# Wrangling and Analyzing Data Project - WeRateDogs

**Name:** Michael Owusu Agyebeng
**Description:** Udacity Student

As part of Udacity's Data Science course, I had to wrangle and analyze WeRateDogs posts on twitter. This involved gathering data from twitter using Twitter API and combining it with two other data from a URL and a csv file to form one grand file called twitter_archive_master.csv.

## Data Gathering

I initially imported some packages that are useful in generating and analyzing the data. After that, I loaded the csv file into my jupyter notebook using pandas then I loaded the image predictions file which is a tab separated file. Thereafter, I used queried twitter API for each tweet in the twitter archive for a JSON file which I saved as a data frame called new_df.

## Assessing Data

After assessing the data, I found ten quality issues and two tidiness issues which I have listed below.

### Quality Issues

1. tweet_id must be a string not an integer
2. retweeted_status_id is not a float but a string
3. retweeted_status_user_id is not a float but a string
4. timestamp is a date not a string
5. in_reply_to_status_id is not a float but a string
6. in_reply_to_user_id is also not a float but a string
7. replies could also mean duplicates so we will have to drop them
8. retweets in df data hence there are duplicates so we have to drop retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp
9. dog names in the img_df data were in different cases
10. some dogs names are "a", "None", "the", etc

### Tidiness Issues

1. html tags in source column of df data
2. doggo, floofer, pupper and puppo columns need to be merged

## Data Cleaning

Tweet_id's were changed from integer to string format in all 3 files. retweeted_status_id, retweeted_status_user_id, in_reply_status_id, in_reply_user_id which are not floats but strings were to be changed but since after changing their formats, it would be tedious to

drop I decided no to change them since I had intentions of dropping them anyway. Since there are retweets in the data, it means there is a possibility of duplicated tweets hence I filtered the retweets which in return removed the duplicates and dropped the replies and retweets. Timestamp was changed into datetime format.

Dog names in the img_df file were in different cases so I capitalized them to put them in the same case. Also, some dog names were invalid so I changed them to “None”. The source column of the df data contained both html tags and URLs which is messy so I had to separate them using the split function and adding them back into the data. Finally, since dog stage names were in different columns, I had to merge them to form one column called stage. There were some dogs who had more than one dog stage, hence I put them under one name, doublestage, signifying they were not classified under two stage names.

## Analyzing and Visualizing data

I merged the data and stored it in the twitter_archive_master.csv. I ended up with 1954 entries with 24 columns. Then I began the analysis to uncover patterns and correlations between the retweet count and favorite count overtime and also the lion’s share of our tweets was posted by iPhone users. And also identified the top 10 dog names and dog breed names.