# Data Wrangling

## We Rate Dogs

Our goal in  wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations.

### Gathering Data :

I gathering the data from three locating, the first it's come from twitter API but unfortunately they decline my application and i take it  with tweet-json file from your web site, the thread file i download from the URL you gave me.

### Assessing Data :

In assessing part we try to find the issue and find solution, we find too many issues but we take some of them and try to prepare the data for analysis.

#### Quality :

#### In twitter-archive-enhanced.csv :

Rating it should not more than 10.
I have a lot of missing values in (in_reply_to_status_id , in_reply_to_user_id , retweeted_status_id , retweeted_status_user_id , retweeted_status_timestamp ).
I don't have any duplicated (This is comfortable ). (:
Timestamp type it’s should datetime not a object.
I can merge these four columns ( doggo , floofer , pupper , puppo ) in to one.
The null values inside the four columns ( doggo , floofer , pupper , puppo ) it should  come as (null) not (None).

In Tweet_json.txt:

I have a lot of missing values(Tweet_json.txt)  in (contributors , coordinates, geo, in_reply_to_screen_name , in_reply_to_status_id , in_reply_to_status_id_str , in_reply_to_user_id , in_reply_to_user_id_str , place , quoted_status , quoted_status_id , quoted_status_id_str , retweeted_status ).

In Image_prediction:

in image_prediction there is no column for most confidence breed of dogs.
 there are missing tweets since the tweets in tweet_archeve are 2356 and in image_prediction are 2075.
we need tweet with images together

Tidiness :

1- All three database it's should be in one dataframe.
2-All columns 'doggo','floof', 'pupper' and 'puppo' it should in one column.
3-some columns like "in_reply_to_status_id and" they have too many missing value, my opinion is deleting them.

## Data cleaning :

after we assessing the data we make a copy and drop unwanted columns and in some columns such as (doggo, floofer, pupper, puppo) the system does not read them it come (None) and put them in one columns (dog_stage), in image_prediction file that don't have column for most confidence breed of dogs so we crating one, after we almost finished cleaning data we merage all the three file in one, in the last we change columns (timestamp) type to datetime, after we clean all the issue we save the database in file the called(final_data.csv), i drop the null (jpg_url) for best analyses.