As this project comes to a close, I will share my insight and visualization processes. The first thing I analyzed the clean data for is the ‘retweets’ mean, maximum, minimum, and standard deviation. I chose retweets because they are generally an indicator of popular a tweet is. After that I used the number given for maximum to query the data for that entry. The post with the maximum retweets is a labrador retriever. From analysis above, I found that labrador retrievers are the mode for the image_predictions set. Further studies are needed to be able to conclude that labrador retrievers are the favorite breed.

The next thing I investigated was the minimum for retweets, which is 11 as found earlier. I queried the data for the minimum number of retweets. The corresponding post is an english setter. I then queried the data for other English setters to see how it compared with the others. The average number of retweets for an English setter is 1,346. That means that this post, with only 11 retweets, is an outlier, and could potentially skew the data.

The last thing I investigated was the distribution of the ratings. I did this by querying the data for each numerator and receiving the length of that list. I found that The most frequently occurring numerator is 12. On either side of it, the frequencies go down like a bell curve. The numerator 10, however, has a high frequency. 13/10 is a common rating, but generally the highest rating they give. Few dogs made 14/10.

For my visualization, I created a scatter plot comparing the number of favorites a dog receives with the rating given to them. As expected, there appears to be a correlation between the number of favorites and the rating given to the dog. Further investigation is needed to conclude a correlation.