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.

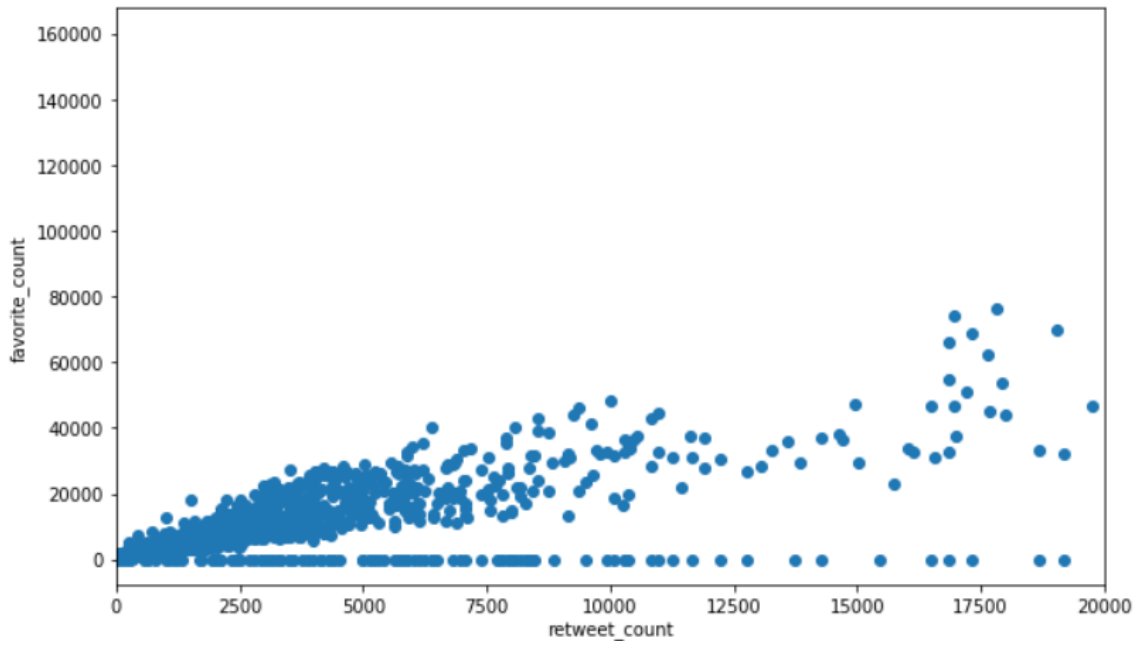
For this project I was given three data sets. The first data set was just given to me via the Udacity link. The data set was 5000+ twitter archive of group WeRateDogs . The data set had information on tweet_id , text posted , image link and some other extracted data like name of the dog and dog stage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

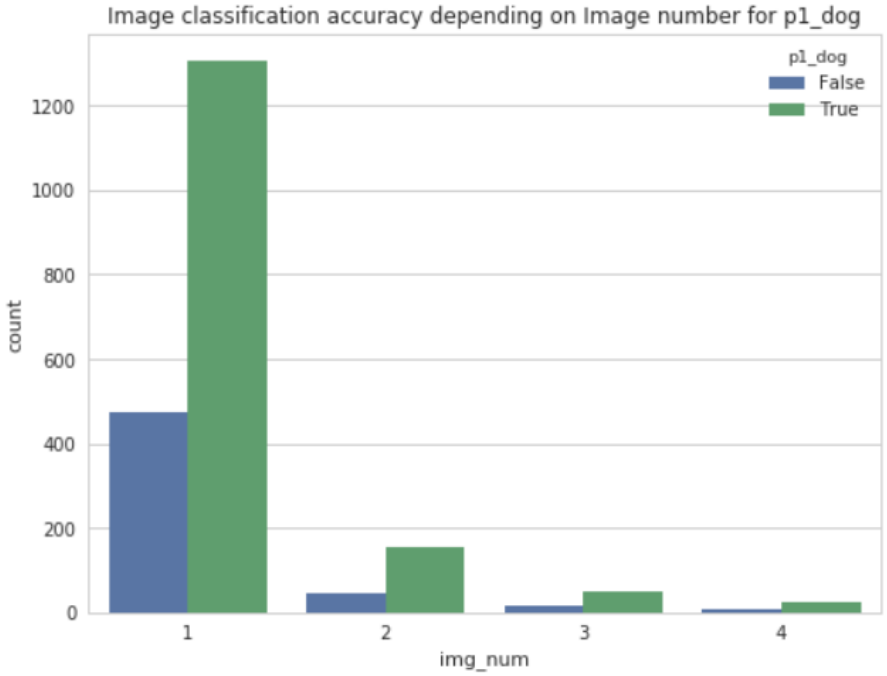
The third dataset I need to download was using Twitter API using the tweepy library. I created a developed account on Twitter and using my unique access keys downloaded favorite count and retweet count data using the tweet_id from the twitter archive from the fist dataset. This was my first experience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

I am sharing some of the insights I drew from the project.

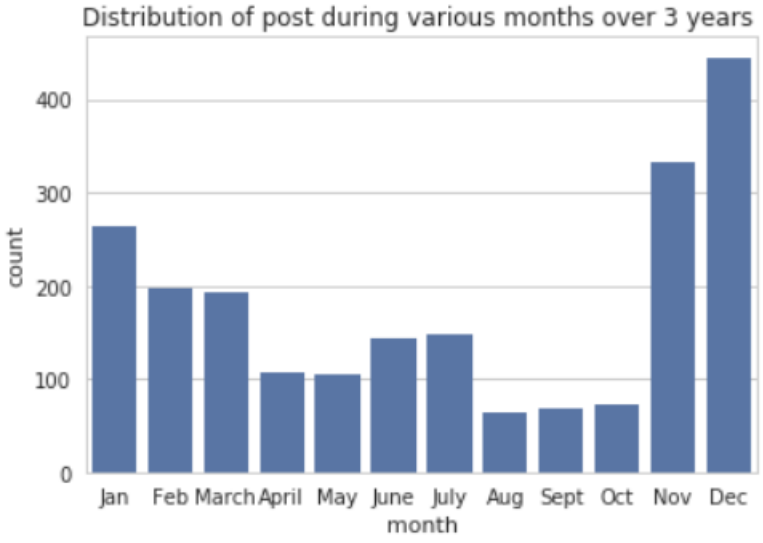
1.I analyzed the relation between retweet count and favorite count. I can see a positive co-relation between them which means chances of favorite tweet getting retweet is high and vice versa. This can help us market any product we want on the page by encouraging user to retweet it more or marking it has favorite.



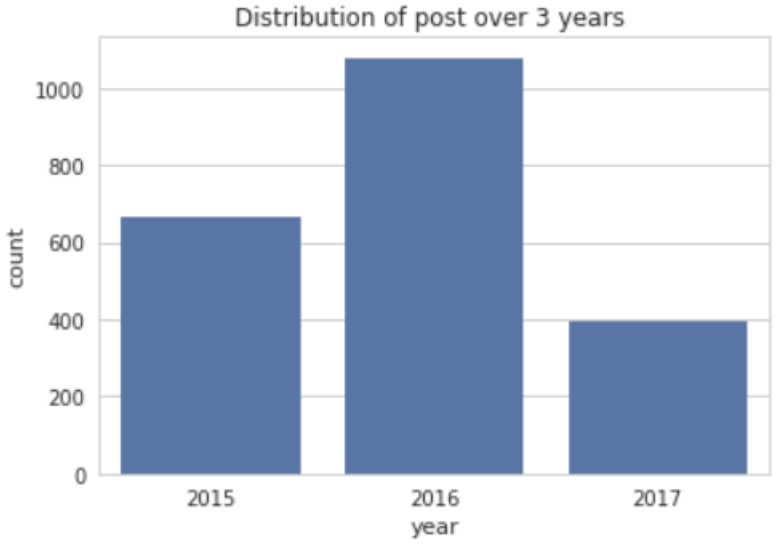
2. After that I proceeded to check how the neural network does on various image numbers. I was not surprised to see that the neural network is not biased towards any image number as such like it does not predict more right for image number 1 vs image number 4



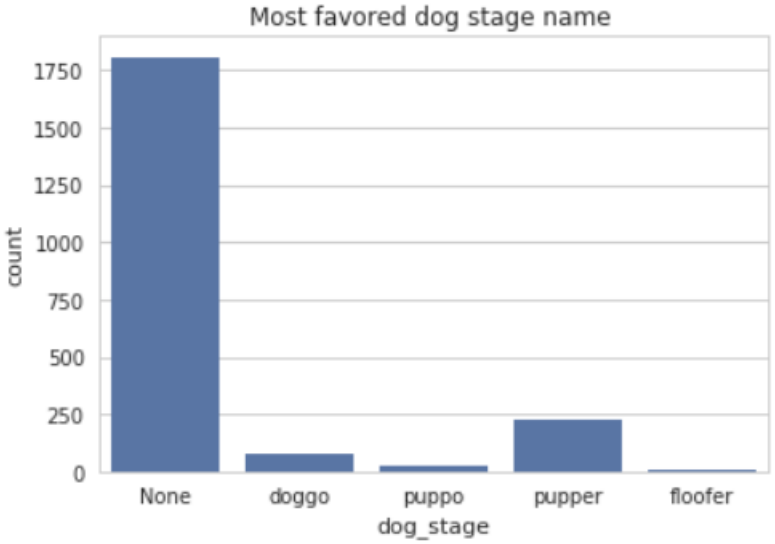
3. I also see that most post are posted during year end like November and December. This show that it is a good time to promote something new in the group due to increased activity



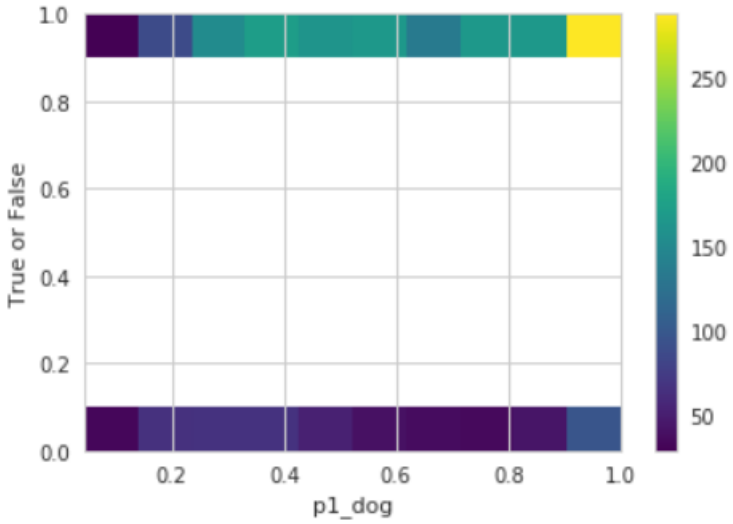
4. Out of the three year of data I had the most post where posted during the year 2016



5. The most favorite dog stage name is pupper.



6. -We see that usually high confidence by neural network is associated with it being right as indicated by the density of points at right side of the graph. That is more points associated with higher probability are generally true


1.0
0.8
0.6
0.4
0.2
0.0
True or False
0.2
0.4
0.6
0.8
1.0
p1_dog
250
200
150
100
50