Act Report for DogRates Data Wrangling Project.

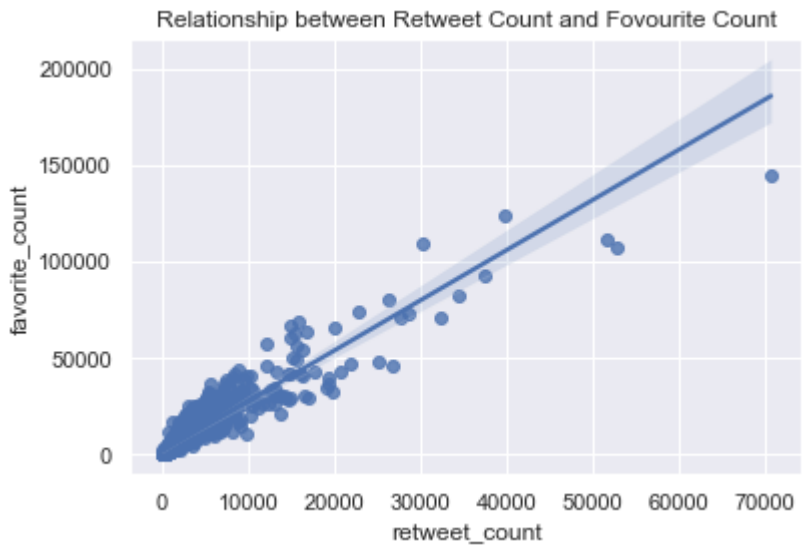
Review of the data imported from the cleaned datasets in the.csv files revealed that integer values that were stored as strings were converted back to integers before exploring the data to determine what topics or questions to investigate. As a result, depending on the data or the project's objectives, the data wrangling steps might need to be repeated.

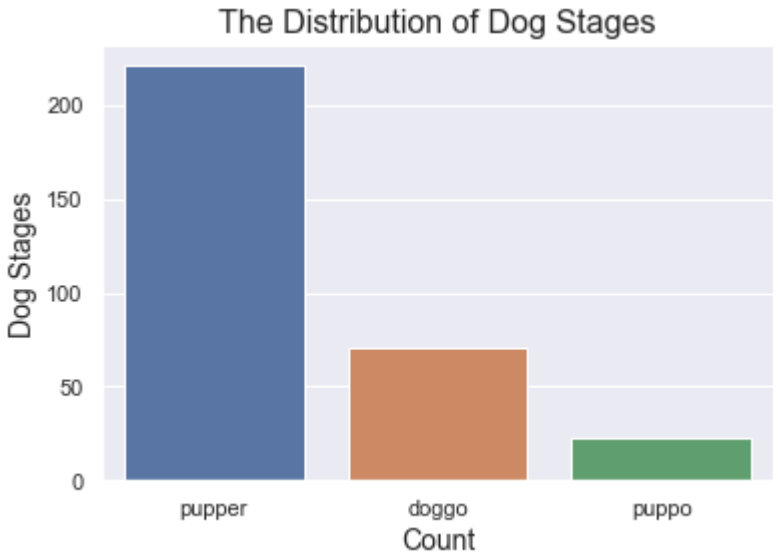
For the first issue, I converted the columns with similar column names to the same data type to ensure data consistency and uniformity. I then proceeded to run some descriptive statistics to just refresh myself with the imported datasets. I grouped the combined datasets by the source from which each tweet emanated, and the pivot table showed that the majority of tweets for @dogrates were sent from the iPhone. I wanted to see if there were other aggregations besides the device by which users tweeted.

I looked at the combined dataset to find a correlation between the variables and discovered that there was some level of relation between favorite_count and retweet_count with a correlation coefficient of 0.925636. This represents a moderate correlation between the variables. Cooper, Charlie, Lucy, and Oliver are the most common dog names as can be seen on the bar plot.

Also, from a dataset containing image predictions for dog breeds, it was discovered that the top 5 dog breeds based on the confidence levels include:

A better understanding of the correlation between favorites and retweets, as well as how that correlation affects image predictions, can allow for new models of information to be created.





The bar chart above shows that the dog stage ‘pupper’ has occurs the most.