Report: act_report

Visualization and Insights

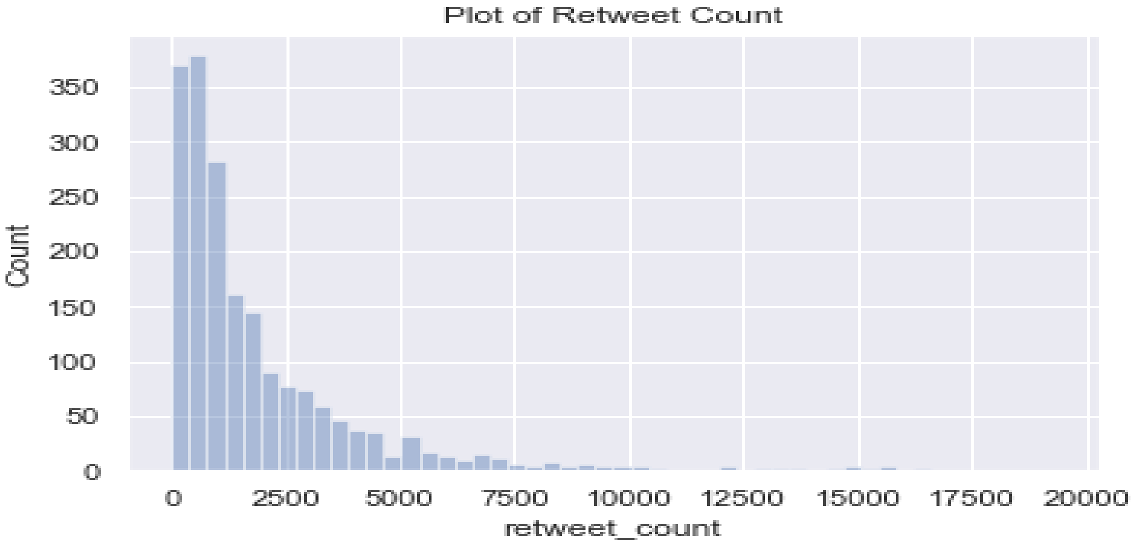
- After the cleaning of the three datasets, I merged them together into a dataframe and named it twitter_archive_master. This was done to make analysis easier.

- The first thing I did before performing an exploratory data analysis was to create a function named “display” that will create a histogram exploring the count of dogs with a specific amount of retweet_count, favoroite_count and rating.

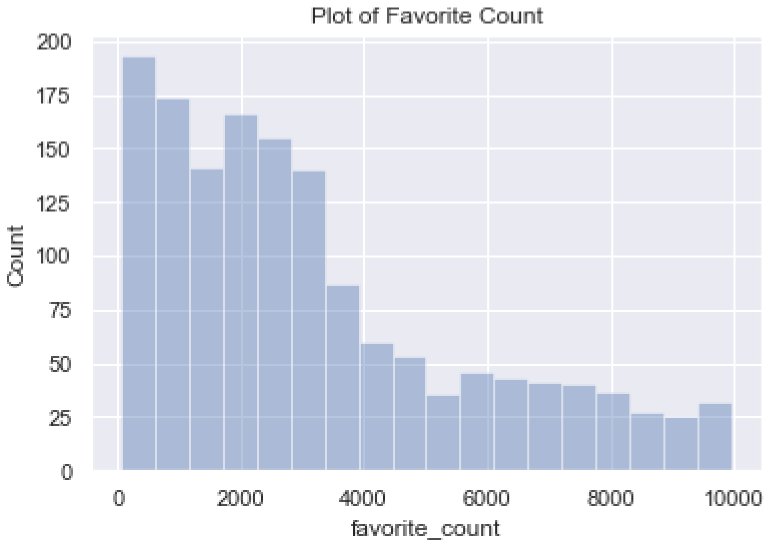
- This function included two python libraries for visualization which are matplotlib and seaborn. I used seaborn’s distplot function so that it will create a multifunctional histogram when called. Matplotlib’s xlabel, ylabel, title and show function was also used. I called the sns.set() function so that the plots created are the default background.

- I then did a summary statistics of all numerical values in the twitter_archive_master dataframe to see the interquartile range of the values of each numerical column and also to check if there are any huge outliers that will affect our analysis and plots. Turns out there were outliers.

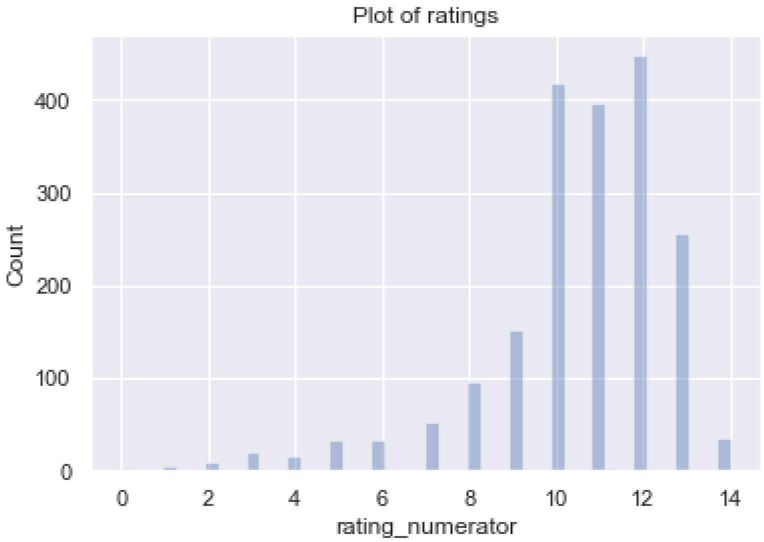
- The first plot I created by calling the display function was used to create a histogram that will display the distribution of retweet count. The arguments that were passed into the function are a subset of the retweet_count column and the title of the plot. The retweet_count column was subsetted to include values from the 0th quantile to the 75th quantile.



- The second plot I created by calling the display function was used to create a histogram that will display the distribution of favorite count. The arguments that were passed into the function are a subset of the favorite_count column and the title of the plot. The favorite_count column was subsetted to include values from the 0th quantile to the 75th quantile.

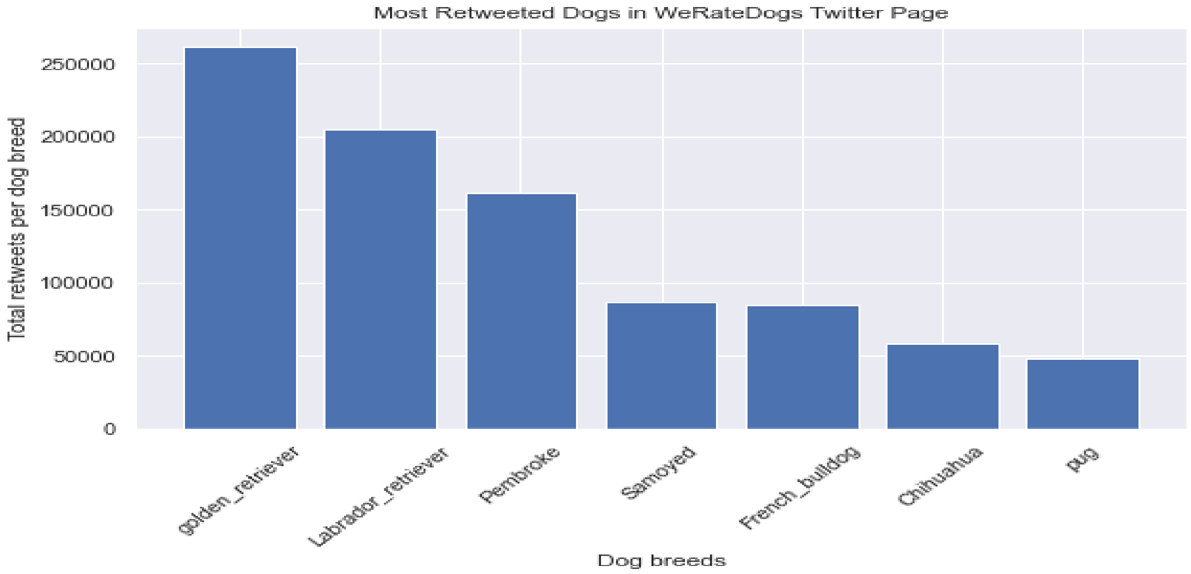


- The third plot I created by calling the display function was used to create a histogram that will display the distribution of rating of dogs. The arguments that were passed into the function are a subset of the rating_numerator column and the title of the plot. The rating_numerator column was subsetted to include values from the 0th quantile to the 75th quantile.

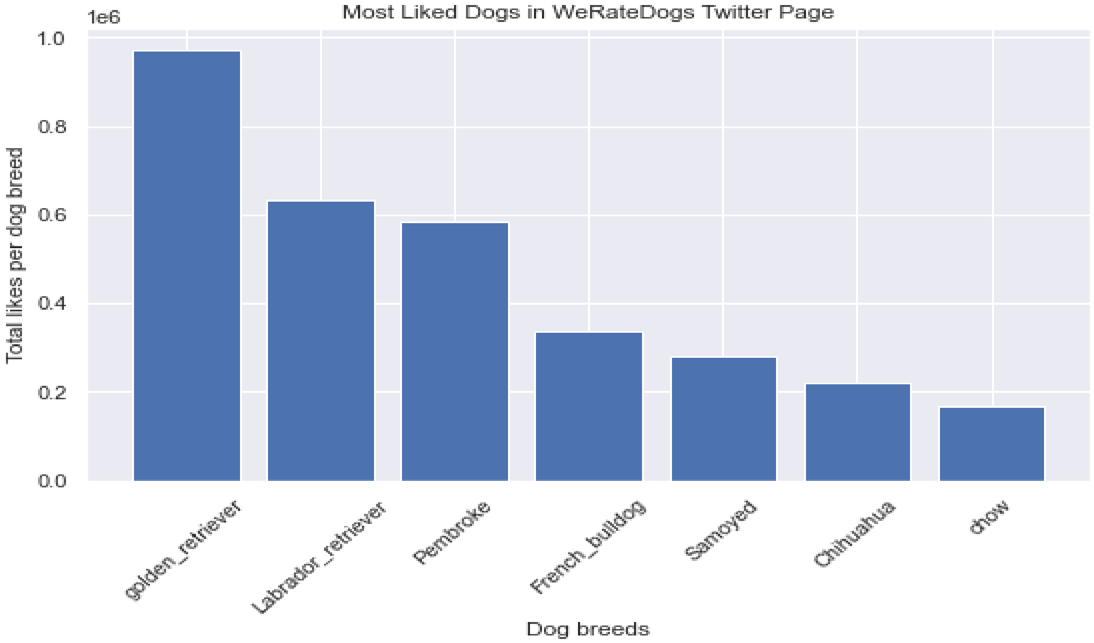


- The distribution of the three histograms created showed that tweets with an average number of total retweet_count and favoritre_count in the dataframe were more than tweets that had high retweets meaning that a specific breed of dogs have different effects on the number of retweets and likes. Also the most popular ratings are 12/10 followed by 10/10 and 11/10.

- After creating histograms, I then proceeded to create a barplot of the 7 most retweeted dog tweets and what breed of dogs were found in those posts. These dog breeds were Golden retrievers, Labrador retriever, Pembroke, French Bulldogs, Samoyed, Chihuahua and Pugs.



- I also created a barplot of the 7 most liked or favorite dog tweets and what breed of dogs were found in those tweets. These dog breeds were Golden retrievers, Labrador retriever, Pembroke, Samoyed, French Bulldogs, Chihuahua and Chows.


1e6
Most Liked Dogs in WeRateDogs Twitter Page
Total likes per dog breed
golden_retriever
Labrador_retriever
Pembroke
French_bulldog
Samoyed
Chihuahua
chow
Dog breeds