Insights and Visualization

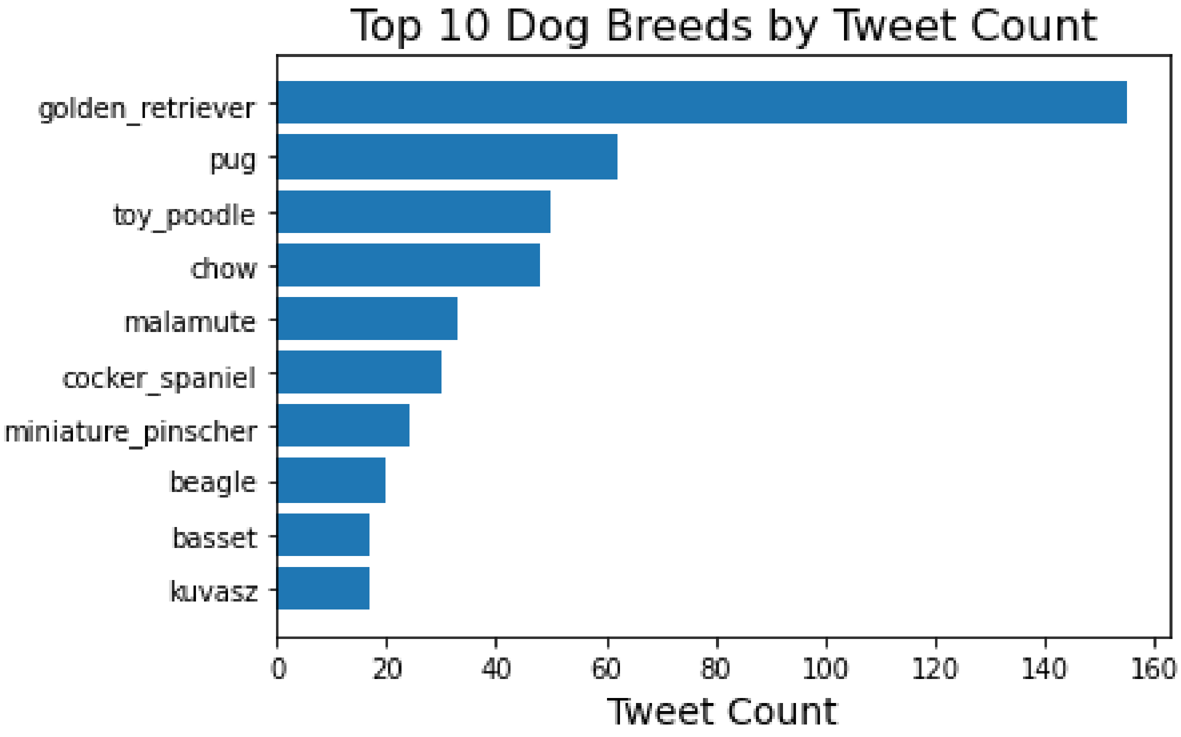
Introduction

The purpose of this project is to practice wrangling data from multiple data sources. The data was assessed, tidied, cleaned and then combined into a final dataset for analysis. The dataset used is the Twitter archive of Twitter account WeRateDogs. WeRateDogs rates images of people's dogs with a humorous comment about the dog. An additional dataset was extracted using the Twitter API in order to obtain favorite counts and retweet counts for the tweets in the Twitter archive. The third dataset being included in the analysis is an image predictions dataset where all of the images from the archive tweets were processed though a neural network in an attempt to classify the dog breed represented in each image. This report includes analysis of the combined dataset

Analysis

Insight 1:

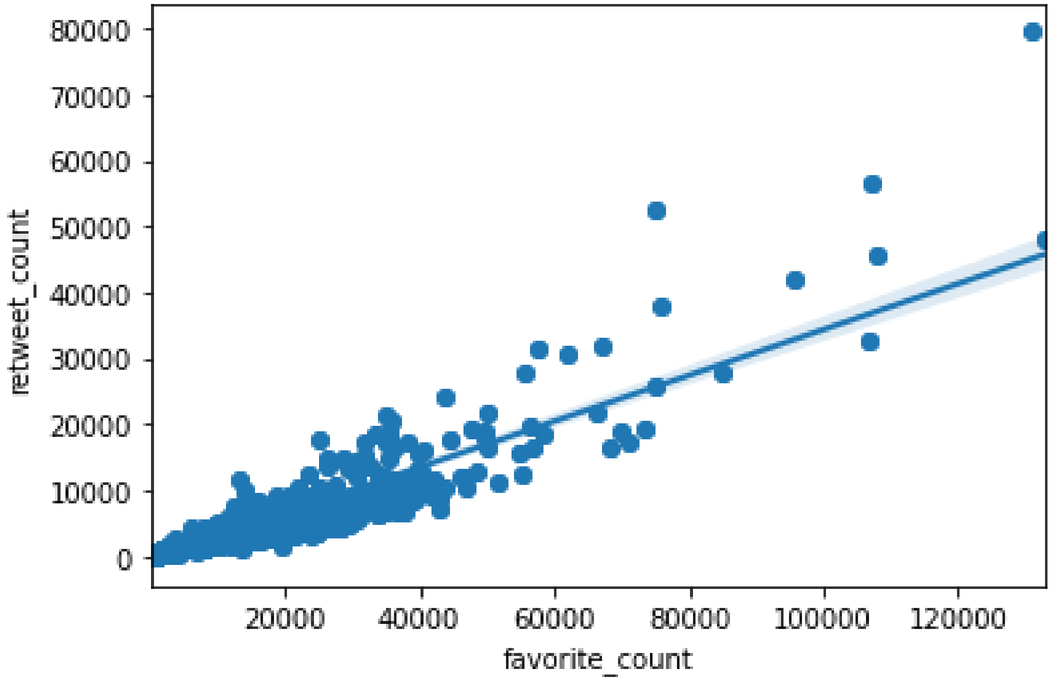
Golden Retriever is the most tweeted breed of dogs.
A dog breed was assigned to each tweet by selecting the most likely correct breed prediction of each tweet between p1, p2 and p3. The data was then grouped by breed and counted to obtain the count of tweets by breed. Golden Retriever was the breed most tweeted.



Insights and Visualization

Insight 2:

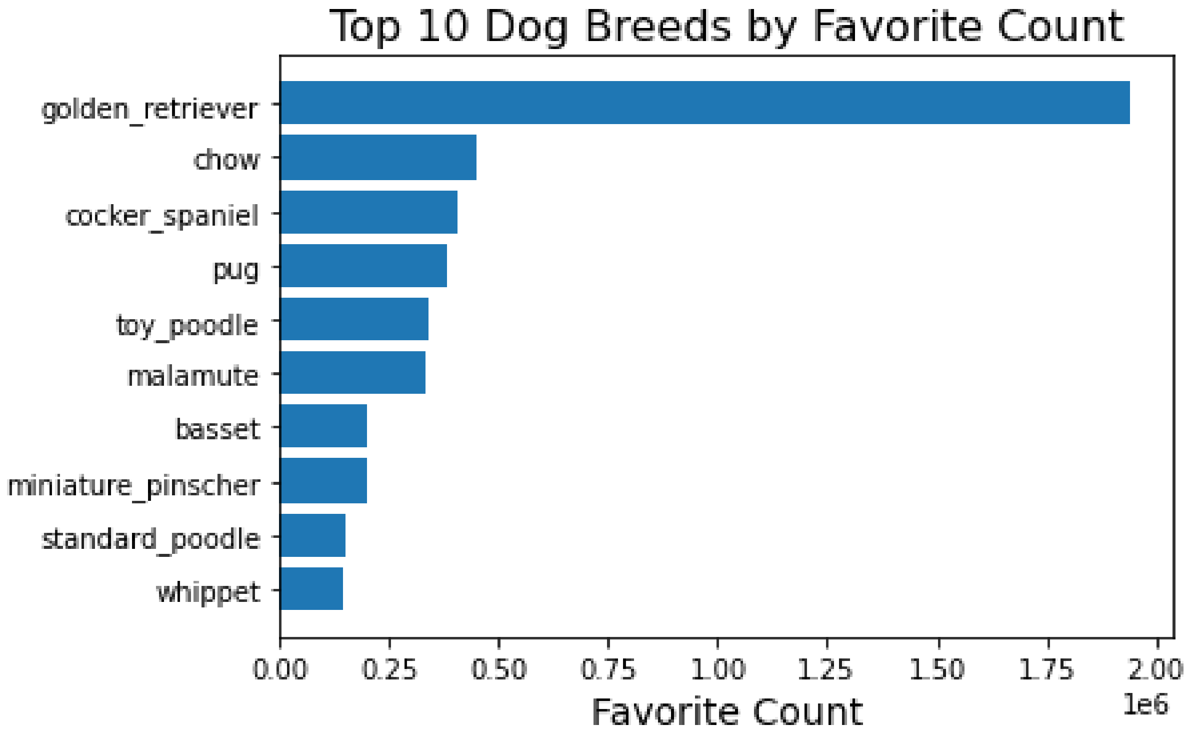
There is strong positive correlation between favorite_count and retweet_count. This means that if there is a high favorite count then there will also be high retweet count. The chart below shows this relationship.



Insights and Visualization

Insight 3:

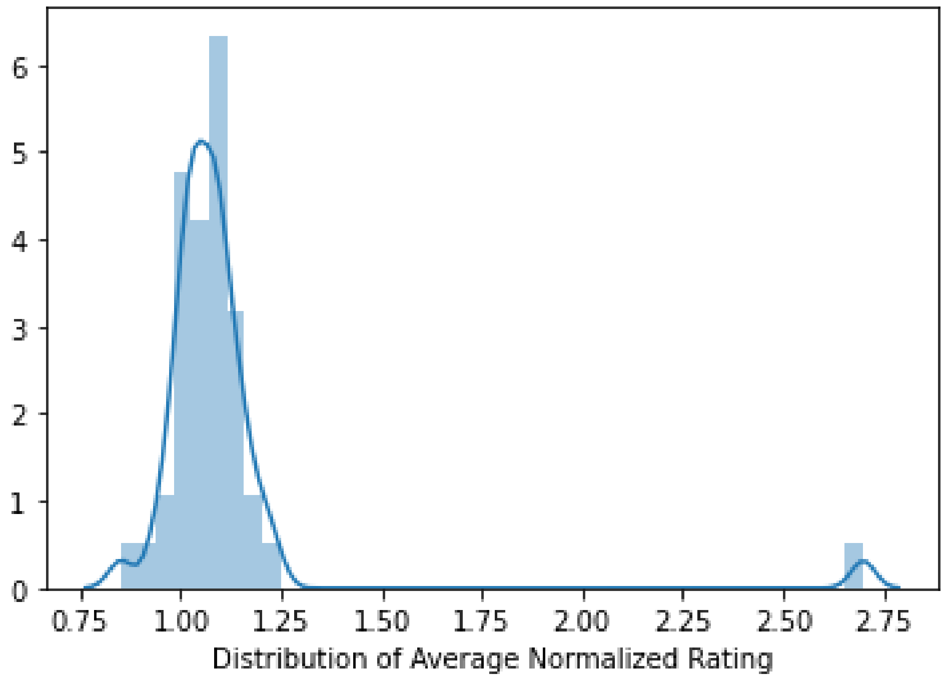
Golden Retriever was the breed with the highest aggregate favorite count.
A dog breed was assigned to each tweet by selecting the most likely correct breed prediction of each tweet between p1, p2 and p3. The data was then grouped by breed and summed over favorite count to obtain the total of favorite count by breed. Golden Retriever had highest aggregate of favorite count.



Insights and Visualization

Insight 4:

Most breeds have a average rating between 1 and 1.2, or between 10/10 and 12/10. The ratings were normalized by dividing numerator by denominator to create a numeric value that could be compared. The plot shows that most ratings fell between 10/10 to 12/10. There was a single outlier of 27/10.


Distribution of Average Normalized Rating