Data Analysis practice

Data Wrangling

We rate dogs analysis

Alejandro Sierra
Email: asierra21@gmail.com

27th December 2020

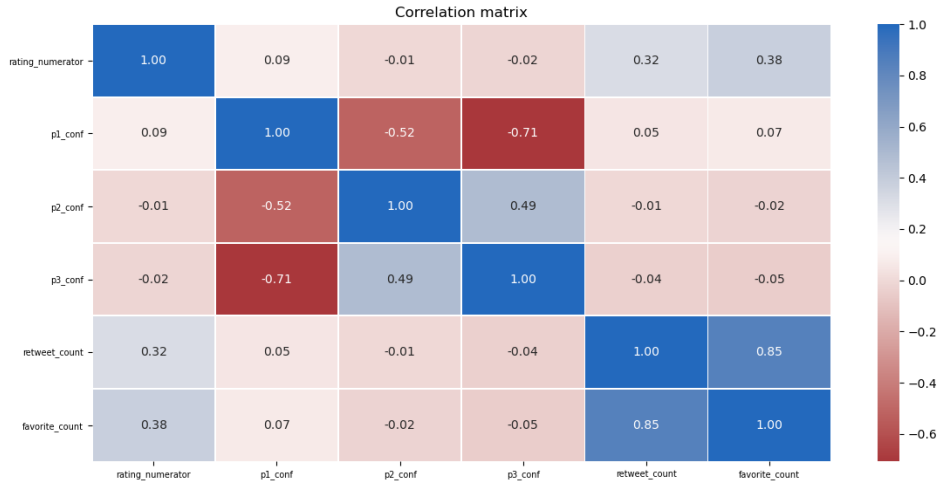
Contents

1 Data exploration

In this chapter we are going to study the data and perform an analysis on the variables rating, favorite_count and retweet_count.
We would like to determine whether rating system and favorites are correlated.
Besides, we are going to explore what breed of dog gets the higher ratings and favorites or retweets.

1.1 Feature correlation

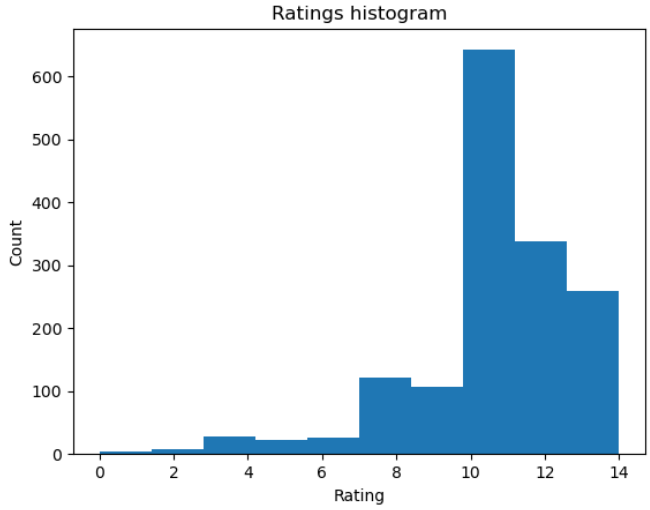
First thing we are going to explore is the correlation matrix.



As we can see in the correlation matrix, the only high interesting relation is, as expected, between favorite counts and retweet counts.
There is a low correlation also between rating and both favorite count and retweet count.

1.2 Ratings

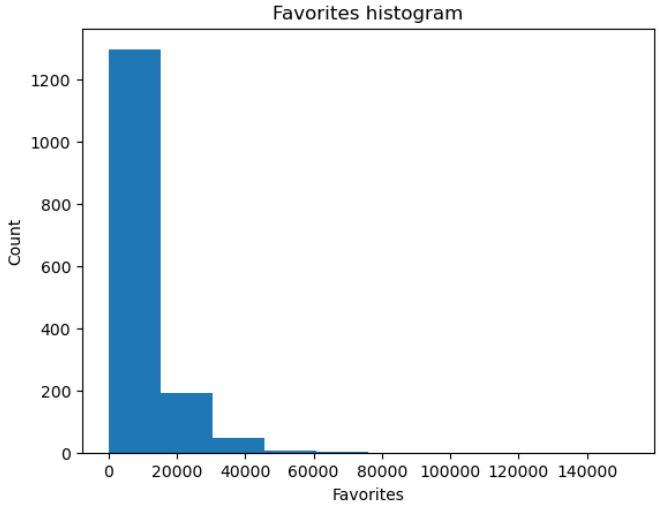
The univariate exploration for numerical variables is best explained by histograms.



As we can see, ratings higher than 10 are really common in this popular site, being the rating 10/10 the most used.

1.3 Favorites

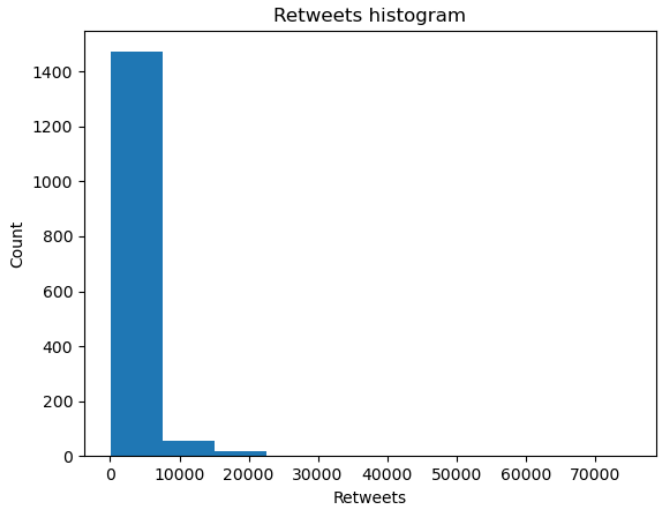
The following plot will show the likes distribution a tweet gets.



We can see in this plot that the ammount of favorites tweets get is commonly in the range 0-1000.
It is not usual to get over 4000 likes for a tweet in this account.

1.4 Retweets

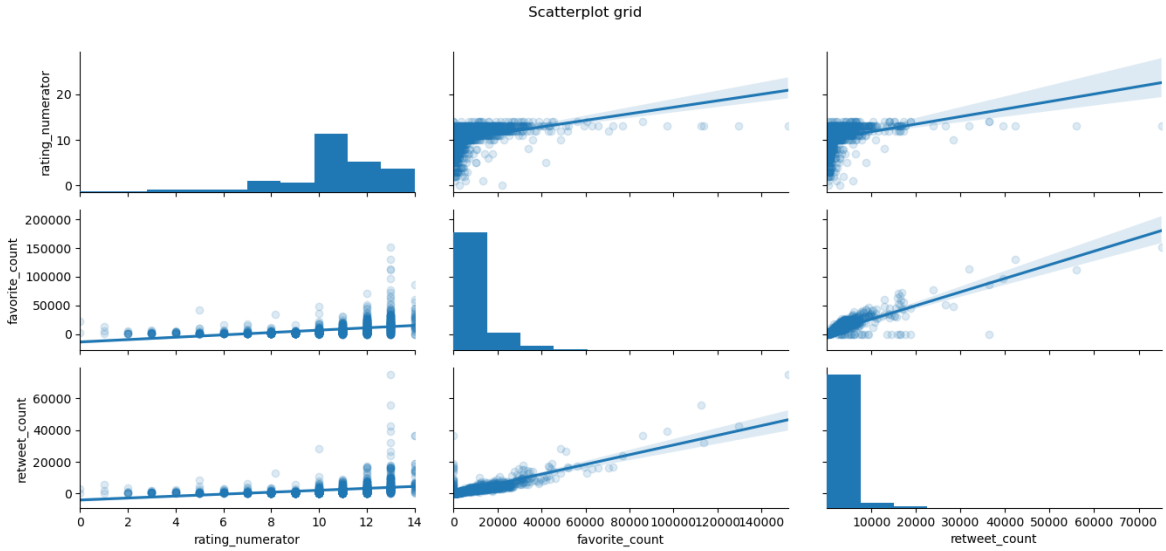
We repeat the analysis for retweets to check their distribution.



As we saw with favorites, retweets behave in the same way. More than half of the dataset gets around 500 retweets or less.

1.5 Bivariate exploration

Ratings, retweets and favorites:
Now we are going to analyze how these variables relate to each other with a scatterplot grid.



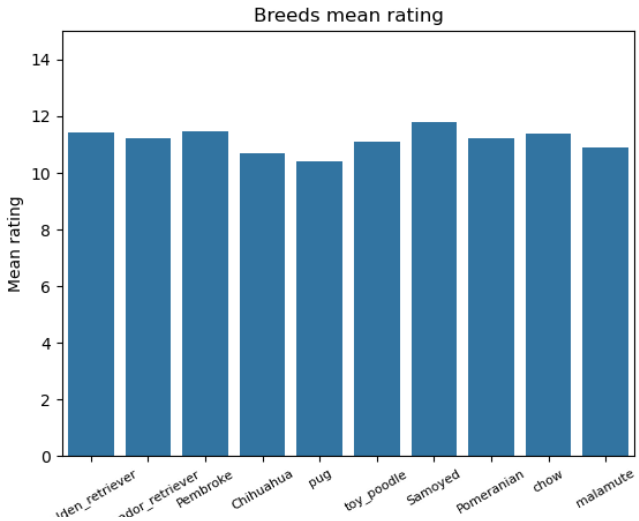
We can see that for ratings higher than 10, both favorites and retweets start to increase. Below this threshold, they show a stable behavior with fewer interactions.
Favorites and retweets seems positively correlated with a linear relation with the key aspect of the slope not going through the origin.

Ratings by dog breed:

We are going to group by the first prediction made for the tweets picture and see what is the top ten dog breeds that show up.

Breed	Count
Golden_retriever	114
Labrador_retriever	66
Pembroke	64
Chihuahua	61
Pug	47
Toy_poodle	34
Samoyed	33
Pomeranian	29
Chow	28
Malamute	20

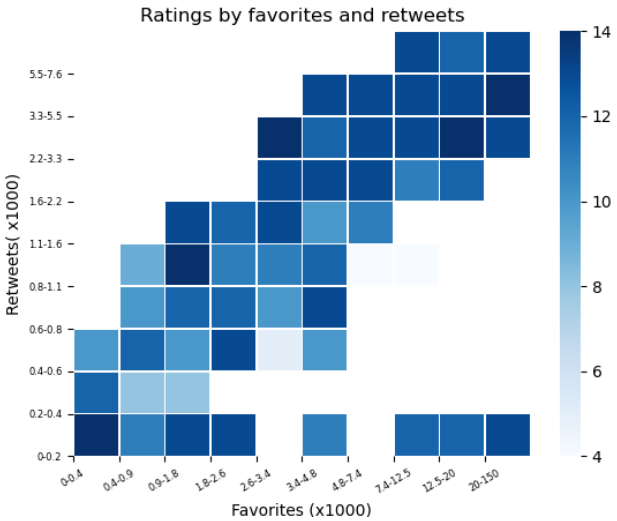
And now we are going to plot the mean rating each of those breeds recived.



This plot that there is not a considerable difference between dog breeds that could determine a rating increase or decrease.

1.6 Multivariate exploration

Ratings by favorites and retweets:
We are going to see how ratings relate to both variables, favorites and retweets.



The heatmap shows, as expected, that with the increase of favorites and retweets, we have darker blues corresponding to higher ratings in the top right corner.
Relation seems linear across the diagonal for the heatmap.

2 Conclusions

The main conclusion for this dataset is regarding the ratings system, dogs usually get a rating over 10, higher than the scale maximum.

Another main finding is that there is not a prevalent breed in terms of rating mean, while there is one in terms of tweet counts, being the Golden Retriever the most popular one.

There is a positive correlation between likes and retweets, and that correlation is accompanied by the ratings, which increase with the other two variables.

Most tweets get between 0-1000 likes and between 0-500 retweets.