Data Wrangling Project – Udacity

Data insights

After completing the wrangling process the data was ready to be explored and to find the patterns within the dataset.
Initially, I did some descriptive analyses where I calculated the mean from some of the columns. I wanted to know what is the average ***retweet_count*** which shows how much the post or the photo in this case was retweeted. This can be a bit confusing knowing that it does not always show a positive result. In this case was interesting to know the mean of ***favorite_count*** which in most cases show positive valuation. The other column from which we calculated the mean was ***rating_numerator***, which shows the rating given to a certain dog –photo.

One thing that we noticed here was the average on the ***favorite_count*** (8080.96) was a lot higher compared to the mean on the ***retweet_count*** (3164.79). On the other hand, rating_numerator did not have an equal spread as it had a mean of 13.21 and in the graph that we generated later, we saw that the count can go higher than 1600.

Visualization

We used scatter plots as they are a good way to see how different columns influence each other. First, it was interesting to see the correlation between ***favorite_count*** and ***retweet_count***. The graph shows that they correlate very well with each other. On the other hand, we used a color map with the ***rating_numerator*** to see if this column can somehow affect favorites and retweets. We see that a high rating equals high favorites and retweets but not necessarily as we also see that there are a lot of data with the rating under 200 that have very high ratings.