Udacity Project 3: We Rate dogs data set

This project was part of the data wrangling chapter of Data Analyst Nanodegree program. Data wrangling is a process where engineer gather the data, asses the gathered data for any data quality and tidiness issues and then process to clean the data to make it fit for analysis.

For this project I was given three data sets. The first data set was just given to me via the Udacity link. The data set was 5000+ twitter archive of group WeRateDogs . The data set had information on tweet id, text posted, image link and some other extracted data like name of the dog and dog stage name.

The second data set was hosted on Udacity server and I downloaded the data set using the request library. This dataset had neural network prediction for all dog images from twitter archive data. The neural network predicted if there was a dog in the picture or not.

The third dataset I need to download was using Twitter API using the tweepy library. I created a developed account on Twitter and using my unique access keys downloaded favorite count and retweet count data using the tweet id from the twitter archive from the fist dataset. This was my first experience using the Twitter API.

After getting all three datasets and saving it to my local machine. I started with the initial assessment. I saw a lot of null values in the twitter archive dataset as I had no additional source of getting this values I proceed to delete those columns. Also to match the tidy data criteria I separated the date- time into separate date and time columns. I also made one single column from four different columns for dog stage names. Some columns with dual dog stage names where deleted. I also observed that the numerator ratings where not correctly extracted in some of the cases which were corrected using regular expression.

The other two data sets where relatively clean. I just converted tweet id from int to string and dropped the null values and columns which I felt did not add much value to the dataset.
Finally, I merged the twitter API dataset to twitter archive dataset to create a master twitter_weratedogs dataset from which graphs and charts where created