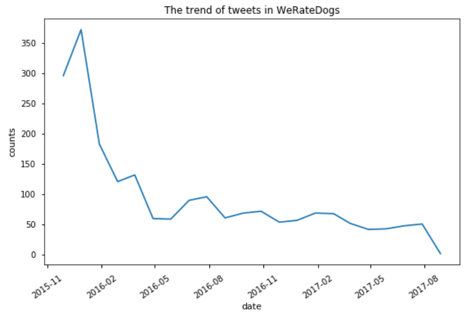
Introduction

We analyze the WeRateDogs data in five domains:

- Tweet activity
- Stages of dog
- Dog rating
- Retweet and favorite counts
- Tweet image prediction

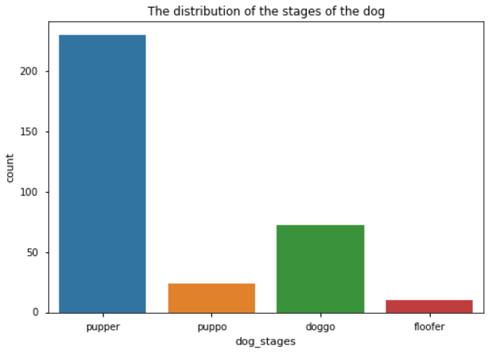
Tweet Activity

The number of tweets are calculated, and the trend is plotted as a time series. The peak of the number of tweets was at the beginning of starting the WeRateDogs account followed by a decay. This may suggest the number of tweets is decreasing. However, this may not be solid since the dataset we have not checked the data completeness.



Stages of Dog

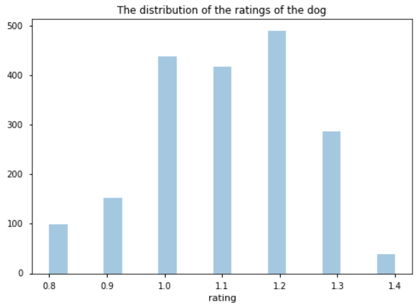
The total sum of the tweets has the information about the stages of the dog is small (< 500), which suggests the stages of dog were not mentioned in most of the tweets or most of the information about the dog stages are lost.

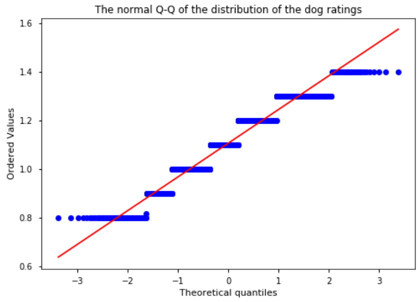


Dog Rating

The rating in the tweets is calculated by dividing the rating numerator to the rating denominator. The outliers are ignored by setting the lower and upper of the data (1.5x IQR from the Q1 and Q3).

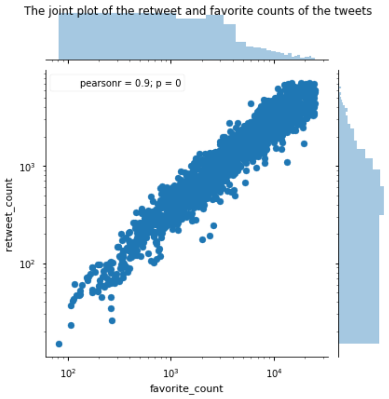
We first plot the distribution of the dog ratings, and it is found that most of the ratings are 10, 11, and 12, which are around the median of the dog ratings. Then, a normal Q-Q plot is generated, and we can see the data distribution is close to a normal distribution.





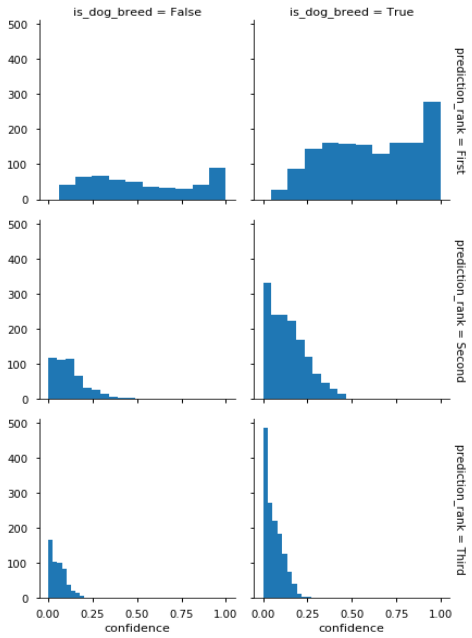
Retweet and Favorite Counts

A joint plot of favorite and retweet counts is generated. The distributions of these counts are not normal. These distributions are skewed towards the left suggesting only small fraction of tweets have a high favorite and retweet counts. Then, we observe there is a strong correlation between the retweet and favorite counts with 0.9 in Pearson correlation coefficient.

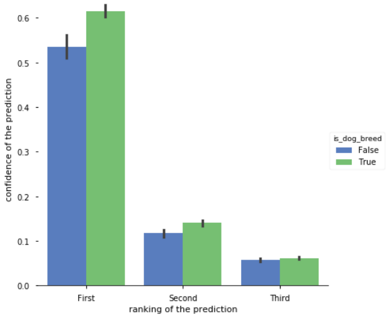


Tweet Image Predictions

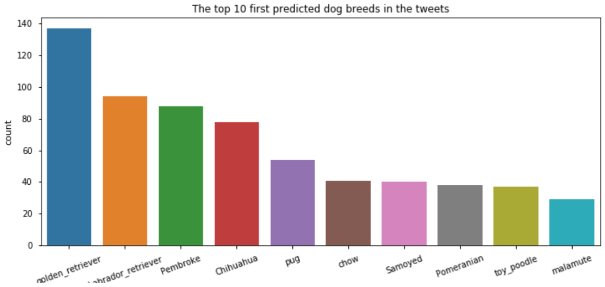
The first predictions have a wide range of confidence (almost the entire space: (0,1)). On the other hand, the confidences on the second and third predictions are always low. Moreover, it is interested that the distribution of the confidences does not depend on the prediction result (is the result a dog breed?).



Since the confidences of second and third predictions are more centralized than the first prediction. The standard derivation on the confidence of second and third predictions are smaller than the first prediction. However, on the other hand, the confidence of the first prediction is usually much higher than the second and third predictions.



Using the first prediction from the neutral network model, we generate a plot of the top 10 predicted dogs in the tweets. If the algorithm of prediction is accurate, this plot may suggest the popular dog breeds among the twitter users.


The top 10 first predicted dog breeds in the tweets
count
golden_retriever
abrador_retriever
Pembroke
Chihuahua
pug
chow
Samoyed
Pomeranian
toy_poodle
malamute