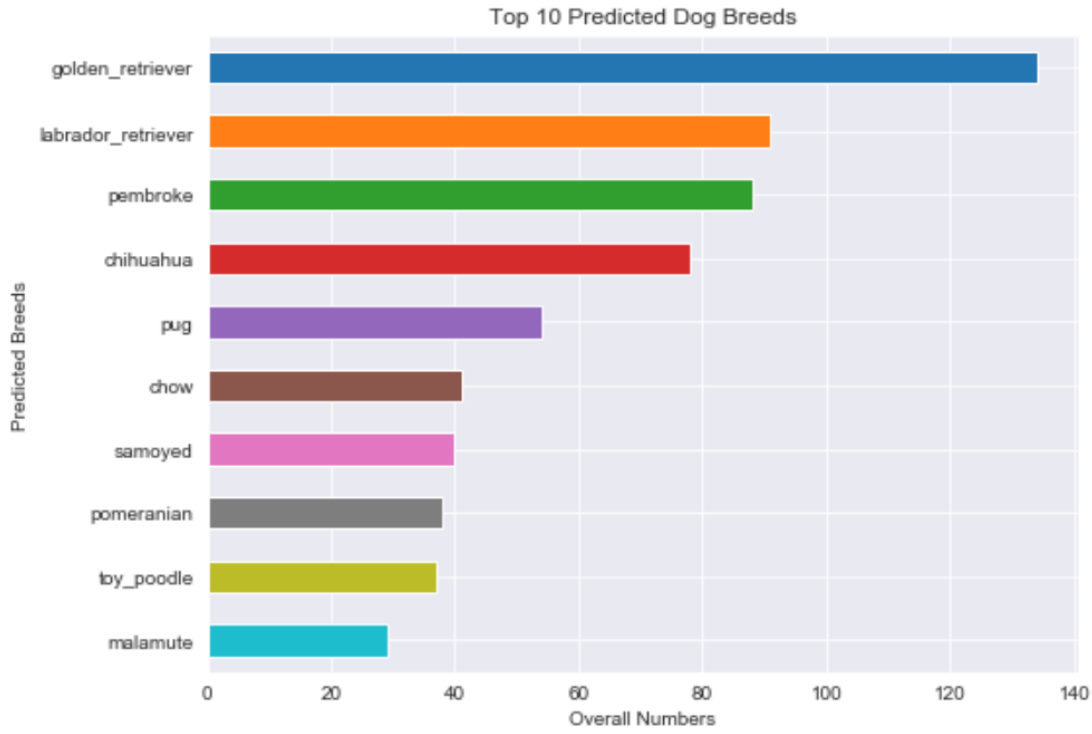
Analyzing and Visualizing Data of WeRateDogs Tweets

WeRateDogs (@dog_rates) is a Twitter user that posts and rates people's dogs. We gathered the 5000+ tweets in WeRateDogs Twitter archive. Also, Udacity provides us the data of the dog breeds, which is classified by neural network based on the pictures in WeRateDogs Twitter archive. Let's analyze and visualize the dog data we have.

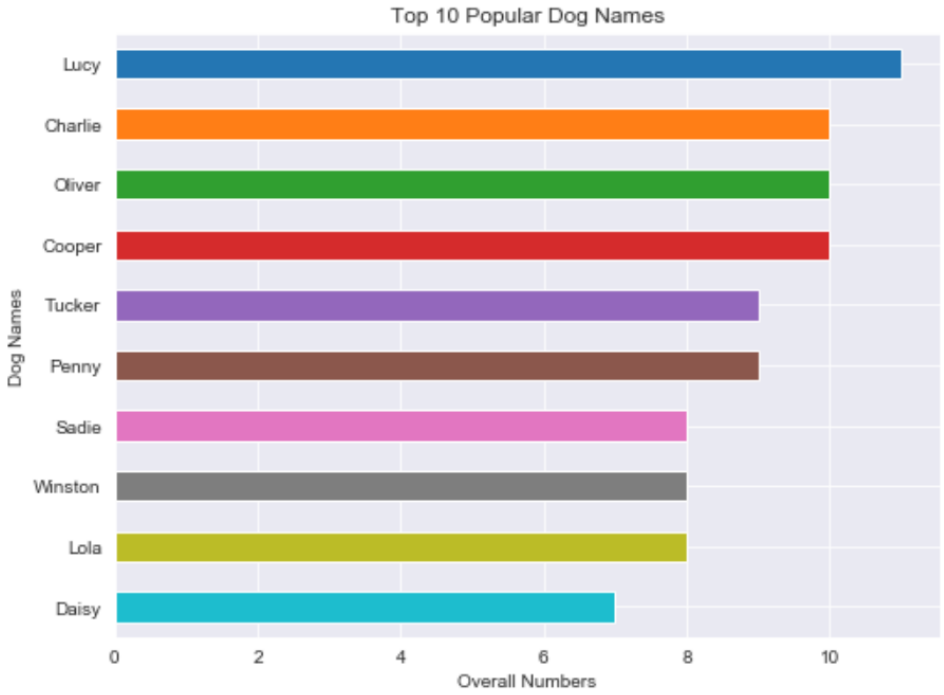
Popular Dog Breeds

The "Top 10 Predicted Dog Breeds" plot is generated based on the prediction from neural network. The Top 3 most popular dog breeds are: Golden Retriever, Labrador Retriever, and Pembroke (Corgi). The prediction results look reasonable. And not surprisingly, the most popular dog is Golden Retriever, which overall number is much higher than the other breeds.



Popular Dog Names

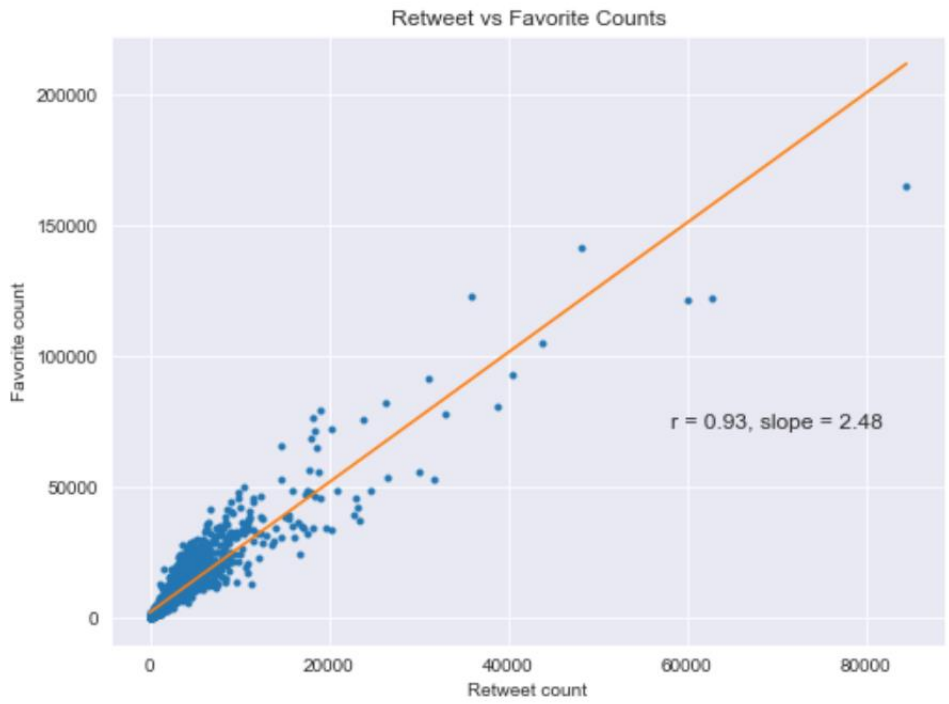
What is the most common dog name in the tweets data we have? From the horizontal bar chart below, we can tell it is "Lucy" is No.1, and "Charlie" is No.2. I actually know some of my friends'' female dog called "Lucy" and male dog called "Charlie". I think I will pick a name from this top 10 list for my future dog!



Favorite counts and Retweet counts

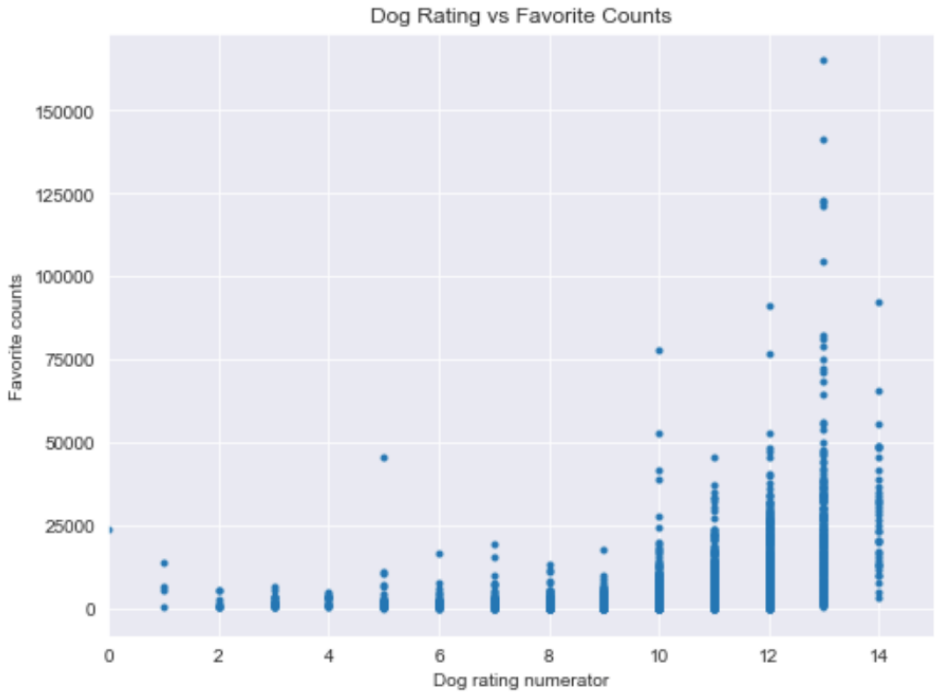
Next, we take a look at the correlation between “favorite counts” and “retweet counts” of each tweet. It does not surprise me that the favorite counts and retweet counts are correlated. Actually they are strong correlated as the r value is 0.93.

For each tweet, the number of the favorites is about 3 times as the number of retweets. It makes sense because people tend to click “favorite” button rather than retweeting.



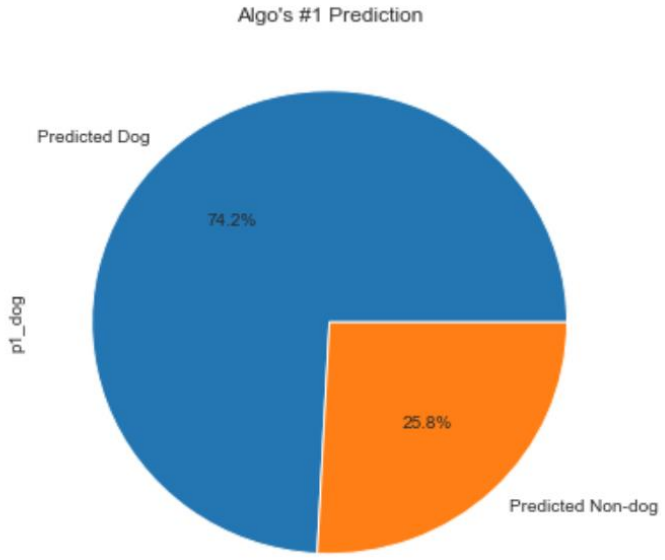
Dog Rating and Favorite counts

Would the dog rating be correlated to favorite counts? It might not be strong correlated, but dogs received higher than 10 rate get more favorites. It is also interesting that the top 5 favorite counts belong to dog rate of 13 rather than 14 (out of 10).

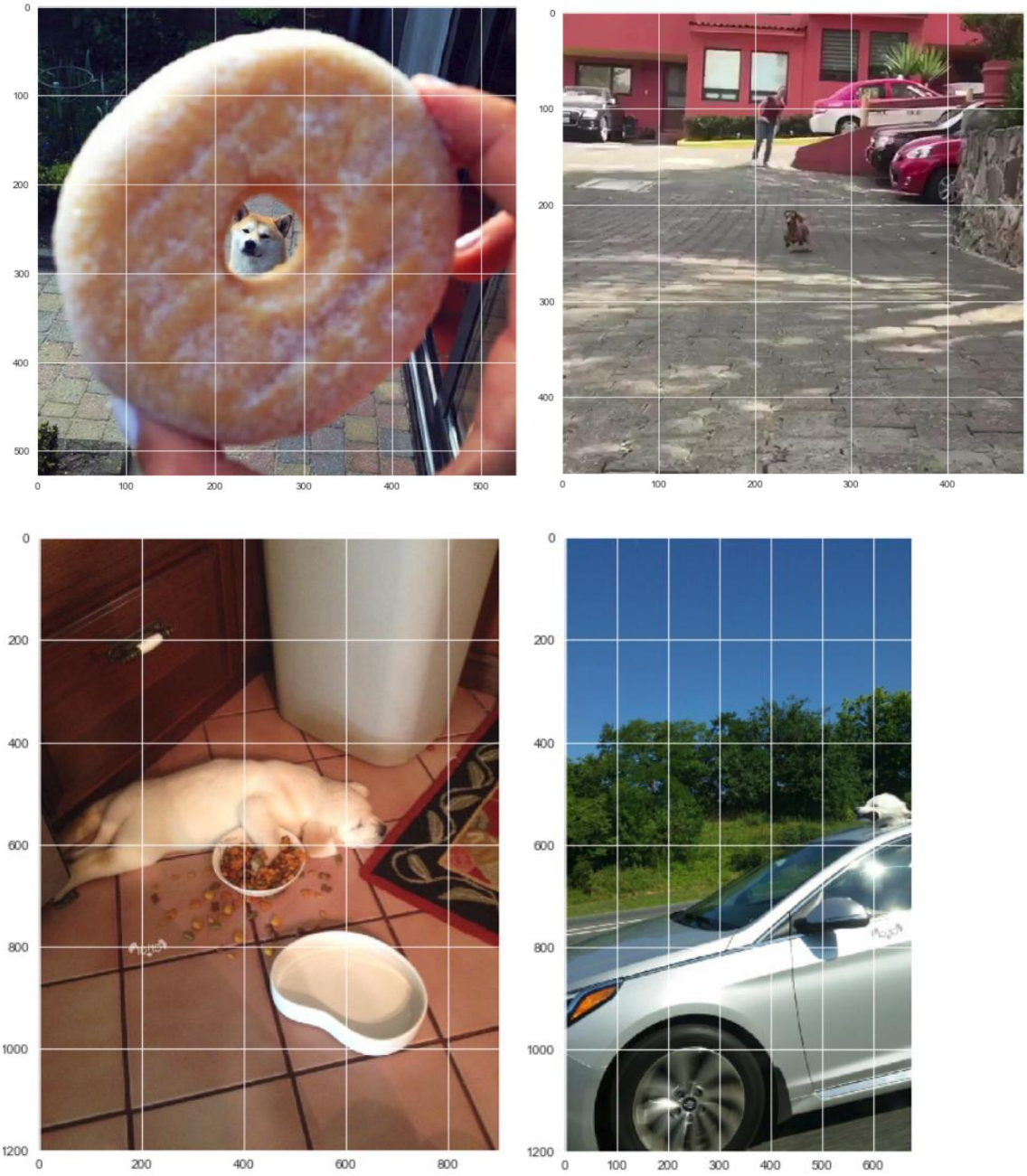


Neural Network Prediction Accuracy

From the pie chart below, we can tell that the recall of the dog / non-dog classification neural network is about 72% (it is quite safe to assume all the pictures are about dogs). 72% recall does not seem to be a very good accuracy.



Let’s take a look at a few mislabeled dogs’ pictures:



A few types of images the dogs tend to be mislabeled include:

- Dog body in an unusual position
- Dog is very small in the picture (scale variation)
- Camera Angle
- Some other items appear at the center of the picture

Likely this neural network was trained base on the online ‘perfect’ animal pictures. However, twitter uses like to take pictures of their dogs in a ‘funny’ way. To improve this neural network dog classification performance, we would suggest to collect more diverse training set.