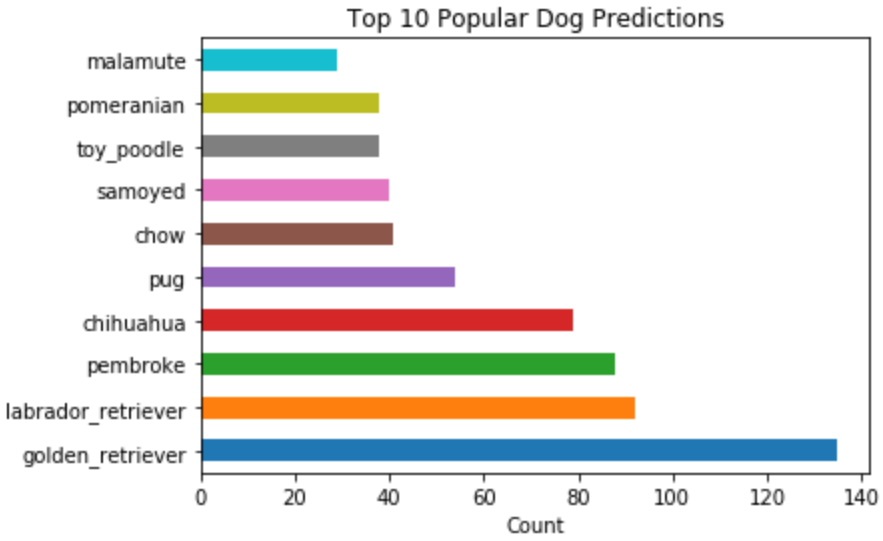
Analyzing, and Visualizing Data

Analyzing the Data

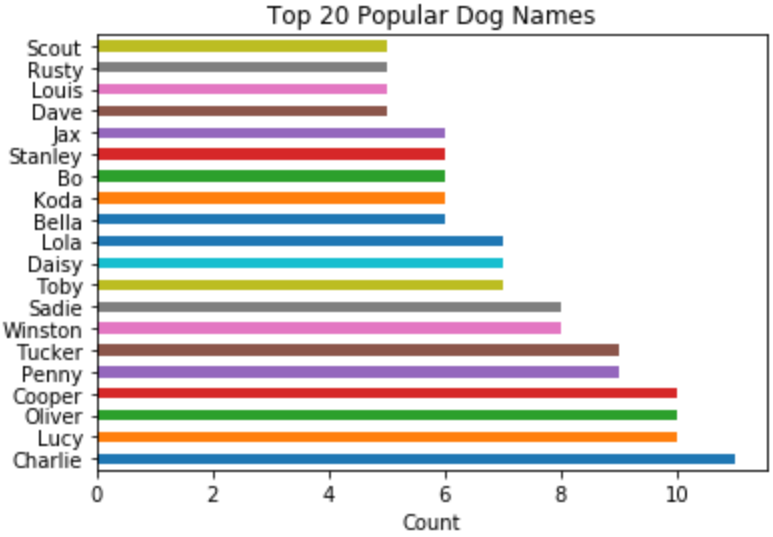
This project involved gathering, assessing and cleaning data from three sources. From there, a clean data set has been created in which further analysis was conducted. The below are some insights gained from the dataset:

1. Categorical Rankings

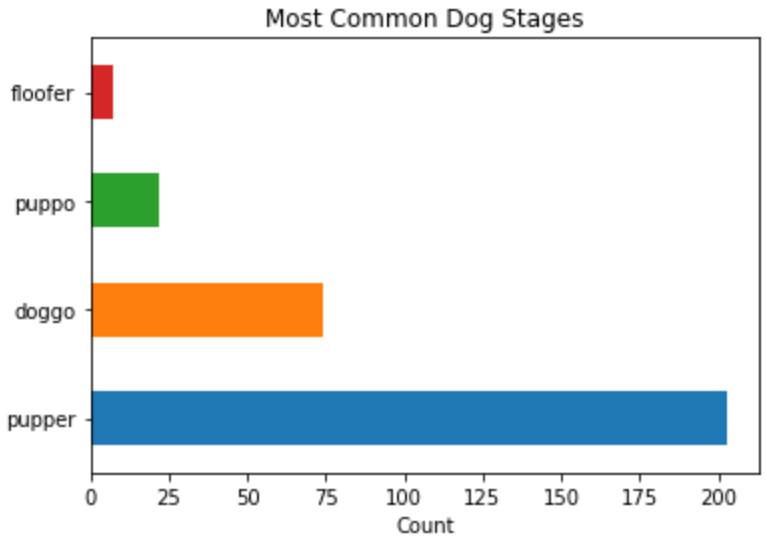
In the chart below, we see that the popular predictions by the machine learning algorithms were Golden Retrievers, Labrador Retrievers, Pembroke (assuming it is referring to Pembroke Welsh Corgis).



In the chart below, we see some of the most popular dog names. The top three were Charlie, Lucy and Cooper.

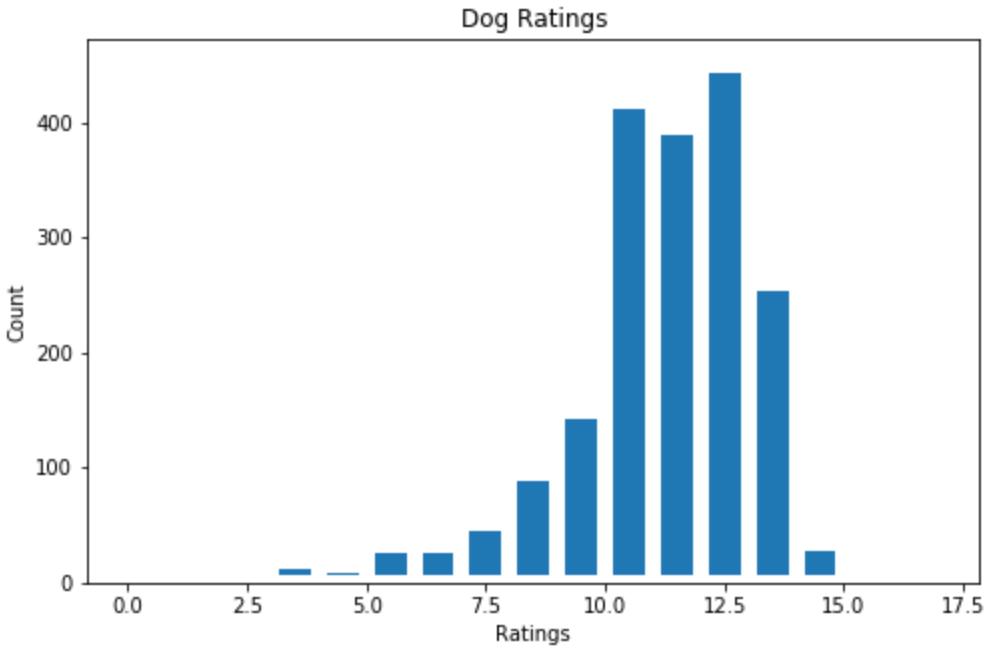


In the chart below, we see the occurence of each Dog Stages. Unsurprisingly, the highest count is with ‘pupper’, as it is the first stage in the dog stage, with the rarest, floofer, being the most rare occurrence.

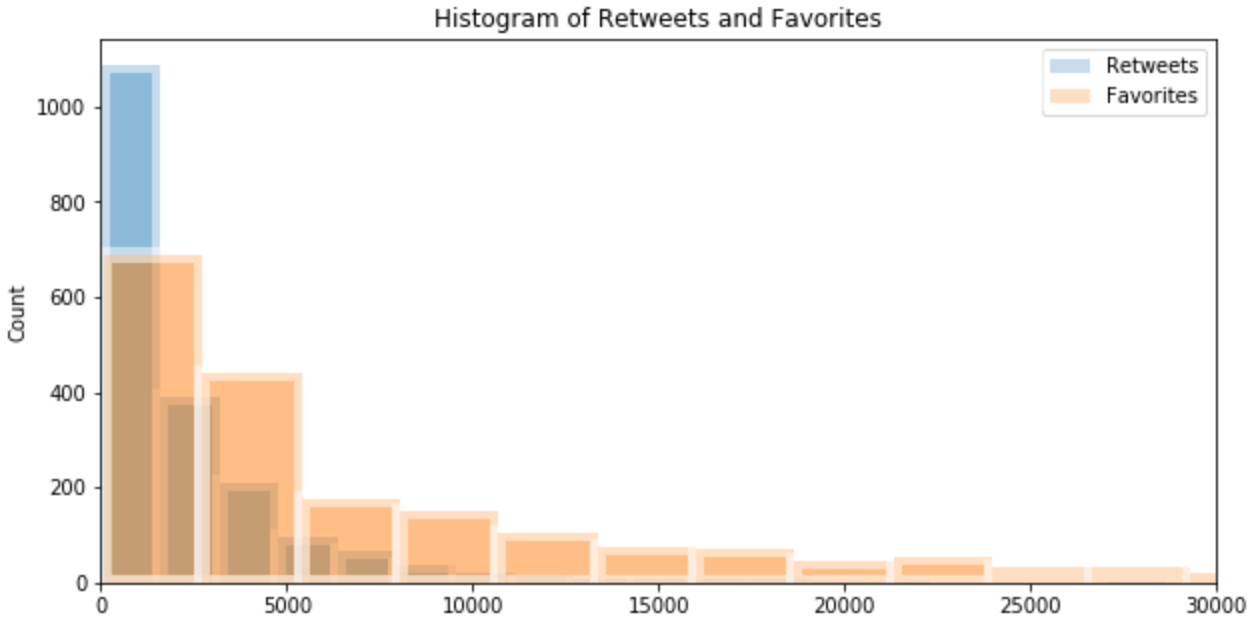


2. Distribution of Data

In the histogram below, we see how the dog ratings are distributed. Overall, the dog rating scores hovered between 10 to 12.5. The distribution has a slightly left skewed bell curve. This is to be expected, as dog ratings above 10 / 10 is perfectly acceptable in this twitter feed.



The histogram below combines both the distribution of retweet counts and favorite counts. The right skewness of the data is to be expected, as higher numbers of retweets and favorites would be rarer to observe.



The histogram below show the distribution of the predictions made by the machine learning algorithm. Overall it appears to be left skewed with a higher concentration of predictions right at the 100% confidence mark.