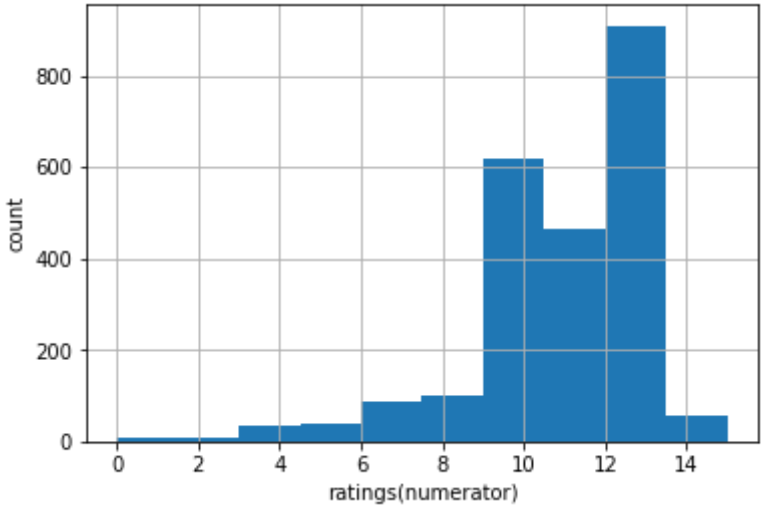
This report documents inspired insights resulting from data wrangling.

1) **What is the distribution of ratings?**

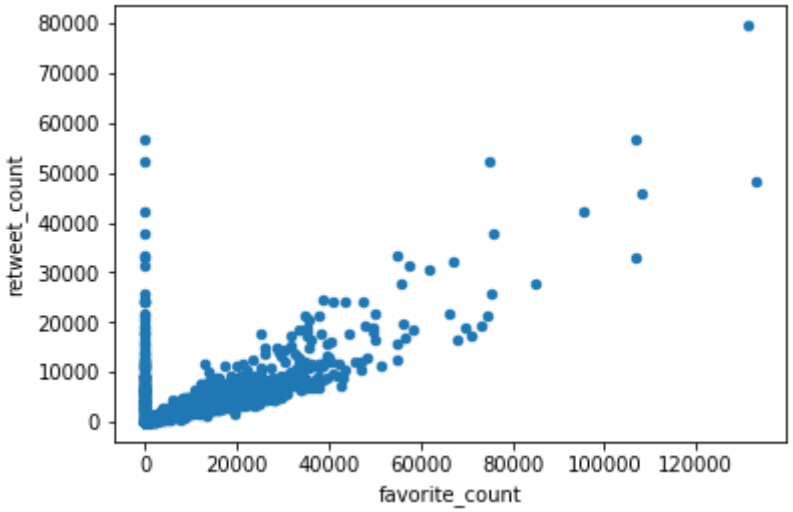
   By filtering out denominators not equaling 10 (according to project introduction, “These ratings almost always have a denominator of 10.”), we find numerators range from 0 to max 1776. If we draw the histogram of these left numerators, it will show only one bar, meaning outliers exist. Using Q3 + 1.5*IQR as an upper limit, the histogram is demonstrated below.



   It indicates most ratings from twitter users are in the range of 9 to 13.

2) **What is the relationship between favorite count and retweet count for these tweets?**

By drawing a scatter plot of favorite count and retweet count, we can see the relationship between these two columns.



The scatter plot shows that favorite_count and retweet_count have a positive correlation. For favorite_count equaling zero, there exist several retweet_counts, ranging from 0 ~ nearly 6000. This shows that users tend to retweet rather than favorite tweets.

3) **How good is prediction of dogs for this neural network model?**

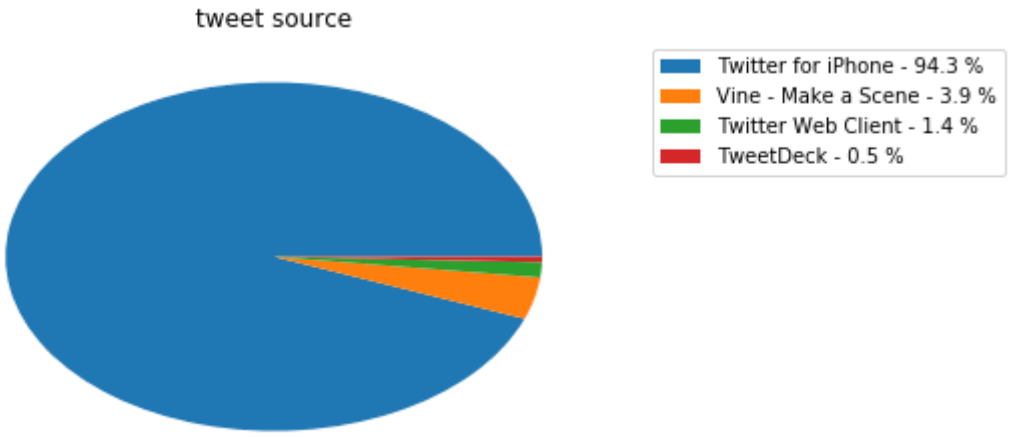
Join two DataFrame twitter and predict by the column ‘tweet_id’. Use the generated column ‘dog_type’ with not-null value as real dog tweets. Then use the ‘p1_dog’ column with True value as predicted dog tweets.

237(dog tweets and predicted correct) / 321(all dog tweets) = 0.738

The accuracy of predicting dogs is 73.8%.

4) Analyze sources of tweets.

The ‘source’ column represents where each tweet is generated. Normalize the counts and draw a pie chart of it.



The pie chart shows more than 90% of tweets are generated from iPhone.