Project
Wrangle And Analyze Data

act_report

Udacity Data Analyst Nanodegree

21.04.2021

By Yalçın FİLİZ

1. Introduction

After I've completed the cleaning process for WeRateDogs twitter account data, I started to analyze data by visualizing it.

First of all, I determined the questions that I will analyze. And then I started the analysis with the type of plot I should use for each question in turn.

2. Questions

I determined 5 questions. The questions are as follows.

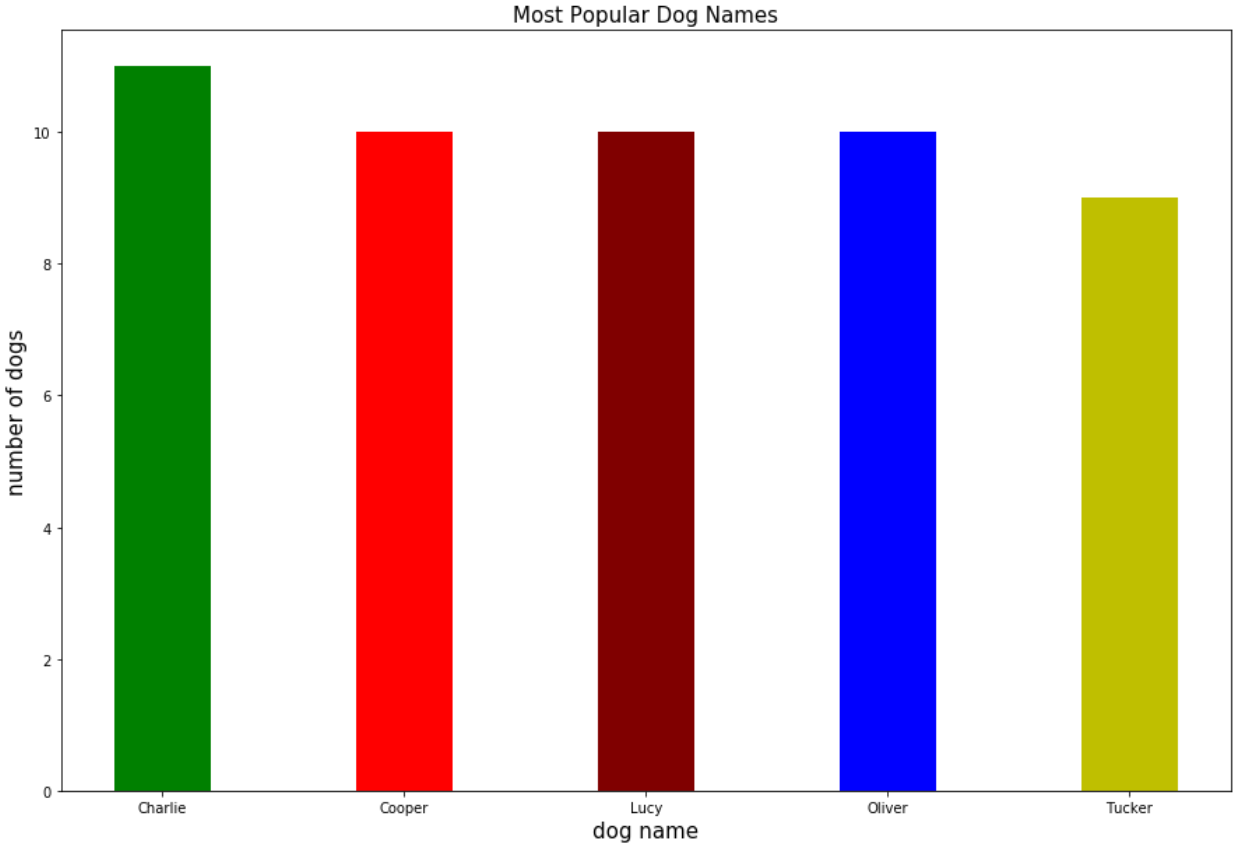
- What is the most popular dog name ?
- What is the most rated dog breed ?
- Which dog breed has the highest average rating ?
- Is there any relationship between Retweets, Favorites and Ratings?
- How to change the Number of Tweets over time ?

3. What is the most popular dog name ?

First I stored value_counts() into a new dataframe as pop_dogs. Then I used a bar plot to figure out which dog name is the most popular one.

As you can see “ Charlie “ is the most popular dog name.

Charlie 11
Oliver 10
Cooper 10
Lucy 10

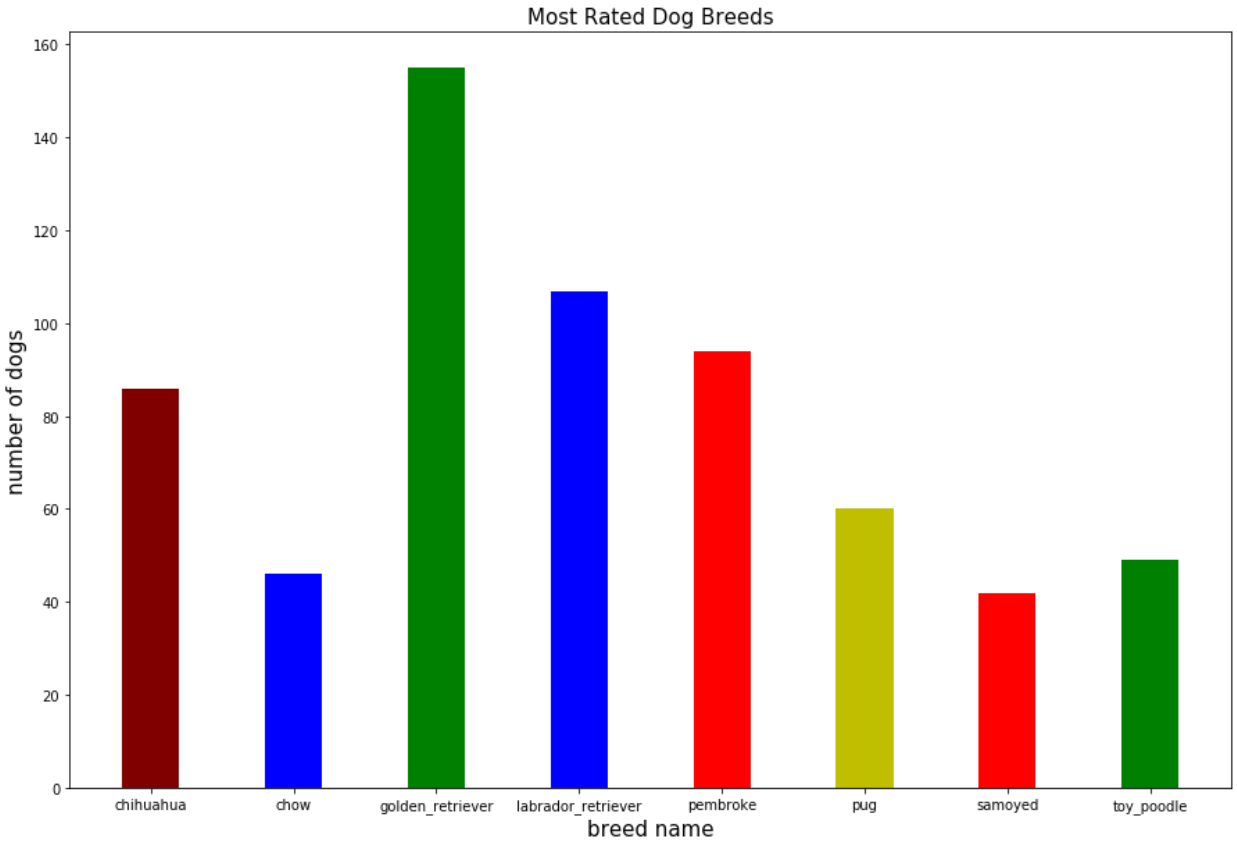


4. What is the most rated dog breed ?

First I created a new column to store dog breeds according to image predictions. Then I found top breeds with value_counts() and stored it into a dataframe as breeds.

I Used a bar plot to compare the number of dogs for each breed.

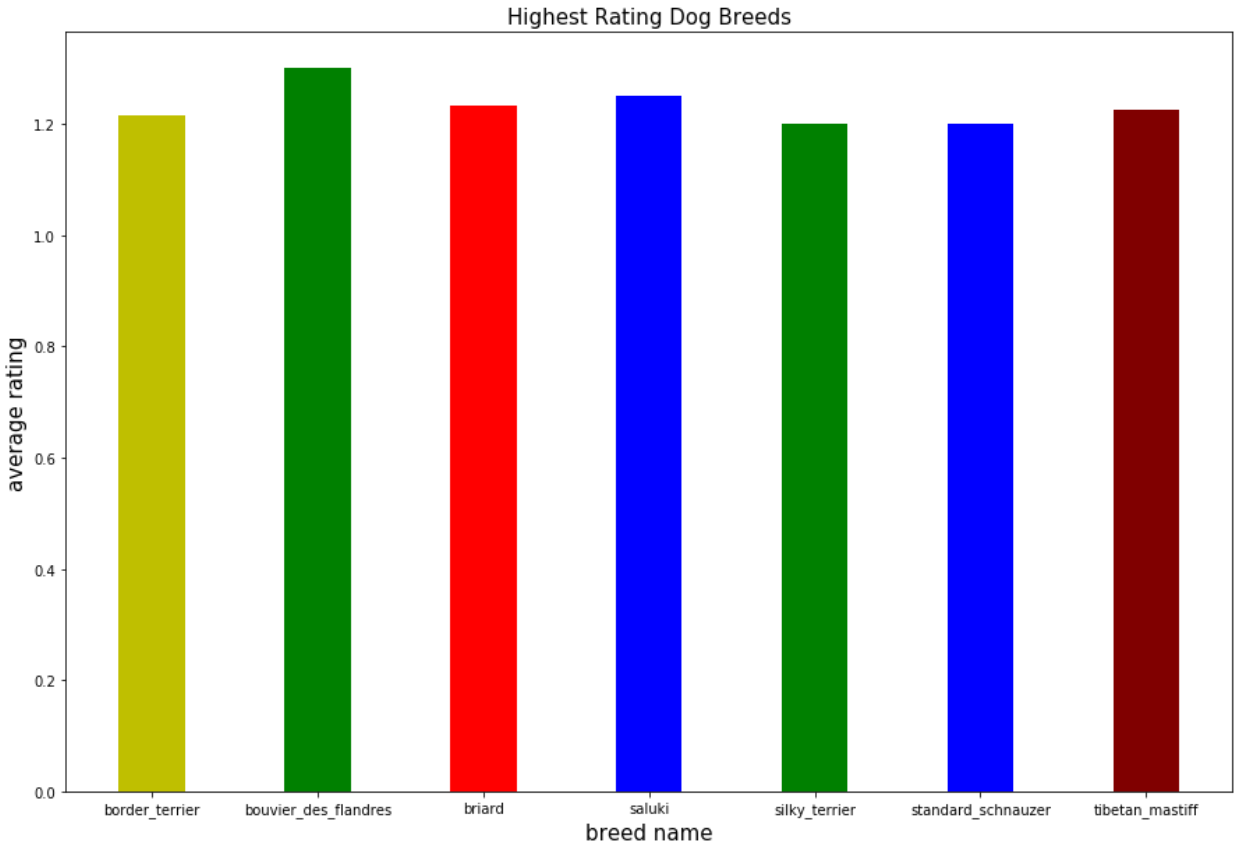
Most rated breed is Golden Retriever.



5. Which dog breed has the highest average rating ?

I used groupby() to calculate the average rating for each breed group and sorted them. I used a bar plot to compare the average rating.

The highest rated dog breed is bouvier_des_flandres.

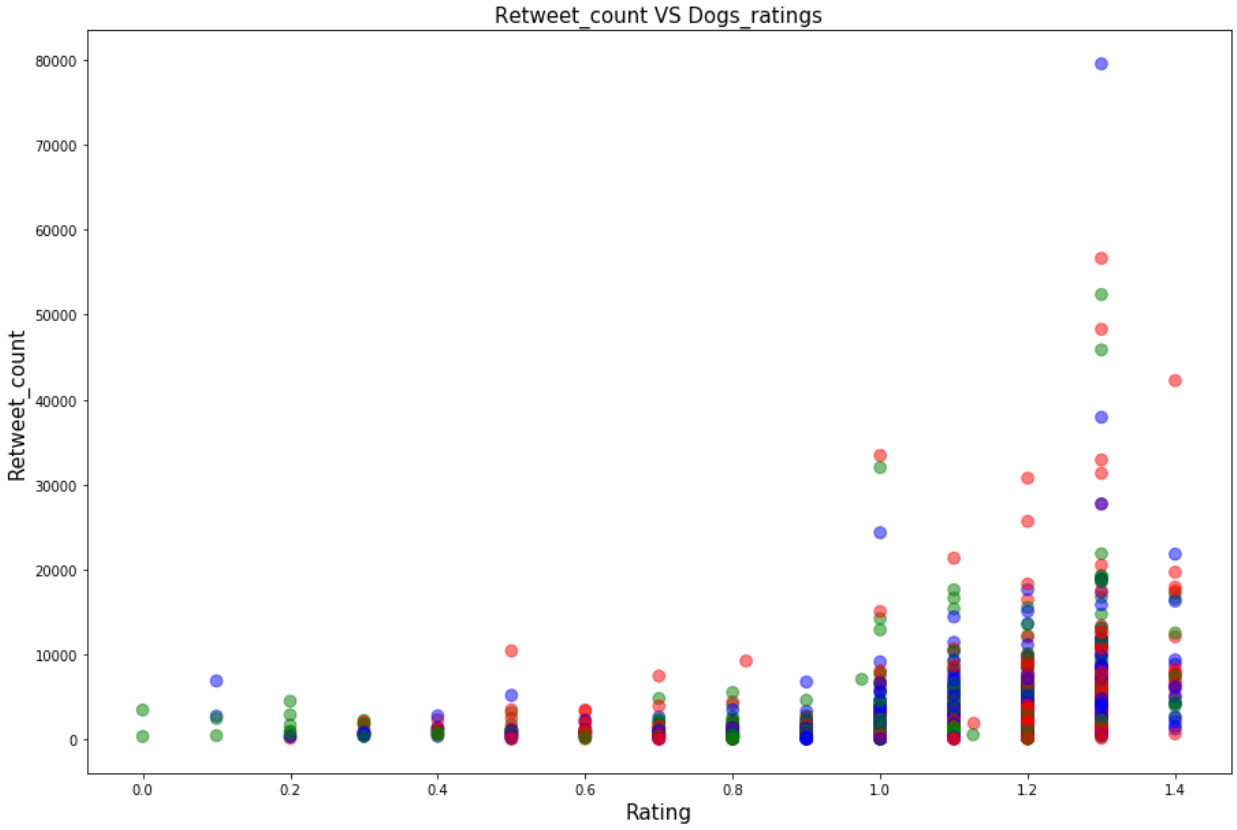


6. Is there correlation between retweets, Favorites and Ratings?

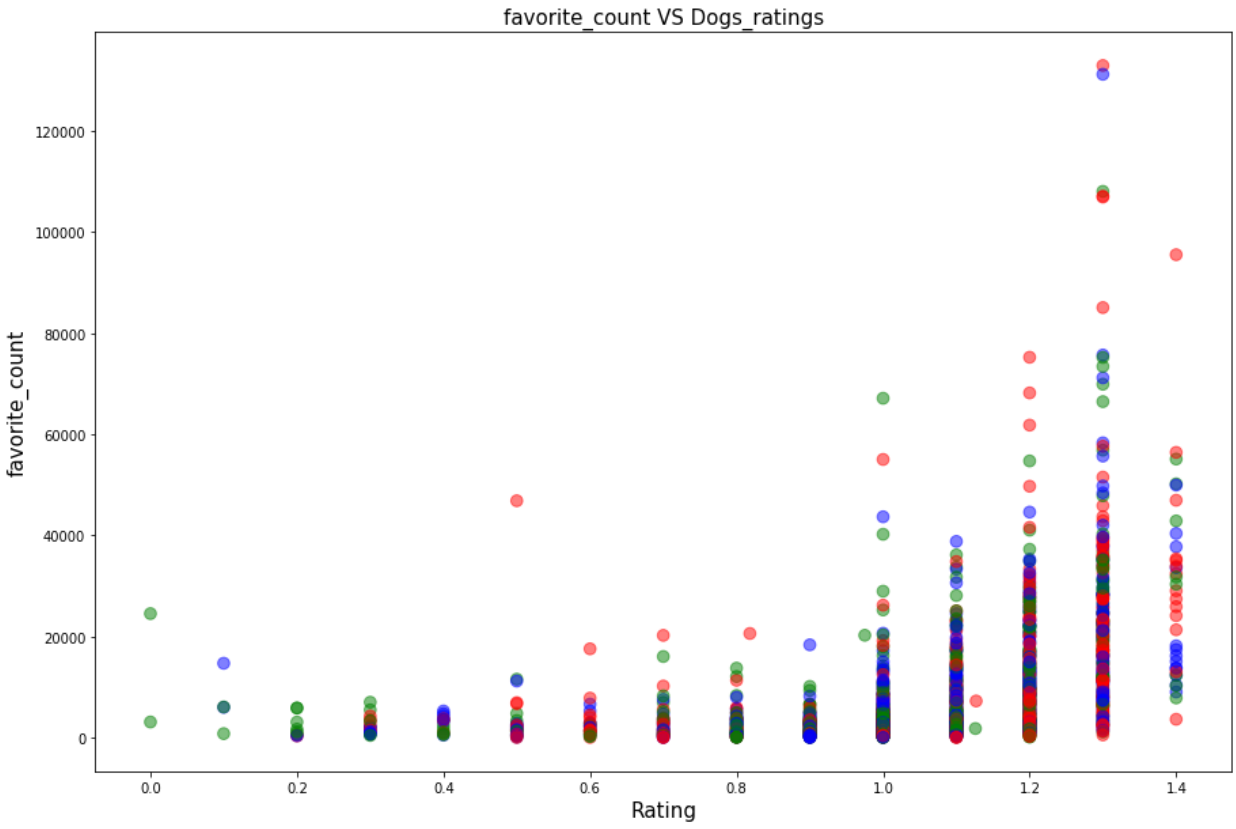
Using a scatter plot in each of the 3 columns, I examined whether there is a positive or negative correlation between them.

Also I calcutaled correlation coefficient for each.

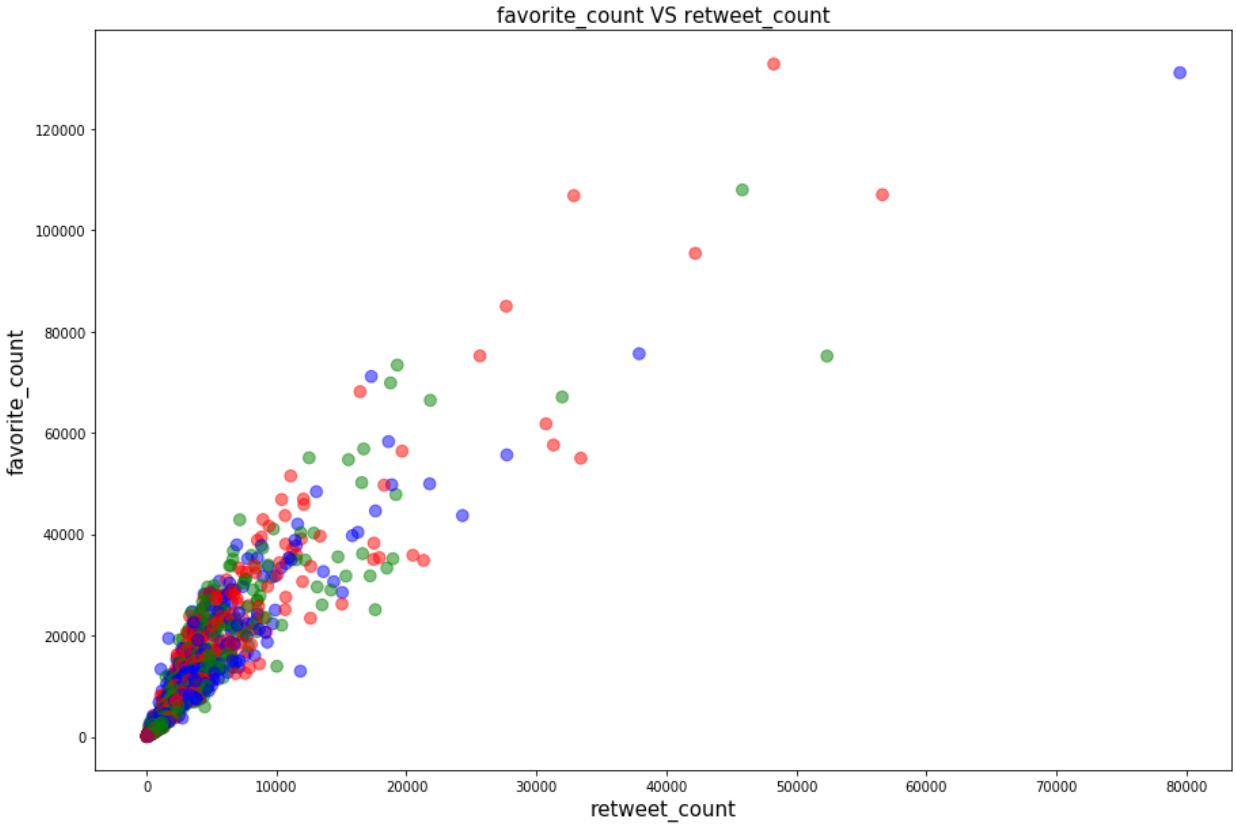
- Dogs_rating VS retweet_count --> **correlation coefficient** is 0.30



- Dogs_rating VS favorite_count --> **correlation coefficient** is 0.40



- retweet_count VS favorite_count --> **correlation coefficient** is 0.91

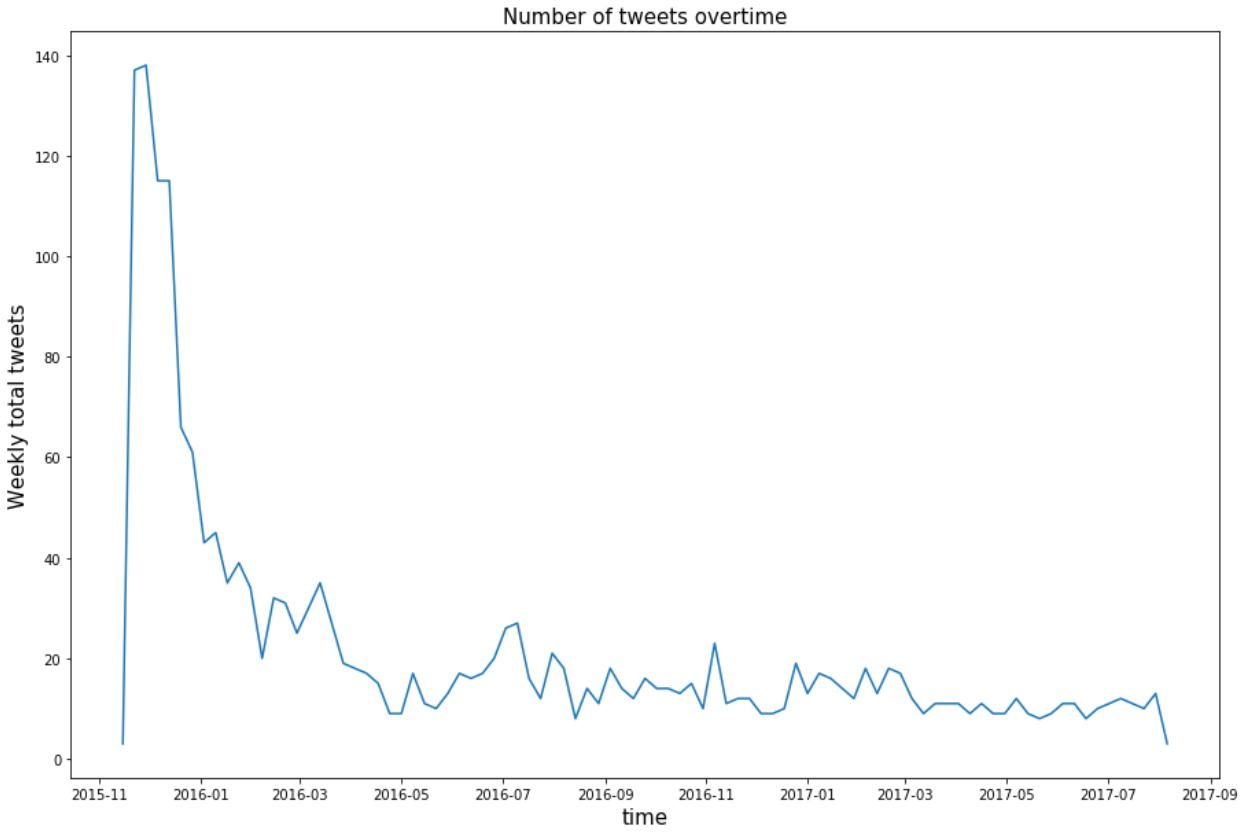


There is a positive and **strong** correlation between Retweet & Favorite counts. Although the other 2 variables also have a positive correlation with dogs_rating, they are much weaker than this.

7. How to change the Number of Tweets over time ?

I first grouped the dataframe by time_stamp weekly and then reached the total number of tweets per week with count ().

The total number of tweets has gradually decreased over time.



8. References

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.iterrows.html

https://pandas.pydata.org/docs/reference/api/pandas.DataFrame.plot.scatter.html

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.DataFrame.corr.html

https://en.wikipedia.org/wiki/Pearson_correlation_coefficient

https://stackoverflow.com/questions/42255458/how-to-group-a-pandas-dataframe-by-a-defined-time-interval

https://pandas.pydata.org/pandas-docs/stable/reference/api/pandas.Grouper.html