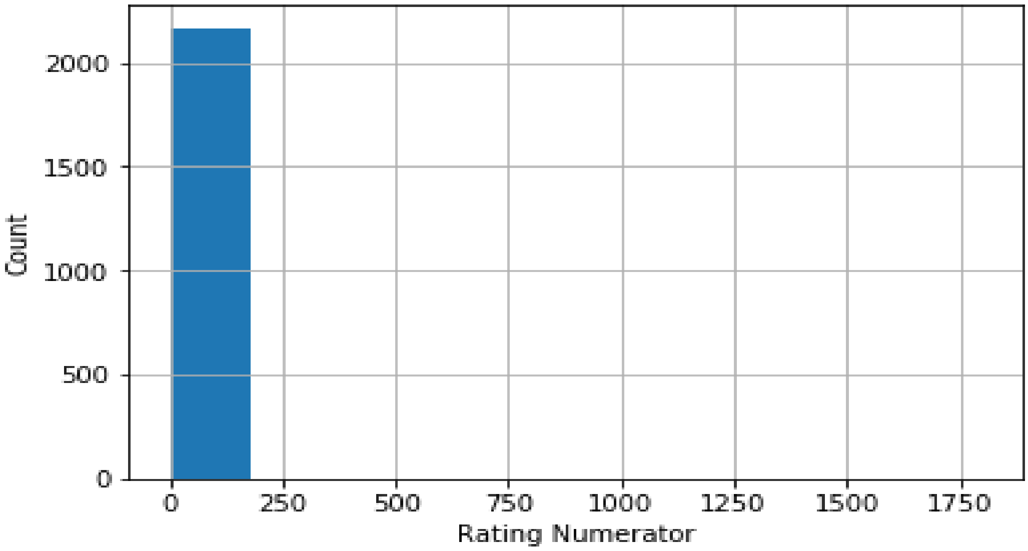
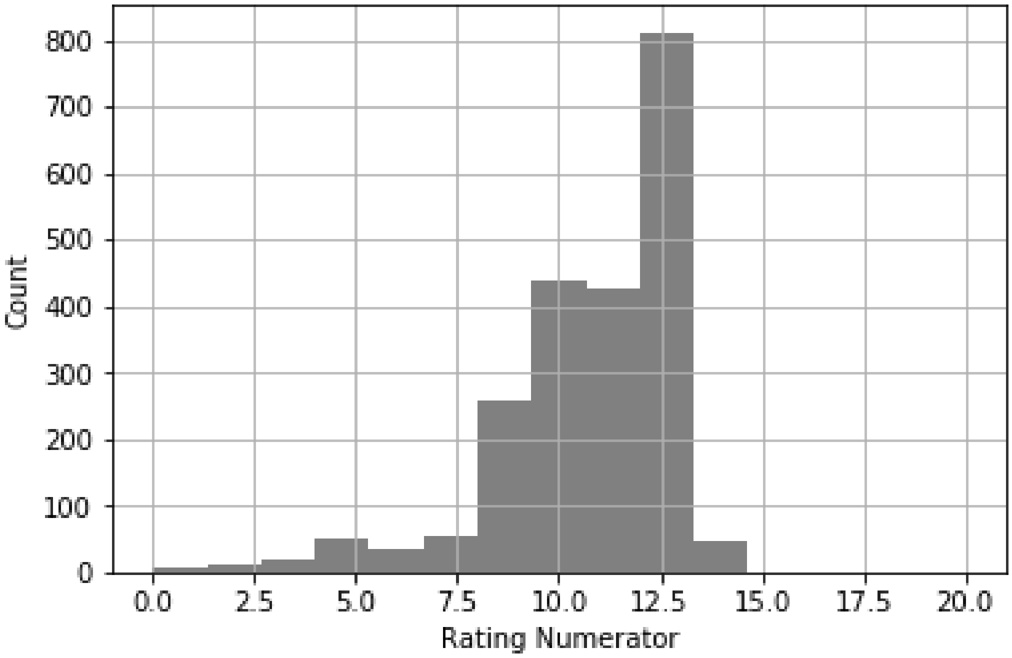
Report

Rating Numerator

Rating numerator is the most important part of the whole dataset since it gives a score to the dogs. While plotting a histogram with a range of the minimum and maximum value of variable rating numerator as observed in it's describe function, we get the following plot.



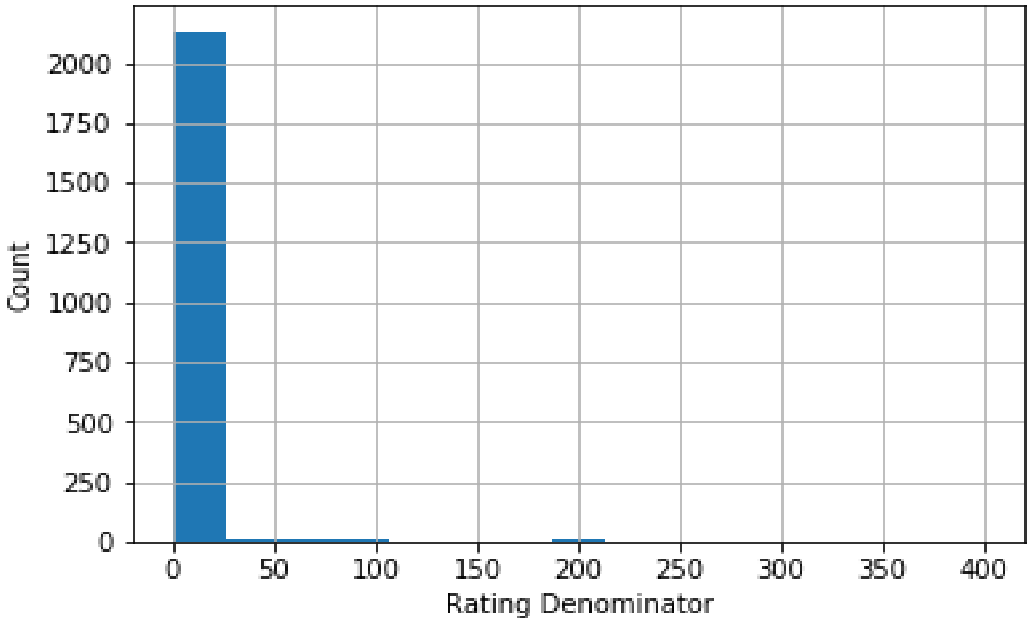
All we see is a single bar. This indicates that there lies some outliers in our dataset. When we observe the describe function, we see that the 75% quantile value of rating numerator is 12. So by limiting the range from the minimum value to 20, we get the following plot.



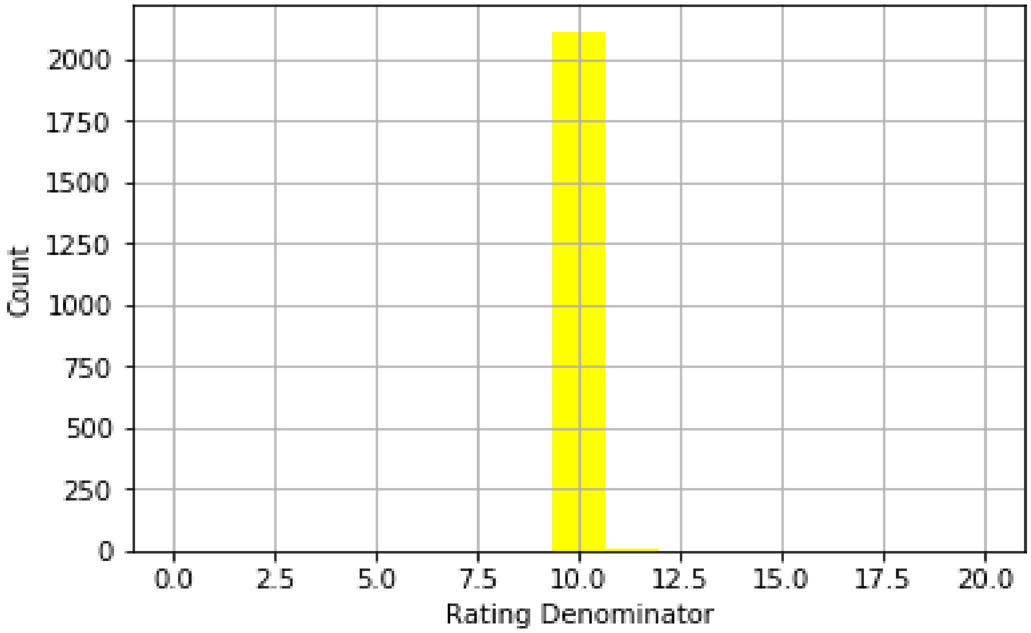
This plot shows more of a normalized plot and we can see that most values lies in between 12 and 13.

Rating Denominator

Since the rating of dogs in the twitter account 'WeRateDogs' is bit different than the conventional methods, it becomes really necessary to find more about the rating denominator. We use the same method as used before. With the help of describe function we find the minimum and maximum value of the rating denominator and then plot a histogram keeping those values as the range. We get the following plot.



What we observe is similar to that of the initial plot of rating numerator which indicates that there are outliers present in the given dataset. So we do the same thing as before and limit the range to 20.

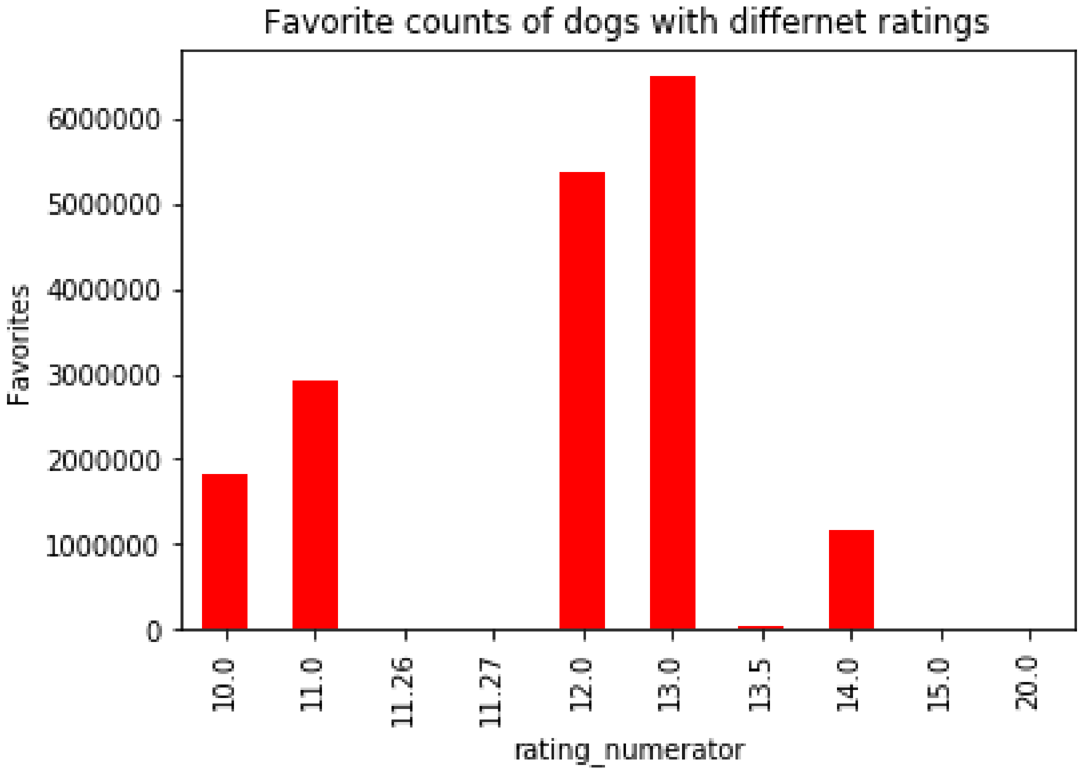


We can see that the common value of rating denominator is 10.

Most Common Rating: rating_numerator and favorite_count

For further analysis, we need a standard rating for the dogs. As we observed previously, there are outliers in the given dataset. So, if we could find the most common rating provided by 'WeRateDogs' we may continue our analysis efficiently.

We started by limiting the number of entries using where function. Initially, we cut off the entries that has a rating numerator values less than its 25% quantile that is 10. Then we cut off the entries that had a rating numerator values greater than 20. Then we grouped them along with sum of favorite counts of each and every rating numerator value in between the range 10 and 20. The following bar chart is what we got at the end.



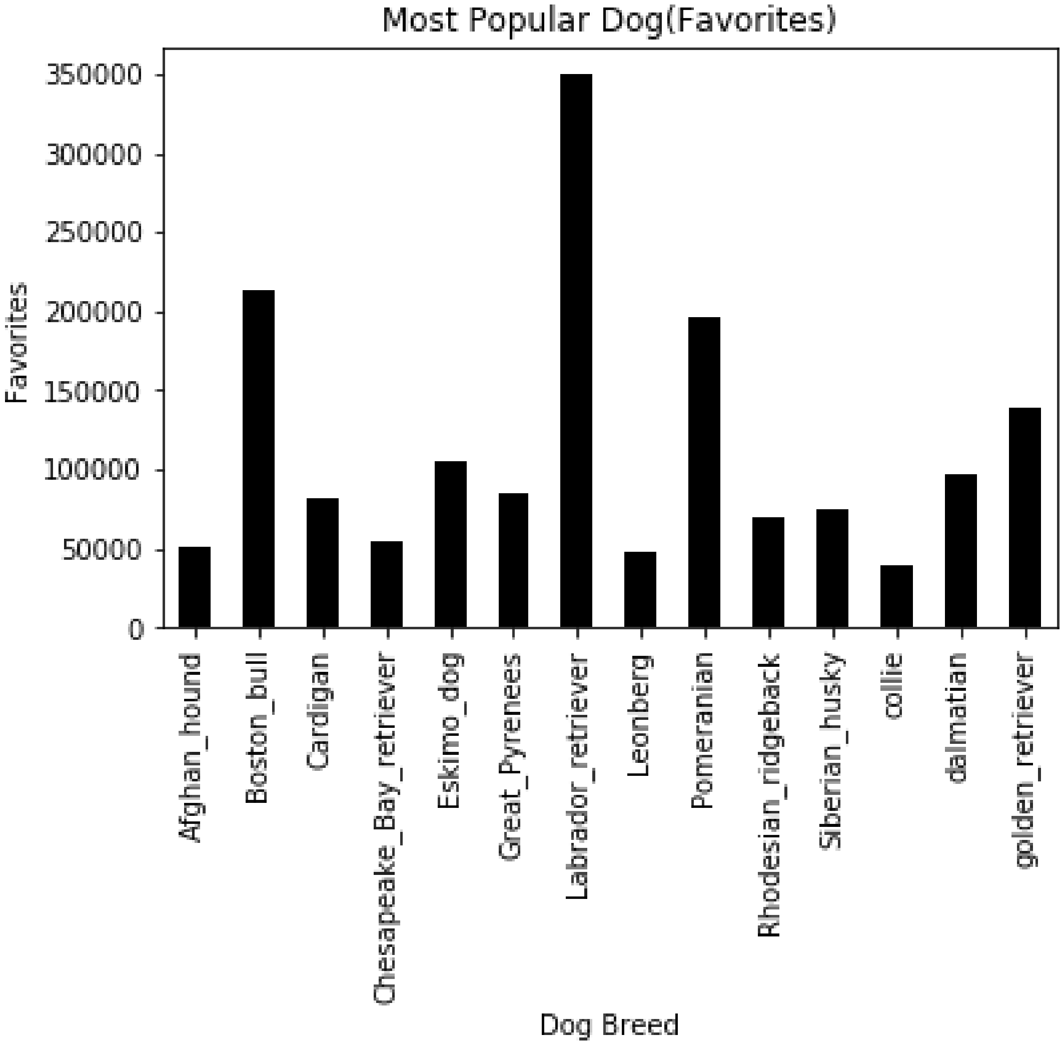
From the plot we can say that 12 and 13 are the most common ratings provided by 'WeRateDogs'. If we have to use a standardized value, we could use 13.

Most Popular Dog Breed: rating_numerator, p_dog, p and favorite_count

Using the number of favorite counts for different dogs we could find the most popular dog breed.

To do so, we started with creating various filtered data frames. We started with cutting off the entries in dataset which didn't have a rating numerator of 13. Then, we cut off those entries whose prediction was false.

With the help of describe function, we found the 75% quantile value of the favorite count of those remaining values to be 28497. Then we dropped the entries whose favorite count value was less than 28497. Then again with the help of describe function, we found the 50% quantile value of the favorite count to be 53327. We performed the same operation as before and dropped some entries and finally we grouped the predictions of dog breed (p) with the sum of favorite counts. The following plot is what we got.



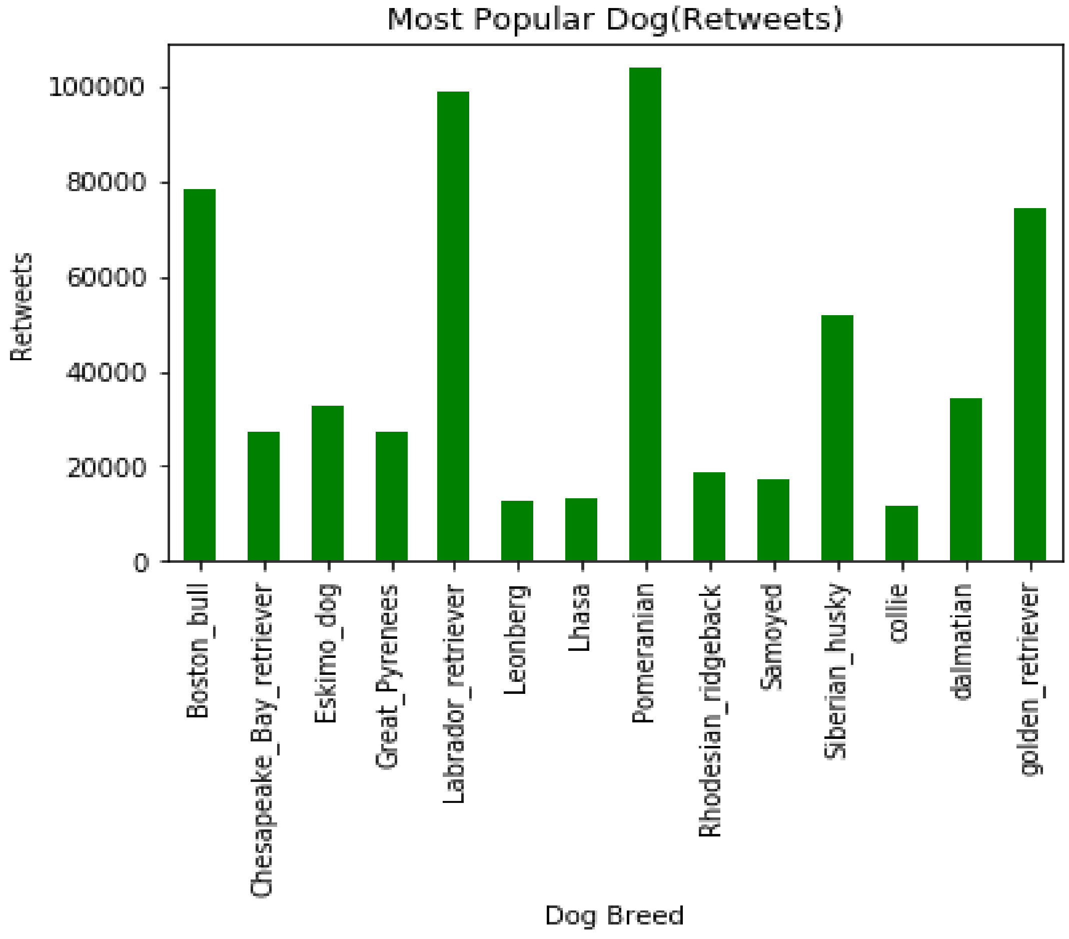
Based on the favorite count, Labrador retriever is clearly the most popular dog which has more the 350,000 favorites. It is followed by Boston bull, Pomeranian and Golden retriever.

Most Popular Dog Breed II: rating_numerator, p_dog, p and retweet_count

Using the number of retweet counts for different dogs we could find the most popular dog breed among them.

We started with cutting off the entries in dataset which didn't have a rating numerator of 13. Then, we cut off those entries whose prediction was false.

By applying describe function on retweet count, we found the 75% quantile value to be 7695. We dropped the entries whose retweet count value was less than 7695. Again with the help of describe function, we found that the 50% quantile value of the retweet count is 11757. We performed the same operation as before and dropped some entries and finally we grouped the predictions of dog breed (p) with the sum of retweet counts. The following plot is what we got.



The results were not exactly what was expected. Pomeranian which took the third place in terms of favorite counts took the position of most popular dog when it comes to retweets. It has a retweet count greater than a 100,000. It is closely followed by the most popular dog in terms of favorites 'Labrador Retriever'. They are followed by Boston bull and Golden retriever.