#WERATEDOG
DATA WRANGLING.

Project Wrangle Data consists of:

1. Gathering
2. Assessing
3. Cleaning

Gathering*:* To begin with my project I collected data from below mentioned three different sources.

1. The WeRateDogs Twitter archive. I Downloaded this file manually by clicking the following link: twitter_archive_enhanced.csv

2. 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) is hosted on Udacity's servers and I downloaded it programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Tweet_json.txt file - Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file then I loaded .txt file line by line into a pandas dataframe with tweet_id , retweet count , and favorite count.

Assessing:

Firstly, I visually assessed the three gathered datasets then I programmatically assessed them one by one. These are some dataset related information:

- Twitter Archive: This is the .csv file having 2356 rows and 17 columns. However, there are several columns which are of no use in our further analysis. Here are the names of some important columns in twitter archive: tweet_id, timestamp, retweeted status id, rating numerator, rating denominator, name, doggo, floofer, pupper , puppo.
- Tweet_json : This is a json.txt file with2330 rows and 4 columns. Columns are tweet_id, retweet_count, favorite_count and retweeted.
- Image Prediction: This is a .tsv file with 2075 rows and 12 columns .Some of the important columns are tweet_id ,p1 , p1_conf , p1_dog, p2, p2_conf ,p2_dog.We are going to merge some of the columns to carry out our analysis properly.

Cleaning: Before I started cleaning I made a copy of all the three datasets.

- **Twitter Archive:** These are quality issues and tidiness issues that I fixed:

1. We only need original ratings of dogs those have images and no retweets. We will delete those entries which have retweets.
   As per the given dataset, there were 181 values in retweeted_status_id and retweeted_status_user_id. I deleted those 181 entries.
2. Missing data is represented by 'None' istead of 'NaN' in name,doggo floofer ,pupper ,puppo. I replaced all the ‘None’ entries with ‘Nan’ from all the five columns.
3. There are many incorrect entries in name column like a ,the, an,just etc I deleted them.
4. Deleted unnecessary column like 'source','in_reply_to_status_id’,'in_reply_to_user', ‘retweeted_status_id','retweeted_status_user_id', 'retweeted_status_timestamp' and 'expanded_urls' which won't be used for analysis.
5. There are lot of incorrect numerator ratings which needs to be fixed.(420,80,50,60,165,75,88,144,182,143).I fixed them manually.
6. There are lot of entries in rating_dinominator other than 10 need to be fixed.
7. Melted doggo,pupper,puppo,floofer into one column.

- **Tweet Json:** These are the data issues that I fixed.

1. Keep original tweets only.
2. Need to change the datatype of the column tweet_id , retweet_count,favorite_count to int.

- **Image Prediction:** The dataset consists of 2075 rows and 12 columnss. These were the data issues that I fixed:

1. Droped 66 jpg_url duplicated entries.
2. Created 1 column for image prediction and 1 column for confidence level.
3. Deleted the columns that won't be used for analysis like p1 , p1_conf , p1_dog p2, p2_conf ,p2_dog .
4. Change the datatype of tweet_id from object to int to merge dataframes.

Key Challenges:

1. I explored the twitter archive dataset in excel where there were lot of incorrect entries in numerator ratings and names columns. I corrected them manually which took lot of time in data cleaning.
2. I melted four column i.e. doggo, pupper, puppo, floofer into one column dog_stage by deleting the four columns and making copy of each column and then merging them together and then concatenating them to original dataframe which was a quite long process.
3. While merging the three dataframe it took me 2 days to resolve issues like fixing null values, data type issues , etc.

Conclusion: I cleaned all the three data sets and merged them into one data frame called Master_df. Then I loaded this dataframe into csv file called twitter_archive_master.csv.