EXPLORING DATA ABOUT DOG RATINGS

Analyzing tweets of @dog_rates Twitter account

Luan Fernandes

Exploratory Data Analysis of @dog_rates tweets

Brainstorm

After the gathering and cleaning of the dataset about dogs, contained in tweets of @dog_rates Twitter account, we are ready to explore the data for insights. Each tweet in the final dataframe refers to a dog (804 different tweets), and contains its name, rating, timestamp, of the stweet, text of the tweet, an url for the dog image, numbers of favorites and retweets, dog classification (as defined by the page owner) and the most likely race of the dog.

Based on these information, 4 questions were formulated to extract new insights of the dataset, and the results are shown below.

HOW MANY DOG RACES DO WE HAVE IN OUR TABLE? LIST THE TOP 5 RACES THAT GOT MORE FAVORITES AND THE TOP 5 FOR THOSE THAT GOT MORE RETWEETS

To answer that, we use the “groupby()” and “nlargest()”methods, and we see that the golden retriever and the labrador retriever are the most favorited dogs.

```
In [57]: twitter_archive_master.groupby('dog_races').agg('count')['tweet_id'].nlargest(5)

Out[57]: dog_races
         golden_retriever      57
         labrador_retriever    50
         pembroke              38
         chihuahua             30
         chow                  20
         Name: tweet_id, dtype: int64
```

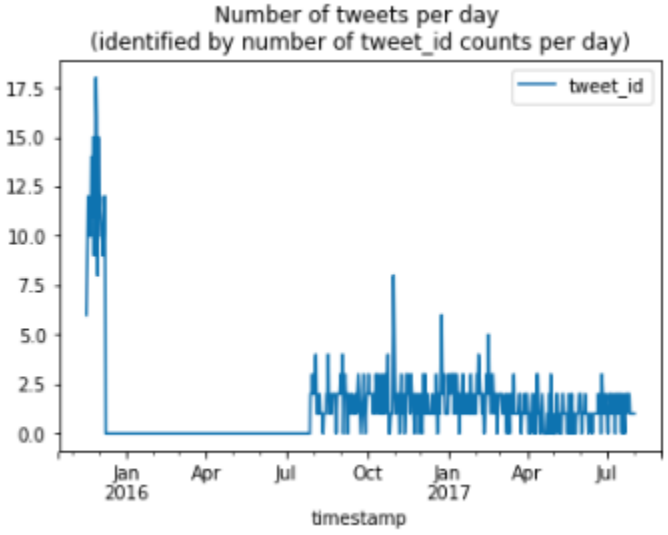
ARE THERE MANY DUPLICATES FOR NAMES? WHAT'S THE MOST COMMON NAME?

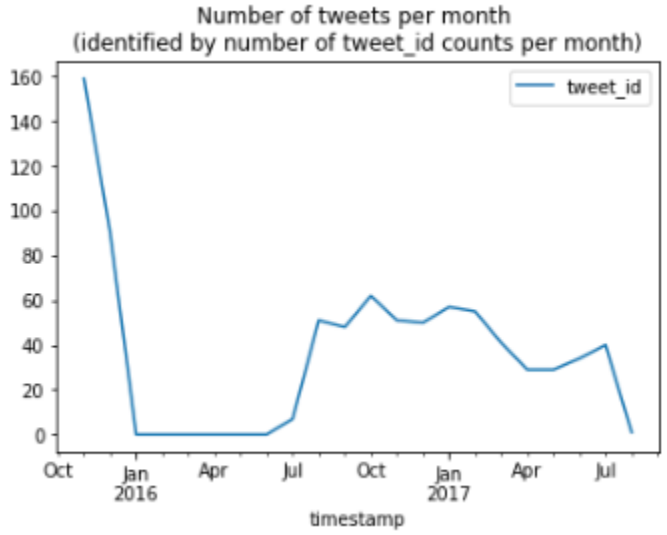
In the image below, we see that we have 212 duplicated names, which represent a large number compared to 804 tweets in our dataframe, while the most common names are Penny and Tucker.

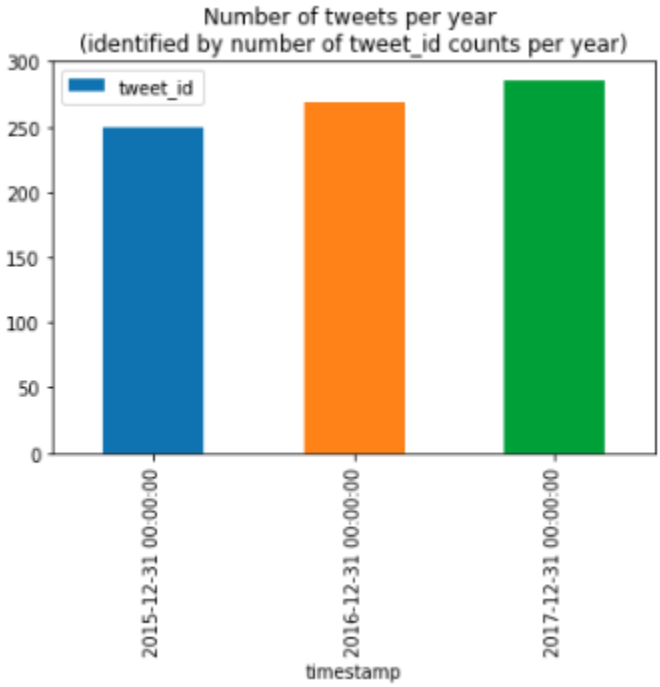
```
In [61]: #twitter_archive_master.groupby('name').name.count()
         twitter_archive_master[twitter_archive_master.name.duplicated()]['name'].count()

Out[61]: 212

In [62]: twitter_archive_master.groupby('name').agg('count')['tweet_id'].nlargest(5)

Out[62]: name
         Penny      7
         Tucker     7
         Charlie    6
         Daisy      6
         Bo         5
         Name: tweet_id, dtype: int64
```

HOW THESE TWEETS WERE POSTED THROUGH TIME? PLOT NUMBER OF TWEETS PER DAY, PER MONTH AND PER YEAR;

The following pictures show the distribution of tweet posting for different time series. For the daily series, the pattern is a bit confusing, but when looking at monthly series, we see that the account had a high frequency of posting in the begging of the time we can evaluate, in October 2015. However, the first semester of 2016 had zero posting, and the following months reestablished a lower frequency compared to 2015, even though the latter graph shows that the accumulated number of tweets did not change much in 3 years.


Number of tweets per day
(identified by number of tweet_id counts per day)
tweet_id
17.5
15.0
12.5
10.0
7.5
5.0
2.5
0.0
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
Jul
timestamp


Number of tweets per month
(identified by number of tweet_id counts per month)
tweet_id
160
140
120
100
80
60
40
20
0
Oct
Jan
2016
Apr
Jul
Oct
Jan
2017
Apr
Jul
timestamp



WHAT'S THE DISTRIBUTION OF `RATING_NUMERATOR` (MIN AND MAX VALUES, AVERAGE ETC)?

By looking at the numbers below, we see that the average rating is actually not high, but close to 10 (which is the rating denominator for all dog rates). 80% of the ratings are above 10 and the standard deviation is small, even with very high values as 75.

In [67]:
```
twitter_archive_master['rating_numerator'].describe()
```

Out[67]:
```
count    804.000000
mean      11.281095
std        2.974879
min        2.000000
25%       10.000000
50%       12.000000
75%       12.000000
max       75.000000
Name: rating_numerator, dtype: float64
```

In [68]:
```
twitter_archive_master['rating_numerator'].plot(kind='hist')
```

Out[68]:
```
<matplotlib.axes._subplots.AxesSubplot at 0x1203becc88>
```