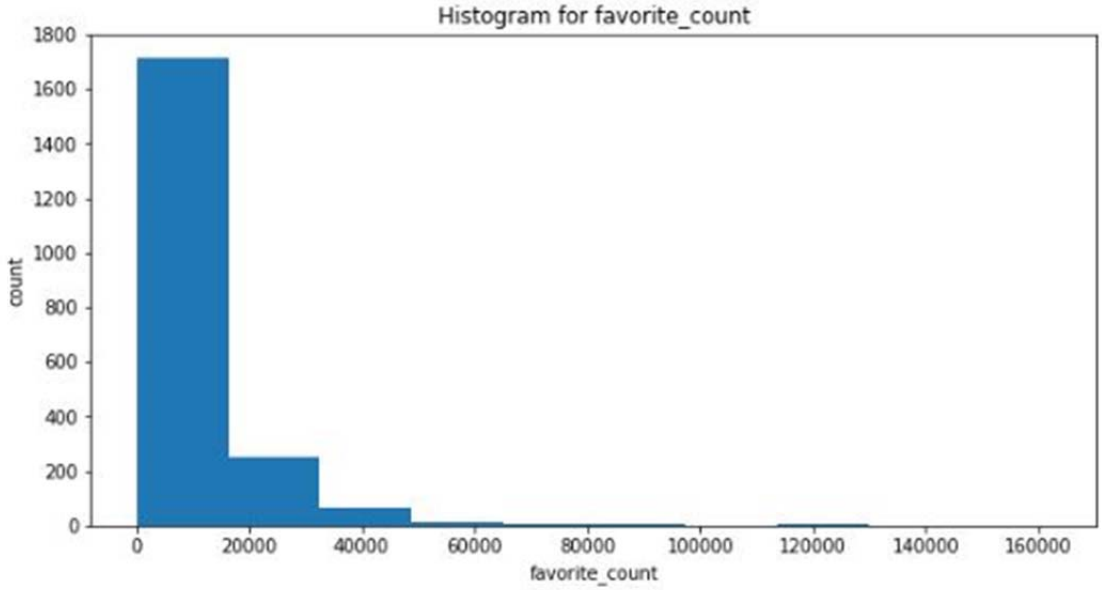
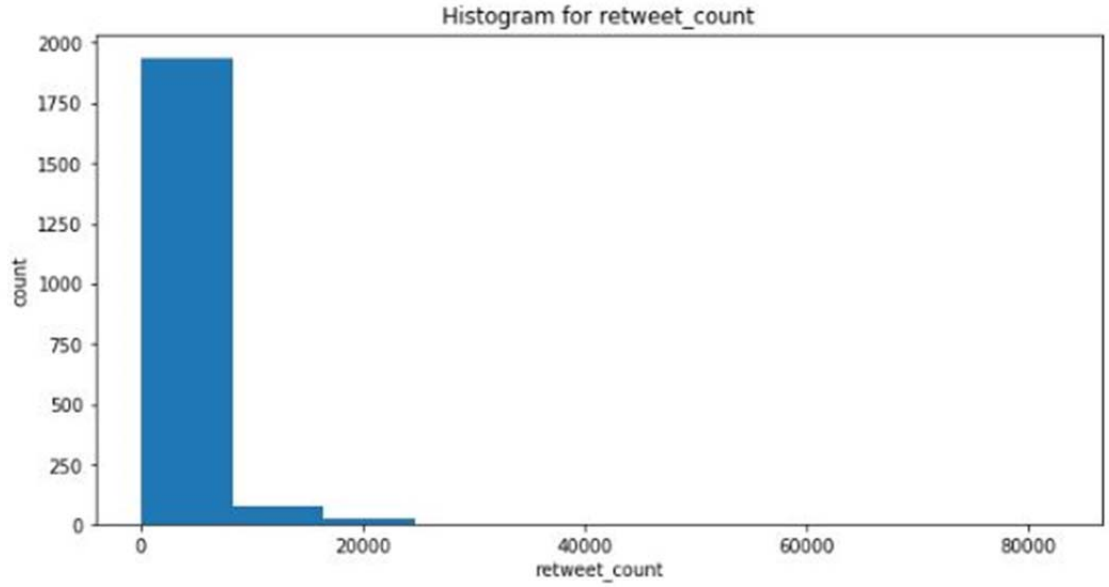
After wrangling the datasets and get it ready to work with, I am going to make some insights and visualization. However, since the focus in the project is on the data wrangling, I will do only some basic insights and visualization.

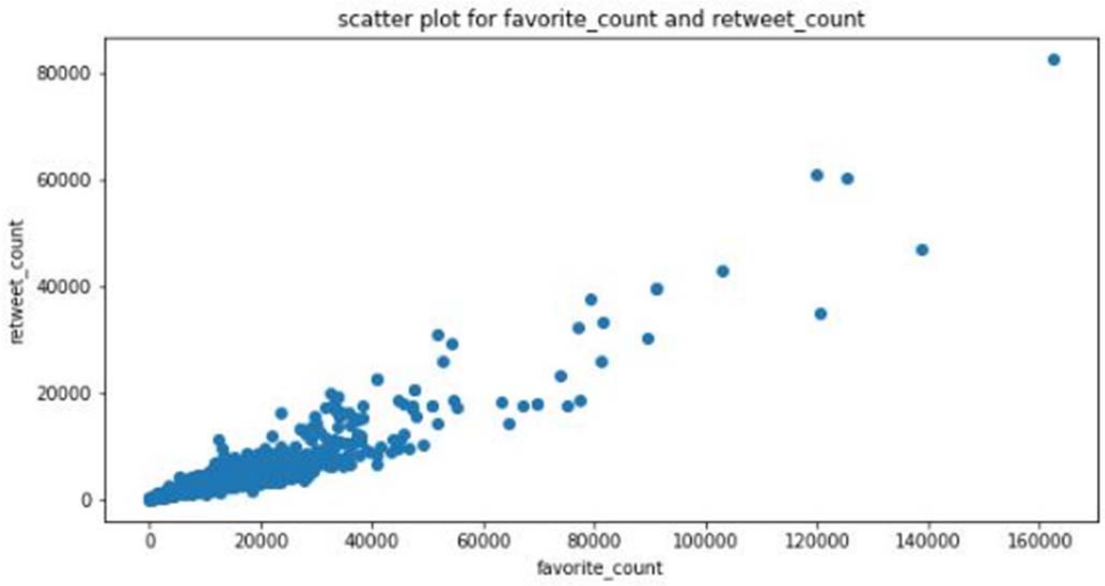
The first thing I am interested to know is whether there is a relationship between `favorite_count` and `retweet_count`. Before running the scatter plot, I checked the distribution of each variable as follows:





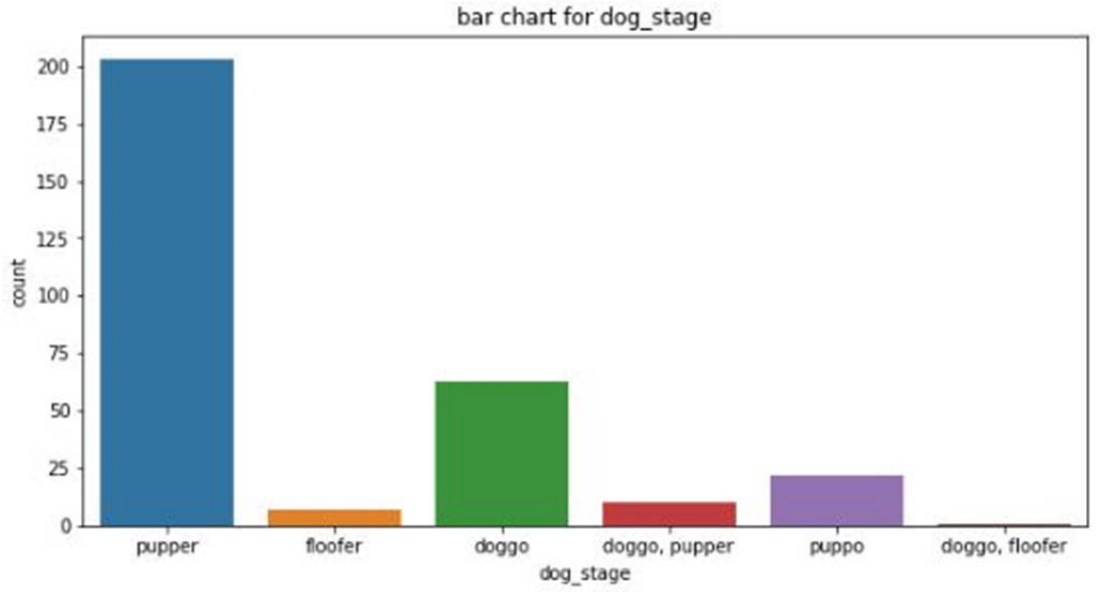
As we can see, both variables are highly skewed to the right, most of the `favorite_count` frequencies ranges between zero and slightly less than 20000, while `retweet_count` between zero and 8000.

To see the the correlation between the two variables, I ran scatter plot using `matplotlib.pyplot.scatter.`



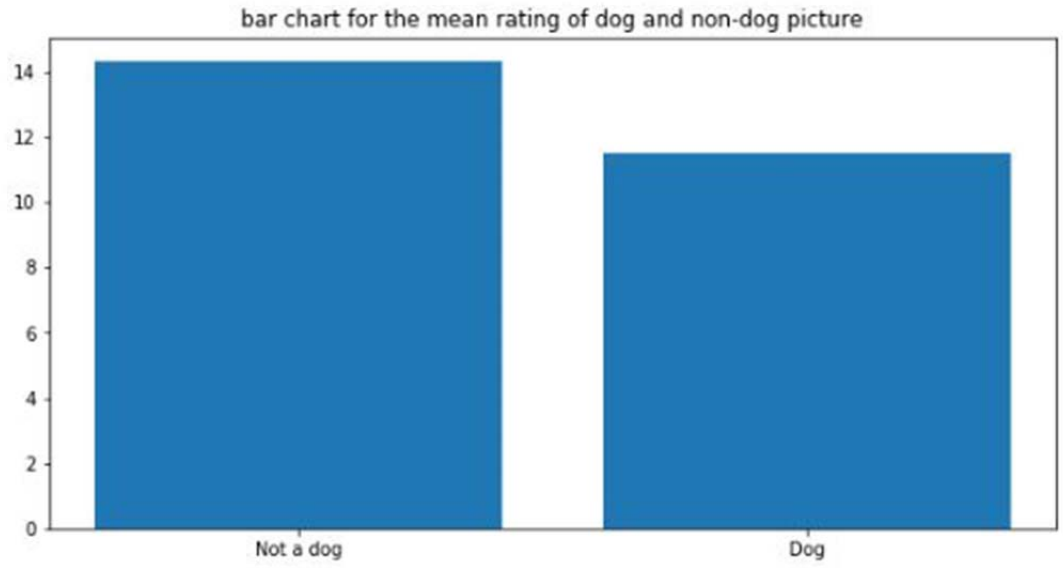
From the scatter plot above, we see that there is a strong positive relationship between the two variables. One explanation could be that the more the tweet was liked the more it was retweeted.

I also was interested to see the first three highest counts of dog stage, so I ran a bar chart using `sns.countplot` for the `dog_stage` that is not equal to 'None'. Here is the bar graph



So from the graph above, we see that the highest three dog stages are pupper, doggo, and puppo.

Another thing, I was interested to see if there was a difference in mean rating between the pictures of dogs and those that are not of dogs. The mean rating for dog pictures was 11.486 and for non-dogs was 14.325. Here is the bar chart



Since this is a dog rating, I was expecting to see higher mean dog rating. However, since this rating is for humor, we cannot take the it seriously, and it won’t be a good idea to include it in any analysis.