Data Analysis

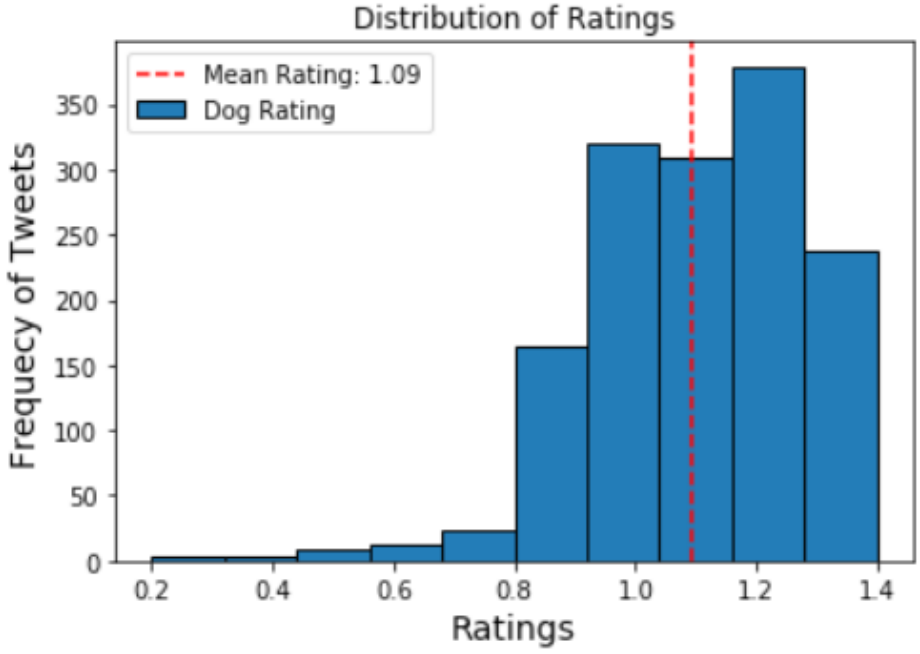
After wrangling the data and saving to twitter_archive_master.txt, it is time to get some insights and trends. The questions we want to ask and get answers of are:

- How are the ratings distributed?
- Which dog breeds are most recurring and see the top 10
- How do retweets and favorites fare against ratings
- Which of the top 10 breeds got how many retweets and favorites

In order to explore and answer above questions, we do not need all the columns from master dataframe. We only need following columns:

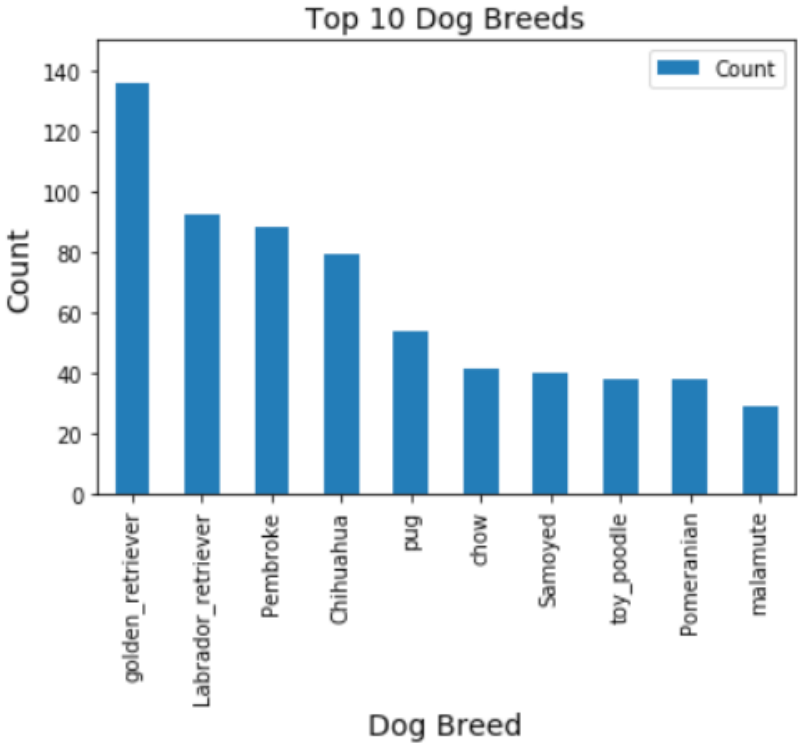
- 'tweet_id'
- 'timestamp'
- 'rating_numerator'
- 'rating_denominator'
- 'retweets','favorites'
- 'dog_stage'
- 'p1' (prediction of dog: e.g. Golden Retriever)
- 'p1_conf' (confidence level of algorithm)
- 'p1_dog' (is it really a dog?)

Once the basic efforts are done, we plot the histogram of ratings:

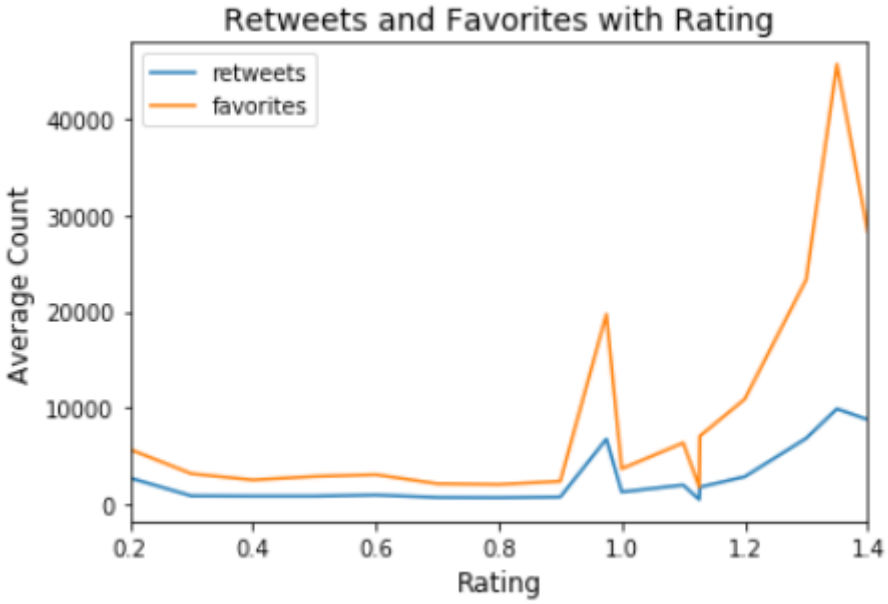


As we know the ratings are usually given out of 10 but one of the reason for popularity of this twitter handle is because the numerator in the ratings is higher than denominator. This is indeed reflected in the mean rating of 1.09.

Let’s see top 10 breeds in our dataset. These top 10 breeds are not necessarily most popular in terms of ratings. One of the many factors could be sheer number of submissions. On top of that, dogs like Golden Retriever are part of many households compared to other breeds.

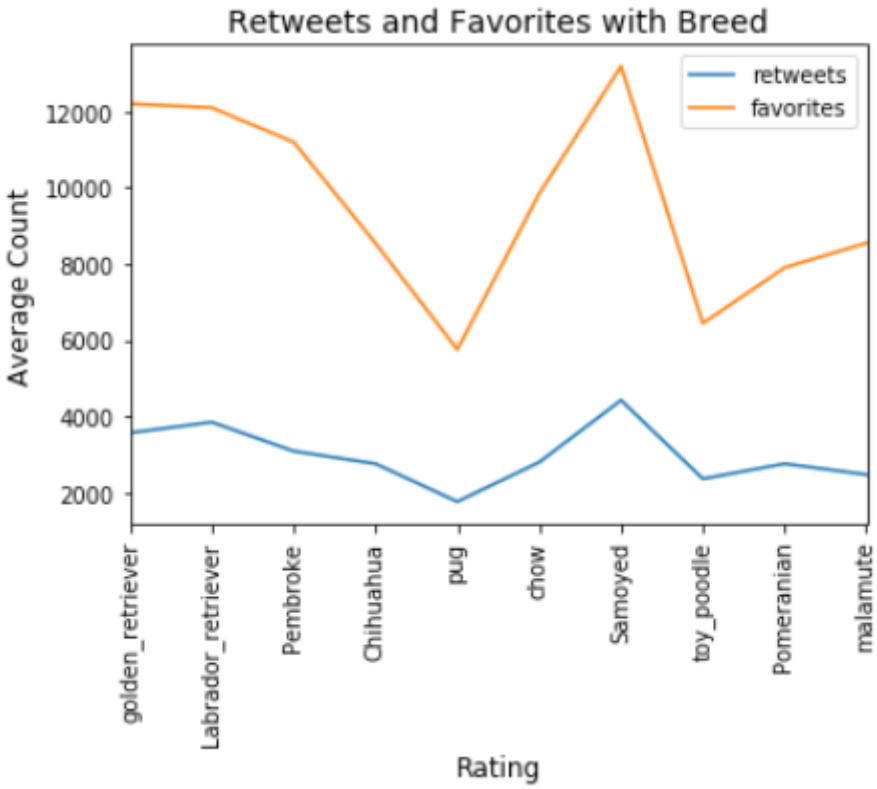


How do ratings fare against retweets and favorites?



We can see here, as the rating increases the popularity of the tweet increases too. Although, we do see a sharp spike at around 0.9 rating.

Which of the top 10 breeds got how many retweets and favorites?



We know that Golden Retriever is highly popular and most submitted dog but Samoyed is giving a tough competition to all the other breeds despite being 7th most popular one. Pug being 5th most rated but it seems that Pug is not a people’s dog! As well as toy puddle.