The wrangle report of WeRateDogs Twitter data

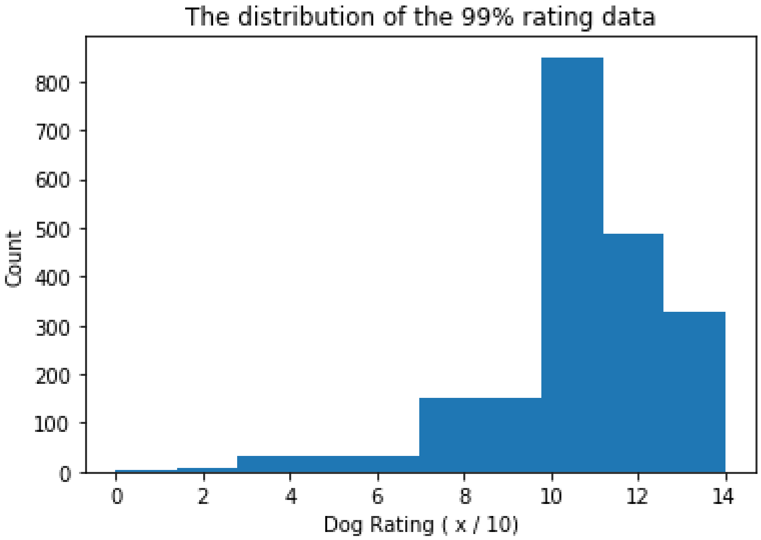
Introduction

The dataset that we wrangled 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. These ratings almost always have a denominator of 10. The numerators are almost always greater than 10. WeRateDogs has over 4 million followers and has received international media coverage.

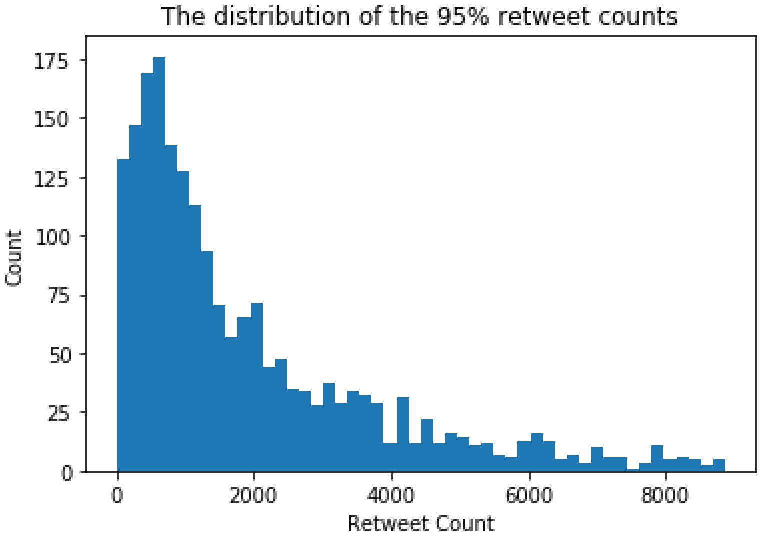
WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for us to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Insights and Visualizations

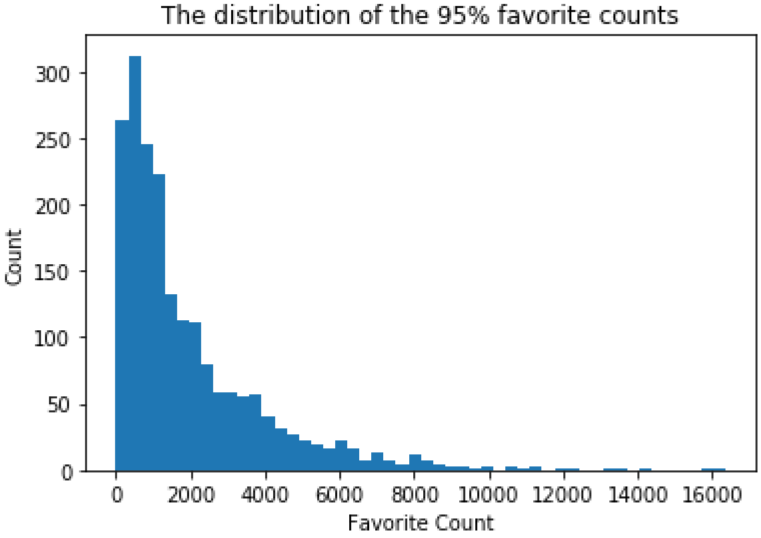
1. Since the major purpose of WeRateDogs is to rate people's dog. I'm interested to see some details about what ratings the dogs got. There are some tweets which rated multiple dogs together with rating like 170/17, etc. Since the majority is one rating for one dog, I removed those tweets with multiple dogs to make it consistent across all the tweets so we can have a better view on the ratings. After cleaning the data, it shows the median dog rating is 11/10. 99% of dogs got rating under 14/11. There is an extreme case of 1776/10, which can been taken as outlier. To plot the distribution of the rating, I chose to exclude the top 1% data.



2. I'm also interested how many retweets these tweets got. It can reflect how many users are reading these tweets or interested in these tweets. It turns out the distribution of retweet count is right skewed. To take a closer look at the data, I chose to select 95% data for the plot. It shows the median retweet count is 1305.5. And 95% user's retweet count is under 8876.15.



3. The distribution of favorite count is also right skewed, very similar to that of retweet count. The median retweet count is 3983.5. And 95% user's retweet count is under 30225.40. Generally the favorite counts are much higher than retweet count. I think it's because it's much easier for people to click favorite than retweet.



4. By checking the dog stage data, it shows not many people share their dogs' stage. There are 2084 tweets in total after the dataset is cleaned. But the total count with dog stage data is only 336. Among this 336, pupper is the most common stage,which is 230.