Originator: Michael Wei
Rev.: B
Date: 3/31/2019

Executive Summary

This report summarizes our Exploratory Data Analysis of the WeRateDogs twitter account. Data were analyzed using the twitter_archive_master.csv file that we generated in the wrangling effort, as outlined in the wrangle_report and the wrangle_act.ipynb files. The following questions were addressed:

1) Are retweets or favorite counts correlated with higher ratings?
2) What types of dogs do we see the most with a high rating?
3) Are low image prediction confidence values indicative of a certain range of dog ratings? Will we see lower numerator ratings associated with low image prediction confidence values?
4) Are higher numerator ratings associated with the presence of a “Dogtionary” term in the tweet text?

In general, we found most dogs were in fact “good dogs, Brent,” and are correlated with several characteristics:

1) If a tweet has a favorite or retweet count, it is more likely to be associated with a higher numerator rating, and therefore, a good dog. However, a limitation of this observation is that some of the tweets with favorite or retweet counts are also associated with some not-so-good dogs.
2) The Golden Retriever seems to top the charts as the dog type with the most high ratings (Defined by Numerator Ratings greater than 10). The Labrador Retriever also makes the top 10 dogs, creating solid grounding for Retriever type dogs.
3) A low image prediction confidence values is not necessarily related to numerator rating in most cases. However, if the image prediction confidence value is sufficiently high (>0.95), it’s more likely to be associated with a good dog.
4) If a tweet showed a “dogtionary” term, it would be more likely to be associated with a higher numerical rating. This suggests an association between cute dogs and good dogs.

Originator: Michael Wei
Rev.: B
Date: 3/31/2019

Analysis

Retweet and Favorite Counts:

The following figure shows how the tweets are distributed amongst numerator ratings. Based on this plot, a substantial number of dogs were in fact good dogs, with a numerator rating of at least 10.

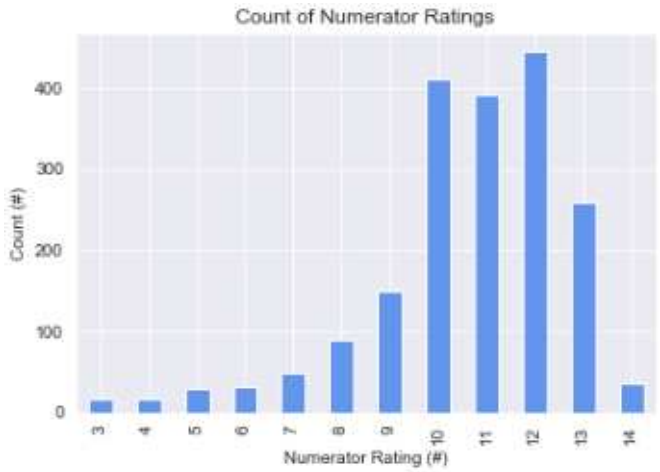

Figure 1: Distribution of dog ratings

Next, Figures 2 and 3 shows scatter plots of favorite count versus numerator rating and retweet count versus numerator rating, respectively.

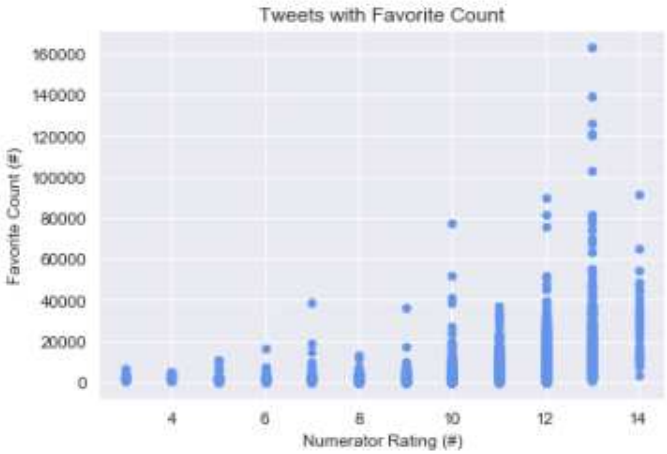

Figure 2: Favorite count

Originator: Michael Wei
Rev.: B
Date: 3/31/2019

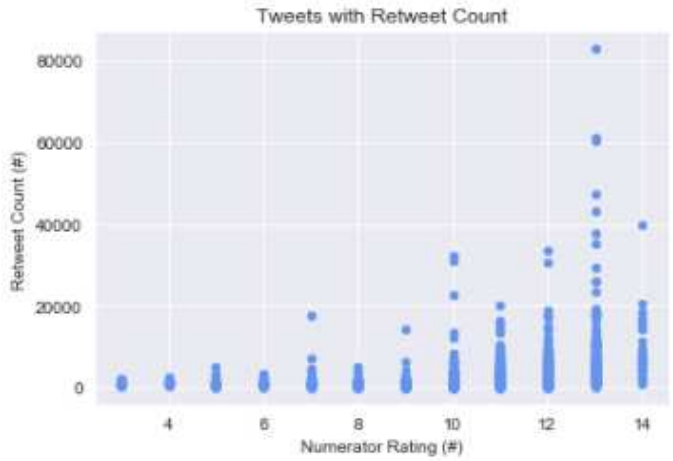

Figure 3: Retweet count

Figures 2 and 3 suggest that a retweet or favorite is more likely to be associated with a higher numerator rating. However, this is not the entire picture, because there are a number of dogs with ratings in a numerator rating range of 3 – 7. Furthermore, this does not imply a direct correlation, because a correlation calculation shows 0.408 and 0.307 for Favorite Count versus Numerator Rating and Retweet Count versus Numerator Rating, respectively.

Dog Types

Next, we analyzed dog types. For this analysis, we selected the type of dog for each tweet based on which dog type was associated with the maximum image prediction confidence value.

Figure 4 shows the top 10 dogs that received a Numerator Rating greater than 11. Figures 5 also shows similar top dog types based on which ones received a retweet count. Because the data were nearly the same between Retweet and Favorite Counts, we only show Retweet Count versus Numerator rating below. One observation of note is that two types of retrievers made it within the top 5 dogs, which suggests that they could be associated with good dogs. Furthermore, the Chihuahua seems to have made the top 10 types of dogs, which conflicts with conventional wisdom that the Chihuahua is an aggressive and mean dog.

Originator: Michael Wei
Rev.: B
Date: 3/31/2019

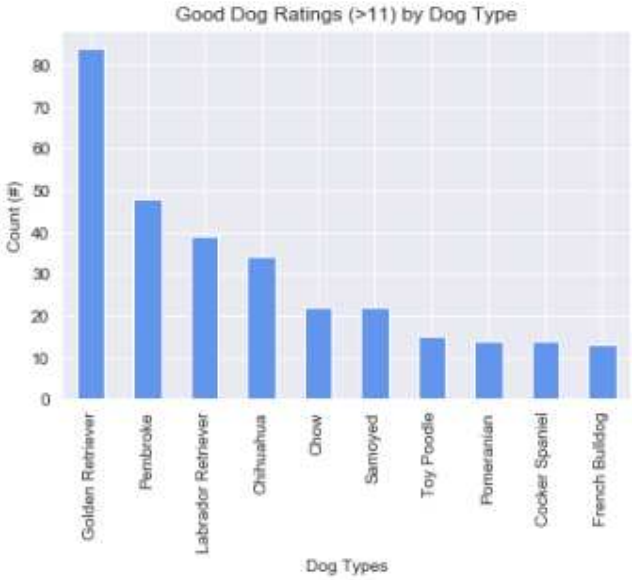

Figure 4: Top dog types (> 11)

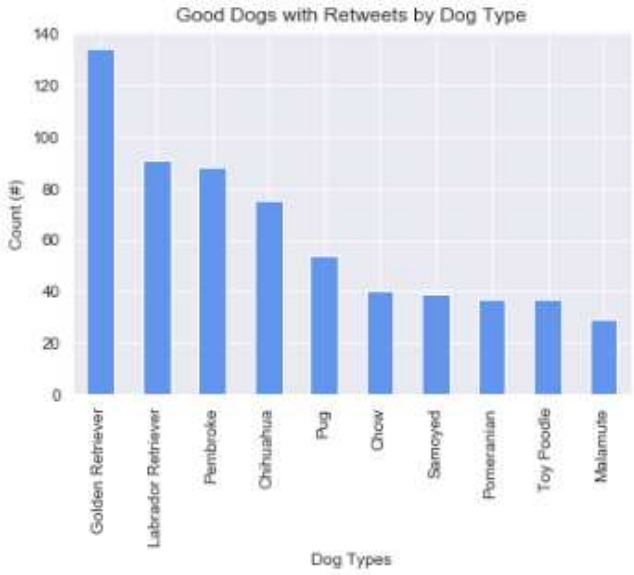

Figure 5: Top dog types (By Retweet count)

Image Prediction Confidence Value

Next, we looked at image prediction confidence values, to determine a possible relationship with numerator ratings. It is possible that if the max image prediction confidence value for a tweet is sufficiently low, then it might be more likely to be a non-dog tweet, or a bad dog. Figure 6 shows how image prediction confidence values are spread out amongst numerator ratings. Figure 7 shows a subset for low image prediction values (< 0.15). Figure 8 shows a subset for high image prediction values (> 0.95).

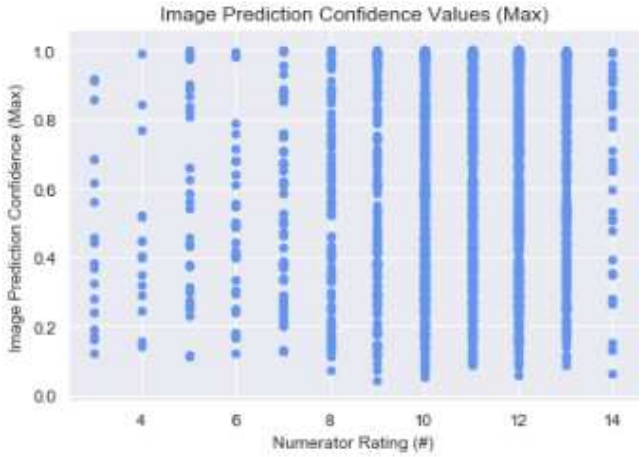

Figure 6: Image prediction confidence value spread

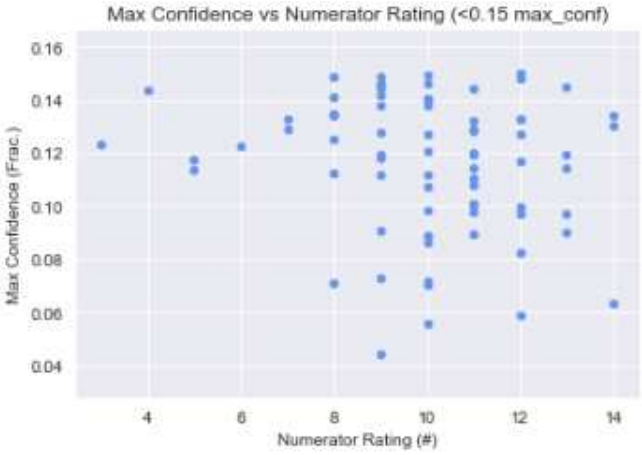

Figure 7: Low image prediction confidence values

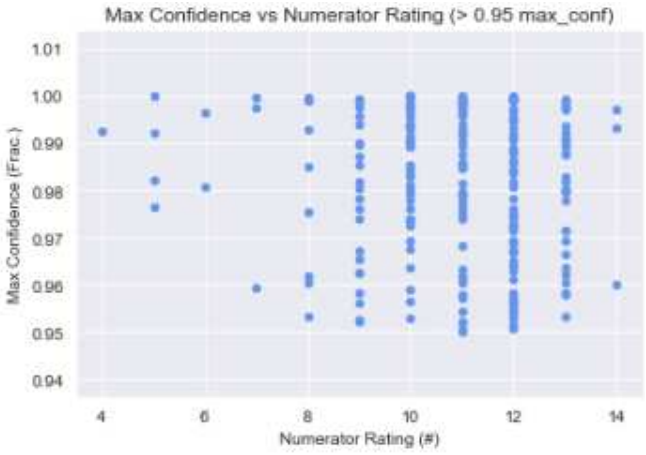

Figure 8: High image prediction confidence values

Figure 6 suggests that the image prediction confidence value is not necessarily indicative of a high numerator rating, because the data are spread out along the entire range. Tweets with a low image prediction confidence value also span a large range. However, figure 8 shows that if the image prediction value is sufficiently high (> 0.95), it's more likely to be associated with a good dog, with most numerator ratings within a range of 9 and 13.

Dogtionary Terms

Lastly, we looked at the impact of having a "dogtionary" term in the tweet text to determine whether that could be associated with a higher numerator rating. Figure 9 shows the count of "dogtionary" terms detected for each numerator rating.

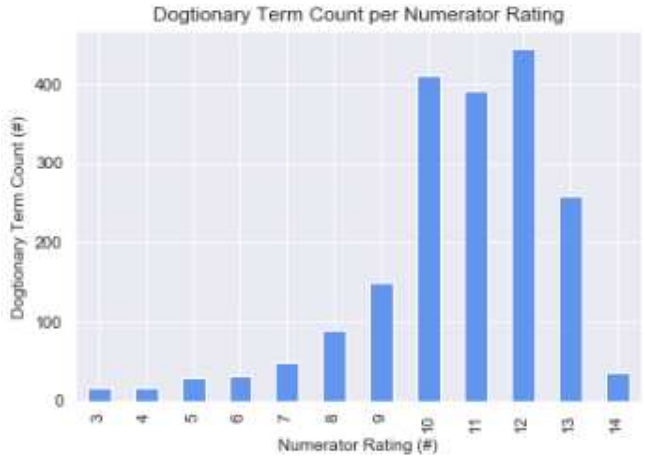

Figure 9: Dogtionary term impact

Although the data subset for tweets with "dogtionary" terms is low, we can see that most of these tweets have a numerator rating above a 10. This suggests that a tweet with a "dogtionary" term is more likely to be associated with a higher numerator rating. A cute dog is more likely to be a good dog.

Conclusion

In this report, we focused on basic data analysis to determine key characteristics of good dogs and any potential relationships. The parameters we compared against numerator ratings include favorite count, retweet count, image prediction confidence values, and "dogtionary" terms.