WeRateDogs Twitter Archive – Analysis and Visualization

Introduction

The WeRateDogs Enhanced Twitter feed contains approximately 2.3K tweets of the 5K tweets (as of the time the data was released) posted by the @dog_rates twitter account between 11/2015 and 08/2017. This initial dataset contains basic information like tweet ID, timestamp, rating, name etc.

However, two very important columns – retweet count and favorite count were not available directly. This was downloaded from the Twitter API directly using the 'tweepy' library.

Along with these two data pieces, there was an additional data piece called Image Predictions that was generated using a neural network algorithm to classify the dogs in the picture into different breeds.

Together, these three data sets formed the basis for my analysis.

Data Wrangling

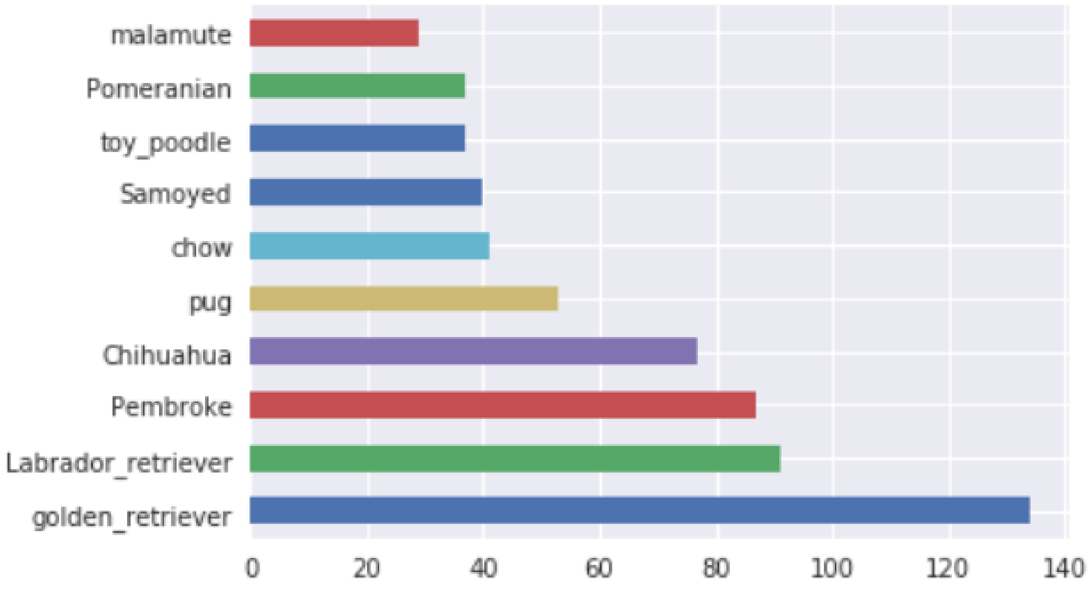
As with any data analysis project, the first step is gathering the data which is explained above. Once I had all the data pieces, I spent a good amount of time reviewing the data pieces visually and programmatically for data quality and tidiness issues.

This was followed by cleaning the issues and merging the tables together to create a 'clean' final table. This final table had around 2K tweets, which are the source of the insights below.

Insights

There was a lot of opportunity to generate multiple insights and graphs here but the three insights that I generated are given below.

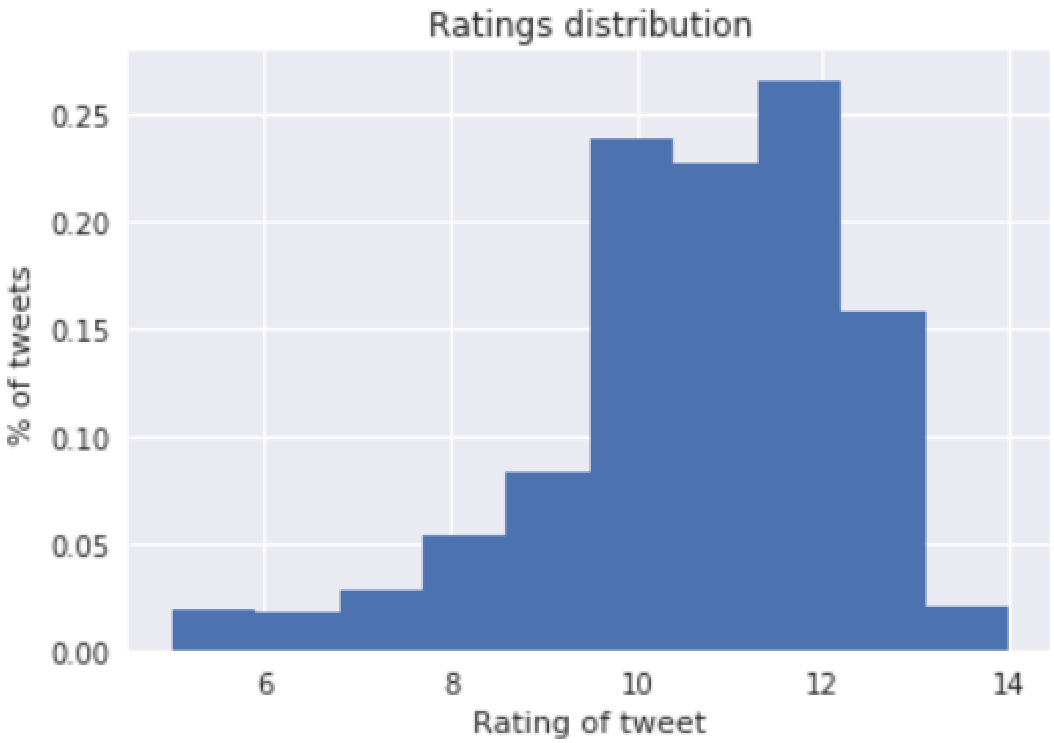
<u>Insight 1 – Retrievers are the top two tweeted dogs</u>



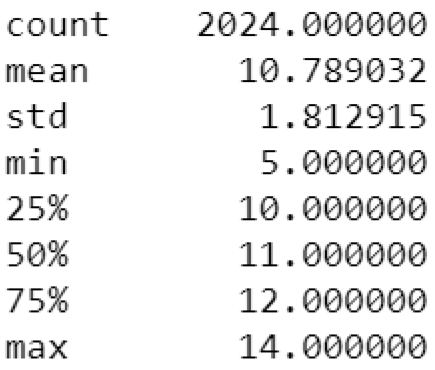
As you can see in the image above, Golden retrievers and Labrador retrievers are the top two dogs with the most number of tweets. In fact, these are the top two breeds in the Americana Kettle Dog website for 2019 as well! Playful dogs sure do get people's attention!

Insight 2 – Most ratings are concentrated between 10 and 12

In our final dataset, we restricted the denominator to be always 10, which is 99% of the records. Looking at the distribution of the numerator here, we can see that most ratings are between 9 and 12.

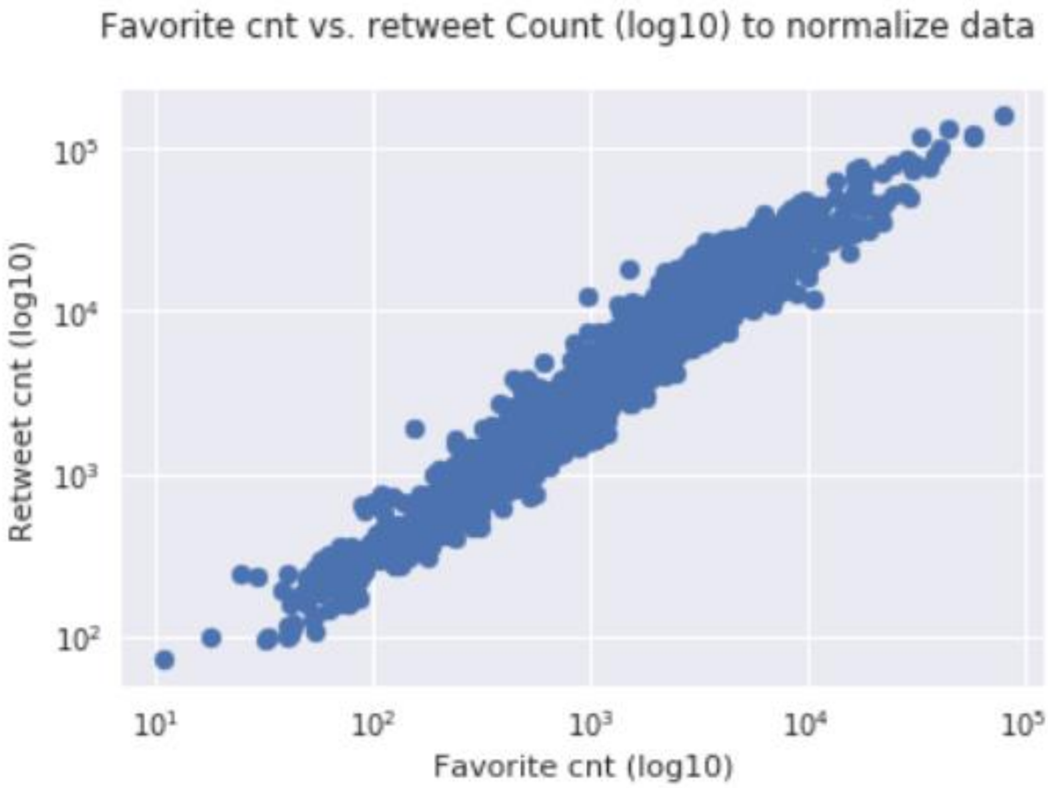


I also used the 'describe' method to get the below table which aligns with our chart above.



```
count    2024.000000
mean       10.789032
std         1.812915
min         5.000000
25%        10.000000
50%        11.000000
75%        12.000000
max        14.000000
```

Insight 3 – There is a heavy positive correlation between retweets and favorites, which is not too surprising



Tweets that are heavily favorited tend to be retweeted heavily and vice versa. This is not very surprising since only interesting tweets are both retweeted and favorited. It was also interesting to see that the most retweeted tweet is also the most favorited tweet. Link below for the curious!

https://twitter.com/dog_rates/status/744234799360020481