Analysis Report

In the wrangle report, I talked about how I gather, aggregate, assess, and clean up Twitter data from the account WeRateDogs. WeRateDogs is a Twitter account that posts pictures of other people's dogs with a comment and gives them a rating out of 10 per dog in the picture. However, the uniqueness comes in the numerator of the rating that can exceed 10 depending on the content. Therefore it will be interesting to investigate the uniqueness of the posts.

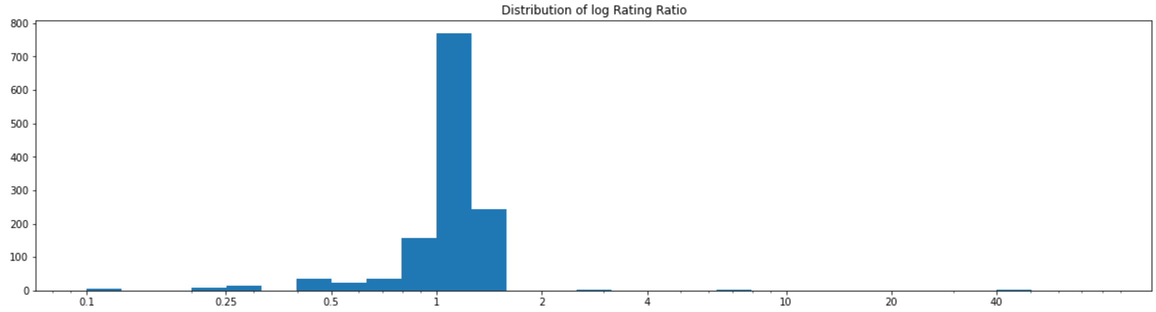
Having the data sent by WeRateDogs to Udacity, we have access to each tweet's ID, text, timestamp, and ratings up to 2017- 8 -1. However, it will be useful to take in account of the favorite counts and retweet counts to analyze the relationships between ratings and the posts. Also, a file is provided by Udacity, image_predictions.tsv, to offer interesting insights on how neural networks perceive the picture in the tweet.

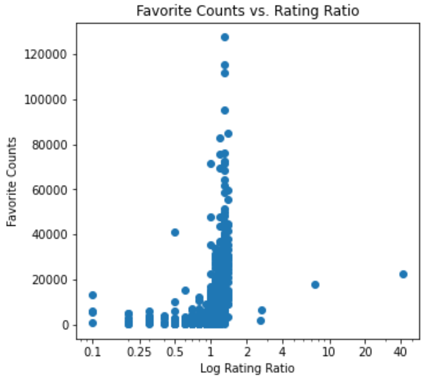
Research Questions:

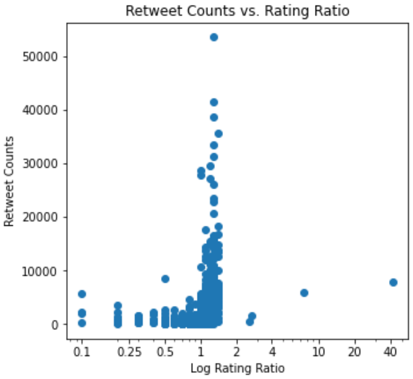
1. Relationship between rating ratio vs. favorite count, img_num, retweet_count
2. How does the number of retweets and likes relate to the number of images?
3. What kind of dogs are predicted the most?
4. Hypothesis Test: Does Rating differ if dog is predicted as a dog?

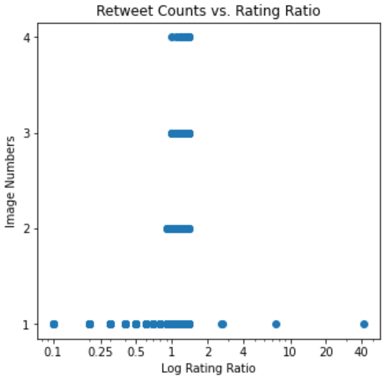
Analysis

1. Relationship between rating ratio vs. favorite count, img_num, retweet_count





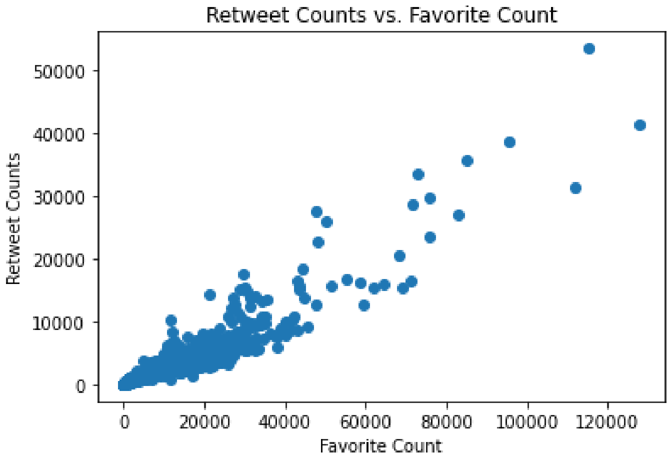




By looking at the rating ratio, it is found that ratios mostly cumulate between 1 to 2 according to the histogram above. However, there are outliers such as 40 and 0.1 as ratios.

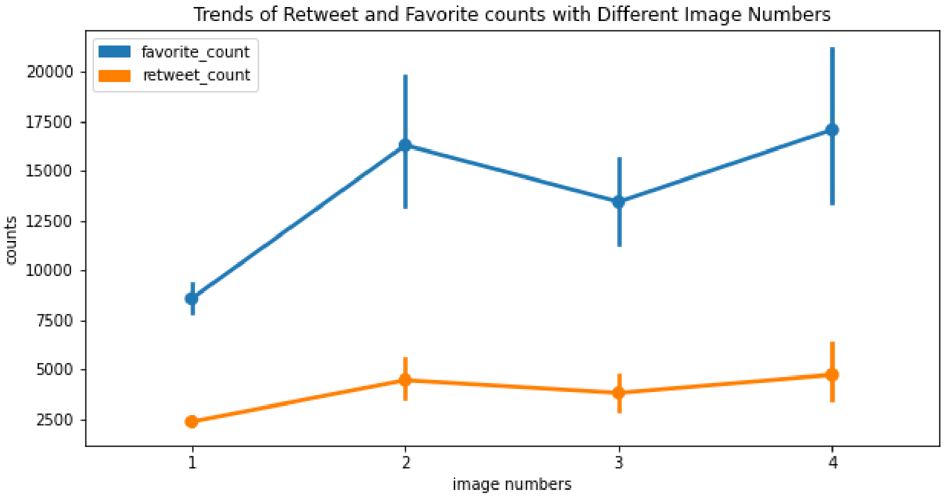
The relationship with image number shows a trend that if people post 2-4 images, they are likely to get rating ratios between 1-2.

From the relationship scatter plots below, it is found that the most number of favorites and retweets are also in the log rating ratio of 1~2. The higher the rating ratio doesn't appear to have a higher number of retweets and likes than some of the tweets with rating ratios 1-2. On the other hand, lower rating ratios doesn't necessarily mean lower retweets and likes. As the distribution for number of retweets and likes look similar we found that they have a positive correlation with each other, the higher the number of retweets most likely the higher the number of likes.



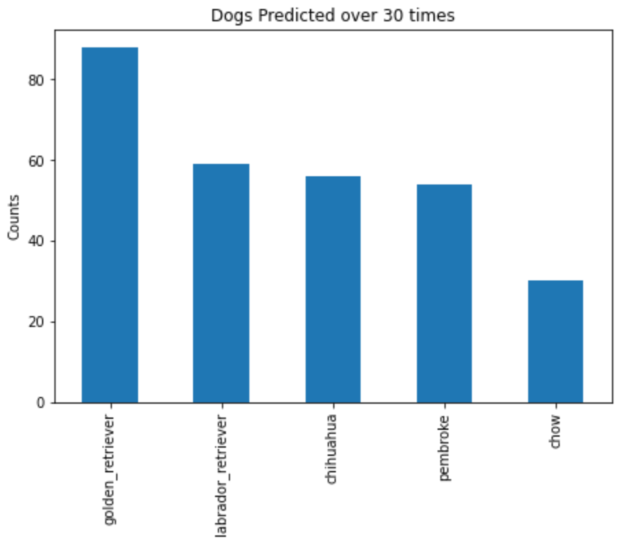
2. How does the number of retweets and likes relate to the number of images?

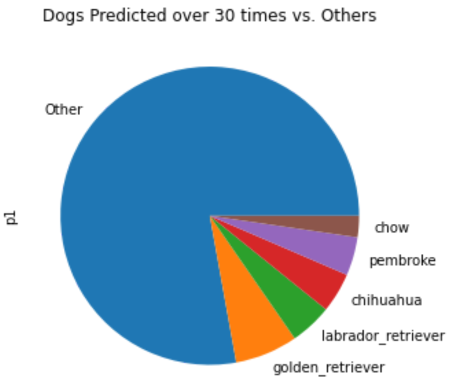
Besides observing the relationships between rating ratios and different variables, it is also insightful to see how these variables interact among themselves.



The trends of likes and retweets across image numbers look similar, although there is a steeper drop in favorite counts when 3 images are posted. With 2 or 4 images, the number of retweets and likes peaked the most, with around counts of 4000 and 16000, whereas 1 and 3 images received less likes and retweets. Posting 1 image may seem to be less favorable to the public in this plot, since its standard deviation and landing point is lower than others.

3. What breed of dogs are predicted the most?

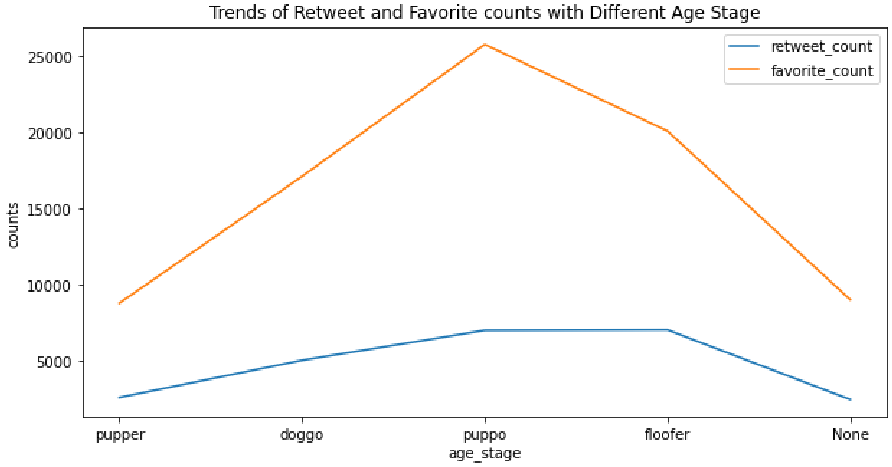




It is found that golden retriever, labrador retriever, chihuahua, pembroke, and chow are popular breeds of dogs to be the most confidently predicted ones. And, compared to others these top 5 occupy almost ¼ of the predictions. It is unknown that if the predictions are correct, however, it may provide an estimate that these 5 breeds of dogs are most likely popular breeds to get.

4. Do people like puppies more or older dogs more?

From the graph, it seems like people like teenage dogs (puppo) most, which received the most retweets and likes among other stages. The ones not labeled or puppers received less retweets and likes. However, the amount of data given is not enough to explain the significance.



5. Hypothesis Test: Does Rating differ if dog is predicted as a dog?

Maybe ratings are based on pictures instead of tweet content? To answer this question, we can use the boolean feature, whether the picture is a dog or not, in image predictions dataframe.

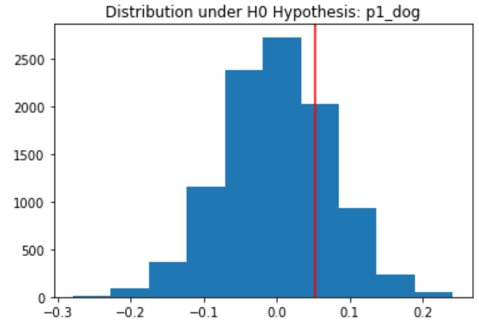
Z-test is performed to examine if p1_dog to p3_dog is useful in helping to determine rating ratio of tweets.

μ_f = rating_ratio for falsely predicted dog

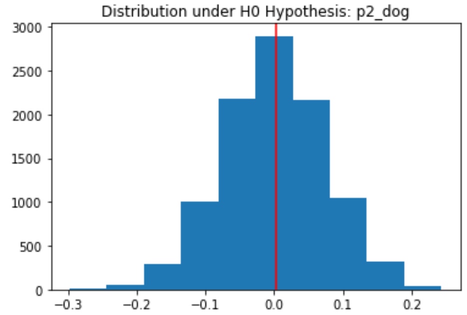
μ_t = rating_ratio for truely predicted dog

$$H_0 : \mu_t - \mu_f = 0$$
$$H_1 : \mu_t - \mu_f \neq 0$$

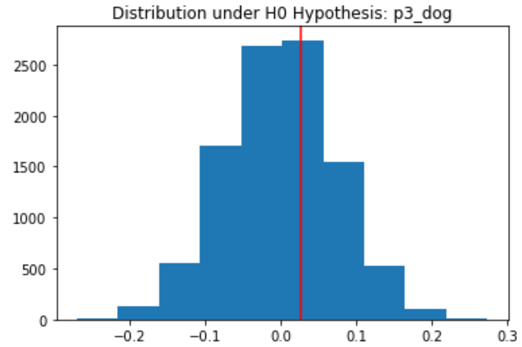
The null hypothesis is not rejected in all 3 p_dogs. It is proven that there is no difference in ratings ratios for pictures predicted as dogs and pictures not predicted as dogs.



```
CI: -0.1445485316613067 ~ 0.14211593812773707
Acutal diff: 0.053316936944582505
Do not Reject Null hypothesis, rating_ratio is not different
```



```
CI: -0.14742515107069631 ~ 0.14558674359594043
Acutal diff: 0.0034996684557595525
Do not Reject Null hypothesis, rating_ratio is not different
```



```
CI: -0.14192711908046923 ~ 0.14255082831947397
Acutal diff: 0.027177105184301276
Do not Reject Null hypothesis, rating_ratio is not different
```

Conclusion:

- We have found that people like to rate with numerators over denominators more, since most rating ratios lie in the range around 1-2.
- For favorite and retweet counts, more counts doesn’t mean higher ratings, again proving the unique rating system of this account. But themselves are positively correlated, the higher number of likes the more retweets
- Posting 1 image may not receive many retweets and likes compared to posting 2 and 4 images.
- Golden retriever, labrador retriever, chihuahua, pembroke, and chow are popular breeds of dogs to keep by owners.

- People like puppo the most according to favorite counts and retweet counts, and the breed with the same retweet counts as puppo is floofer, but it's favorite count does not exceed puppo.
- Maybe ratings are based on pictures instead of tweet content? The answer is no, from the hypothesis testing of predicted pictures, but more information should be given.

Future directions:

- More data can be given in the image predictions set, for example its true label will be a valuable information to answer my last research question. Also, if dog stages can be less sparse, meaning that more posts can be categorized in these 4 categories, puppo, doggo, pupper, floofer, it will help for the analysis of likes and retweets, since there is a drastic difference.