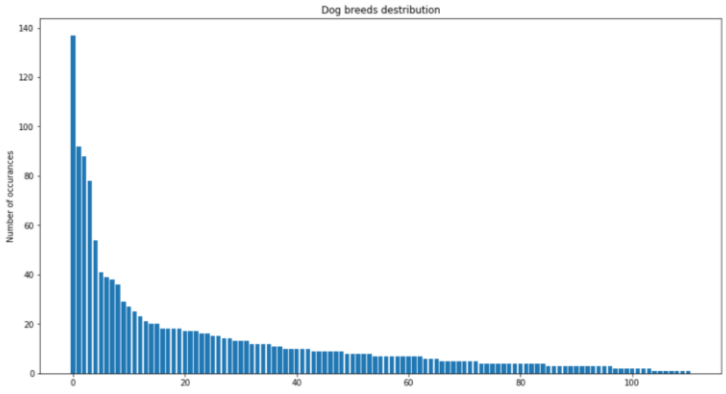
ANALYSIS AND VISUALIZATION

A report

This dataset can be a lot of fun to play with, a very basic analysis can yield very interesting results. For starters one may wonder what is the most common dog breed tweeted about. The answer seems to be th e Golden retriever with 137 occurrences, significantly higher than the mean value of 13 occurrences. The distribution of the number of occurrences for each breed is shown in the following figure, though it doesn't give us exact values, it helps us paint a better picture of the distribution.



Another question to consider is which dog breed got the highest ranking. This question might not reflect deep insights about the data or user behavior due to the special ranking system, but it is fun to explore nonetheless. The winner is the 'clumber'.

There is another interesting aspect of the data we can look at, which is the tweet time. It might seem reasonable to assume that there is no pattern as to when the tweets are written, after all, there are no restrictions at all. But the data shows otherwise. In the following figure we see the tweet-hour on the left, and the number of times this tweet-hour was shown in the data. We can clearly see that it is far from random. The user of the account seems to have a tendency to tweet after midnight or in the afternoon and almost none at all in the morning.

```
In [93]: time_counts

Out[93]: 1     275
         0     248
         2     209
         16    189
         3     164
         17    157
         23    108
         18     98
         4      95
         15     90
         19     81
         20     75
         21     69
         22     67
         5      26
         14      8
         6       3
         13      1
         Name: timestamp, dtype: int64
```

We can also look at and fairly basic question is ‘does the number of retweets increase the favorite count on average?’ the intuitive answer is yes. The following plot does show that, but the huge jump suggests a threshold, after which the number of favorites and increase at a much faster rate, or a tweet goes ‘viral’ if you will. With better analysis this threshold can be determined.