

The data was cleaned according to the issues listed above. However, before the cleaning process, a copy was made for each of the dataframes. For each quality or tidiness issue, the cleaning process followed the Define, Code, Test rule.

**STORING DATA**

The entire cleaned data was stored in a twitter_achive_master.csv file