# **PROJECT: WeRateDogs**

## Wrangling Procedures

The wrangling of the data were carried out using the three pronged approach of gathering, assessing and cleaning the dataset. The project required the use of the following modules:

- *pandas*
- *numpy*
- *requests*
- *tweepy*
- *OAuthHandler*
- *json*
- *timeit*

### **Gathering:**

The data set were gathered from three sources, namely;

1. *"twitter_archive_enhanced.csv"* was downloaded directly from Udacity platform. This data is a tweeter archive on the dogs rating with over 5000 entries.
2. *"image_predictions.tsv"* was scrapped from the web using a provided URL. This held the prediction certainty of each dog breed when ran across a machine model.
3. *"tweet_json.txt"* was accessed from Twitter using a tweeter API. This held tweet's retweet count and favorite count alongside the corresponding tweet ID of the source file. The data was saved to *.csv* file extension.

The following are the names of the associated dataframes;

*"twitter_archive_enhanced.csv" = we_rate_dogs*

*"image_predictions.tsv" = image_p*

*"tweet_json.csv" = tweet_data*

### **Assessing:**

Assessing was done visually using a spreadsheet and programmatically. All three dataframes were perused using this technique. Much emphasis was laid on the *we_rate_dogs* dataframe. This sprung out 8 qualitative issue and 1 tidiness issue. Another tidiness issue was seen in the *image_p* dataframe. The following are the list of qualitative and tidiness issues spotted;

*Qualitative issues*

*we_rate_dogs table:*

- *The retweet columns are not needed in the scope of the analysis.*
- *The reply columns are not needed in the scope of the analysis.*
- *Some data from the text column were not contents associated to an original tweet but a retweeted post.*
- *Incomplete or missing data.*
- *The timestamp column data type was ascribed an object.*
- *The dog breeds occupied multiple columns.*
- *Some uploaded pictures were not that of a dog.*
- *The source column contains html tags which can be cleaned in other to access the values.*

*Tidiness issues*

- *There are multiple columns for dog stages.*
- *All three datasets are part of same observational unit.*

Copies of the data were made with the following description;

*ratings = we_rate_dogs.copy()*

*image = image_p.copy()*

*tdata = tweet_data.copy()*

**Cleaning:**

I commenced cleaning the data by first of all resolving tidiness issues which led to merging the dataframes into one observational unit under the name "*tweet_df*". The cleaning of the data was done in line with the define-code-test protocol. All qualitative issues were subsequently resolved, producing one harmonious clean data set save to file as *"twitter_archive_master.csv".*