ACT_REPORT BY CHINEDU UZORUE

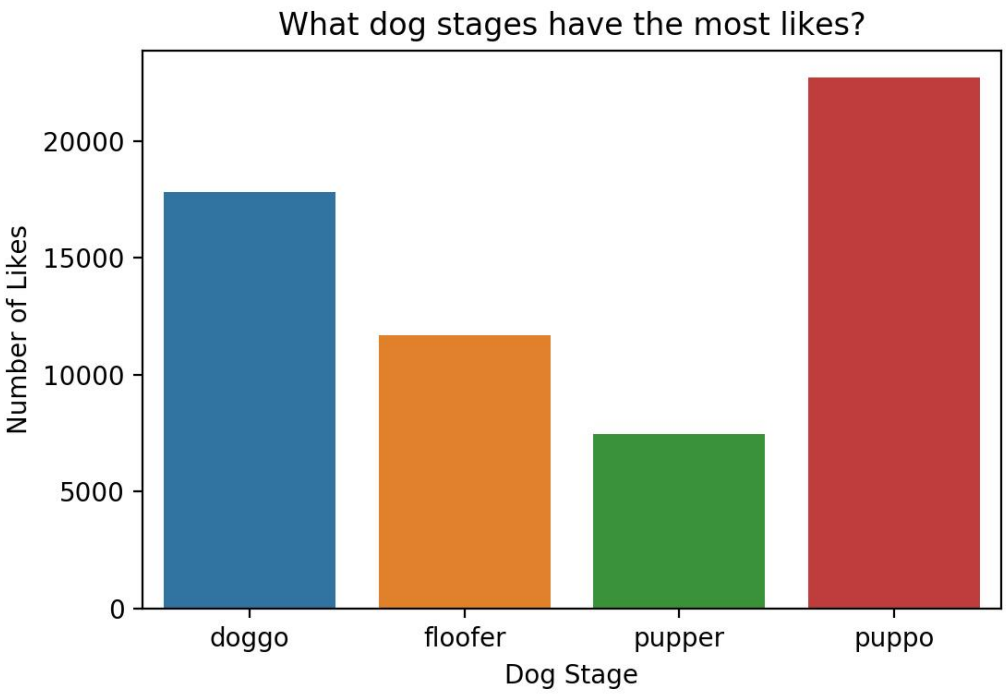
The ultimate purpose of any data analysis process is to draw meaning insights from data using colorful visualizations. This project is not any different, as I was able to draw four (4) insights from the cleaned data and make visualizations from them as well.

To begin, I asked myself 4 questions which I needed to answer, the questions are:

1. Which dog stage has the highest likes (favorite_count)?
2. What is the dog stage with the highest retweets?
3. Which dog stage has the highest average rating?
4. What is the name of the most famous dog?

1. Which dog stage has the highest likes (favorite_count)?

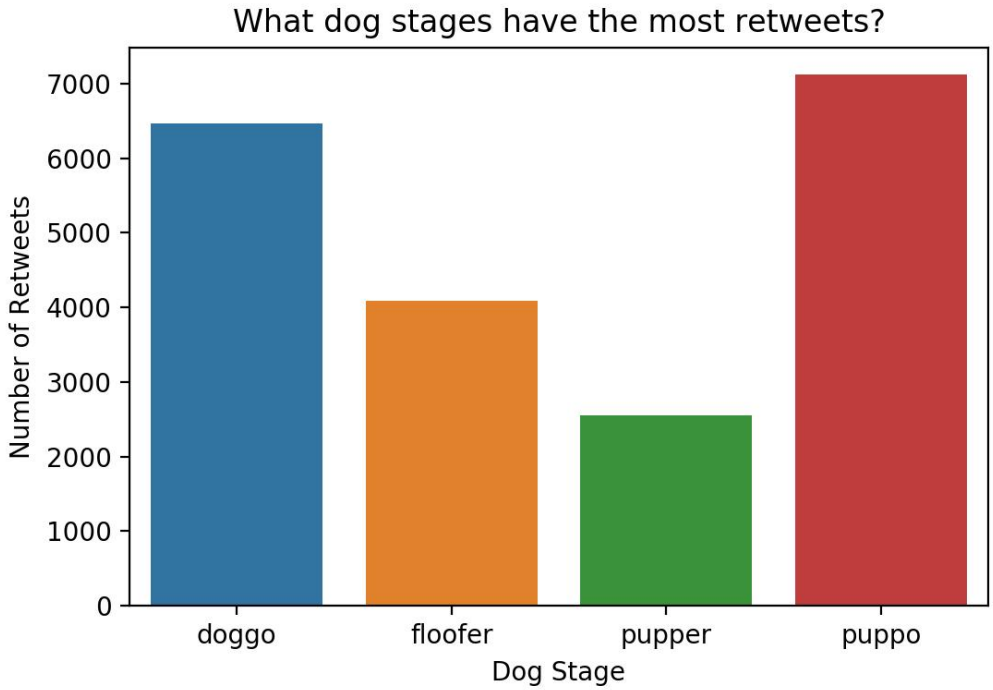
To answer question 1, I had to plot a vertical bar chart using the barplot() method of seaborn. The chart below shows the ranking of the dog stages.



From the figure above, the puppo stage is the most favorited/liked of the four dog stages.

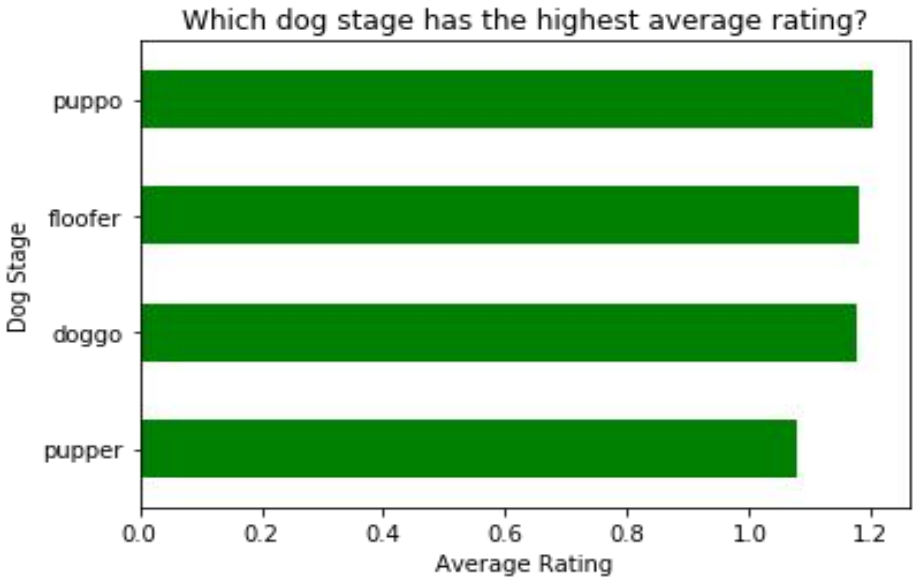
2. What is the dog stage with the highest retweets?

A similar approach was used to answer this question as question 1, that is, using a vertical bar chart. The chart below shows that dogs in the "puppo" dog stage had the most retweets:



3. Which dog stage has the highest average rating?

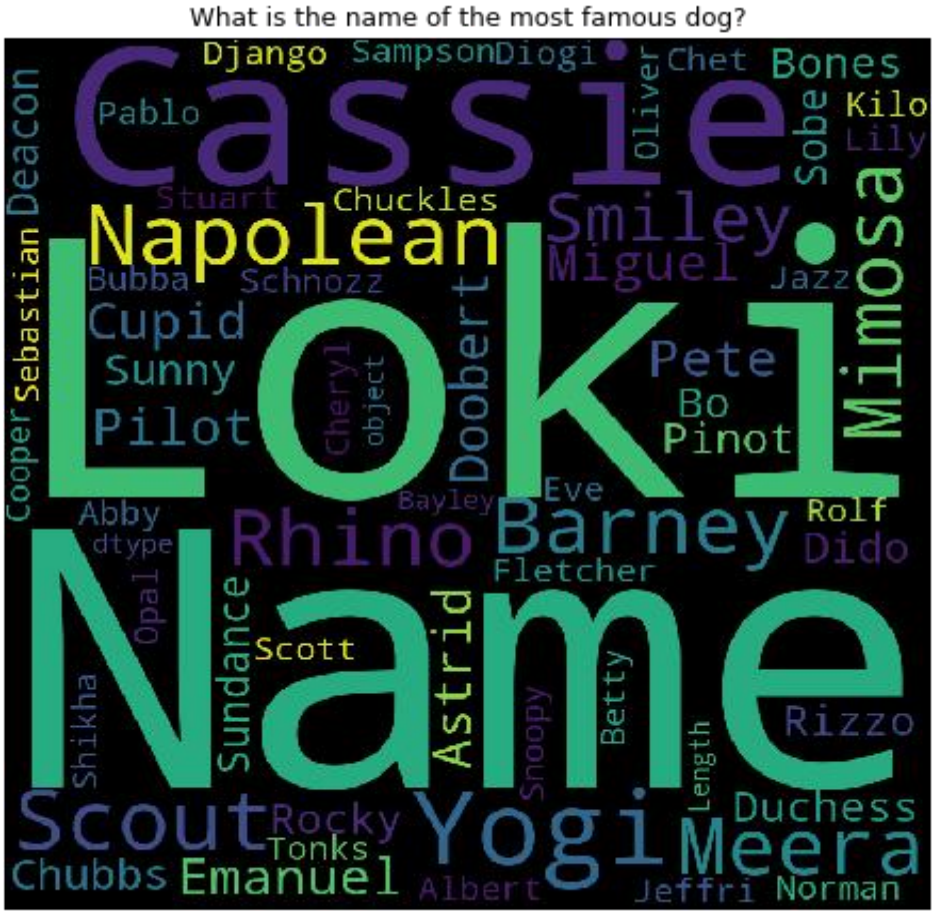
To answer this question, I had to create a new column named "rating" by dividing the "rating_numerator" column by the "rating_denominator" column. After this I had to find the average rating of each dog stage using the pandas mean() function before plotting on a horizontal barchart as seen below:



The chart showed that the dogs "puppo" in the stage had the highest average rating, followed closely by "floofer" stage and "doggo" stage before the "puppers."

4. What is the name of the most famous dog?

Here a wordcloud was used to identify the name of the most famous dog.



There were so many entries with invalid names like “a”, “an”, “by”, which had to be replaced with NaN. The final result is not clean as it showed “Loki” to be the most famous dog name, followed by “Name”.

LIMITATION

In collapsing the four columns “doggo, floofer, pupper, and puppo” to one, my dataset reduced to only 347 entries, meaning that more than 1000 entries had no classification, hence, were not considered.