Wrangle report :For @WeRateDogs_rates data

- Firstly, I imported all the needed libraries into my jupyter notebook, read the first data from a csv file" twitter-archive-enhanced.csv" into a data frame 't_archive', using the requests library I save the prediction data set into a .tsv file,read then to a second data frame 't_image'. The third data set I used was gotten from WE RATE DOG twitter page which needed to be assessed through twitter ,I connect Twitter API to download json format text file and use pandas to read into data frame as 'r_json' then extracted three columns need for my analysis.

Assessing Data:

Here I assessed all my three data frames both Programmatic assessment and visual assessment.

- The twitter enhanced data set: the data set contains 2354 rows and 17 columns
- The image predictions data set: the data set contained 2075 row and 12 columns
- The additional Data via the Twitter API

Assessment issues:

- Not all column are needed for this analysis which needs to be droped
- The denominator of the ratings should not be 0.
- The doggo,floofer,pupper,and puppo should be a single column and a category type of data
- The "retweeted_status_id,in_reply_to_status_id" column have lot of null values
- The name column has lot of undefined values
- Timestamp is datetime instead of string and the timestamp later than August 1st, 2017 should be removed.
- Some images and false and are not dogs
- The url has duplicated links
- the id columns is not the same with other values

Cleaning data: Before cleaning I made copies of each of my data frame

Quality issues AND Tidiness issue

Twitter archive table:

- Here for the enhanced data set I dropped the columns in the retweet_ and reply_ columns that are not original tweet this are tweet that have values in them which made the column empty.
- Dropped columns that are not needed for my analysis which are in_reply_to_status_id,
- in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id,retweeted_status_timestamp, expanded_urls
- Changed data types then created a single column for all the dog stages

- Cleaned the ratings column into one and used '/' as the delimiter
- Since the name column contain may words that are undefined, cleaned the undefined words into null

image prediction:

- Here I removed all the p images that are false which mean the row do not contain any dog image. For my tidiness I extracted the highest confidence level for the p1,p2,p3 which are true along with their corresponding breed name into a new column using the np.select() statement and dropped the old columns.

Additional Data via the Twitter API table:

- Changed the id name to tweet_id to make my merging possible.

For my final cleaning I merged all the data set into a single master table and dropped empty columns.It contained 2088 row and 15 columns.