DATA ANALYSIS NANODEGREE

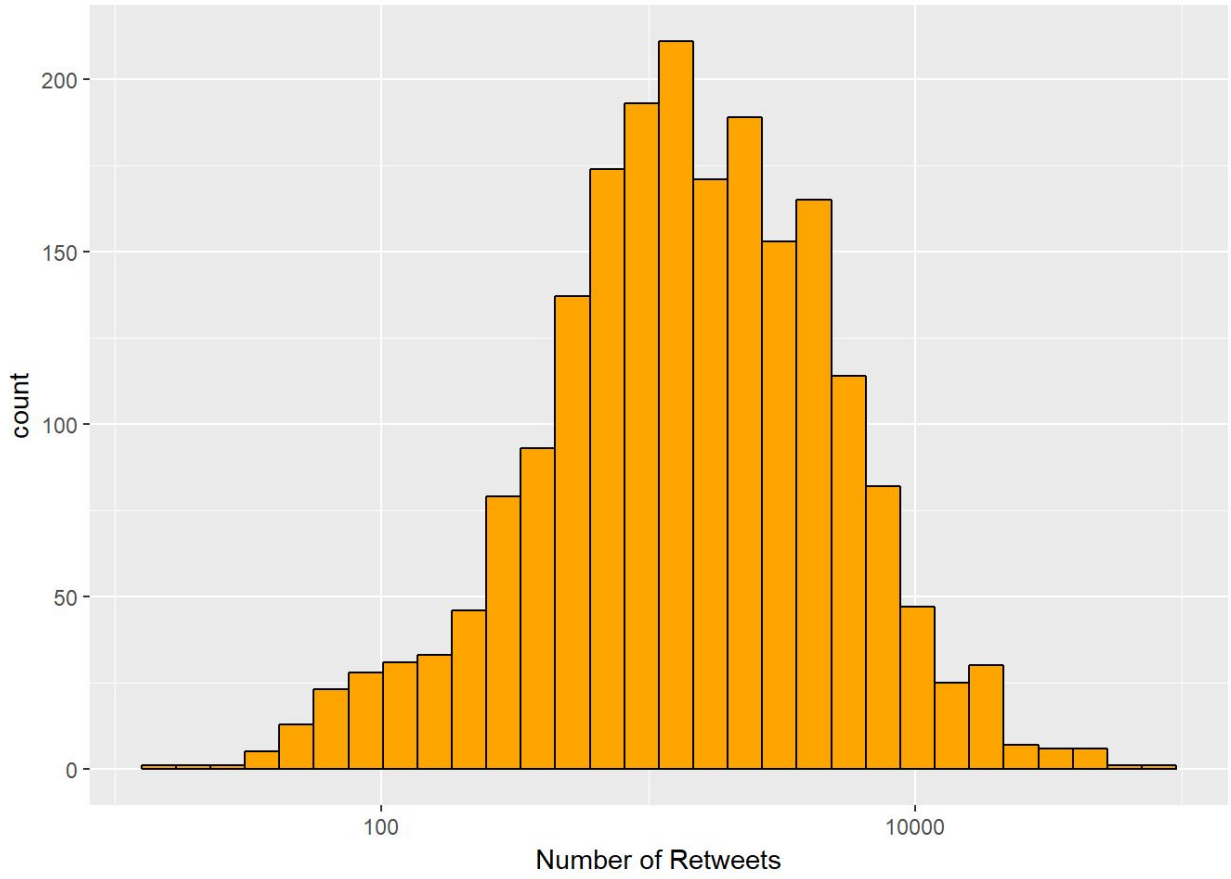
WeRateDogs - Data Wrangling Project

By Kai Xi

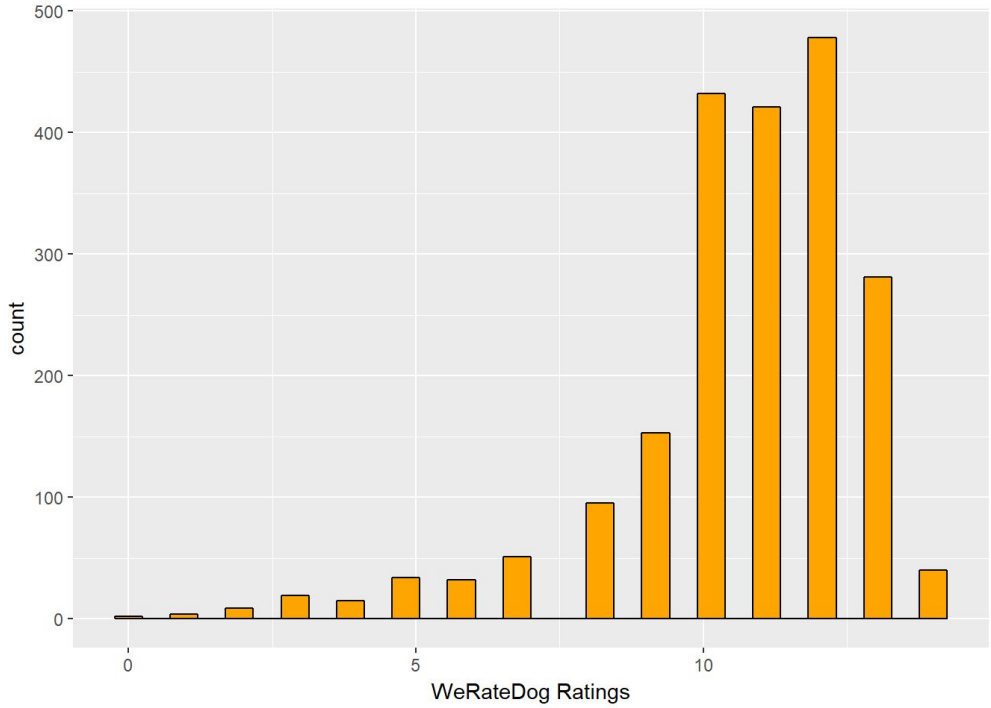
This project examines more than 2000 tweets from the account WeRateDogs, where pictures of dogs are rated on a 10 point scale. Some of the dogs are also categorized into puppo, pupper, doggo and floofer. The dataset contains the unique tweet IDs, the tweet texts, retweet counts, favorite counts, the timestamp, the dog names & types (extracted from the text body) and the breed (as determined by a machine learning model).

Insights

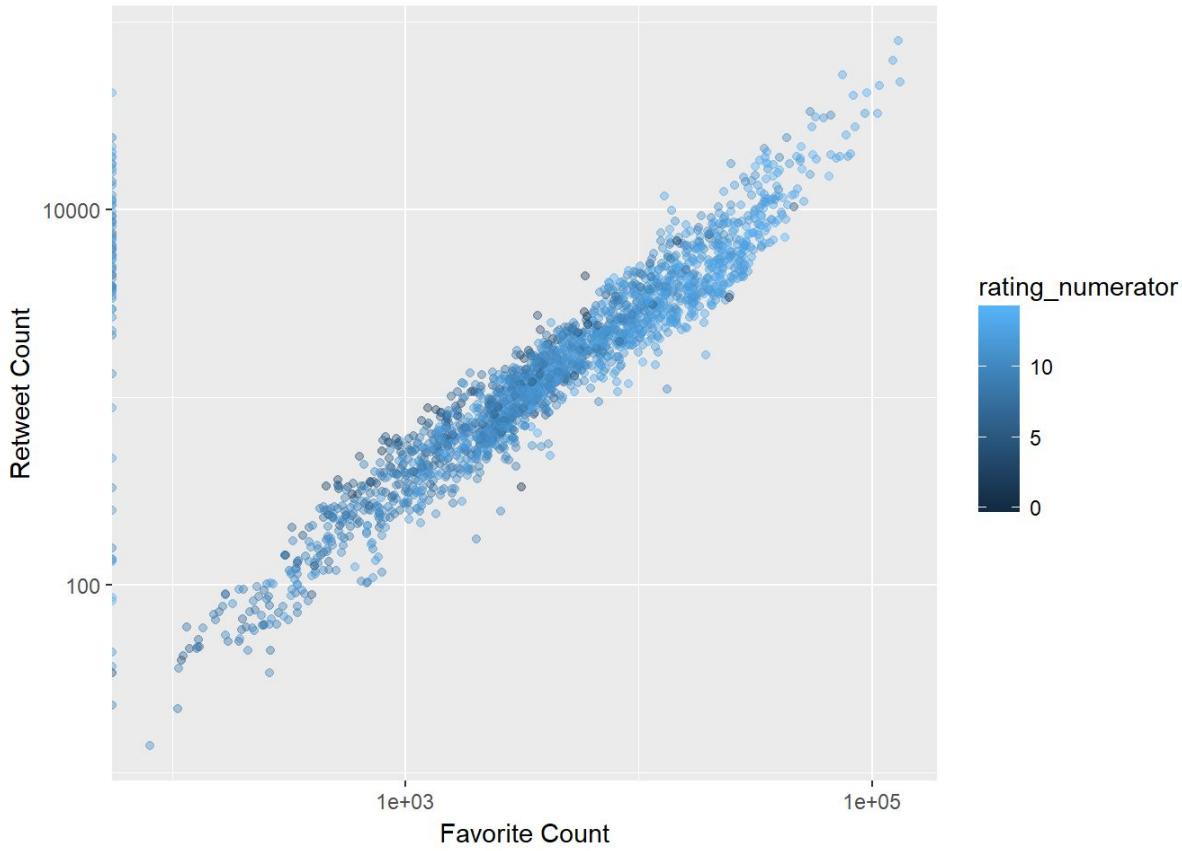
The distribution by retweet counts, after applying a log 10 scale, looks pretty close to a normal distribution with mean around 3200.



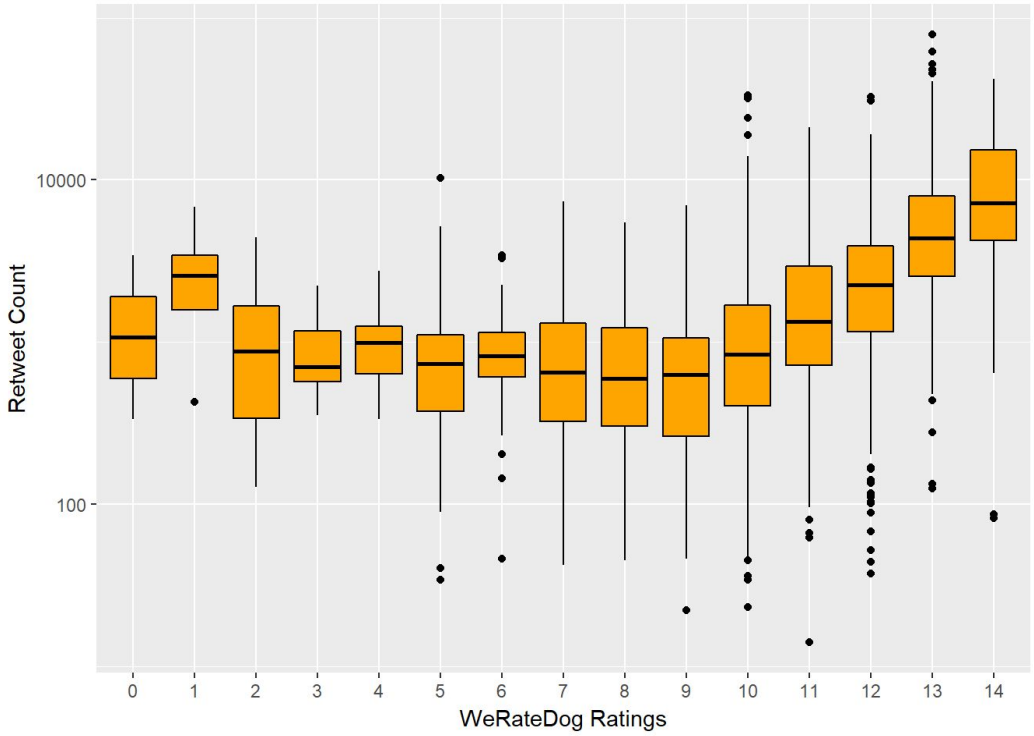
Most ratings are 10, 11, 12 and 13.



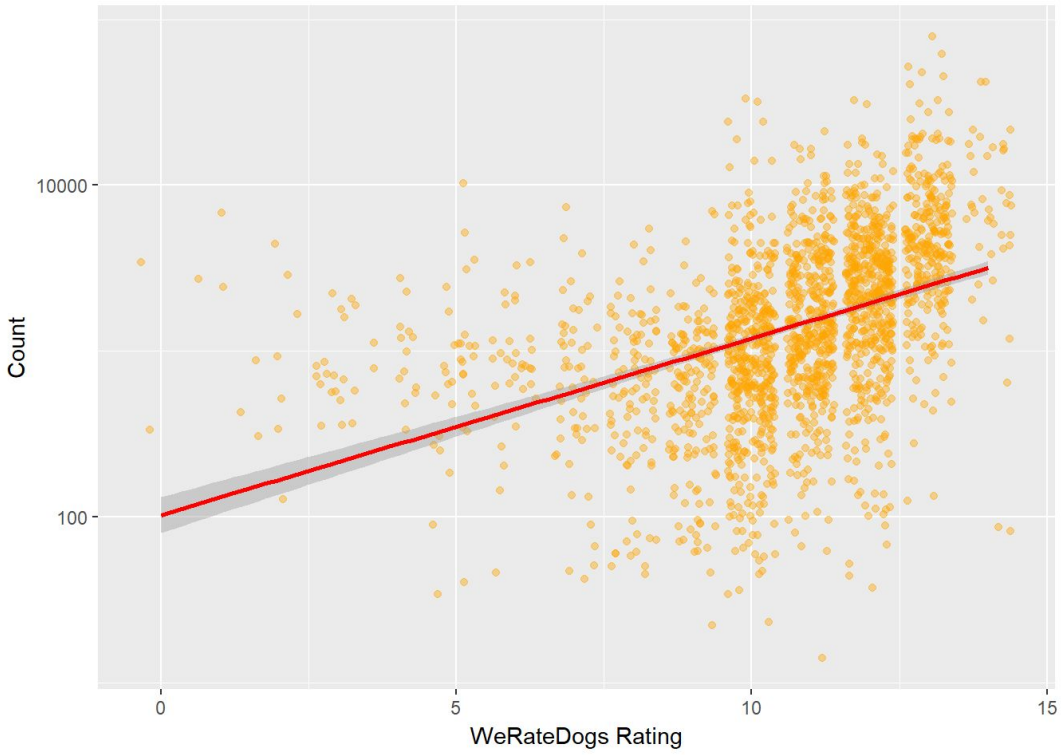
Not surprisingly, favorite counts and retweet counts have almost perfect positive correlation. After applying a color scale to the rating, it looks like the more highly rated (lighter colors) tweets received higher retweets and favorites.



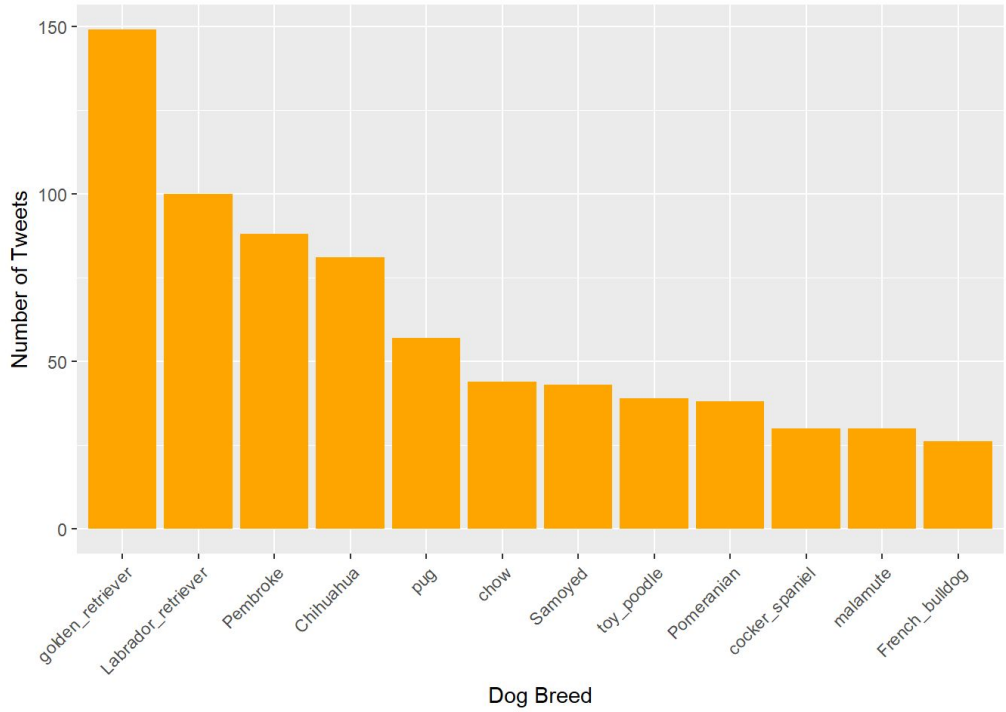
It looks like higher ratings do get more average retweets and favorites



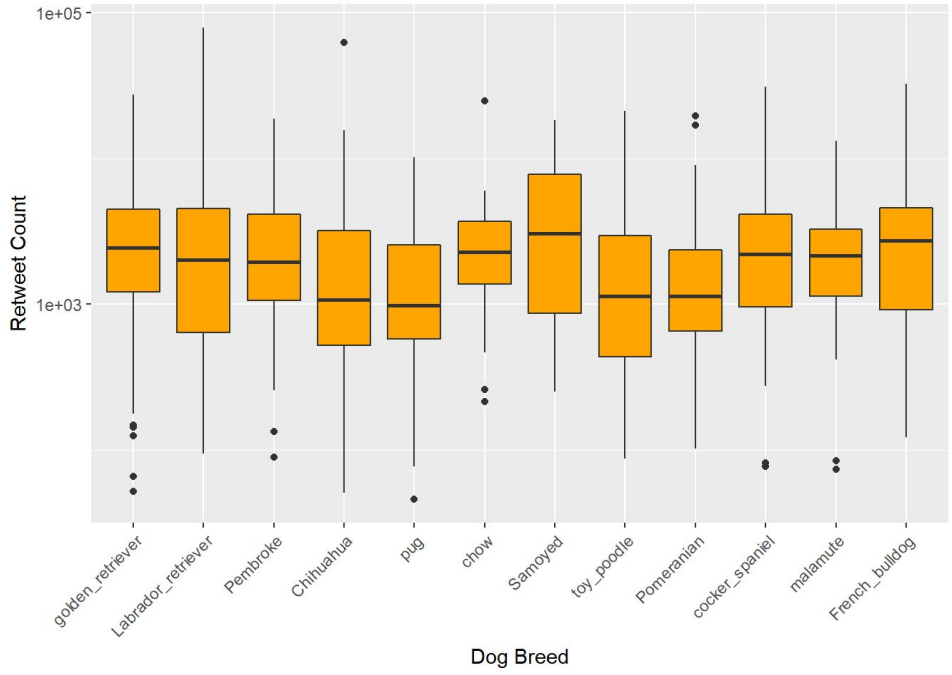
To explore the relationship between dog ratings and retweet/favorite counts, I plotted dog ratings against retweet counts (scaled with log 10) and added a regression line. There is indeed a positive correlation between ratings and retweets, as indicated by the upward sloped regression line.



I also examined tweets and rating by breeds. Subsetting the pictures identified as dogs (such that we exclude the ones unidentifiable), I rank all the breeds by number of times tweeted. The plot below shows the 12 most tweeted breeds of dogs.



There is no noticeable difference in terms of retweet counts across different breeds of dogs.



Lastly I looked at the retweets and ratings by dog types. It looks like puppos, floofers and doggos have significantly more retweets and higher ratings comparing to puppers and all the other dogs.