

Analysis and Visualization of Data Set WeRateDogs

by Laila Shahreen(November 2020)

Name a dog that you would call a family!

Do you love dogs? I do. If you want to know more about dogs, let me take you to the twitter page[https://twitter.com/dog_rates?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor]. This page tells a story about dogs of several stages and you can't leave without laughing! Also a great resource for finding your friend if you ever feel to call one a family.

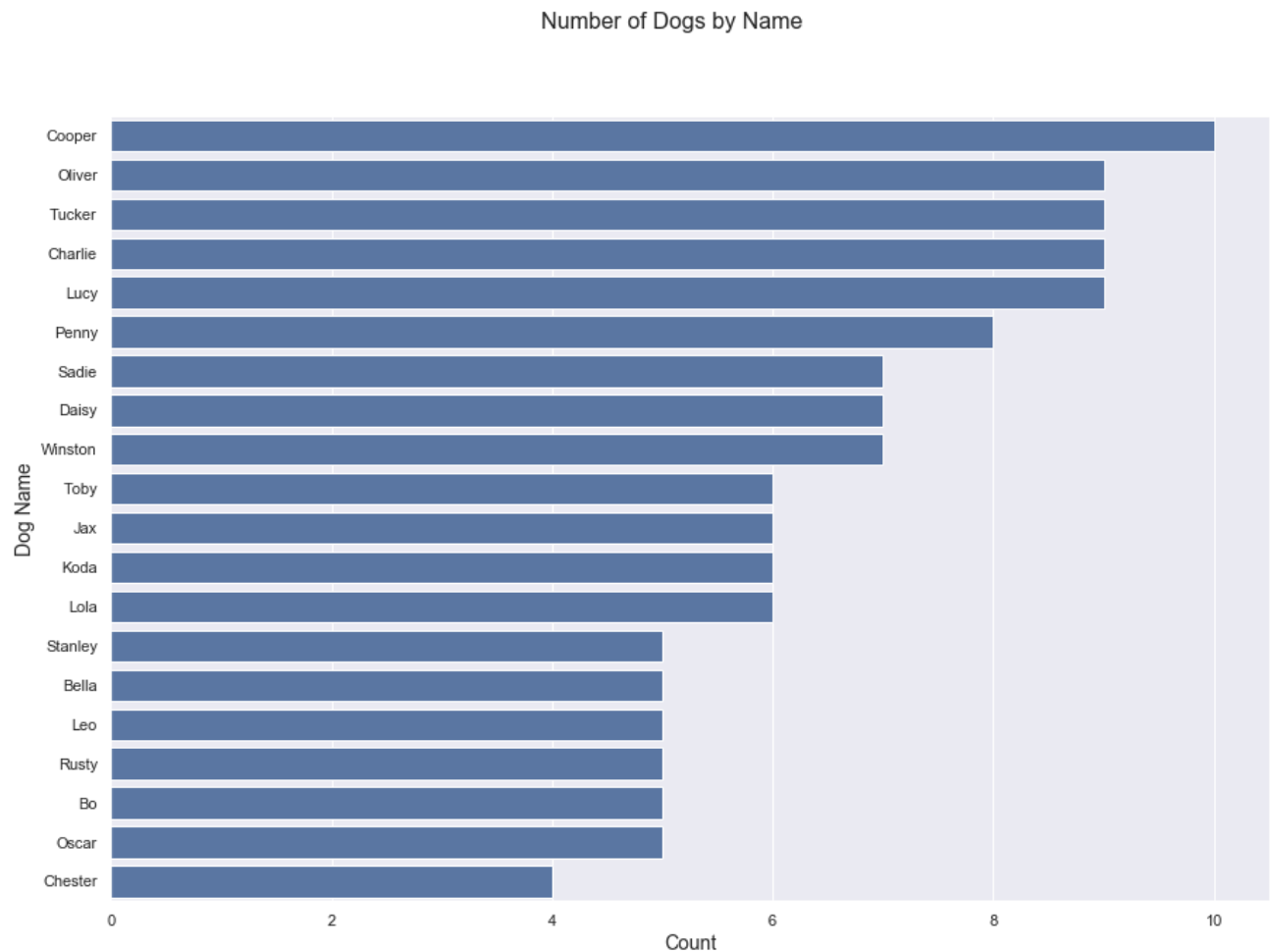


This is Jaeger. He tried to get up the steps. He really did. Does not want your help. Just gonna take a nap and figure it out later. 12/10

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. The twitter data set archive from Udacity has provided an excellent challenge to analyse several tweets' data for wrangling and analysing various information about dogs. WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. While the Twitter archive is great, it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations. Here I will take you to the important findings and insights in regard to dogs' rating, breed and also their fame. After preliminary visual inspection, programmatic data screening, assessing, and cleaning, I stored those tweet data in a master data frame called `twitter_archive_master.csv`. For our analysis and visualization we created a copy of each dataset i.e `twitter_archive_clean`, `predictions_clean` and `json_data_clean`. Since the

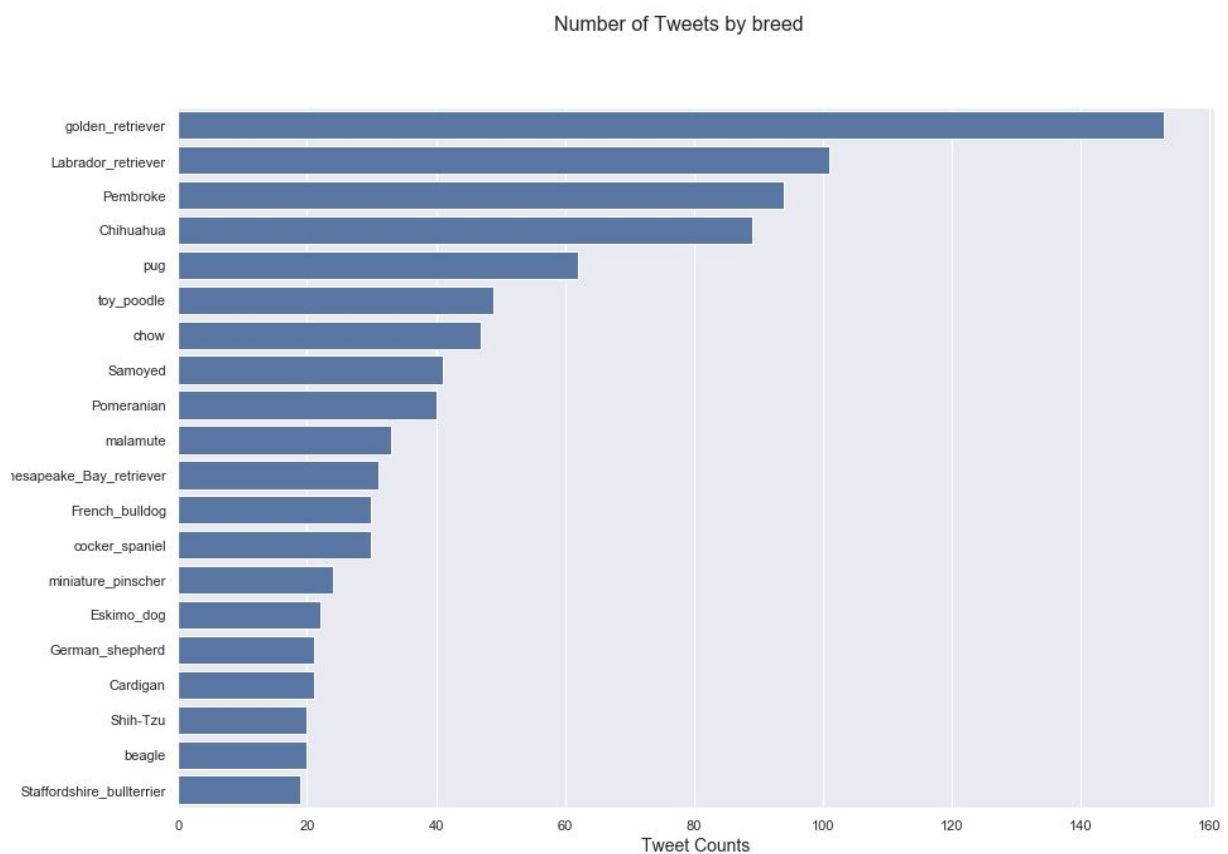
twitter_archive_clean data contains most of our variables of interest, we saved a copy of it as we_rate_dogs(1942 tweets) and then used it for further analysis. We did univariate, bivariate and multivariate feature analysis that provided the results about my favorite dogs and a potential friend in future I can pick.

What is the most popular name for a dog we rated?



From the above visualization, Oliver, Charlie and Cooper are the three dog names with 10 counts sharing top position. Lucy, Pennie and Tucker are at second position with 9 counts while Sadie and Winston secured third position. So, maybe I can get an idea of my future buddy's name. Wait, let's check the breed type which is most popular as well.

Which breed got most tweets?

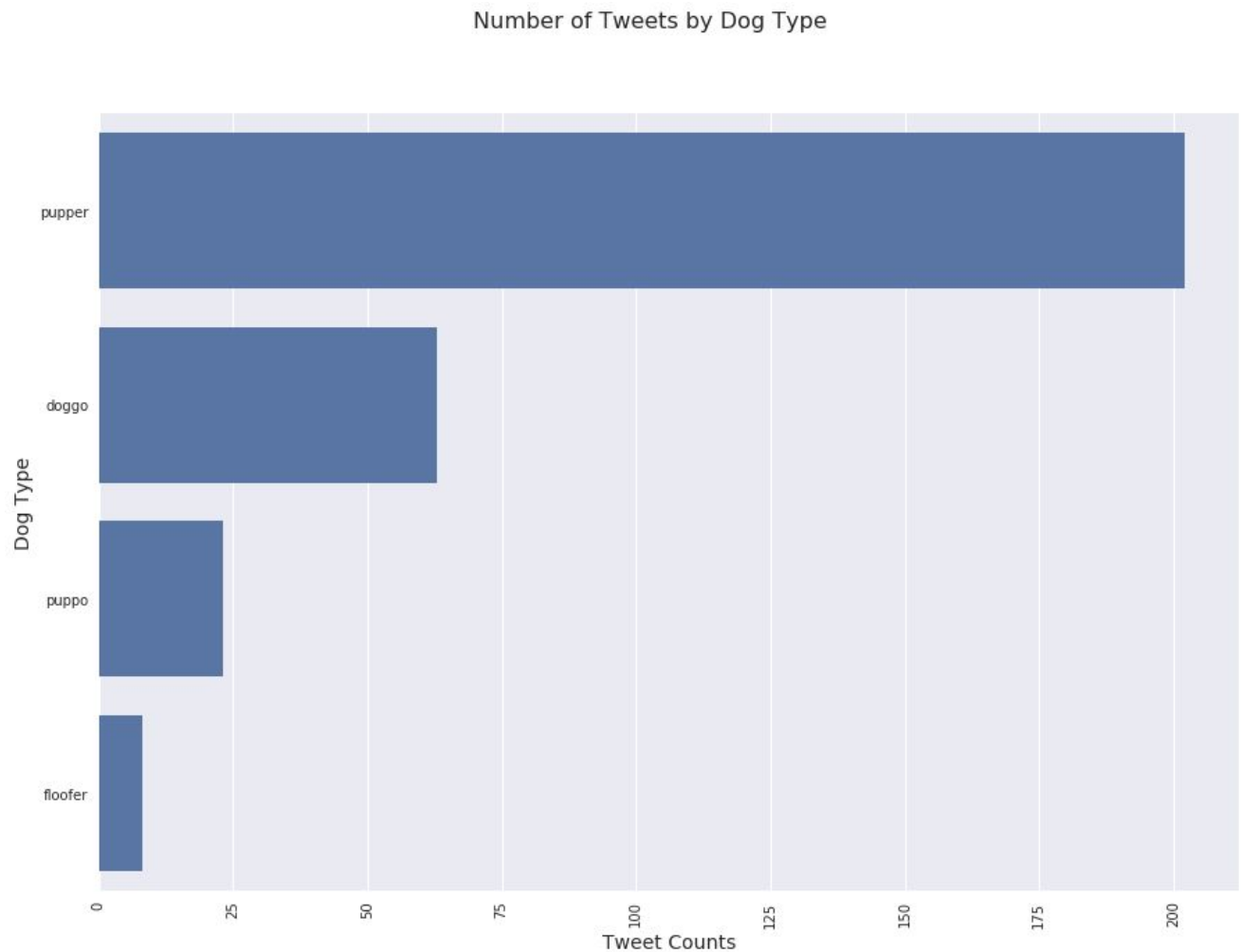


Golden Retriever is the most tweeted dog with 156 counts. Labrador_Retriever placed second with 102 counts and Pembroke got third place with 90 counts. Let's dive into other characteristics of dogs. Below is Zeke, a Golden Retriever.



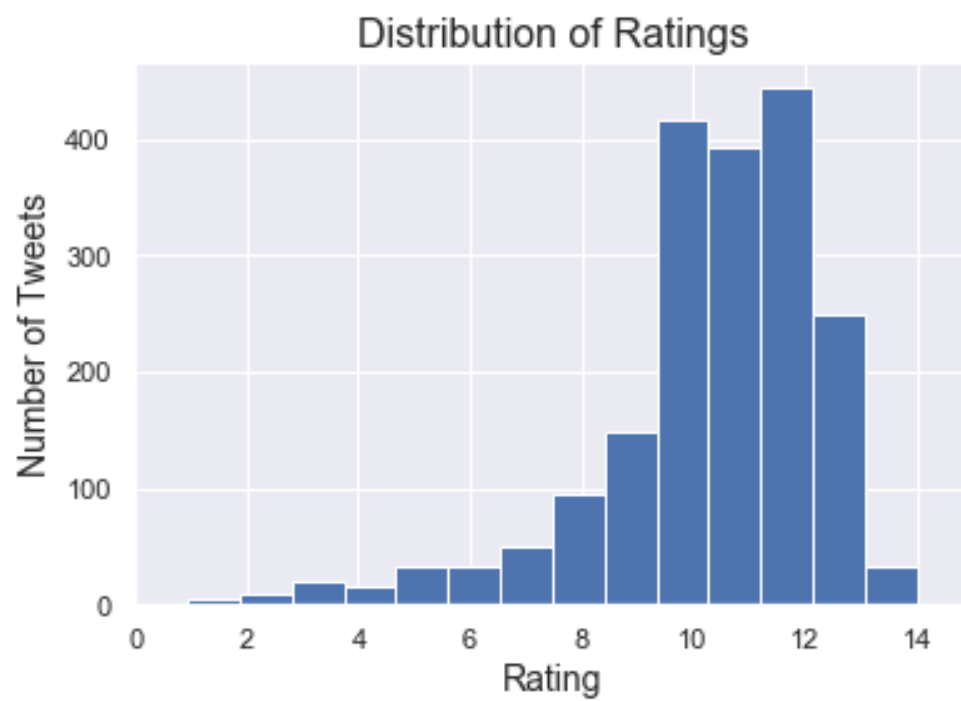
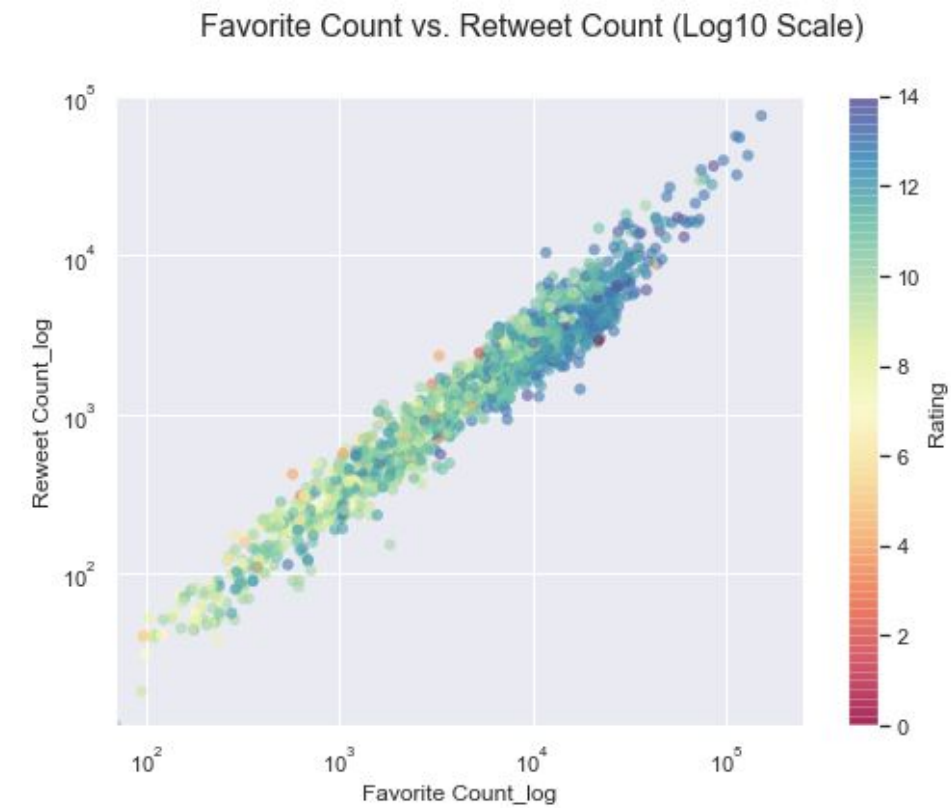
Can you guess the most common dog stage?

Looks like pupper is the most common dog stage among four types i.e. doggo, pupper, floofer and puppo. It has a significant percentage of tweets.



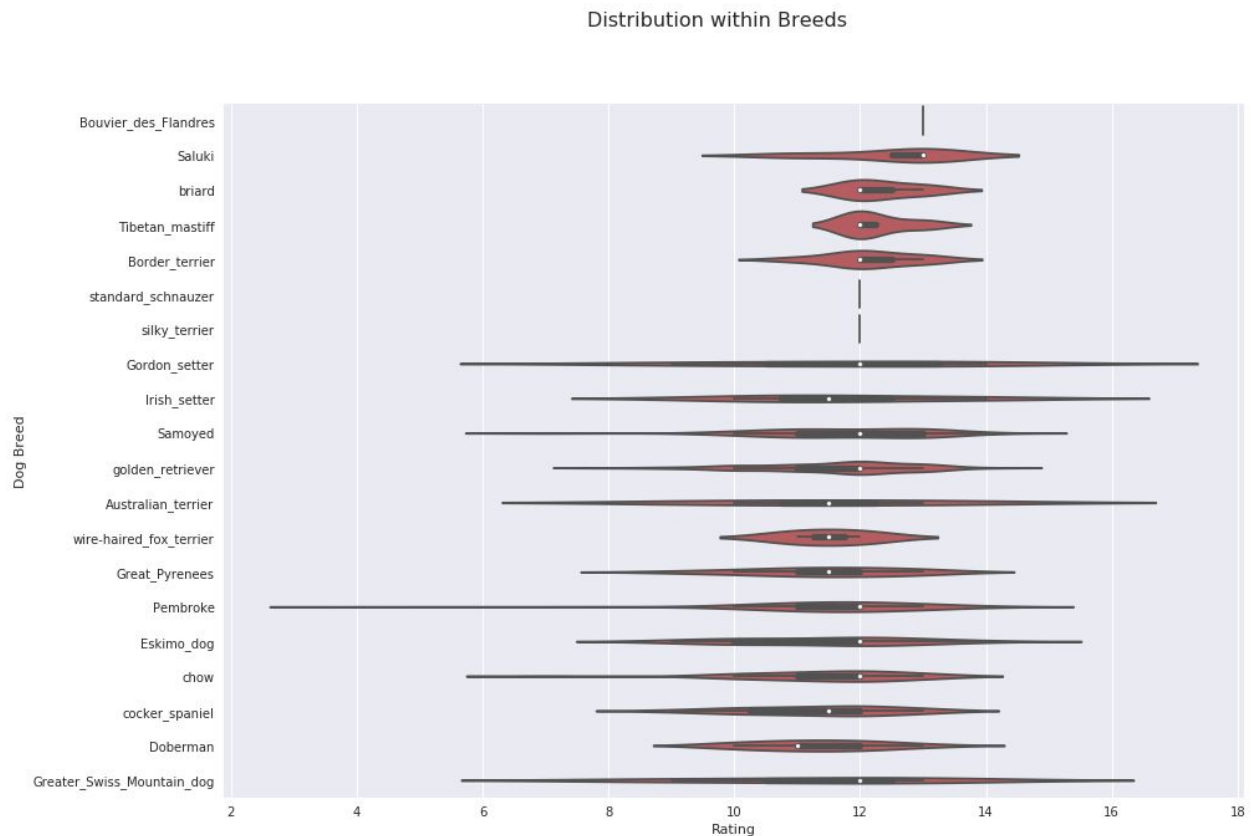
So, the most tweeted dog type is pupper. Obviously those young and cute pupper can win your heart anytime.

Did the rater do justice to the rating?



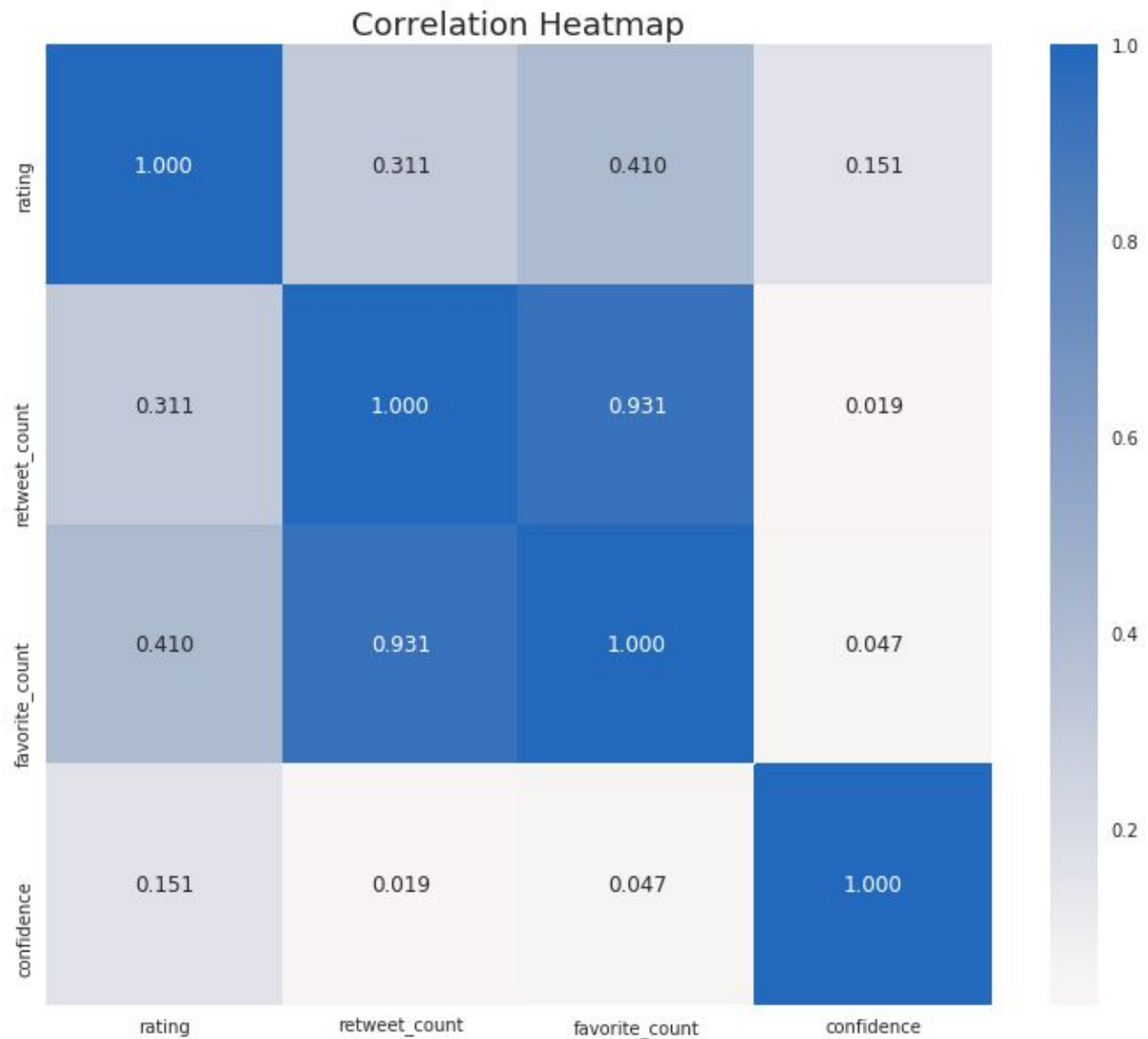
The distribution looks normal and mostly skewed towards the left. Majority of tweets got ratings that fell between 10-13.

How is the average rating distribution among breeds?



Golden Retriever also falls in the category of higher average rating and this can be another point to sort list my potential friend. Other high average rated dogs like Pembroke has lowest of around 2.7 and highest of 15.6 (wide range) whereas Briard, Tibetan Mastiff are of ratings 11.5-14 (short range).

Are retweet counts and favorite counts related to dog rating?



This is very clear from both heatmap and the colormap for retweet count and favorite count against rating that favorite count and retweet count are correlated with positive strength. Rating increased with higher number of retweets and favorite count which is expected. People marked their loved dogs as their favorite more than retweets. Confidence has shown very weak correlation coefficient value against retweets and favorite counts.

Finally I can pick a golden retriever that is favorite to the community and highly rated and tweeted by the dog lovers.