# Data Wrangling - WeRateDogs

In this project we have been asked to wrangle @weratedogs data.  The effort that went into wrangling this data has been summarized in this document. Data wrangling has three tasks associated with it.

1) Gathering data
2) Assessing data
3) Cleaning data

## **Gathering data**:

As a first step I gathered the required data from three different datasources to create analysis and visualization for WeRateDogs.

1) Enhanced twitter archive was already created by the trainer and available as twitter-archive-enhanced.csv.
2) The retweet and favorite count were acquired from twitter API tweepy. Connecting to twitter API required developer access to twitter. Since data was large and due to rate limit setup by twitter, downloading this data took some time (~20 minutes).
3) Results of image predictions from Udacity servers were programmatically downloaded to local machine. Web Scraping section in lesson 8 Gathering Data was very useful in achieving this.

## **Assessing data**:

This step required identifying data quality and tidiness issues in the WeRateDogs dataset put together. There were several issues noted. They are listed as below,

### DataQuality Issues:¶

1) Non descriptive column headers in df2 (p1, p1_conf, p1_dog and so on)
2) expanded_urls in df1 missing data which means image predictions couldnt be possible.
3) Missing data in df1 columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id)
4) tweet_id in twitter-archive-enhanced dataset and image-predictions dataset are integer whereas it is string in API dataset.
5) Several dog names are incorrectly identified. Some are not even dog photos. Examples are (a, an, all, by, His, such, not, one, very, O,my,this,unacceptable,the,life,Jo, quite) - Visual assessment
6) Some of the records in the data set are retweets.
7) Image Prediction dataset: P1, P2, P3 columns start with Upper case and Lower case. And there are underscores in between where there should be spaces.
8) Timestamp is stored as object(str) should be datetime.
9) Rating_Numerator has float values. - Visual + Programmatic assessment

### Tidiness Issues:

1) There are 4 columns of dog stages, doggo, puppo, pupper and floofer. These violate the rule 1 of the tidy data. doggo, floofer, pupper, puppo should be 4 categorical values in 1 column(stages of dog).
2) All the datasets can be merged into 1 table(datset) as 1 observational unit (tweets info); Unnecessary columns can be removed.

## Cleaning Data:

This process involves three steps.

A. **Define**: convert our assessments into defined cleaning tasks so that they serve as an instruction list.
B. **Code**: convert the definitions to code and run that code.
C. **Test**: test the dataset to make sure the cleaning operations worked.

### *Define: (Quality issues & fix)*

1. Issue: Non descriptive column headers in df2 (p1, p1_conf, p1_dog and so on)
   Fix: Replace it descriptive column headers (example: prediction1, prediction1_confidence, prediction1_isdog)
2. Issue: expanded_urls in df1 missing data which means image predictions couldnt be possible.
   Fix: Remove records that have expanded URL's missing.
3. Issue: Missing data in df1 columns (in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id)
   Fix: After fixing issue 6, remove these columns. They are not needed anymore.
4. Issue: tweet_id in twitter-archive-enhanced dataset and image-predictions dataset are integer whereas it is string in API dataset.
   Fix: Modify tweet_id in twitter-archive-enhanced dataset and image-predictions dataset to String so that we can merge all datasets together.
5. Issue: Several dog names are incorrectly identified. Some are not even dog photos. Examples are (a, an, all, by, His, such, not, one, very, O,my,this,unacceptable,the,life,Jo, quite) - Visual assesment
   Fix: Replace the incorrectly identified dog names above with None.
6. Issue: Some of the records in the data set are retweets.
   Fix: Remove records with Retweets as we need only original ratings.
7. Issue: Image Prediction dataset: P1, P2, P3 (dogbreed) columns start with Upper case and Lower case. And there are underscores in between where there should be spaces.
   Fix: Make all these columns start with Upper case and replace underscore with spaces.
8. Issue: Timestamp is stored as object should be datetime.
   Fix: Convert timestamp column datatype to datetime.
9. Issue: Rating_Numerator has float values
   Fix: Convert the column to float datatype and cleanup the data

### *Define:* (Tidiness issues & fix)

1. Issue: There are 4 columns of dog stages, doggo, puppo, pupper and floofer. These violate the rule 1 of the tidy data. doggo, floofer, pupper, puppo should be 4 categorical values in 1 column (stages of dog).
   Fix: Convert these 4 columns to 1 column with 4 categorical values.
2. Issue: Multiple columns can be removed from twitter-archive-enhanced and image_predictions datasets; and all the datasets can be merged into one dataset because they are at same granularity (TweetID)
   Fix: Merge all the cleansed datasets to one.

### Code & Test:

All the quality and tidiness issues were fixed and tested in the Jupyter notebook. Cleansed data is stored in twitter_archive_master.csv.