# Introduction

In weRateDog project we menially focused on data wrangling which I will describe the steps and the work I did. Data wrangling, consists of:

1. Gathering data
2. Assessing data
3. Cleaning data

## 1- Gathering data

I Gather data in a Jupyter Notebook from three different resources as described below:

- The WeRateDogs Twitter archive. I downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv , and saved it to my desktop , then I used the library panda to download the file using the function read_csv and stored the data frame to twitter_archive .
- The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) hosted on Udacity's servers and we downloaded it programmatically using python Requests library (requests.get) on the following (URL of the file: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv). Then I read the file using function read_csv with the attribute separator (sep ='\t')
- Each tweet's retweet count and favorite (i.e. "like") count and any additional data we found interesting. Using the tweet IDs in the WeRateDogs Twitter archive, we could query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data stored in a line.

## 2- Assessing data

using the notebook Jupiter and the function. head , I inspect visually missing data (none. Nan),and programmatically using the function info , describe I checked the datatype and get the following quality and tidiness issue :

- **Tidiness issue**

  1-In to twitter_archive file: Combine each dog stage column into a single column named "dog_stage".
  2-In all files Drop unnecessary columns.
  3- Merge tweet data table into the twitter_archive table.

- **Quality issues for the following files:**

- **twitter_archive file:**

  1-In the column “rating_numirator” there are extreme values greater than 10, will not affect the analysis.
  2-In the column "rating denominator" there are some value less than 10, I think its incorrect data and should be checked
  3-There are missing values in (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp) also not needed in this analysis because the key point is original tweets not retweets.
  4-There are missing values in_reply_to_status_id, in_reply_to_user_id also not needed in this analysis because the key point is original tweets not retweets.
  5-There are missing data in expanded_urls .
  6-There is text in the ‘source' should be removed.
  7-The timestamp stored as object which should be date

  - **image_predictions file:**

  1-Change the names of columns p1, p2 and p3 to reasonable names.
  2-There are some strange values in the p1 columns should be checked.
  3-Prditicton of dog should be capitalize letter for consistency.

  - **tweet_data file:**

  1-rename the column 'id' to 'tweet_id'

## 3- Cleaning data

Cleaning our data is where we fixed the quality and tidiness issues that we identified in the assess step. First I take a copy of all three file I used programmatic data cleaning process which is:

1. Define: convert our assessments into defined cleaning tasks. These definitions also serve as an instruction list so others (or yourself in the future) can look at your work and reproduce it.
2. Code: convert those definitions to code and run that code.
3. Test: test your dataset, visually or with code, to make sure your cleaning operations worked.

- Cleaning quality issues:

- **image_predictions file:**
  - Change the names of columns p1, p2 and p3 to reasonable names, I used rename function.
  - Capitalize the first letter of the first prediction for consistency
- **twitter_archive file:**
  - In the column "rating denominator" there are some value less than 10, I think its incorrect data and should be checked
  - 10 is the default value of 'rating_denominator', then replace all value below 10 based to the information in the text of the tweet it shows that its wrong data entry .
  - Remove all retweeted data because the key point is original tweets not retweets (retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, in_reply_to_status_id, in_reply_to_user_id)
  - remove missing data in expanded_urls using dropna .
  - remove text from source for clarity using lambda .
  - change timestamp to date using to_datetime
- **tweet_data file :**
  - rename the id to tweet_id using rename .

- Cleaning tidiness issues:
  - Create a new variable 'dog_stage' to show the four dog stages, drop the four columns, and fill the empty with NaN.
  - Merge the tweet_data into the twitter_archive file using inner join.
  - Remove column contain retweet data in twitter_archive file.

Last step was storing combined and cleaned data in twitter_archive_master.csv using to_csv function for visualization and analysis our insight on the data .