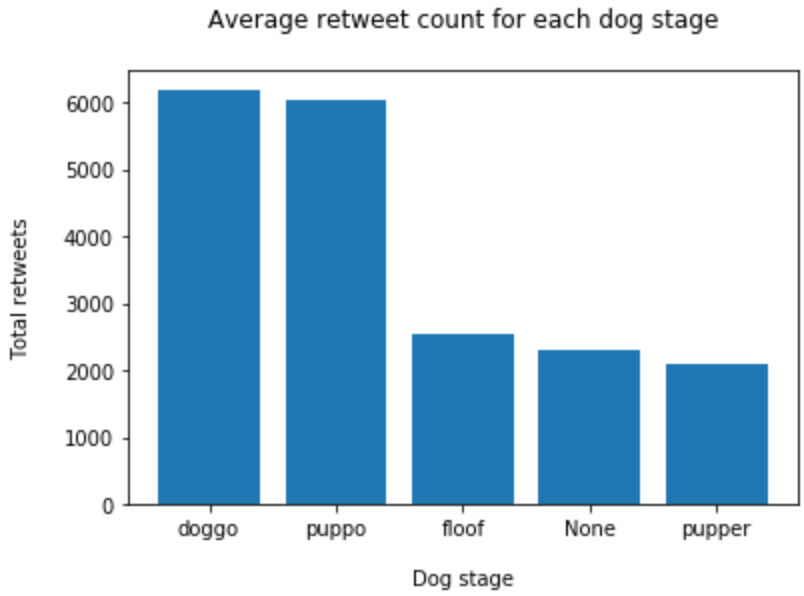
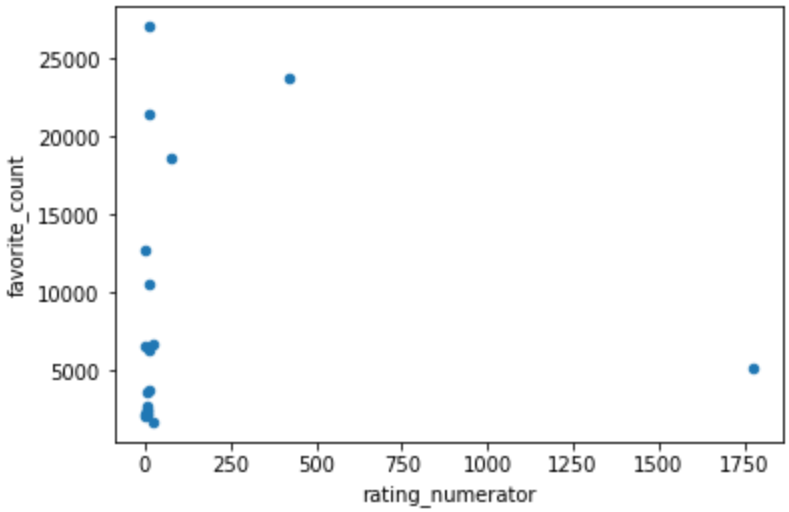
Acting on Wrangled Data Report

After completing the data wrangling process, I moved on to act on my wrangled data, and completed a brief exploratory data analysis. I chose to investigate the following 3 insights: I looked at which dog stage is most commonly mentioned in the datset's tweets, and then looked to see if there was a correlation between dog stage and number of retweets. Next I looked to see if there was any correlation between the rating numerator and number of favorites: did tweets with higher rating numerators have more favorites than those without? Finally, I wanted to look into some of the information from the image_predictions.tsv document, namely the first predictions made by the neural network. Was there a difference in number of retweets or favorites, when looking at whether the neural network was able to predict a dog in the first try or not?

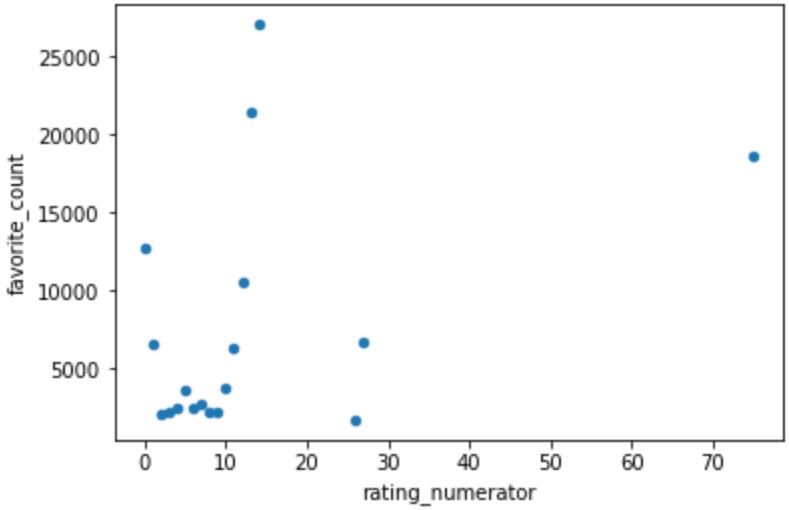
In looking at the first insight, "Which dog stage is the most common, and does the dog stage seem to have a correlation with the number of retweets?", I found that a significant number of tweets do not mention the dog stage, even after scanning the tweet texts for this information. None was mentioned nearly 8x more in the tweets than the next runner up pupper, and doggo, floof, and puppo were mentioned significantly less. Regardless, I still wanted to look the correlation between dog stage and number of retweets, and found that overall it seems that tweets mentioning a dog's stage were retweeted significantly more than tweets with no dog stage (except in the case of puppers). I created a bar graph (below) comparing the average retweet count across dog types.



Next, I looked to answer the question "Is there a correlation between the rating numerator and number of favorites?". I grouped rating numerators and looked at their average favorite counts, and found that there was a wide range between the min and max rating numerators (which had been adjusted during wrangling). At least one tweet had given a dog a 0/10 (unbelievable!) and at least one had given a dog a 1776/10 (that must have been an extremely impressive dog!). When plotting the first scatterplot of the rating numerators and favorite counts, it was clear to see that the very large ratings (420 and 1776) caused the data points to be in an almost completely straight line, as the x-axis was so spread out (below).

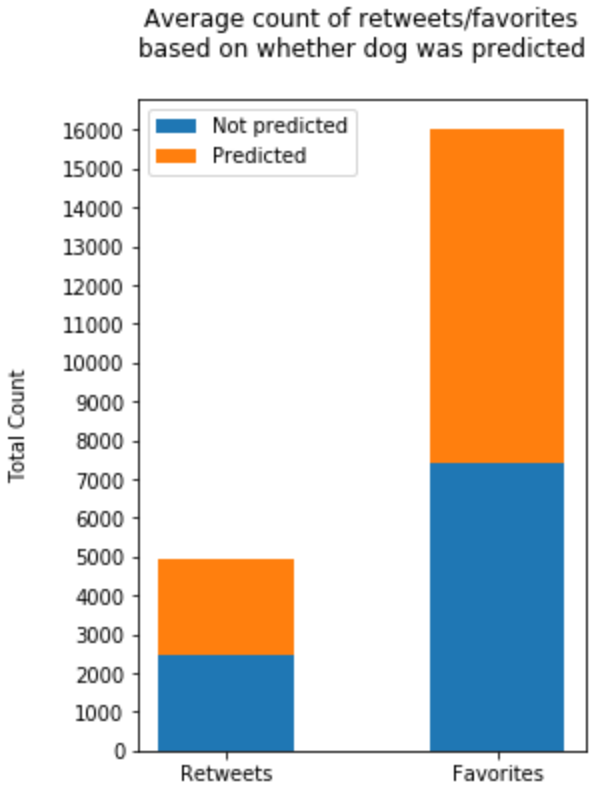


I decided I wanted to take a look at what the graph would look like if I dropped these 2 large values, and the 2nd graph I created showed a bit more detail:



I then decided to look at the correlation coeffecient of both graphs (with and without the 2 large values) in order to have a more in-depth statistical look at the correlation between these values. With the 2 large values, the correlation is a very weak positive one (≈ 0.029) and after dropping these, is a moderator positive one (0.394). These results show that it may be worthwhile to iterate and revisit the cleaning stage, to potentially remove these 2 large values.

Finally, I chose to look at the neural network information, and sought to answer the following questions: “For how many tweets was the neural network unable to determine that the image was of a dog? Did these images have fewer retweets than those in which a dog type was identified in the first go?” I found that there were nearly 3x the number of True values in the p1_dog column as False values, which shows that the neural network predicted a dog about 74% of the time in the first prediction. I then grouped by first the retweet count, and then by the favorite count, and saw that the average favorite count was substantially higher than the retweet average count. There also appeared to be a very minimal difference between True and False for the first predictions in terms of retweets (≈ 27), but there was a more significant difference in terms of favorites (≈ 1207). I then plotted these results in a stacked bar graph which more clearly shows these differences (below).



It seems possible that tweets are more popular when the dog in the image is more clearly recognizable as a dog/when there are less background objects that may distract from the dog itself. This information was quite interesting to look at and think about, and it made the various data wrangling and gathering of sources worth the effort, as without having taken these steps, this insight would not have been available for interpretation.

In closing, I also identified several other insights that could be looked at in this dataset, such as whether certain months have more posts than others, and if the number of tweets on the WeRateDogs page has changed over time, as well as whether certain dog names were shared more often than others, which kind of dog type has the most favorites, does the length of a tweet have an impact on the number of retweets/favorites it gets, and so on. I realize that my analysis could be more thorough, and that the dataset could be cleaner. I also made the assumption that if a row has a value in the retweet status id column, it means it is a retweet, but this could have been incorrect, which would have led to skewed data.