Act report

In terms of insights I have first of all used the value_counts() function as a means to find the most common names in our name column for my final master dataframe (df). This call returned the following top ten:

None 603
a 55
Charlie 11
Lucy 11
Cooper 10
Oliver 10
Tucker 9
Penny 9
Lola 8
Winston 8

As we can see ‘None’ is the most mentioned name. Now that I have seen this in this part of the process rather than the clean I would be able to go back to the cleaning section and replace all of those 603 ‘None’ values with np.nan which would in turn mean that the value_counts() function would not return ‘None’ as a value to count at all and this would be more fitting for our dataset. I would do the same with ‘a’ and ‘the’ as well as any other instances of clear mistakes that quite possibly come from a parser system that is acting on the tweet data.

Next I used the same function value_counts() to look at the image predictions from the neural network. In this I found that the golden retriever was the most common breed with 137 appearances and next the labrador retriever which are both very similar dogs which differ only in size and therefore it would be interesting to view how sure the network was in making predictions of these two breeds and also then to check with the second place prediction which I predict would be the one of this pair that is not predicted as first choice. This finding may make me go back and add back the second choice prediction as it clearly has some use here for investigating the integrity of the data on the basis of the image predictions.

Next I used the same function to retrieve the highest favorite counts and found that 30696 was the highest count present in the dataset.

Lastly as a visualization I wanted to show the effect of dog type on rating and plotted a barplot with dog_stages on the x-axis and rating on the y-axis, after observing the order in the first run of this function I then reran and imputed a list of the order to make the graph more clear. The graph helps us see that the difference in means of rating are small and in the range of 1.0-1.2. Using the visualization I produced at the start of the analysis section I can see the respective counts of dog_types to give us a broader picture on the dataset.

Below are the two visualizations I used in my investigation: