5/9/2022

Act_report

Dihia Mezghiche

Report: act_report

In this project we will be studying the tweets of the twitter account @dog_rates.

For the analysis of this dataset, I asked myself the following questions:

1. What is the most common name?
   To answer this question we will just use a value_counts on the name column it will let us know how many time a name occurred in the dataset.

```
Out[108]: None        609
          Charlie      11
          Oliver       10
          Cooper       10
          Lucy         10
          Penny         9
          Tucker        9
          Winston       8
          Sadie         8
          Toby          7
          Lola          7
          Daisy         7
          Koda          6
          Bella         6
          Stanley       6
          Bo            6
          Jax           6
          Oscar         5
          Buddy         5
```

   * Since we already know that None is just a way to fill missing values and wrong names that makes Charlie the most common name with 11 dogs named that way in our dataset.

2. What is the breed of the highest rating?

```
]: tweet_id              778027034220126208
   img_num                                1
   Prediction_1                     clumber
   P1_accuracy                     0.946718
   p1_dog                              True
   Prediction_2             cocker_spaniel
   P2_accuracy                      0.01595
   p2_dog                              True
   Prediction_3                       Lhasa
   P3_accuracy                     0.006519
   p3_dog                              True
   rating_numerator                      27
   rating_denominator                    10
   name                              Sophie
   dog_age                           pupper
   favorite_count                      7320
   retweet_count                       1885
   Name: 1424, dtype: object
```

   * The breed of dogs with the highest rating is Clumber

3. What is the breed of the dog with the highest retweet?

As you can see below , using the describe function we got the highest retweet that has been done on a tweet which is 56625, and we want to know what is the breed of the dog that got that high number of retweet, to do that we only need to create a query that will select only the tweet where the number of retweets is the same as this one and the result is shown in the second figure.

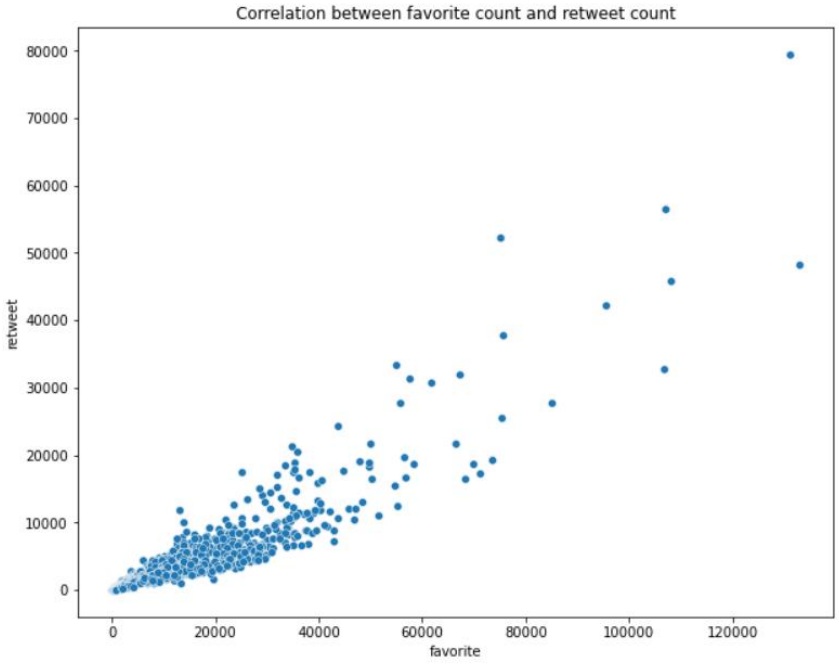
	tweet_id	img_num	P1_accuracy	P2_accuracy	P3_accuracy	rating_numerator	rating_denominator	favorite_count	retweet_count
count	1.189000e+03	1189.000000	1189.000000	1189.000000	1189.000000	1189.000000	1189.0	1189.000000	1189.000000
mean	7.420630e+17	1.224558	0.627774	0.143487	0.061188	10.970563	10.0	9340.322119	2763.436501
std	6.937515e+16	0.582462	0.253556	0.104115	0.052762	1.801601	0.0	12604.508454	4533.376621
min	6.660209e+17	1.000000	0.044333	0.000056	0.000011	2.000000	10.0	81.000000	16.000000
25%	6.783968e+17	1.000000	0.417107	0.055902	0.014763	10.000000	10.0	2236.000000	653.000000
50%	7.156808e+17	1.000000	0.633037	0.130726	0.048572	11.000000	10.0	4578.000000	1425.000000
75%	8.003883e+17	1.000000	0.853407	0.207753	0.094933	12.000000	10.0	11746.000000	3224.000000
max	8.921774e+17	4.000000	0.999876	0.467678	0.271042	27.000000	10.0	132810.000000	56625.000000

	Prediction_1	retweet_count
1577	Chihuahua	56625

* And so the breed of the dog with the highest retweet is a chihuahua that has a rating of 13/10

4. Are the favorite count and retweet count linked to one another?

To answer this question, we are gone need to plot the correlation of these two attributes and calculate its rate. The Image bellow on its own clearly shows a strong link between the two attribute and since the correlation plot seems to go higher as the number of favorite tweets and retweets increases, we can suppose that the correlation type is a positive one. To confirm that I calculated the correlation coefficient and the results was of 0.91 which confirms are supposition that the correlation is positive.




```
array([[1.        , 0.91304379],
       [0.91304379, 1.        ]])
```

* We can clearly see that the correlation coefficient between favorite count and retweet count is positive as its rate is of 0.91 which mean that the value of the two attributes is linked to one another.