WRANGLE AND ANALYZE TWITTER DATA

By Kalpana Srinivasan

When you sit back and think about how far Twitter has come since it launched in 2006, its rise to glory is impressive. It’s difficult to make your way through a day without seeing a tweet referenced on television, the radio, or on a news website. It doesn’t mean that Twitter is the biggest or most popular company or service in the world, but it does prove that its social impact has reached a level that not many technology companies have reached.

The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user **@dog_rates**, also known as **WeRateDogs**. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. WeRateDogs has over 4 million followers and has received international media coverage. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent."

The data has to be gathered from three different sources. WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. The images in each tweet was run through a neural network in order to predict the breed of the dog. These predictions are hosted on Udacity servers and have to be downloaded programmatically using the Requests library. Using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and I stored each tweet's entire set of JSON data in a file. This can be used for getting each tweet's retweet count and favorite ("like") count, which would be used for analysis.

The final master dataset would be the merge of all 3 data frames. It has the main columns to consider when we are analysing.

On checking the statistical summary, we can observe that the rating numerator has an outlier of 1776. This means, a dog that has got this big rating from people, must be something extra special.

	rating_numerator	rating_denominator	retweet_count	favorite_count
count	2175.000000	2175.000000	2175.000000	2175.000000
mean	13.215172	10.492874	2703.164598	8685.298851
std	47.725696	7.019084	4652.832236	12367.888591
min	0.000000	0.000000	0.000000	51.000000
25%	10.000000	10.000000	587.500000	1865.500000
50%	11.000000	10.000000	1300.000000	3948.000000
75%	12.000000	10.000000	3112.000000	10862.000000
max	1776.000000	170.000000	77234.000000	143195.000000

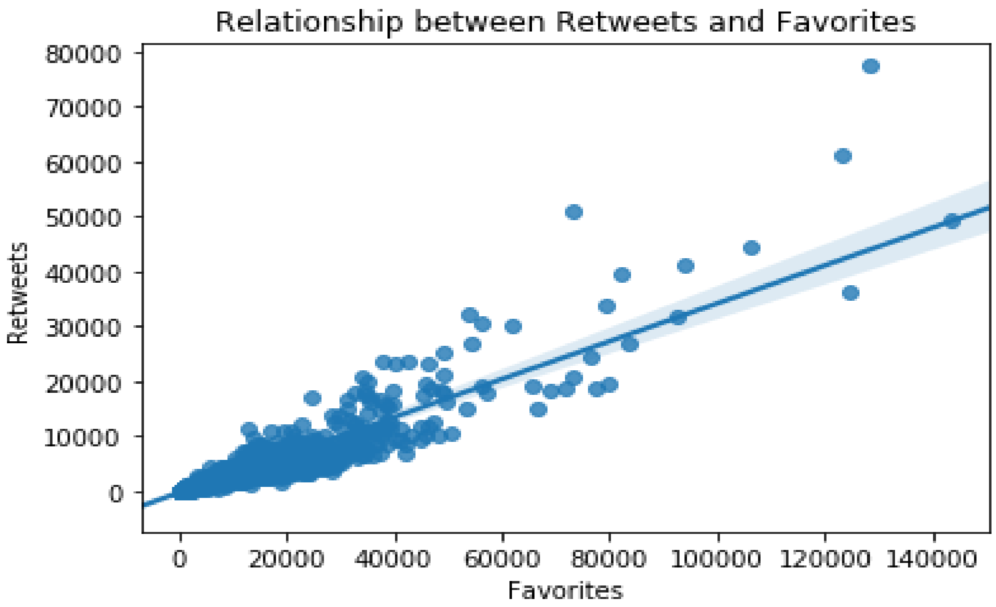
It turns out to be the dog named Atticus (shown below) had posed for America's Birthday with sun glasses and bowtie. No wonder she has got the highest rating. The neural network missed recognising it as a dog because of its accessories!

The dog that has the highest favorite count has 143195 likes.

The most favorited dog is not the most retweeted one. There can be two explanations for this. One is that people tend to favorite more often than retweeting the original tweet. Second one is people might have liked the dog but the tweet itself was not catchy (maybe!) which means it is not about just the dog, there is more to it.

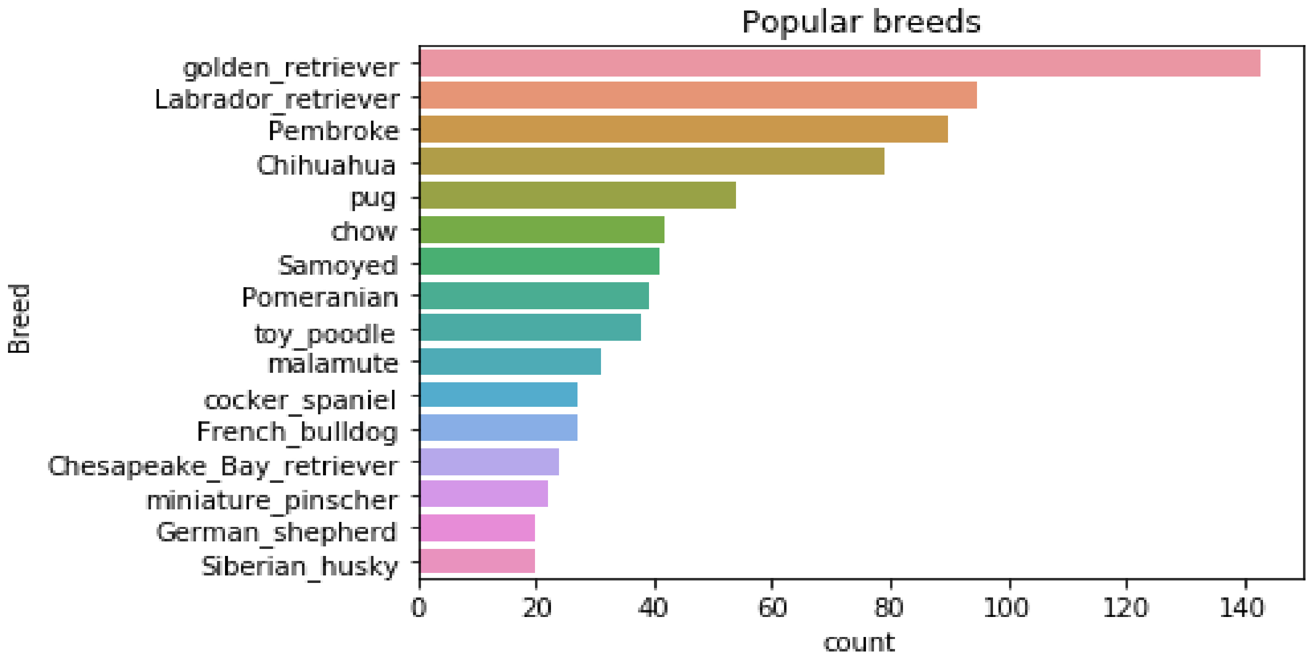
Let us look at the relationship between retweet count and favorite count.

Seeing the correlation coefficients, we can see the highest positive correlation is between retweet count and favorite count. I plotted them to understand the relationship better.



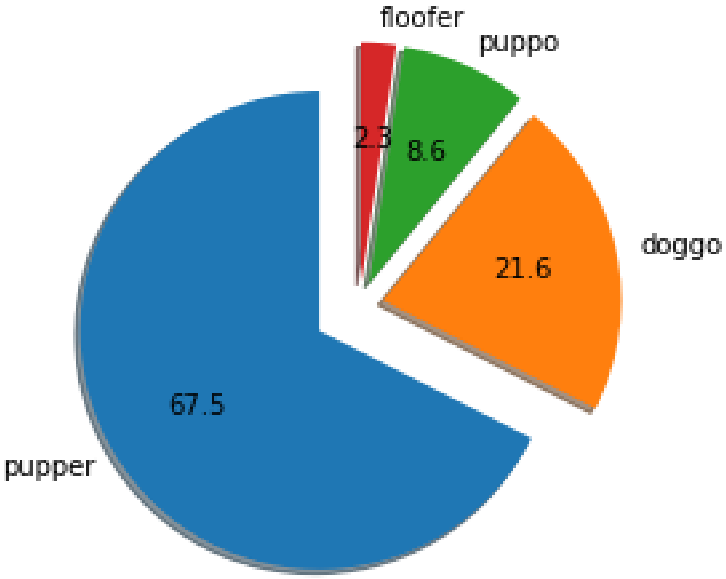
It is evident from the large favourite count that people favorited the tweet more often than retweeting it.

Next question I was thinking was" which is the popular breed? "On examining the given data, we can make out that Golden retriever and Labrador retriever are the popular breeds. This might be because many people would have tweeted about them. Or the image prediction neural network was able to accurately identify these breeds compared to the other breeds.



There are different dog stages as described in WeRateDogs: doggo, pupper, puppo, floof(er).

Let us see how well it is distributed in the dataset. There are many missing data in this column. With the data available, the distribution is as follows.



SUMMARY:

On the whole, the analysis of WeRateDogs shows that the retweet count and favorite count are strongly correlated. The dog with highest favorite count is a cute Lakeland terrier. It is not the most retweeted tweet. One dog Atticus, though it got the highest rating, was not identified by the neural network as dog because of its stylish outfit. Most of the dogs in the dataset are pupper.