Project: WeRateDogs Data wrangling

Introduction

After cleaning data, we are interested in answering the following questions:

- What are the proportion of best prediction per algorithm and the best algorithm?
- Is the length of Tweets's text log-normal distributed?
- How does retweet_count relate to favorite_count?
- How does dogs' stage relate to favorite count?

Analysis and visualizations

Insight 1: What are the proportion of best prediction per algorithm and the best algorithm?

Among the four algorithms run through the neural network, the algorithm P1 has 85.8% proportion of best predictions, followed by the algorithm P2 with 9.5%, then P3 algorithm with 3.2% and finally P4 algorithm with 1.5%. Overall, P1 is the best prediction algorithm. Figure 1 is a bar chart that depicts our findings.

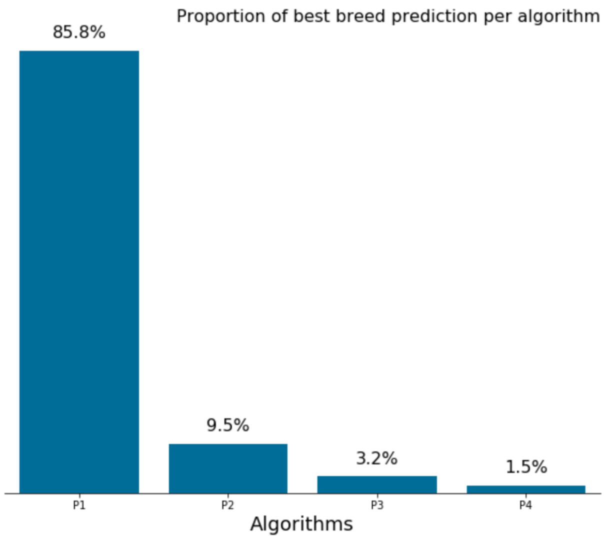

Figure 1: Insight 1

Insight 2: Is the length of Tweets's text log-normal distributed?

The will to know if the Tweets's text is log-normal distributed comes from this Wikipedia article about log-normal occurrence and applications. The article states that "*the length of comments posted in Internet discussion forums follows a log-normal distribution*". However, when plotted the distribution does not appeared to be log-normal.

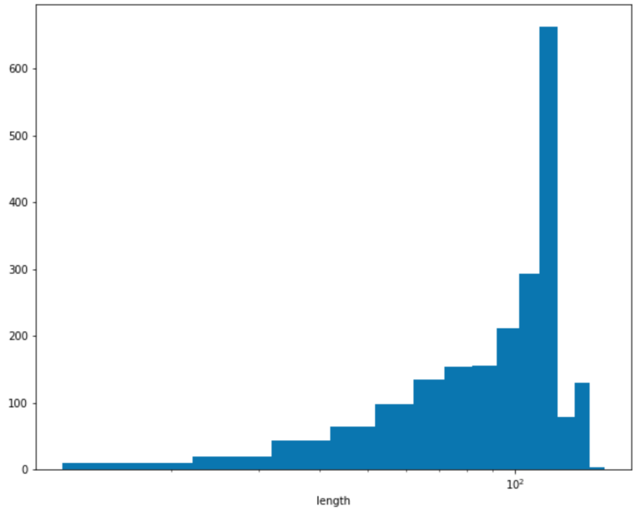

Figure 2: Insight 2

Insight 3: How does retweet_count relate to favorite_count?

Basically, we want to know if more liked tweets are the most retweeted. For this purpose, we first plotted the distribution of retweet_count and favorite_count (Figure 3). However, they are highly skewed. Then, we plotted the squared-root transformation to have a more clear plot. The square-root transformation helped have a good view of both distributions, especially favorite_count (Figure 4). Finally, to find the relationship between retweet_count and favorite_count, we plotted both a scatter plot and a heat map (Figure 5). When we look at the scatter plot, we can firmly say that retweet_count and favorite_count are positively correlated. So, the more a tweet is liked, the more it is retweeted.

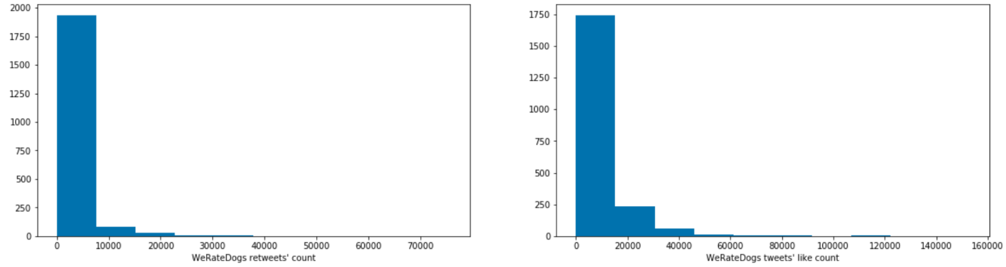

Figure 3: Insight 3-1

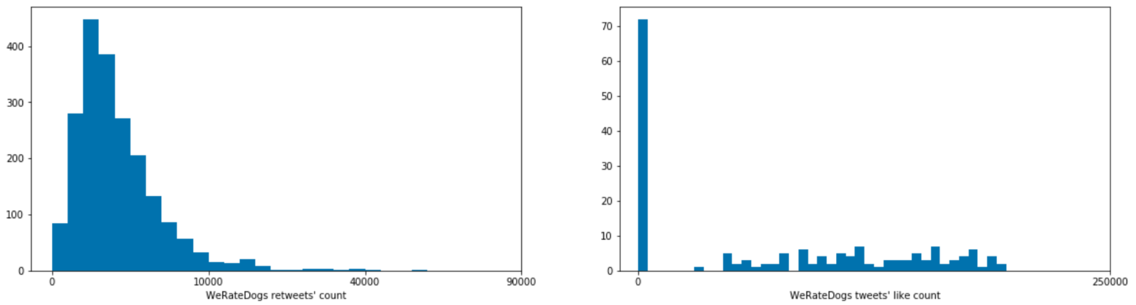

Figure 4: Insight 3-2

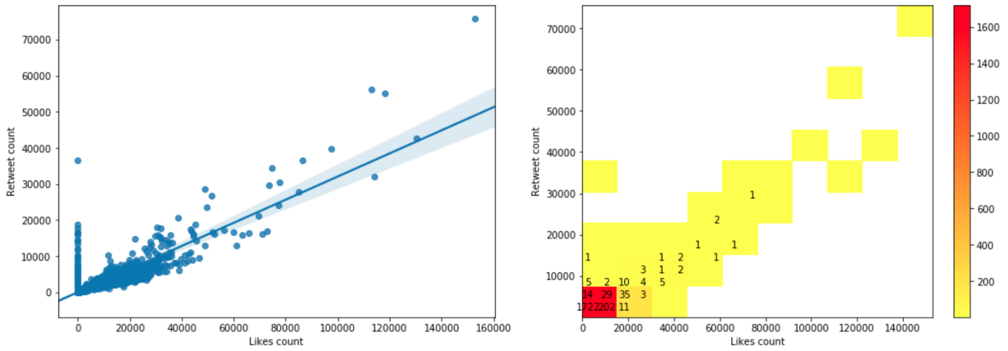

Figure 5: Insight 3-3

Insight 4: How does dogs stage relate to favorite count?

Essentially, we want to see if dogs stage were determinant in how likely people are to like a tweet. We use violin plots to portray this (Figure 6). The plot reveals that there is not a monotonic relationship between dogs stage and favorite count. Tweets with **doggo** and **puppo** as dog stage have the highest favorite count (more than 150000) and dislike (favorite count less than 0). Moreover, their longest tails suggest that they have the highest number of outliers. Tweets with no specified dog stage and **pupper** as dog stage have the next longest tails and numbers of dislikes. Though, their curves are more wider. Thus, the greater number of tweets fall into both of these categories. Finally, **floofer** unlike the other dog stages, have no outliers, no dislike and more tweets fall in it than **doggo** and **puppo**. Considering all the above people tends to like more **floofer** than other dog stages.

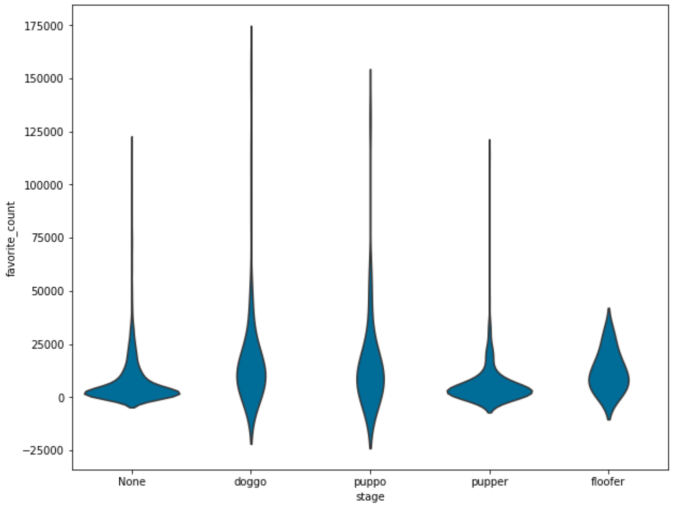

Figure 6: Insight 4

Conclusion

Our analysis and visualizations efforts consists of five (4) insights and ten (9) plots. To be more specific, we performed two univariate explorations and two bivariate exploration.