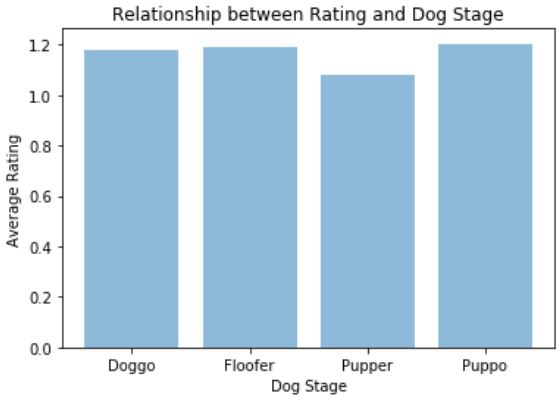
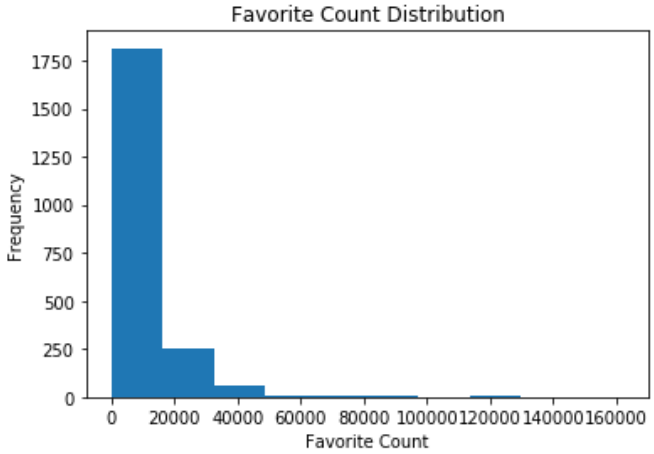
Data Analysis and Visualization

- **How does the dog stage have impacts on the rating?**



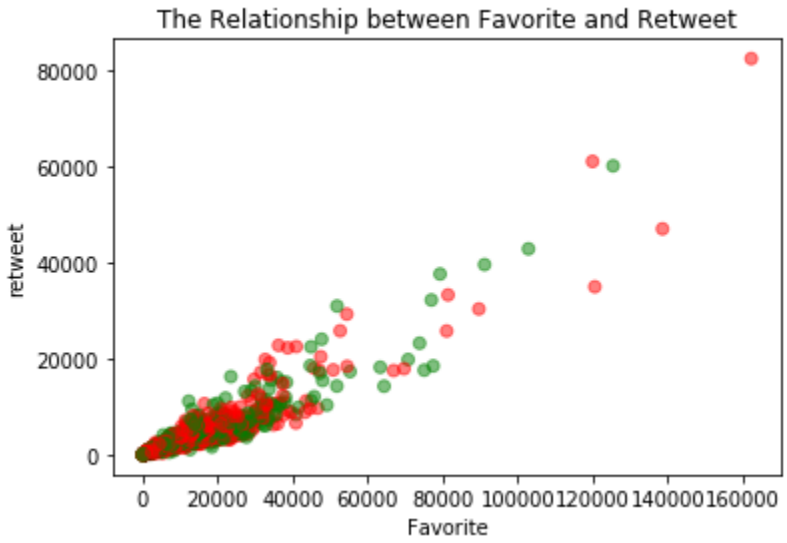
Based on the bar chart above, Pupper received the lowest average rating among these four stages. On the other hand, Puppo has the highest average rating than any other dog stages. Checking the definition of dog stages, I assume the transitional behavior of Puppo could be appealing to many people, because its one of the main traits you can observe from Puppo. Pupper is a dog stage in which dogs which are smaller or less mature than Doggo. People rate big dogs more highly than small dogs.

- **How does the favorite count look like?**



According to the histogram, most of favorite counts received by the dog are below 20,000. The histogram is right-skewed and has a long tail. When I check the dataset, the most likes, 162,213, received by a Doggo, a result which does not contradict to what we found above. The fewest, 51, likes received by a dog of unknown stages.

- **How does favorite count correlate with retweet?**



It is apparent that these two variables on the scatter plot have strong positive correlation. It is actually a logical finding because the more likes you receive from people, the more likely the post will be retweeted. Consistent with what we found in the preceding question, the majority of favorite counts and retweet counts concentrate on the number below 20000.

- **Conclusion:**
  Our data first shows Puppo is the highest-rated dog stage in this dataset. Then we also see how favorite counts are distributed in our dataset. The distribution is right-skewed, meaning the majority of dogs receive fewer likes. One last interesting thing we find in this dataset is the positive correlation between favorite counts and retweet. This finding is also logic and serves as a good indication on how many retweets a dog will receive given how many retweet would be received. We can simply build a regression model for our prediction.

- **Limitation:**
  There are several limitations for this analysis. Firstly, the data is limited to information from 2017. Two years have already elapsed, so the trend might already differ and our conclusion is not valid any longer. Secondly, the conclusion here is based on visual assessment. More statistical testing should be done in order to prove our findings are statistically significant. Lastly, the humongous data in our dataset prevents me from going through the table line by line. I just clean the information I find the most relevant for my analysis. Although data wrangling is already performed, there might be more quality and tidiness issues in our dataset.