# Wrangling Report : WeRateDogs Data Analysis

I wrangled WeeRateDogs Twitter data to create an analyses and visualizations. I will briefly describe the detail of wrangling.

## 1. Gathering data

There are three pieces of data as follows.

* 1. The WeRateDogs twitter archive   
     I manually downloaded from the Udacity website.
  2. The tweet image predictions  
     I downloaded it programmatically using the Requests library from the Udacity’s servers.
  3. Additional data from tweeter

I downloaded additional data from twitter such as retweet and favourite count. By using the tweet IDs in the WeRateDogs twitter archive and query the twitter API by Tweepy library, set of JSON data is downloaded and a pandas DataFrame is created based on this.

## Accessing and cleaning data

* 1. The WeRateDogs twitter archive  
     1. These retweet/reply should be removed from this table as we want only original ratings. Instead, I will make additional columns of the number of retweet based on twitter\_count.csv.
     2. Dtype:

- Timestamp is not time series data.

- Each dog stage is not boolean.

- Five columns relates to ID are not string/object.

* + 1. Summary statistics  
       Some of the denominator which is larger than 10 are incorrect data. And in most of these cases, numerator is also incorrect.
    2. Dog stage  
       Some Floof is not categorised as Floofer.
    3. Dog name  
       Pronoun, short one. (tends to start in lower case)
  1. Image prediction data
     1. Prediction is not a dog
     2. Data type: Tweet ID should be object.
  2. Retweet/favourite count data

#### Quality issues

##### ◆The WeRateDogs Twitter archive

 -Wrong Data type:

 Timestamp (->time series data)  
 Dog stage (-> Boolean)  
 Tweet\_id (-> string/object)

 -Some denominators are incorrect.  
 -Some numerators are incorrect.  
 -Some of the "Floof" are not included in floofer stage.  
 -Some dog names are incorrect.

##### ◆Image prediction

 -Some predisctions are not a dog.  
 -Tweet ID is not string/object.

#### Tidiness issues

##### ◆Twitter archive data

 -Text column includes URL of the tweets.  
 -Retweets / replies are included.  
 -The number of retweet / reply is in a separate dataframe.

## Cleaning of data

* 1. The WeRateDogs twitter archive
     1. Tidiness issue: Retweet and replieRemove rows of non-original tweets(retweet or reply) and retweet/reply related columns.
     2. Quality issue: wrong data type  
        Time stamp should be time series data. As I will remove retweeted\_status\_timestamp later, I will keep this as it is for now.

Each dog stage should be boolean.

Tweet ID should be string/object. Other ID related columns will be removed later so that I will keep it as it is for now.

* + 1. Define: Assess and correct visually where denominators are bigger than 10. As there is higher chance that when denominator is wrong, the numerator is also wrong, I will check and correct both of them. Filter rows >10, make another dataframe with correct denometor/numerator and overwrite.