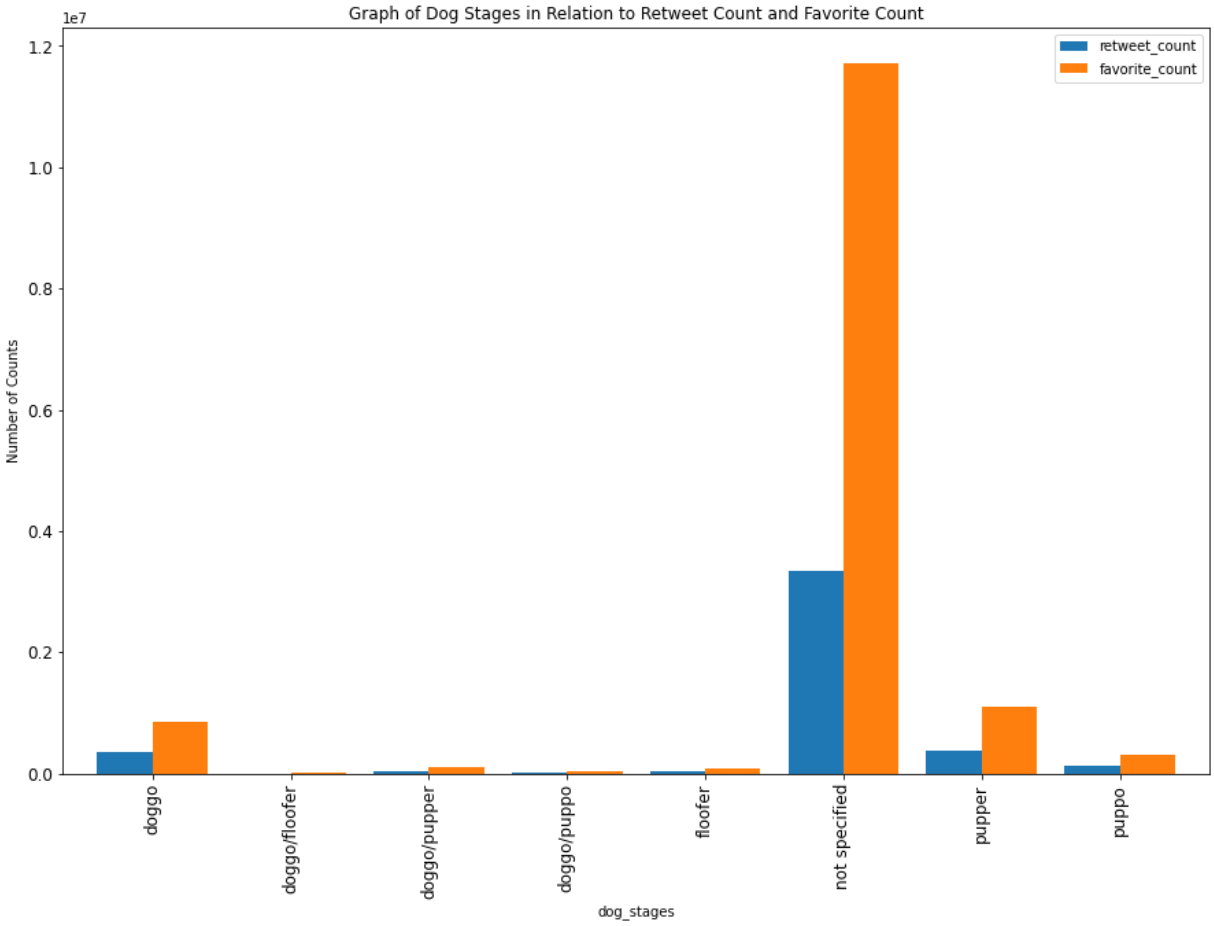
WERATEDOGS DATA WRANGLING PROJECT

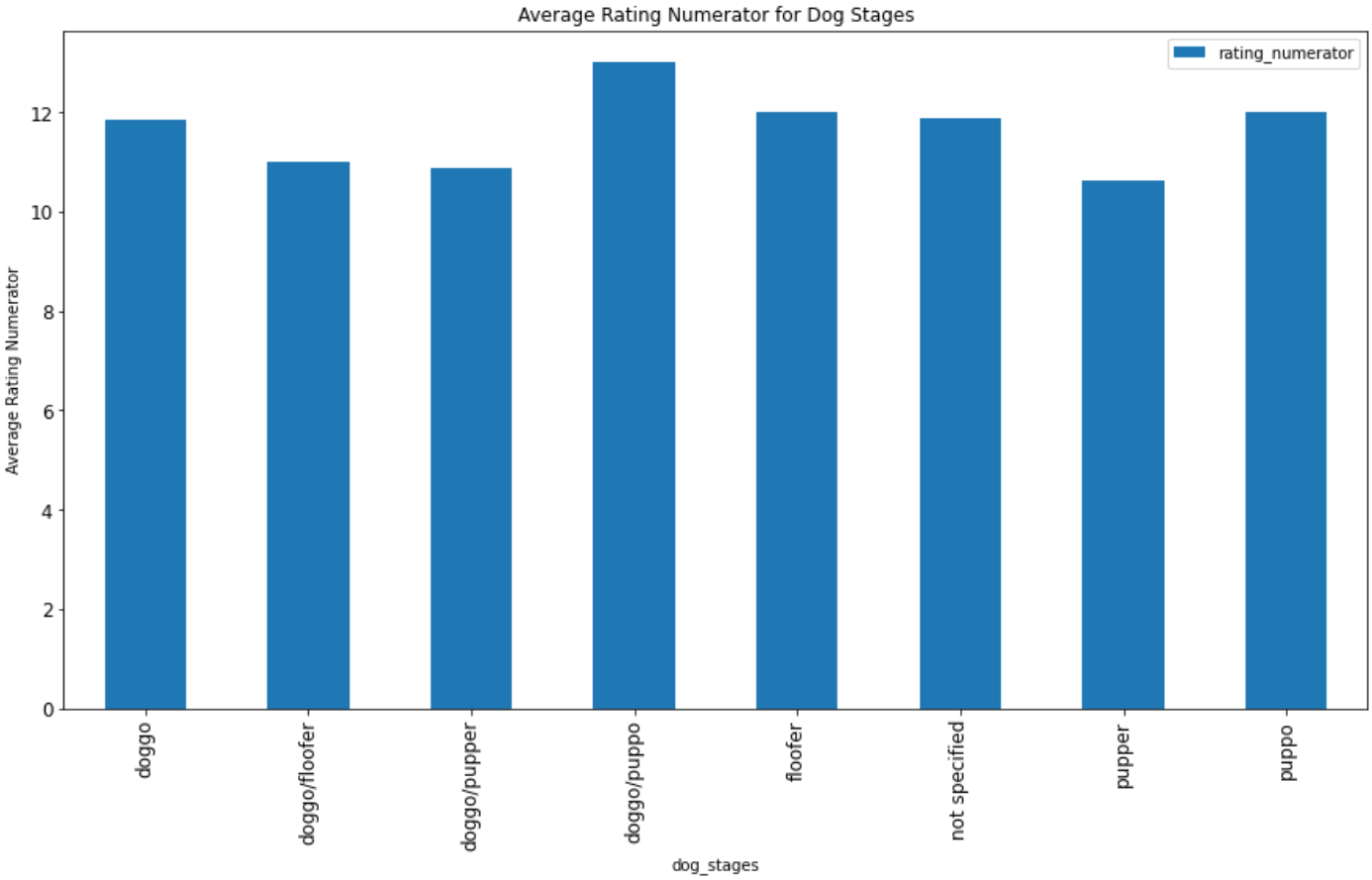
The dataset that I **wrangled**, analyzed and visualized is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent.

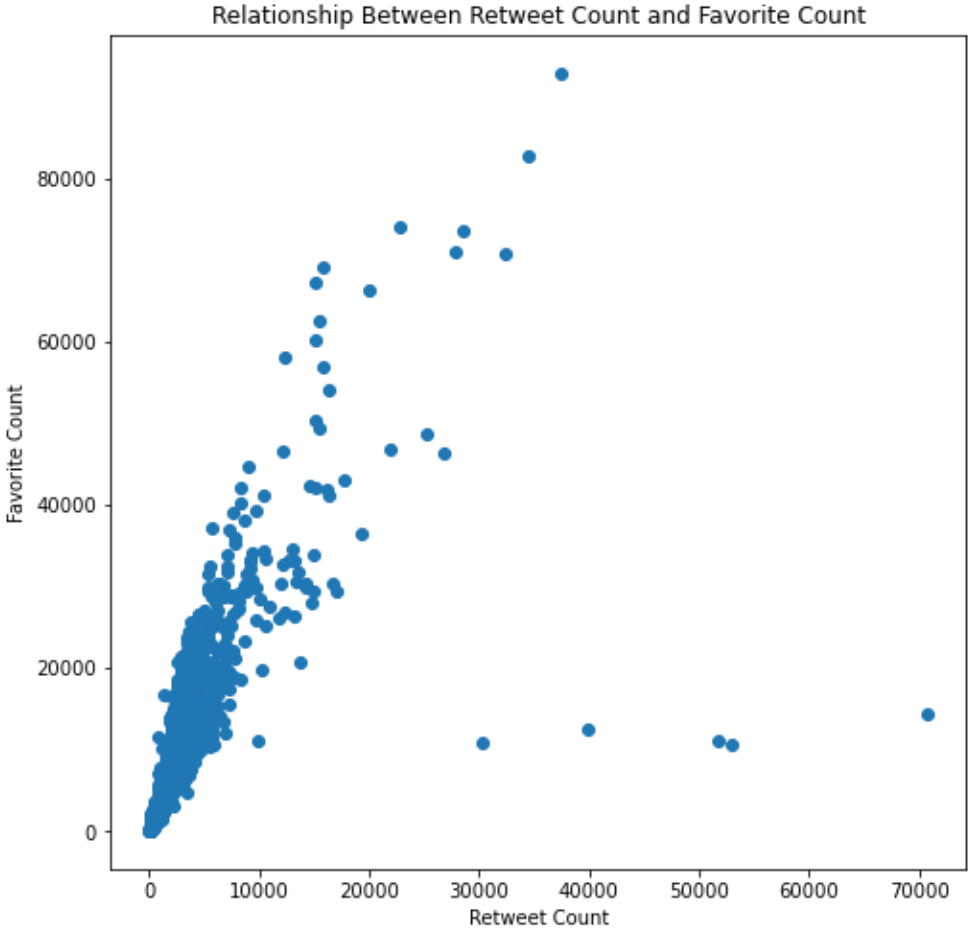
In a bit to compare retweet count and favorite count for the various dog stages in the dataset, the data frame is group by dog_stages, then selected only retweet_count and favorite_count to check the total retweet count and favorite count for each dog stage. This was later show as a graph which make it easier to analyze. The graph below is the bar chart plotted.



The graph above shows that dogs where the dog_stages is not specified has the highest favorite and retweet count, which make it not logical to really determine the dog_stage with the highest retweet count and favorite count. But for the specified, the graph shows that **'pupper'** has the highest total favorite count and same as the highest retweet count, followed by **'doggo'**, then **'puppo'** and so on.

Also, the average rating_numerator for the various dog_stages was computed as a table then a bar chart is plotted using the table gotten. The graph below shows that dog_stage **'doggo/puppo'** have the highest average rating_numerator while **'pupper'** has the lowest rating_numerator.


Average Rating Numerator for Dog Stages
rating_numerator
Average Rating Numerator
0
2
4
6
8
10
12
doggo
doggo/floofer
doggo/pupper
doggo/puppo
floofer
not specified
pupper
puppo
dog_stages


Relationship Between Retweet Count and Favorite Count
Favorite Count
0
20000
40000
60000
80000
0
10000
20000
30000
40000
50000
60000
70000
Retweet Count

The figure above is a scatter plot that shows the relationship between the retweet count and favorite count. This shows that there is a positive correlation between retweet count and favorite count, which means a dog with high favorite count will have high retweet count and vice versa.

Some of the insights gathered from the analysis are:

- Dogs whose dog_stages are not specified has the highest total favorite and retweet count, which make it not logical to really determine the dog_stage with the highest retweet count and favorite count. Hence, @WeRateDogs should mandate clients to indicate dog stages for every submission for a better analysis. But for the specified, the analysis shows that **'pupper'** has the highest total favorite count and same as the highest retweet count, followed by **'doggo'**, then **'puppo'** and so on.
- 'doggo/puppo' have the highest average rating_numerator while 'pupper' has the lowest average rating_numerator.
- There is a slightly positive correlation between retweet count and favorite count, because the correlation is not evenly distributed across the values. This implies that a dog with high favorite count will have high retweet count and vice versa, but this only hold for dogs whose retweet count is less than 20,000 and its favorite count is not more than 40,000. To gain more insight in to this, a predictive analysis can be carried out on the data or gather more tweet ids with relatively high retweet and favorite count.