Analysis and Visualization

In the analysis and visualization section of this project, we answer five questions:

- What is the average length of the tweet text and what is the distribution?
- What is the dog that gets the highest rating (by @WeRateDogs), and the dog that gets the highest favorate count (by tweeters)?
- What is the time most tweeter tweet?
- Is there strong correlation between any two variables?
- How accurate is the image prediction through the neural network?

What is the average length of the tweet text and what is the distribution?

As shown in Figure 1, we have found the average length of text is around 120 characters. And it's left skewed.

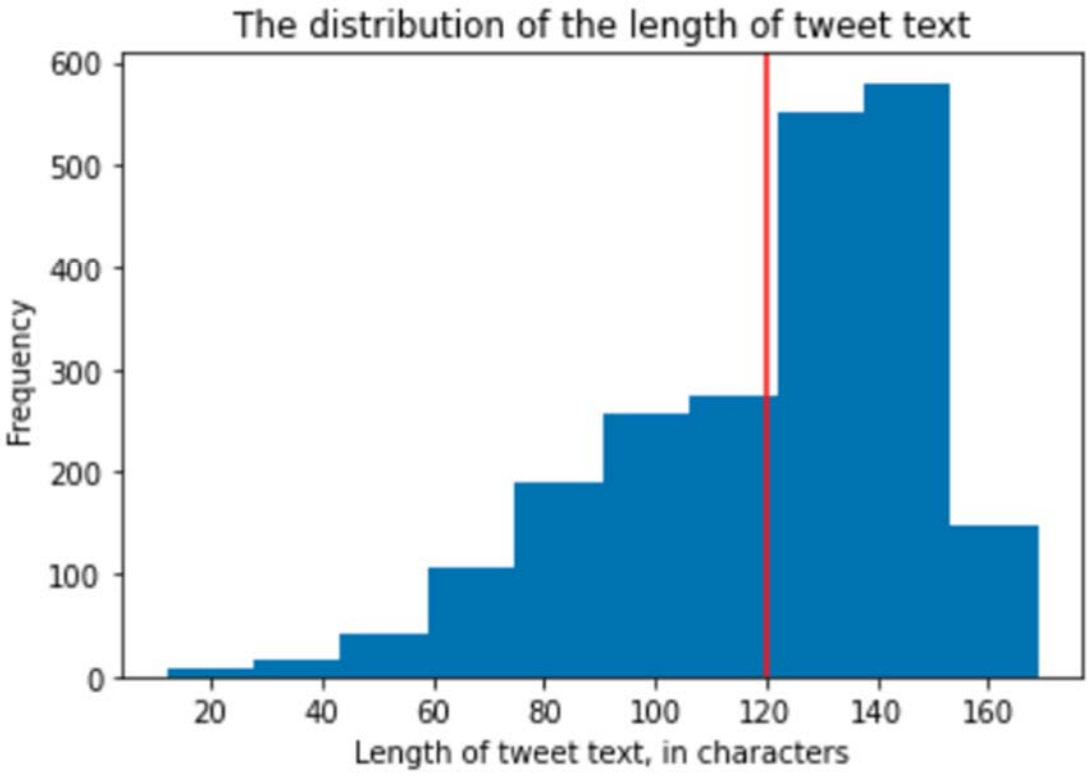

Figure 1: The distribution of the length of tweet text

What is the dog that gets the highest rating (by @WeRateDogs), and the dog that gets the highest favorate count (by tweeters)?

The dog that received highest rating, at 1776/10, was named Atticus, by @WeRateDogs. Unfortunately, the neural network was unable to idenfity this dog. Figure 1 is her picture that we requested from Tweeter.

Figure 2: The dog that receives the highest rating by @WeRateDogs

The dog that receives highest favorate count was a doggo, of which her name was not provided. Although this dog received only 13 from @WeRateDogs, she receives 148,987 favorate counts from the tweeters. The neural network was able to predict this dog as a Labrador Retriever with 0.8253 confidence level. Since a video was uploaded by the tweeter, therefore, a link instead of a picture is provided:

https://video.twimg.com/ext_tw_video/744234667679821824/pu/vid/360x640/aLoem87jSUDyshiY.mp4

What is the time most tweeter tweet?

Surprisingly, as shown in Figure 3, and based on the data, the time period most tweeters tweet is between mid-night and 6:00am. Although this data does not represent the whole population, it does show some interesting and surprising results.

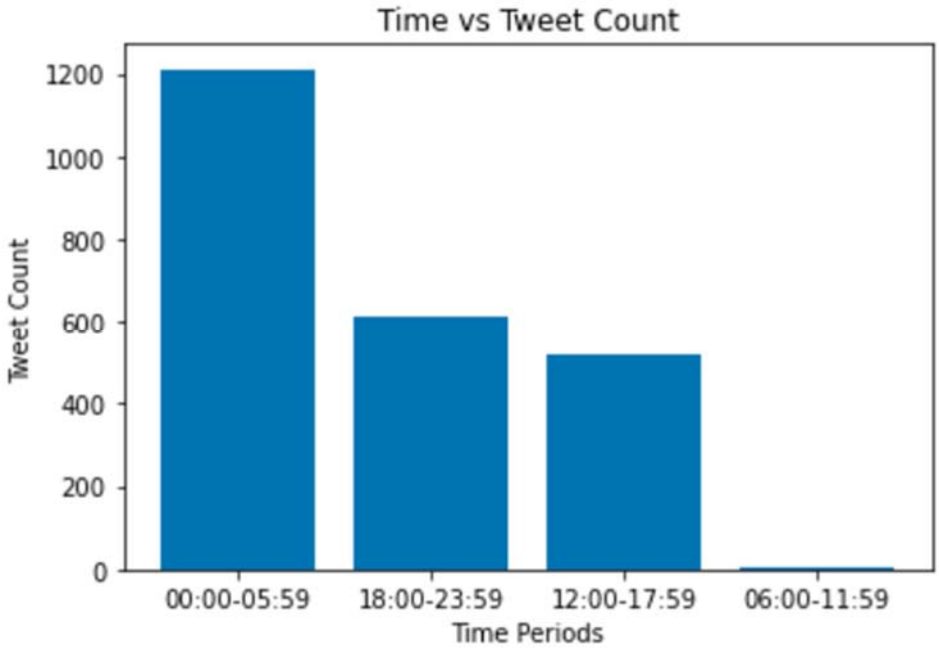

Figure 3: The tweeter count vs. time

Is there strong correlation between any two variables?

Based on the correlation matrix shown in Figure 5, since p1_dog, p2_dog, and p3_dog are not variables, they are predictions, they are not taken into account. However, The heatmap shows that favorite_count and retweet_count are strongly correlated at the coefficient of 0.93.

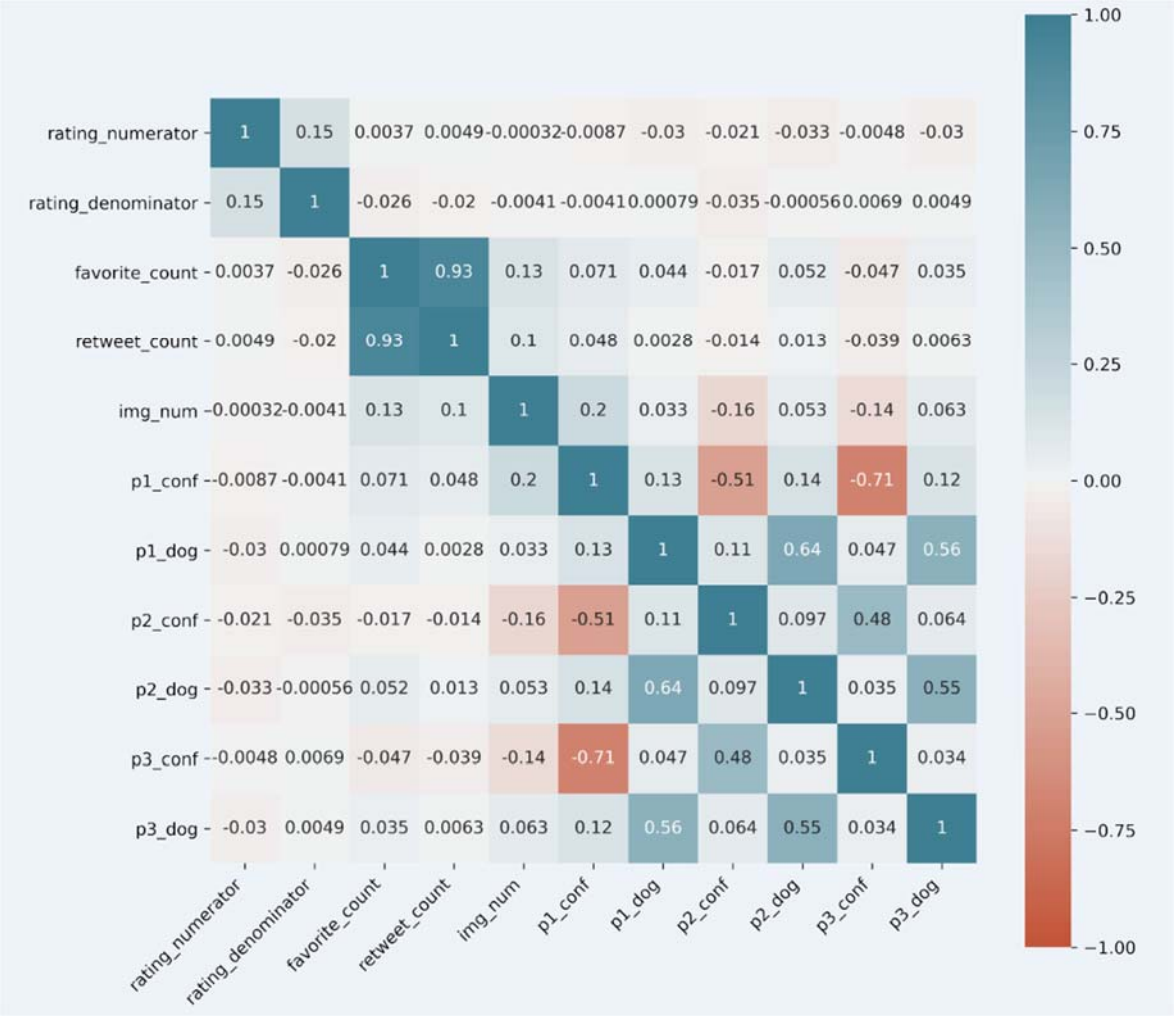

Figure 4: Correlation Matrix for Numerical columns.

How accurate is the image prediction through the neural network?

To answer this question, we averged out the three confidence levels, which are weighted based on our estimated assumptions by the weights (1.625, 1.25, 1.15), which is in ratio 8:2:1, according to the level of confident of the algorithms. And the estimate accuracy of the prediction for the neural network is 0.306, or 30.6%.