# WRANGLE REPORT

The data set that was analyzed is the tweet archive of Twitter users @dog rates, also known as WeRateDogs. This is a Twitter account that rates people's dogs by writing funny comments about the dogs.

This report briefly describes the data wrangling efforts exterted in this project. The entire project was created the on the Udacity's workspace. However, the reports were created and exported as PDFs using Microsoft word.

The wrangling process involves the following three stages:

1. Gathering data

2. Accessing data

3. Cleaning data

Each step is further explained below:

**1. Gathering data**

The project was started by downloading the ‘twitter-archive-enhanced.csv’ file manually and uploading it to the work space. Then, a folder named 'image_predictions' was created before the ‘image_predictions.csv’ file was downloaded programmatically from Udacity's server using the requests library and the following URL:https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'. It was then written into the ‘image_predictions.csv’.

The ‘twitter_data’ was created by accessing and downloading Twitter’s JSON data using the tweepy library. This was done by extracting the list of tweet IDs from the ‘twitter-archive-enhanced.csv’ file each ID was looped through and Twitter’s API was queried with the ID to get each tweet’s JSON data. The resulting data was recorded in a text file named ‘tweet-json.txt’, with each tweet’s data written in a new line. A ready-made file was provided for those who had difficulties creating a developer’s account and that was what was used in this analysis.

After completing the query and writing all data in the text file, the text file was read line by line into a pandas data frame and each tweet's information (tweet ID, retweet count, favorite count, and followers count) was obtained using the json library, and appended the information into an empty list. It was later saved into a ‘twitter_data.csv’ file for future use.

**2. Accessing the data**

After gathering the data, they were accessed visually and programmatically for quality and tidiness issues. This resulted in the following findings (Only the issues resolved in the analysis are mentioned below):

**A. Tidiness**

1. There are 4 columns meant for dogs. They are to be combined into one column called dog stage. Then the individual columns will be deleted.
2. Change the timestamp from string to datetime. Create new columns for year, month and day of week.
3. All data (twitter archive, tweet data and tweet image predictions) is related but seperated into three columns so they need to be merged.

**B. Quality**

***1) Twitter archive enhanced***

1. Make the rating denominator column standard to be 10. The rating numerators can be greater than the denominators.
2. Renaming column from 'text' to 'tweet'
3. Creating a new column for 'rating'
4. Delete unwanted columns not need in the analysis.
5. Remove all wrong values for dogs in the name column such as 'a' and 'none'. Make the first letter capital for all names and replace 'none' with 'nan'.
6. Change tweet id data type from integer to string.
7. Keep original tweets (no retweets) that have images.
8. Excluding uneeded data

***2) Image predictions***

1. Delete duplicate values for images
2. Delete rows with missing photos
3. Some p names start with upper case and some lower case which should be changed.
4. P1, P2, and P3 contains underscores instead of spaces in the string.

**3) *Tweet data from twitter api***

1. Delete rows without retweet count entries

**3. Cleaning data**

The previous problems we're cleaned appropriately which resulted in a high quality and tidy Master Pandas DataFrame.