# Introduction

We are going to describe the data wrangling work done as per required in the project wrangling weRateDog

Data wrangling is split in 3 steps:

- Gathering data
- Assessing data
- Cleaning data

# Gathering

For the project we use 3 source files as per below:

- Twitter archive CSV file: downloaded from the link twitter_archive_enhanced.csv (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archiveenhanced/twitter-archive-enhanced.csv)
- Tweet image predictions:downloaded from predictions file hosted on Udacity's servers programmatically using Python's Requests library and saved it locally to image_predictions.tsv file
- data from Twitter API: access to Twitter API and stored every tweet's entire set of JSON data in a file tweet_json.txt

# Assessing

After gathering each of data source the next step would be assess the files looking for quality issues and tidiness as per suggested in Udacity course. The following quality issues have been found out:

*Quality*

1. Convert *rating_numerator* to fload
2. Delete all rows without extended url
3. Incorrect datatypes in in*_reply_to_status_id*, *in_reply_to_user_id* and *timestamp* columns. *Timestamp* needs to be converted to timestamp datatype and *in_reply_to_status_id* and *in_reply_to_user_id* to int64.
4. Unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone
5. Remove tweets with large rating_numerator that seems to be incorrect
6. Remove duplicate tweets. The duplicates are retweets and we only need the original values
7. Incorrect dog names starting with lowercase characters and unrecognize values
8. Some tweets of archive df are missing in prediction images df

*Tidiness*

1. All rows about dogs stage or phase can be merge in one column
2. The json_data table should be part of the archive table and at the end combine all three dataframes
3. Store prediction algorithm and level of confidence with the first value True

## Cleaning

to sort out all quality and tidiness issues I have created a copy of all the tables and rename as per below:

- archive
- predictions
- tweet_data

For each quality/tidiness issue, I follow the structure of the programmatic data cleaning process in 3 stages:

1. Define
2. Code
3. Test.

During the cleaning process I have updated the datatypes of source as per required according with the field and I have created phase columns in archive table defined to category datatype. Also, I have merged the columns Doggo, Floofer and Pupper. I have stored prediction algorithm and level of confidence with the first value True

### Conclusion

Data wrangling is one of the most important parts in data analysis projects. It is a key requirement that all data analytics should be familiar with. The data cleaning is essential step to produce feasible insights of dataset.