Act Report

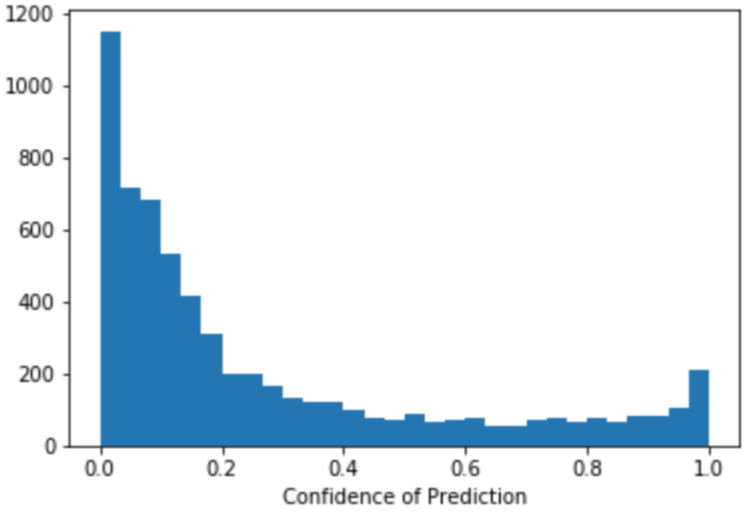
After cleaning the data, now it is time to do a little analysis to gain some basic insights.

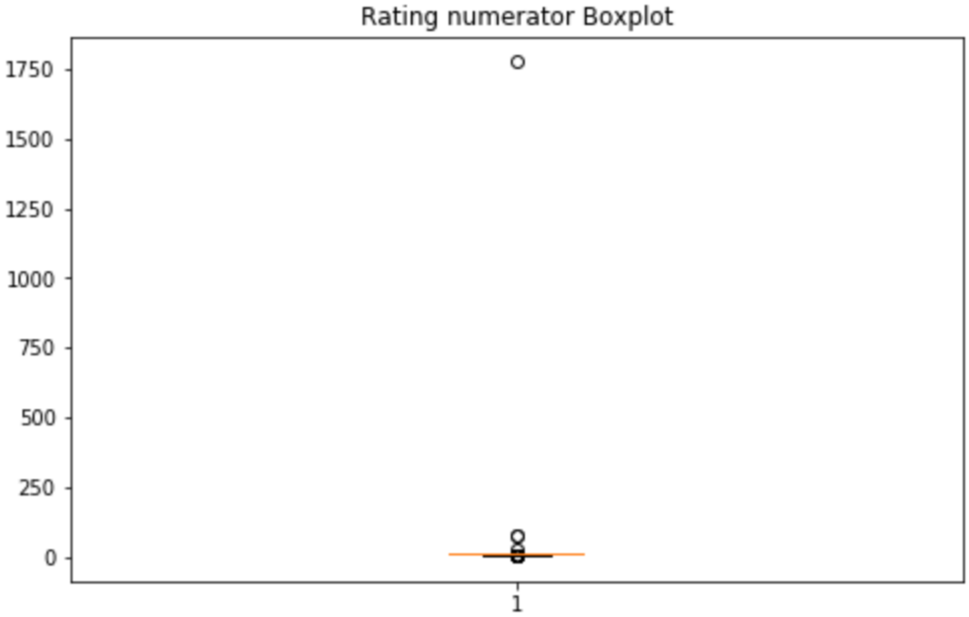
The first insight I gained is about the highest of counting. The highest favorite count is 123132, while the highest retweet count is 60269. We can make educated guess that these two numbers belong to the same tweet and this might be the "hottest" tweet of all the dog tweet database. Moreover, compared to the median tweet count of 1661, the 60269 is an outlier in the retweet count. Likewise, compared to the median of favorite count is 4059.5 , the highest favorite count is also an outlier in the dataset.

The second insight is about the rating. The median rating of dog in the tweet is 12, and the highest is 1776. This gives us a big gap between the majority of ratings and the outlier, as justified by the boxplot below. Since we need to restore the unique system of dog ratings in WeRateDogs, I decided to leave this outlier in the database. Still, somebody might type a random high number, so user should take it with discretion. The lowest rating is 2, which has three tweets, all of which are negative.

Last but not least, the distribution of prediction confidence is right-skewed, which means on average the prediction rate tends to stay on the lower side. However, as shown in the visualization, there are still many numbers of tweets that have confidence of near one. This distribution makes sense since there are still limitations in classifying dogs using the CNN algorithm.

Below are the two visualizations generated to support the insights.


1200
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Confidence of Prediction


Rating numerator Boxplot
1750
1500
1250
1000
750
500
250
0
1