Analysing and visualising WeRateDongs

Introduction:

WeRateDogs is a very popular Twitter account with over 4 million followers and has received international media coverage. WeRateDogs gained its popularity by rating people's dogs with a good-natured comment about the dog.

About the data:

To analyze the tweets from WeRateDogs, three different sources has been used. The first source is an archive file of the past tweets from @dog_rates (https://twitter.com/dog_rates) in a CSV file

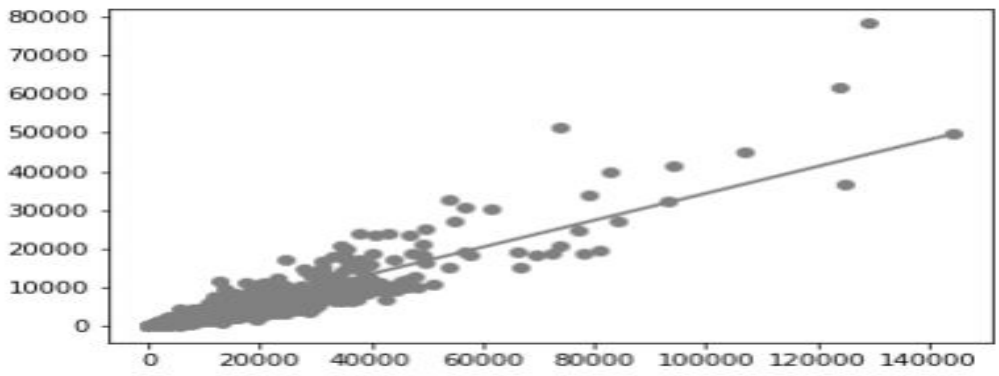
The second source is from the Twitter API used to retrieve more information about the tweets like number of tweets that were retweeted.

The third data source provides us the predicted dog breed in each tweet's image programmatically determined from a neural network.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us.
Additional gathering, then assessing and cleaning was required to present this analyses and Visualizations.

Favorites and Retweets of Dog Stages:

I first observed the relationship of favorites and retweets as I wanted to know how these metrics later related to the most popular dog. So a plot was made (shown below) that most dog tweets had more favorites than retweets.

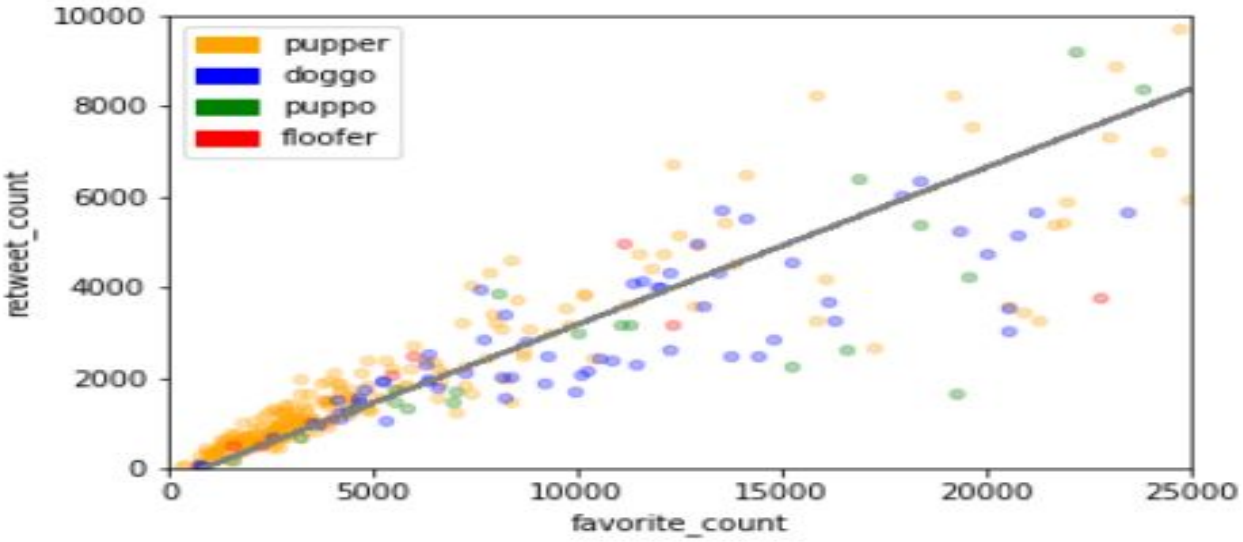


Relationship between the number of favorites and retweets

with x-axis as the number of favorites and y-axis as the number of retweets

This makes sense since more people will favorite a tweet instead of taking ownership of a retweet, so this information wasn't unexpected. A line of best fit was made (equation: retweets = (0.35)*favorites + -293) and had a correlation r of about 0.92. It should be noted though, that the data may be more quadratic especially for a high number of retweets and favorites. But this line will be good enough for now since there are not many high valued data points.

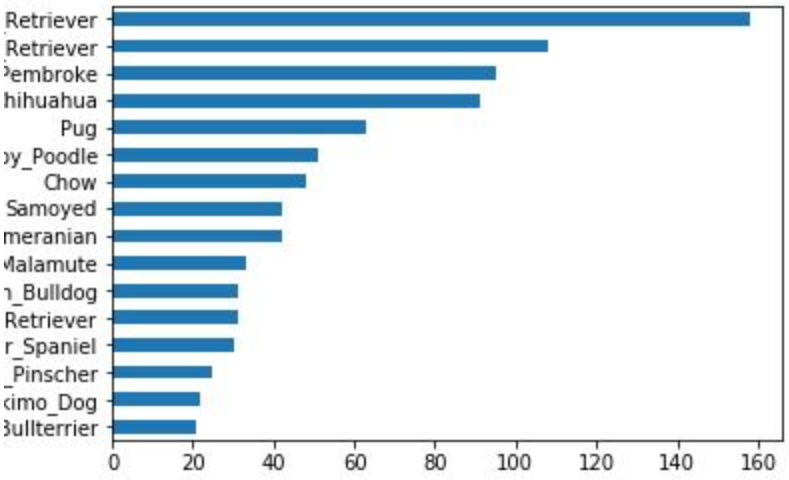
I next plotted the different dog stages to see if there is a relationship between the dog stages and a high number of favorites and retweets. I had to zoom in closer to the origin to show the majority of dog stages and I also didn't include the times the dog stage wasn't classified.



It can be seen that of all the dog stages tweeted, pupper was the most common. However, we see that doggos (the next most common) and puppos were favorited more than retweeted compared to the average tweet. This might be because these stages are less common than pupper or even not being classified as anything. Perhaps having a less common attribute contributes to being loved more? Maybe, but I think we can break our data more finely and better observe how dog breeds play a role. Most Tweeted Dog Breed

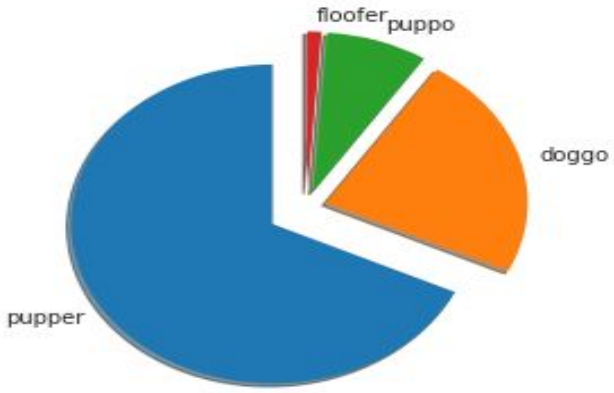
Coming from our previous graph and observeing that perhaps popularity of a tweet is related to its commonality amongst its fellow tweets. When sorting for the most common dog breeds, I found that nearly 500 tweets were not labelled as a dog breed. This might be perplexing. Were so many tweets not of dogs?

Though it is possible some of these tweets weren't dogs, it is more likely that the data for the dog breeds is not accurate. This could come from the fact the neural network did not recognize some breeds. Maybe some dog breeds were harder to classify for the neural network. But since we don't have access to the neural network and I think it would be a better use of our time to use what we have, we'll ignore this issue. We'll just keep this in mind as we investigate further and maybe revisit this another day to improve the data's accuracy.
So for now, we'll just sweep this small issue under the rug (which I never do while cleaning the house of course). So we march forward and plot our fifteen most common dog breeds



We see that the most popular dog breed is the golden retriever, followed by the labrador retriever, the pembroke (aka the much internet beloved corgi), the chihuahua, the pug, and so on. Those five most common dog breeds seem to me to match what you see on the internet. So I would say that this distribution of dog breeds seems pretty representable of what I would expect.

Then i have partitioned my data by dog stages in a pie chart. This pie chart will represent the participation of each dog stages



Through this pie chart we can see that pupper holds major part of our pie.