Insights and Visualizations for Twitter Archive of WeRateDogs

Reem Alansary

September 10, 2020

After wrangling a dataset based on the Twitter account WeRateDogs, we move on to the data analysis phase with two sufficiently clean tables named *ratings_clean* and *image_predictions_clean*. In the data analysis phase, all attributes of both tables were considered and questions were derived for each dataset within the tables to generate meaningful insights into the available data as well as create visualizations to communicate the findings. The rest of this article will showcase each question asked and its answer separately along with a brief explanation of the relations within the data.

ratings_clean contains general information pertaining to a tweet. *image_predictions_clean* contains information pertaining to the output of machine learning algorithm designed to predict the object or being in an image. This table holds various predictions("attempts") for a single image with different degrees of confidence in their credibility.

Columns of *ratings_clean*:

- tweet_id
- timestamp
- source
- text
- expanded_urls
- rating_numerator
- rating_denominator
- name
- dog_stage
- retweet_count
- favorite_count

Columns of *image_predictions_clean*:

- tweet_id
- jpg_url
- img_num
- prediction_level
- prediction
- confidence
- dog_breed

1 ratings_clean

Helper-1 is a boxplot to help us understand how the presence of outliers may affect the main comparison between ratings and retweet counts. There are some outliers but they are not significant enough to cause a problem with the main graph.

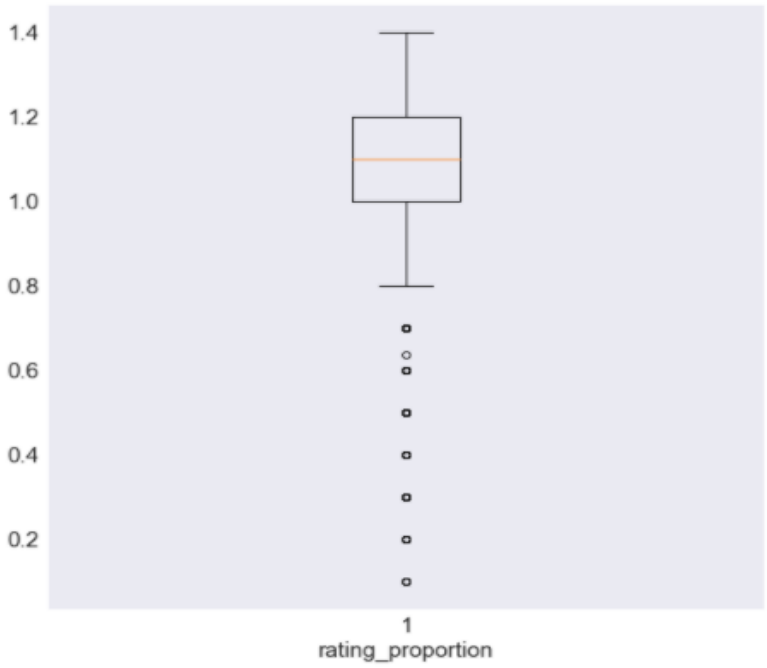

Figure 1: Helper-1

Q: Are dog tweets with higher ratings retweeted more often than others?
Figure 2 is a main comparison graph that is based on rating proportion (rating_numerator \s rating_denominator); we can see immediately that the graph is skewed to the left which means that as the rating increases a tweet is more likely to be retweeted multiple times.

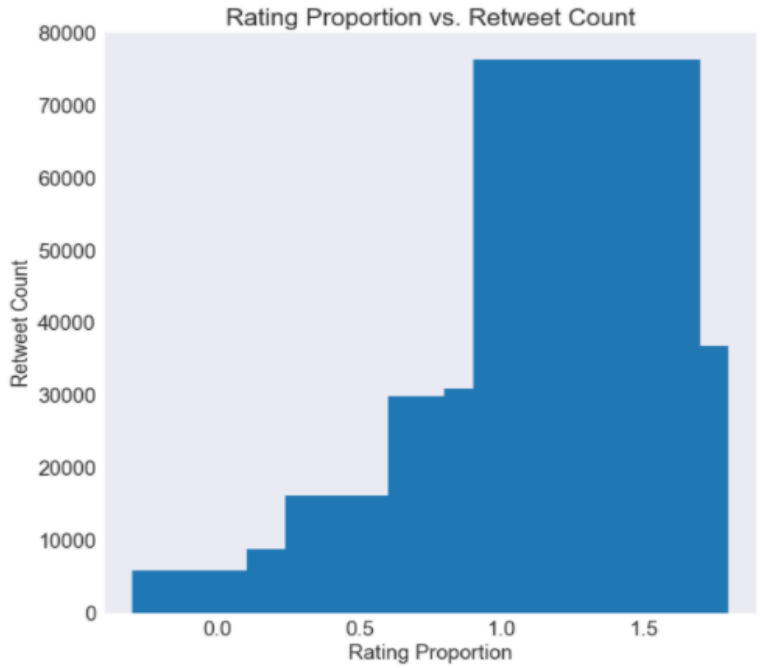

Figure 2: Rating Proportion vs. Retweet Count

Q: Are dog tweets with high ratings favorited by more people than dog tweets with low ratings?
We can see the skewness to the left of Figure 3 as well, which gives the impression that tweets with high ratings are not only retweeted more often, but they also more likely to to be favorited by a larger number of people.

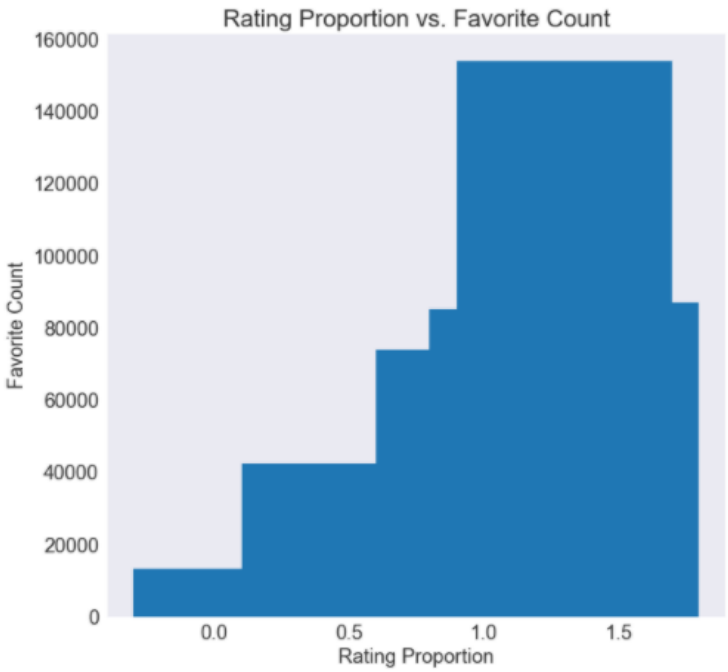

Figure 3: Rating Proportion vs. Favorite Count

Q: Does the number of retweets affect the number of people who prefer a certain dog tweet?

Helper-2 is comprised of two boxplots for analyzing the effect of outliers on further data visualization attempts. In this case it would seem that there are too many outliers and that most probably they would damage the credibility of any visualizations we may create on retweet_count and favorite_count. However, for our next question, these outliers would constitute no risk on the outcome and we will see shortly why.

Due to the similarity of the outlier patterns in Helper-2, the positive correlation seen through the scatter plot at the top of Figure 5 between retweet_count and favorite_count is assumed to be a strong positive correlation. The scatter plot at the bottom of Figure 5, where we cut off outlliers and zoom in on the dense region of the plot, corroborates this assumption and we are now sure of the strong positive correlation between the two variables without having to remove all outliers in this case.

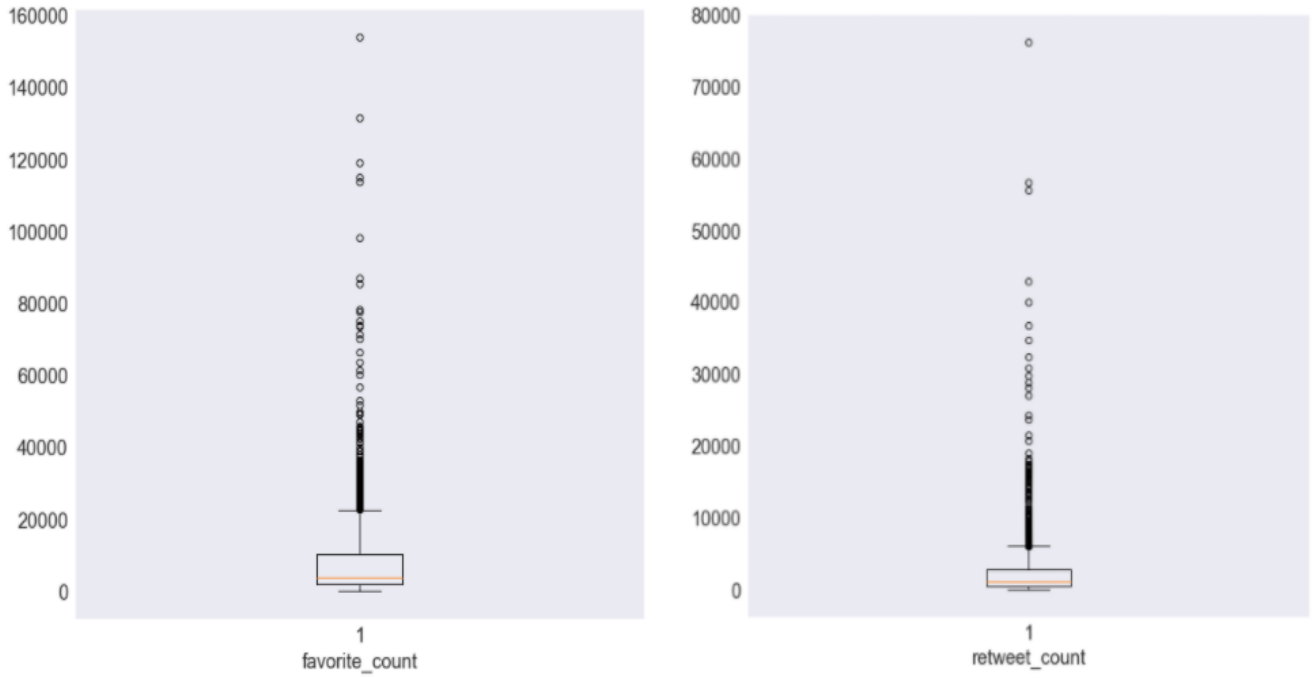

Figure 4: Helper-2

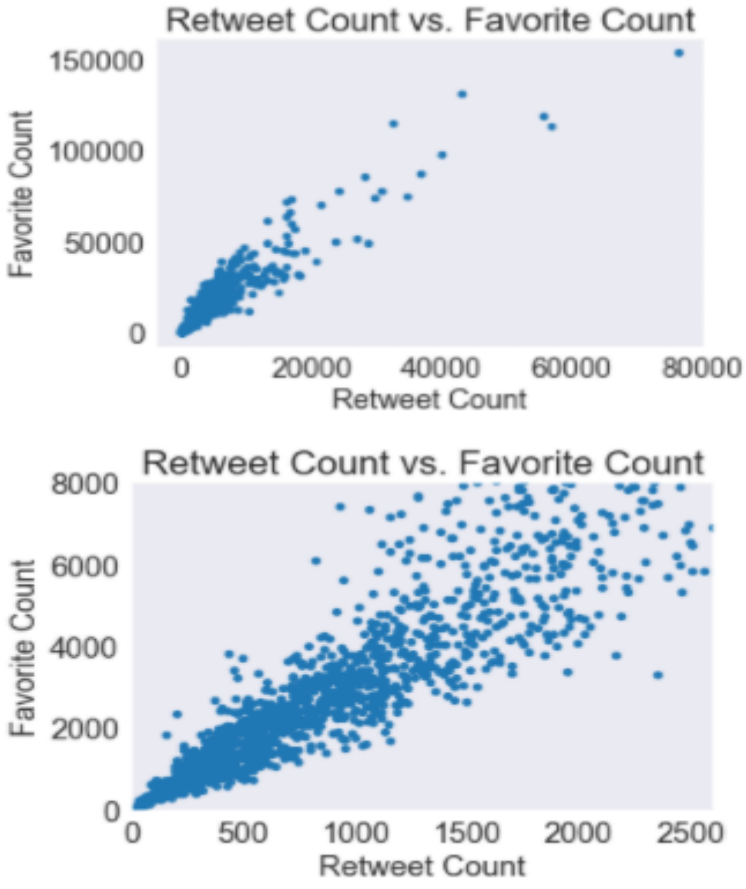

Figure 5: Retweet Count vs. Favourite Count

Q: What is the average rating for each of the 3 most common dog names?

This question may seem irrelevant at first glance, however its answer is unanticipated as shown by Figure 6. The name frequency is actually related to average rating; more precisely, the more popular a name the higher

the average rating. We could speculate that a reason for this is the tendency of humans to unconsciously choose dog names that they have heard often or the that they think may appeal to a large audience, but this is merely a speculation and proving it is beyond the scope of this article.

	name	rating_numerator	rating_denominator	rating_proportion
0	Charlie	11.6	10.0	1.16
1	Cooper	11.3	10.0	1.13
2	Oliver	11.3	10.0	1.13

Figure 6: Top 3 Names and Their Average Ratings

2 image_predictions_clean

Q: How many predictions are actually breeds of dogs?
No one can deny that massive breakthroughs in fields like image recognition have been made but the machine learning algorithms used sometimes are still error-prone. By looking at the pie chart in Figure 7, we could see that a little more than a quarter of predictions were not even dogs. Frankly, this revelation is not only due to the fact some errors may be made by the image recognition algorithm. In fact, some tweets on WeRateDogs intentionally contain dogs that are partially covered under a piece of cloth or dressed in regular clothing, which may have added to the confusion.

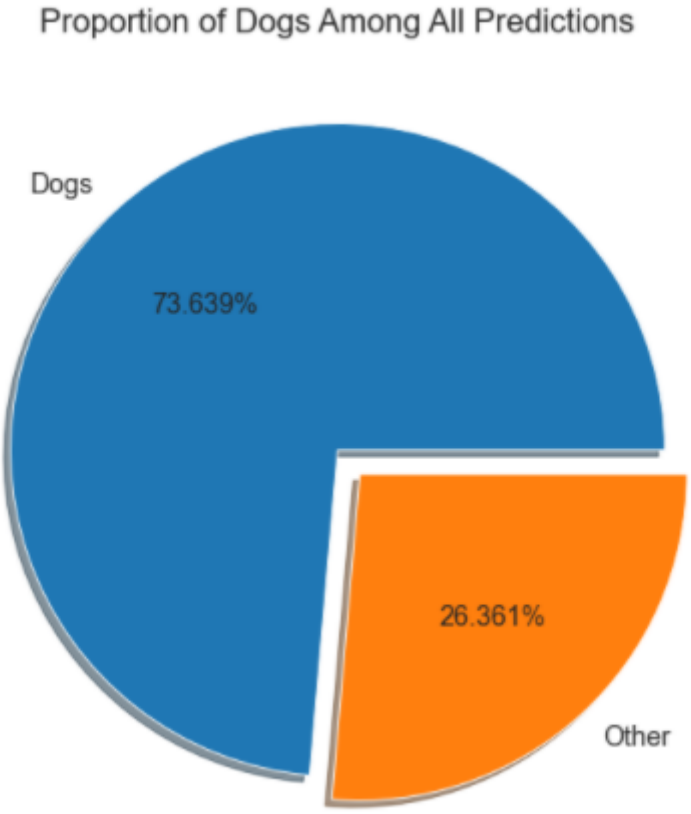

Figure 7: Proportion of Dog Predictions to Other Predictions

Q: Which prediction level has the highest average confidence value?
As we can see in Figure 8 category 1 prediction level is the most confident prediction on average and average confidence decreases as the prediction level category number increases.

	prediction_level	confidence
0	1	0.594548
1	2	0.134589
2	3	0.060324

Figure 8: Prediction Level vs. Confidence

Q: What is the most common prediction for each prediction level?
Although this may be debatable according to the validity of each prediction as we argued above, however Figure 9 shows that most people who send their dogs' photo to WeRateDogs are Labrador owners.

	prediction_level	prediction
0	1	golden_retriever
1	2	Labrador_retriever
2	3	Labrador_retriever

Figure 9: Modal Predictions for Each Prediction Level