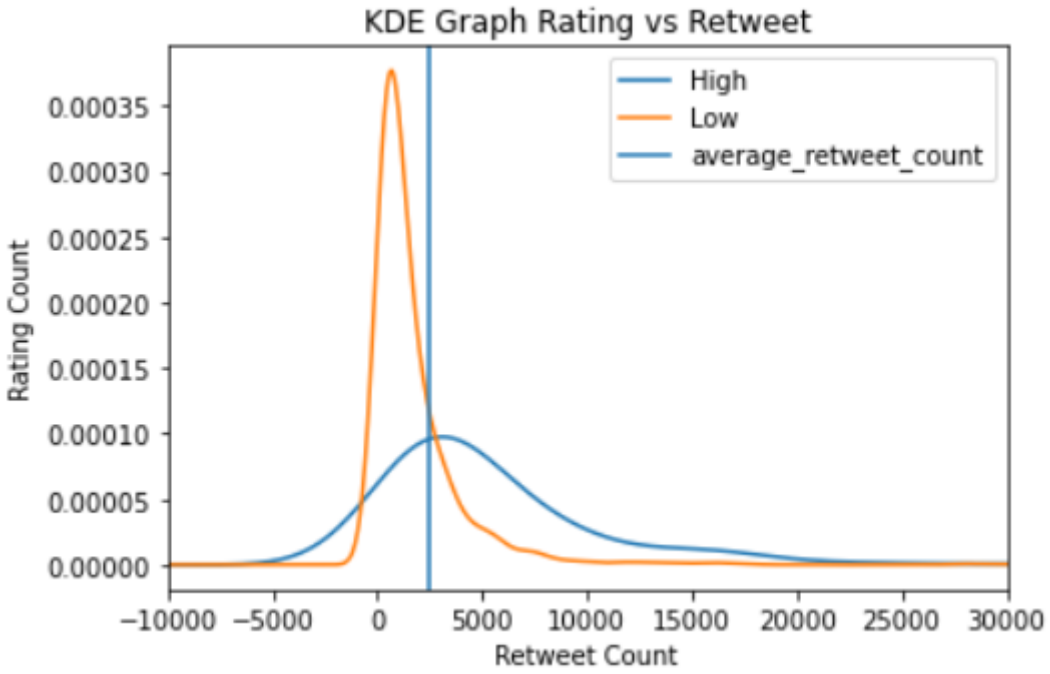
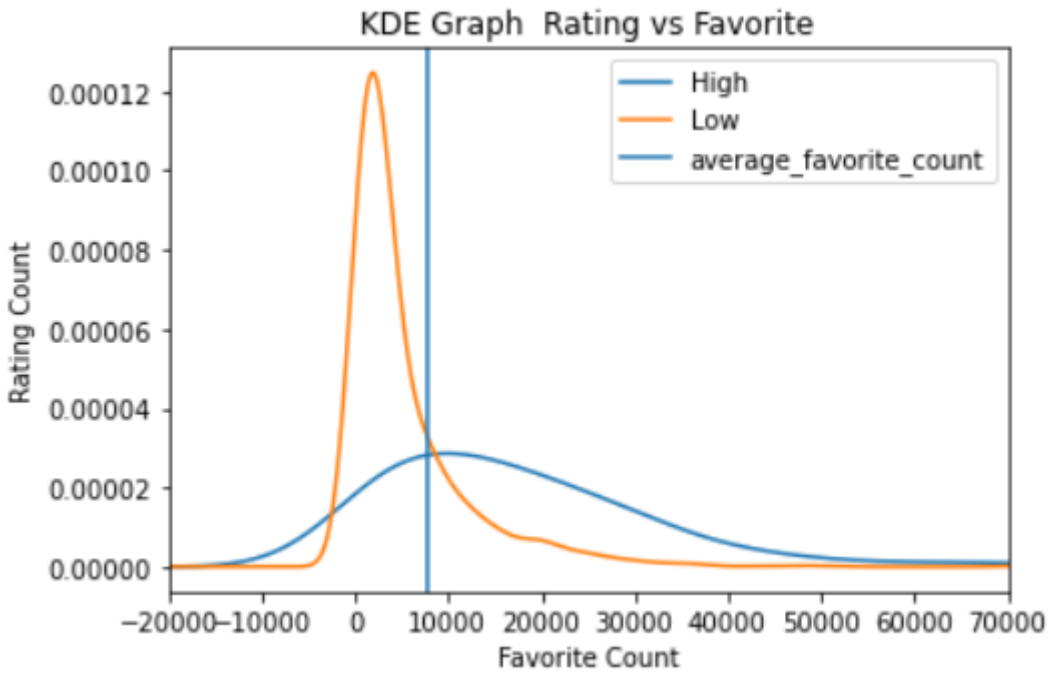
Act Report of WeRateDogs

In this report we will analyze tree insights of the master archive ‘twitter_archive_master.csv’ of WeRateDogs. This is a cleaned dataset, but still it was necessary to do some small tasks such as change ‘timestamp’ from string to date and ‘dog_stage’ from string to categorical variable due to the fact that when saved as csv and when loaded some dataframe information is lost.

The first analysis related the rating that the dogs received with the number of retweets and if they had been marked as favorites. Then what one wondered was if a tweet having more retweets would have a better or worse rating? Or if a tweet got or received a better rating how was the favorite count? This can be seen better in the following figures.

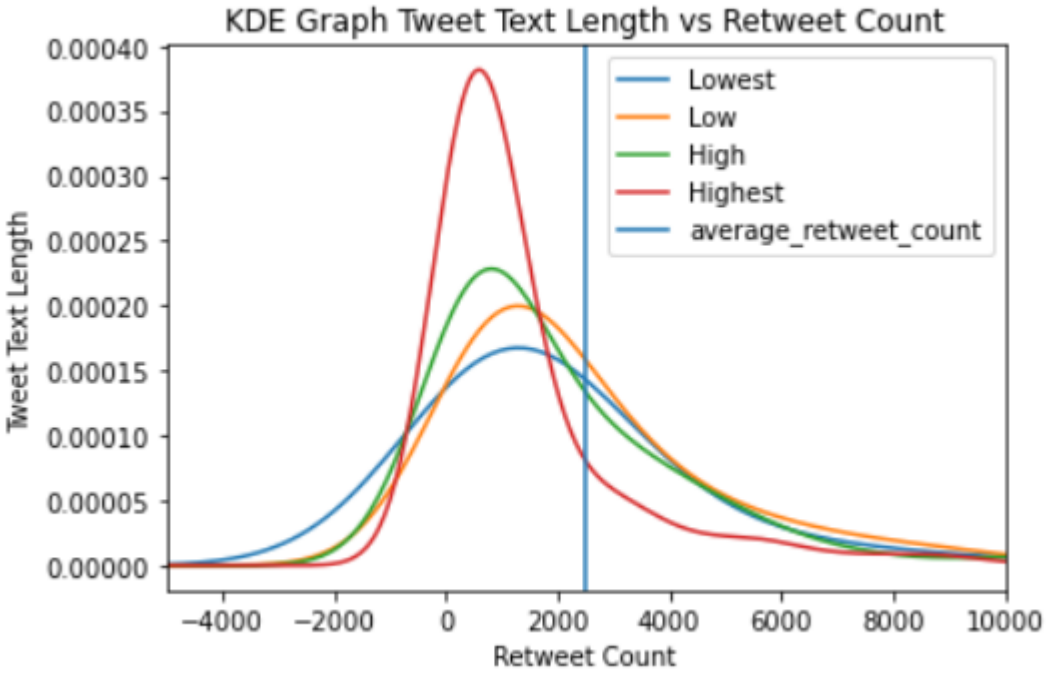


Similarly to the histograms in this KDE graph the more retweets a tweet had the more rating it achieved

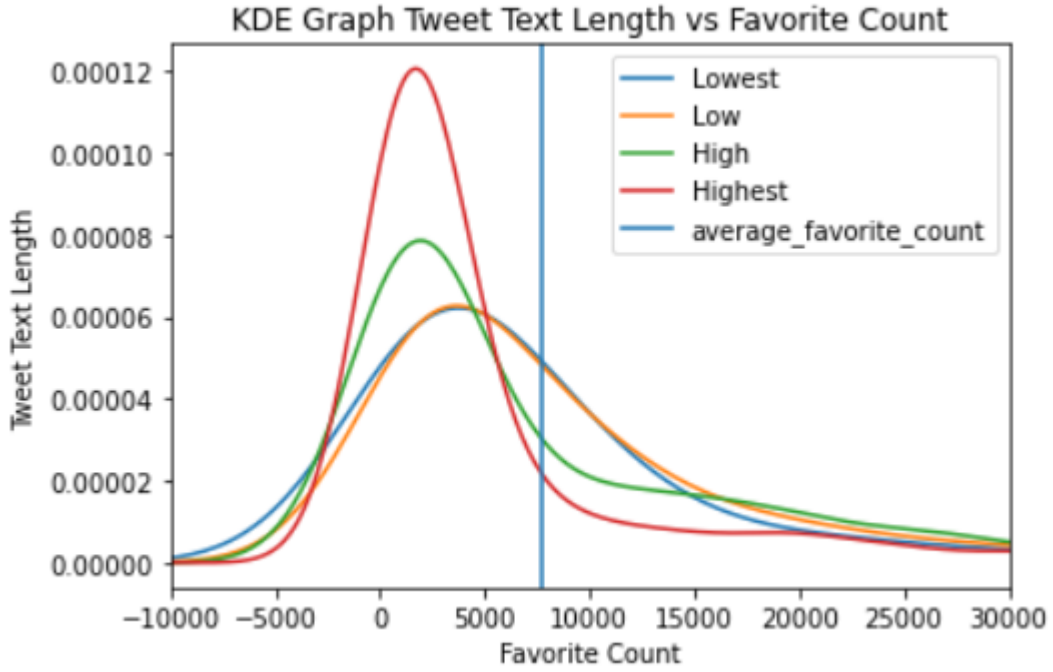


In this KDE graph the more favorite count a tweet had the more rating it achieved. On the contrary the lower rating a tweet received the lower favorite count received.

The second insight dealt with the number of characters or length or a tweet and if they received more or less retweets or if they were more favorited. Here it was important to convert the tweet's length into a variable easier to see the results. That is why it was converted into quarters so that the shortest in the 25% average were LOWEST followed by LOW(25%-50%), HIGH(50%-75%) and HIGHEST(75%-100%) representing the longest tweets with more characters.

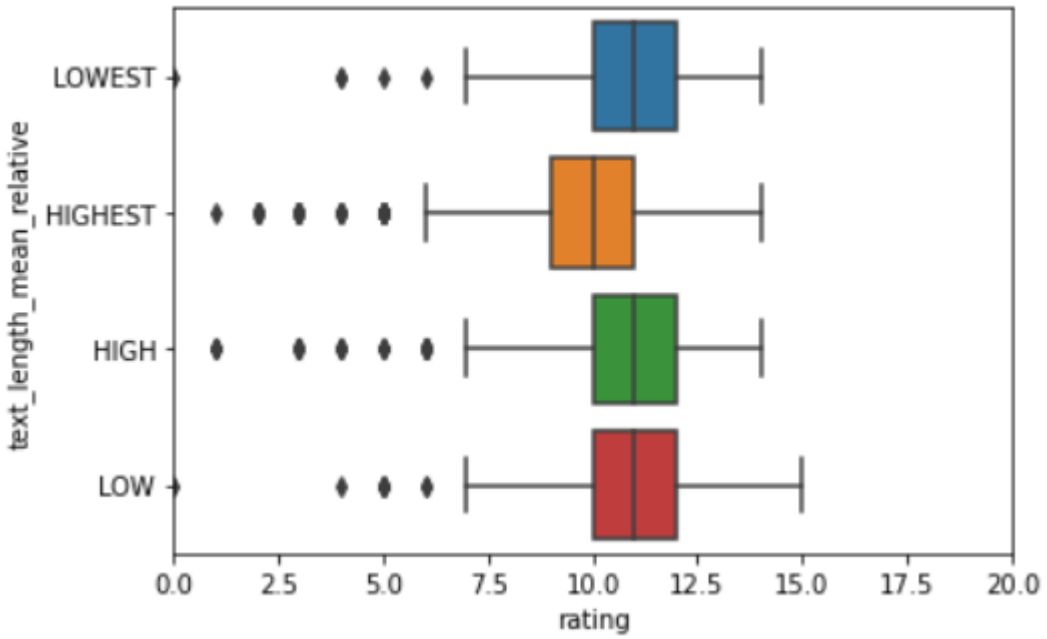


This figure essentially shows that there were more retweets of the tweets in the length range 0-75% because the lengthy tweets in the upper quarter ‘Highest’ had visibly a lot less tweets especially above the average of 2200 reteets.



This graph shows that the most favorite tweets were the ones in the lowest and low length quarter, that is when they have more than around 5000 favorites. The least favorite tweets were the lengthy ones depicted by Highest and High in the figure.

Finally, the third insight cared about if the lengthy tweet had a higher rating. At first the results in the KDE graph were quite similar because the curves were quite close so that a boxplot was created.



The graph above basically showed that the lengthier tweets achieved a lower rating in fact the minimum was reached by the box in orange that represented the upper quarter(75%-100%) of tweet's length. The maximum rating was reached by tweet in the LOW quarter(25%-50%) tweets with length less than average.