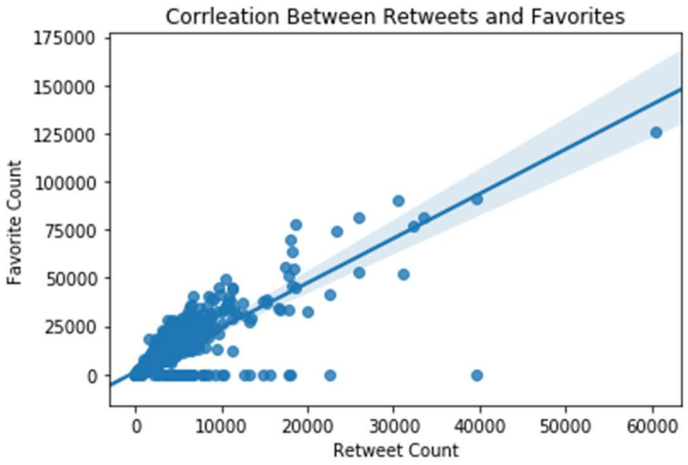
WeRateDogs Data Analysis

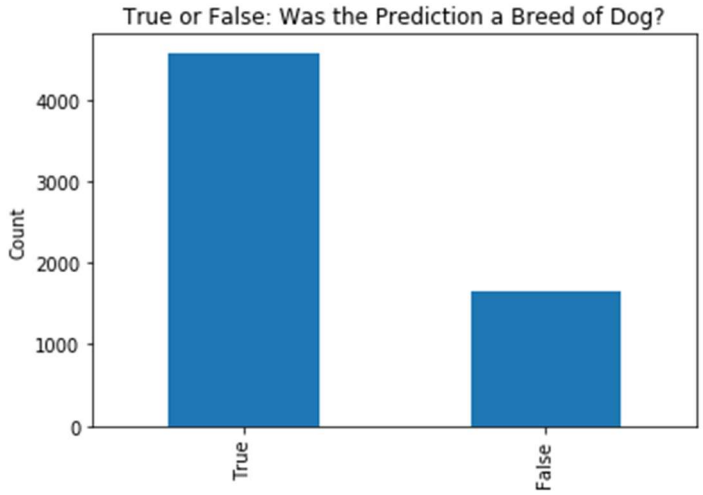
WeRateDogs is a Twitter account to rate people's dogs. After wrangling tweet data from multiple sources into an (acceptably) usable dataset, I set out to analyze the ratings of the dogs. Considering the humorous, tongue-in-cheek style of the ratings, the following probably not be taken too seriously!

To begin, I checked for correlations between variables. After running the Pandas corr() function (to find the Pearson correlation coefficient) I found the highest correlation between the retweet_count and favorite_count variables. Here is what that looks like:



Okay, so there is a relationship between favorites and retweets, but that will most likely be true for any set of tweets, right? Can we learn something about the dogs?

Before looking at the dog breeds, I looked the at the images. How many of the images were actually dogs?

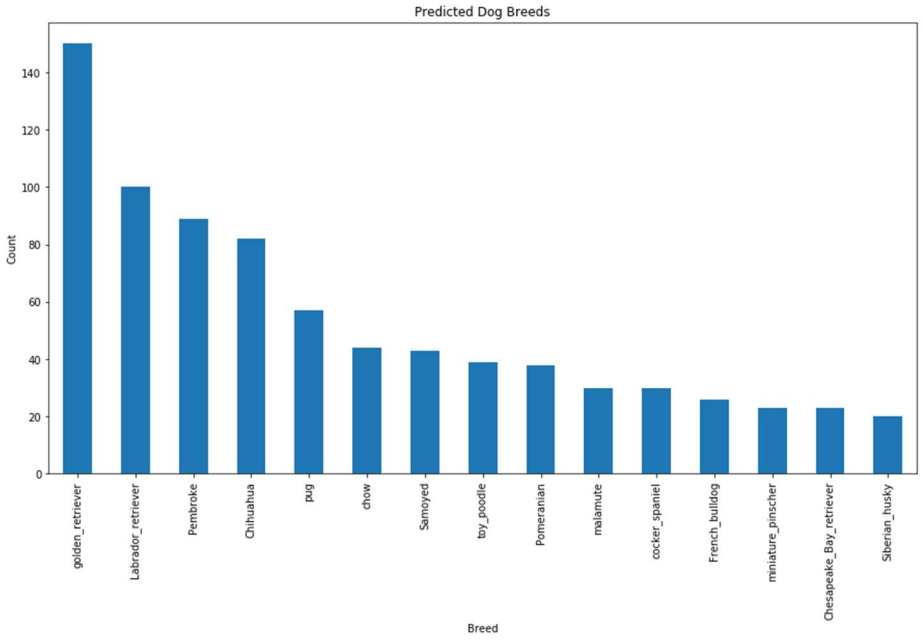


4581 of the images were actually dogs; 1641 were something else.

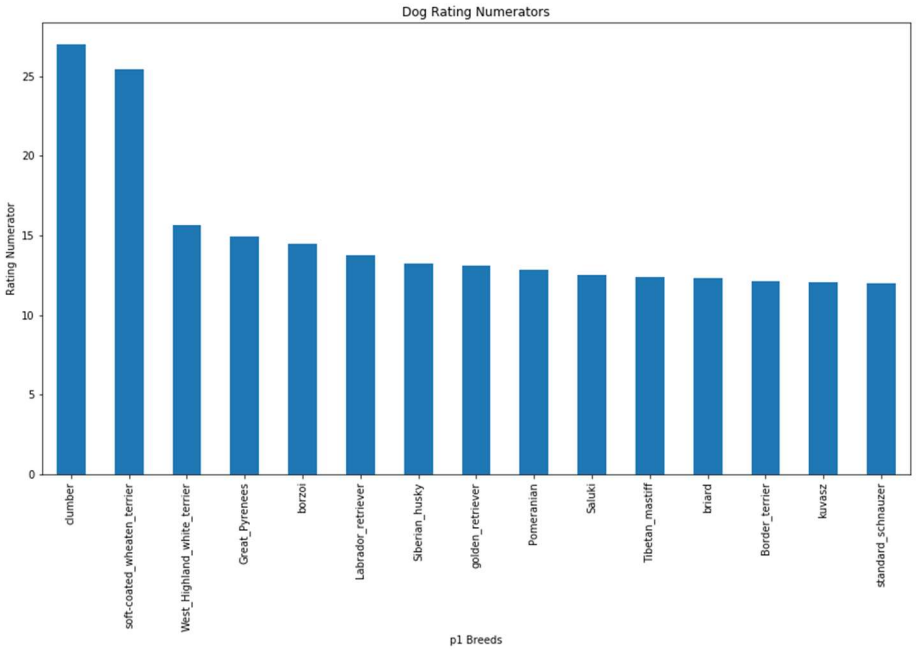
Next, in order to choose a prediction of breed to work with, I looked at how confident the prediction algorithm was with each prediction:

- Prediction 1: **59% confident**
- Prediction 2: **13% confident**
- Prediction 3: **6% confident**

I chose to work with prediction 1 to find the most commonly predicted dogs and the highest rated dogs. Here are the most predicted dogs:



And here are the highest rated dogs:



So golden retriever was the most predicted dog and clumber was the highest rated dog. Keep in mind, WeRateDogs uses denominators of 10 and an anything-goes policy for numerators. As such, I only used the numerators to select the highest rated dogs. These ratings are fun, not a strict analysis of what breed of dog is most loved.

I was not familiar with the breed of dog clumber. I though the algorithm had made some mistake or that I had missed something important in the data cleaning; however, there is actually a real breed of dog called the clumber spaniel!