act_report

February 10, 2019

0.1 PROJECT: WRANGLE AND ANALYZE DATA

BY: SAURABH KULKARNI

0.2 Insights and Visualization

This document presents insights and visualizations created after data wrangling was completed
The following questions were explored and visualized:

1) Which dog stage is most popular in terms of favorite and retweet counts?

2) Which breed is most popular in terms of favorite and retweet counts?

3) Which breed is the most highly rated?

Master CSV tweet file from cleaning was used for analysis

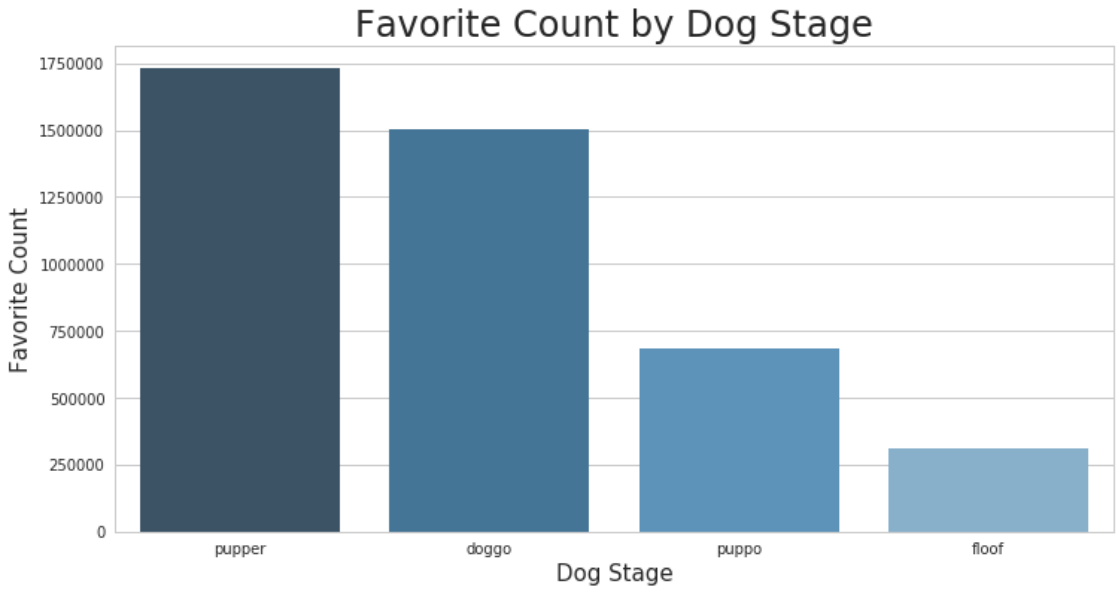
0.2.1 Dog stage popularity: Which dog stage is more popular?

To answer this questions the following steps were executed:

- Data was first grouped by dog stage
- Favorite count was calculated for each dog stage
- A series with each breed and favorite counts was obtained
- Since we are interested in most popular breed, we sorted values in descending order
- The top five popular breeds from this series was then plotted using Seaborn bar plot

The plot is shown below.

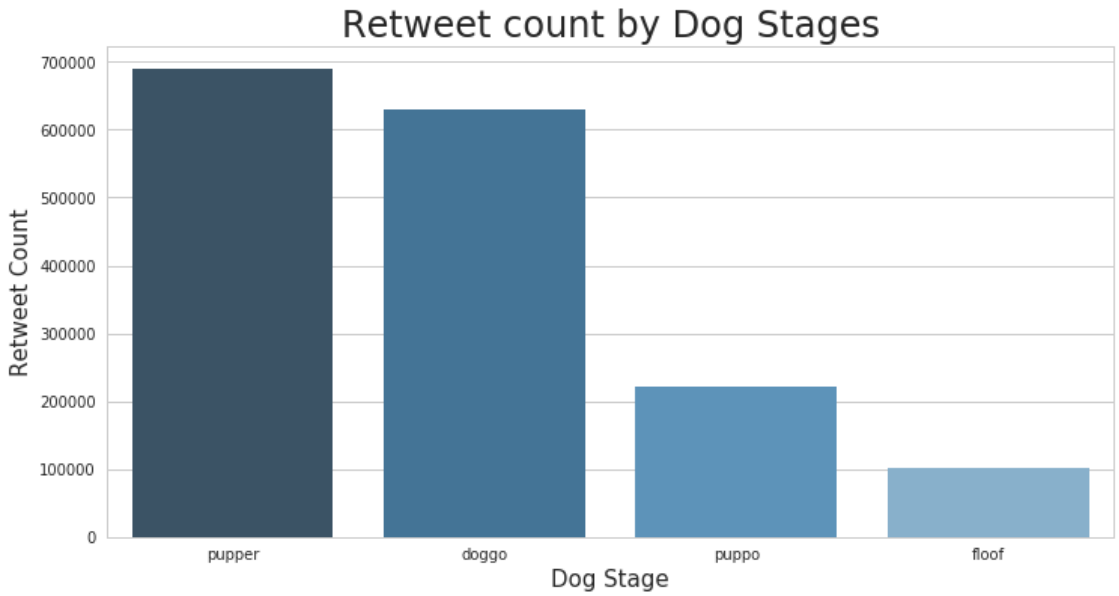
In [76]:



The chart above shows that pupper is the most popular dog stage closely followed by doggo.

To explore further, the same process was repeated using retweet counts. The bar plot is shown below.

In [80]:



The charts above show that pupper dog stage has been more popular in terms of retweets and favorite counts closely followed by doggo. Floof got the lowest retweet and favorite counts.

0.2.2 Which breed is the most popular?

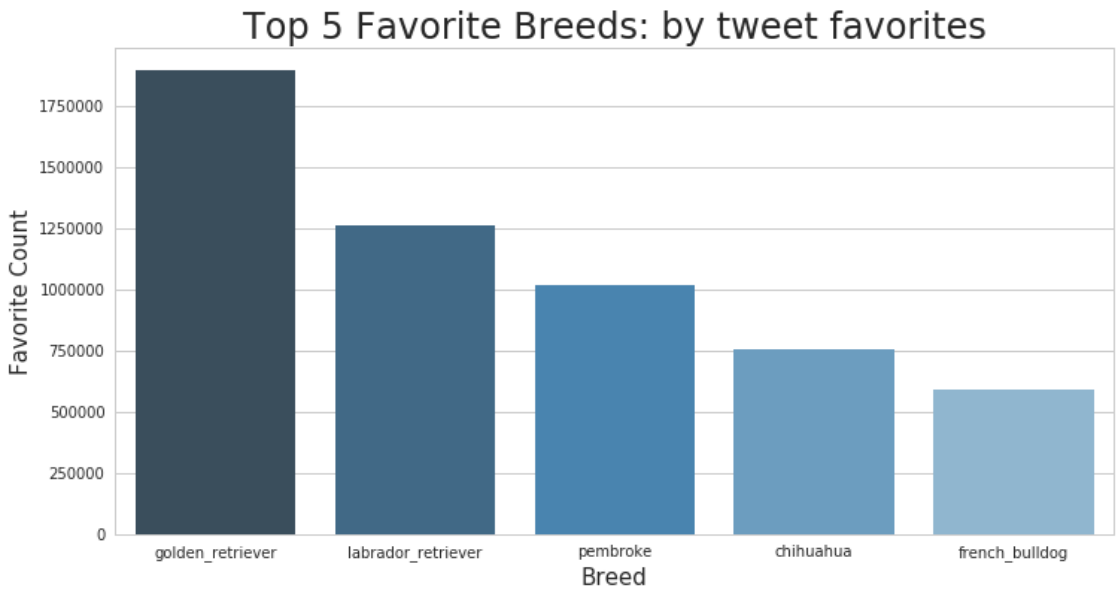
To answer this question, we had to bridge image prediction dataframe with tweet data. Image prediction data file has three predictions for each tweet. There are three challenges with this data:

1) Three predictions exist with each having confidence levels. We need to select the highest confidence.

2) Not all predictions are dogs, but there is a column for each prediction that tells us whether or not it is a dog

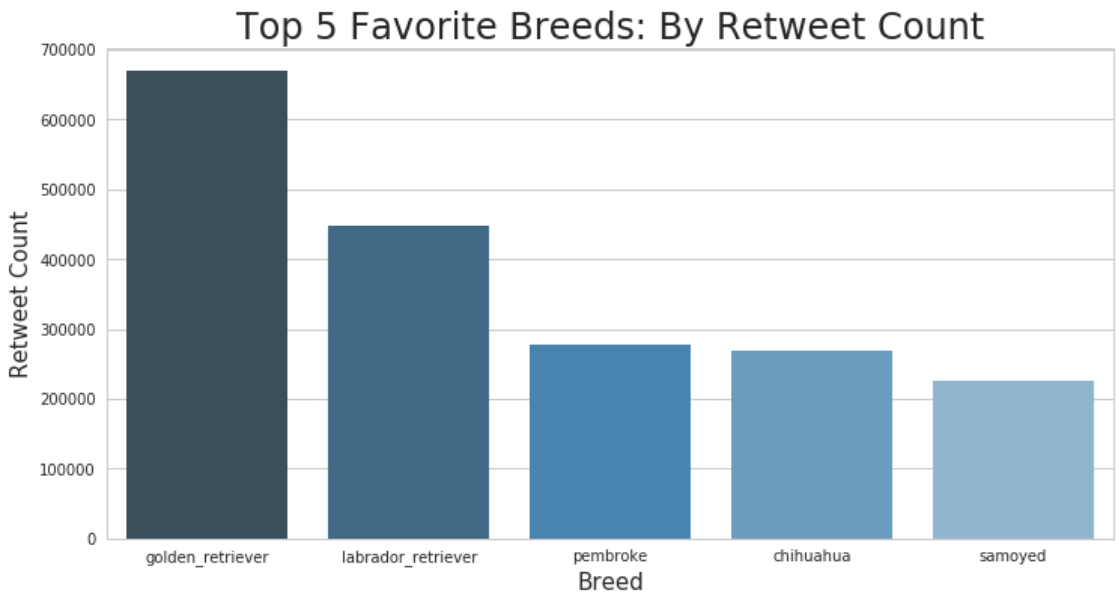
3) All this data are in separate columns, we need to create a function that will identify which breed is the right one for each tweet id (if breed information exists).

In [189]:



The plot above shows that Golden Retriever was the most popular breed in in terms of favorite count.

In [194]:



The charts above show that Golden Retriever is the most popular breed according to retweet and favorite counts.

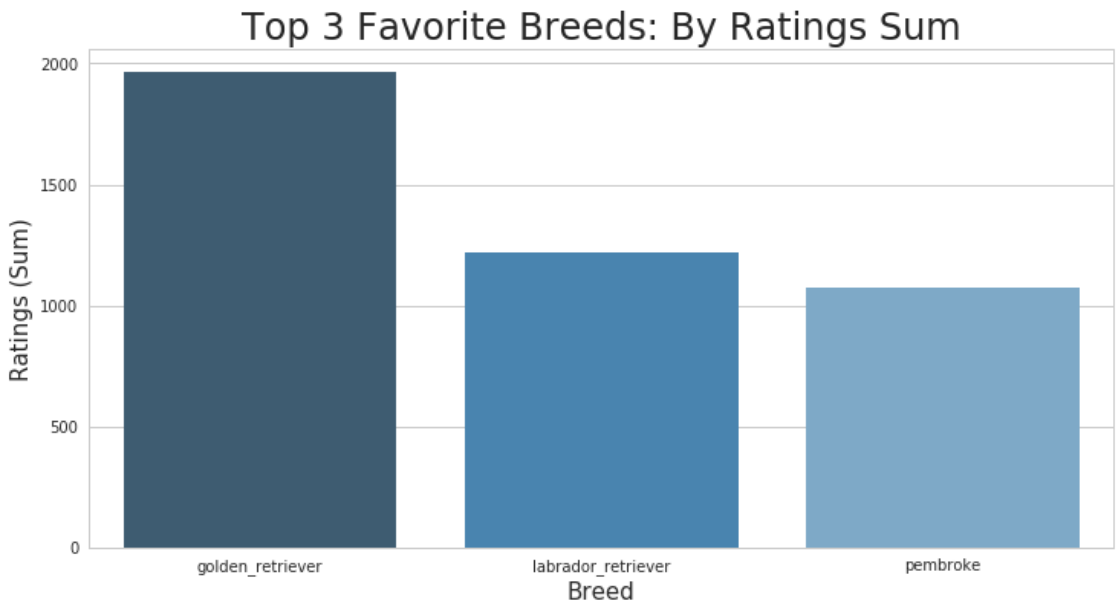
0.2.3 Which breed got the highest ratings?

Here we want to know which breed got the best ratings. We can use the same dataframe df_breed_tweets since it has all the data we need. We will use a bar chart to show the top three breeds with highest ratings.

SCALE OF RATINGS

- When we were cleaning data, we found that there were some ratings that were not given out of 10 (like 87/90).
- We will ignore these ratings for this analysis.
- To do that, we will filter out only those rows which have denominator "10".
- We will then calculate the ratings for every breed and determine which top 3 breeds got the highest rating

In [244]:



According to the chart above, Golden Retriever is the most popular breed.It got the highest ratings as compared to other breeds.