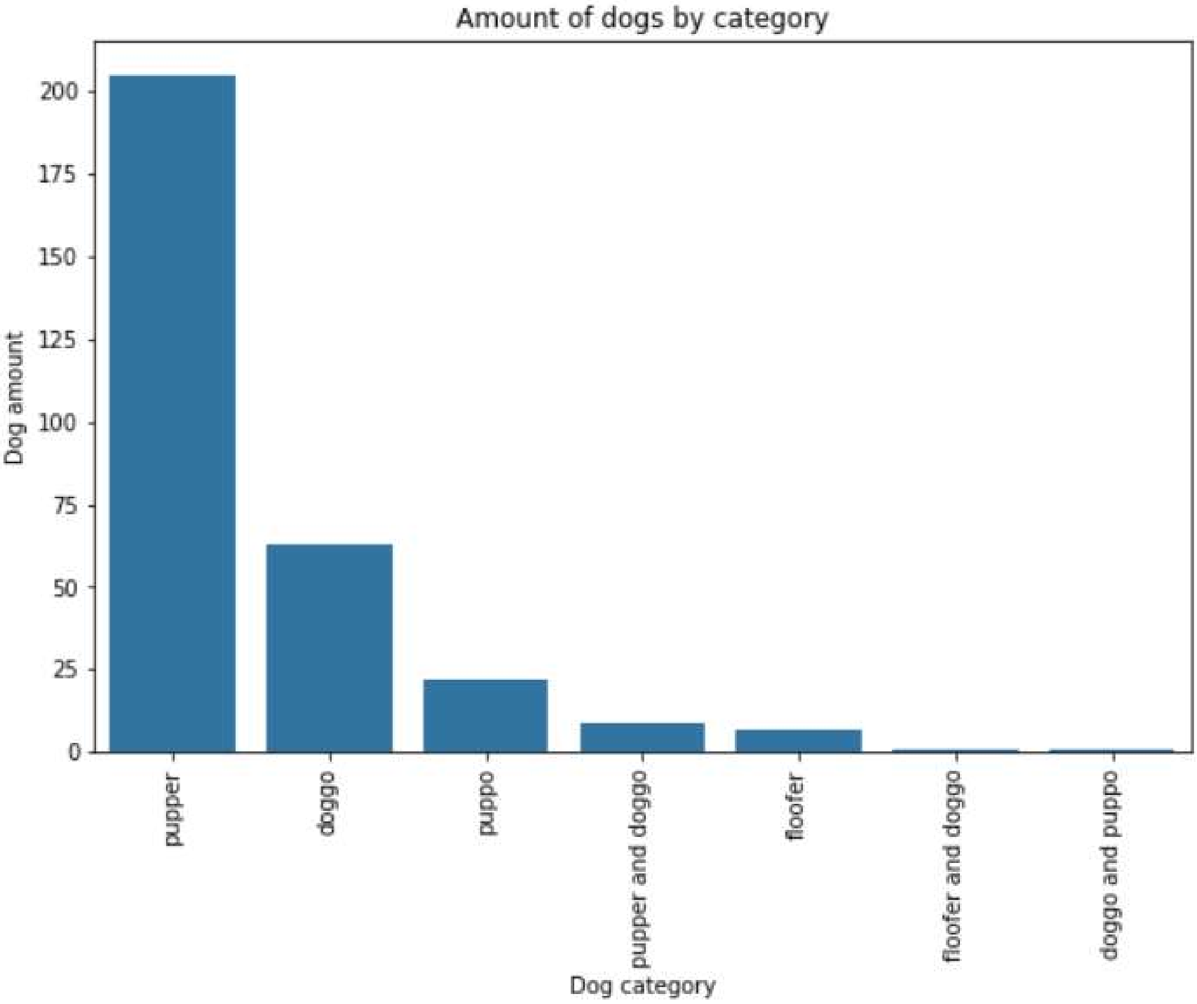
ANALYSIS AND INSIGHTS

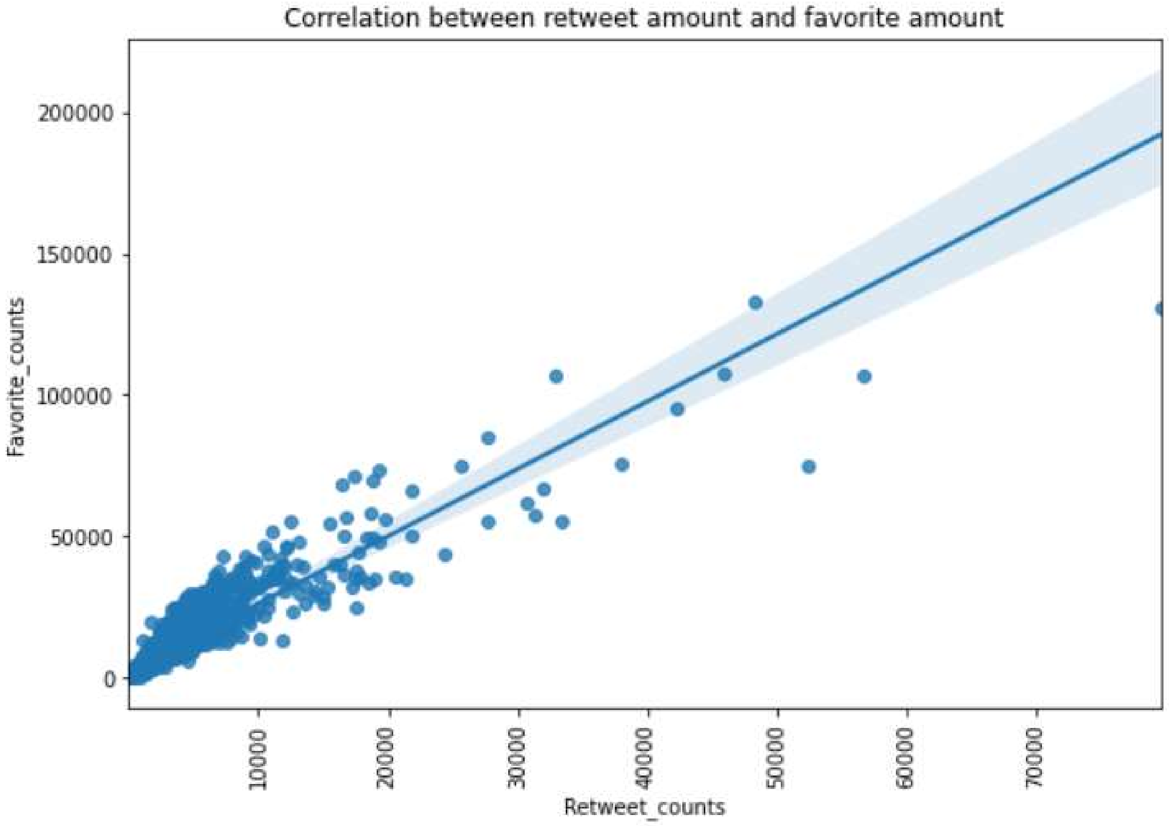
First of all, we have assessed the data so we could identify the tidiness and quality problems to be able to continue with a dataset that would permit to work with it correctly and efficiently. We have cleaned the identified issues, and then we have tested them in order to make sure that the data was ready.

After the above-mentioned data wrangling process, some data visualizations were made in order to understand the content of the twitter information we have and get valuable information.

In order to begin, and so we can have a global perspective of the data, we made a bar chart which shows the dog amount by category. This will give as an overview of the information we have.

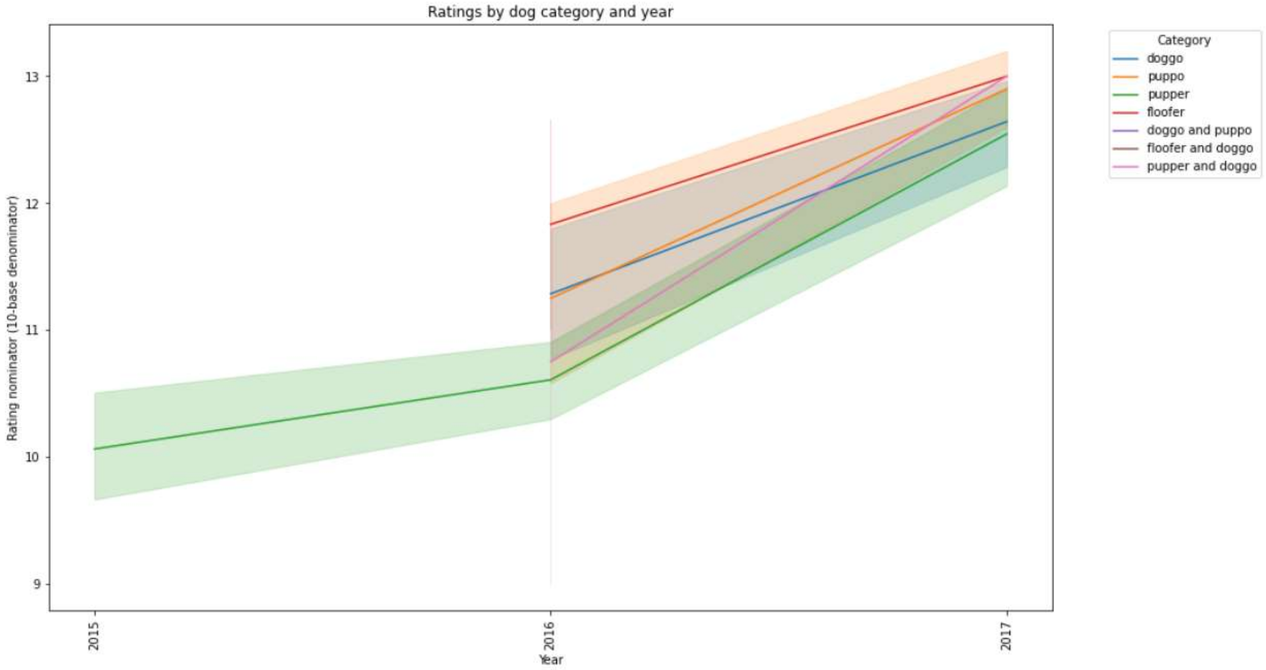


After this, we have plotted the correlation between the retweets and the corresponding favourite amount using a scatter plot:



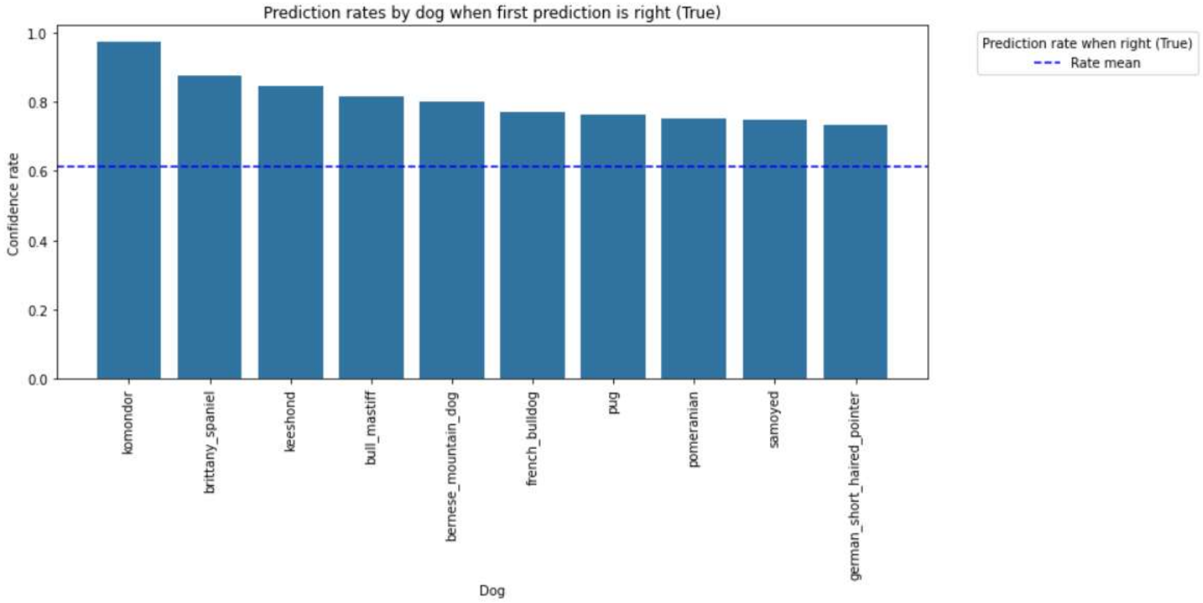
As it is possible to see in the pic above, there is a lineal correlation between the retweets and the favourites' amount. However, almost all the data is concentrated on the tweets that have less than 10,000 retweets and in another lower proportion on the tweets that have between 20,000 and 10,000 retweets.

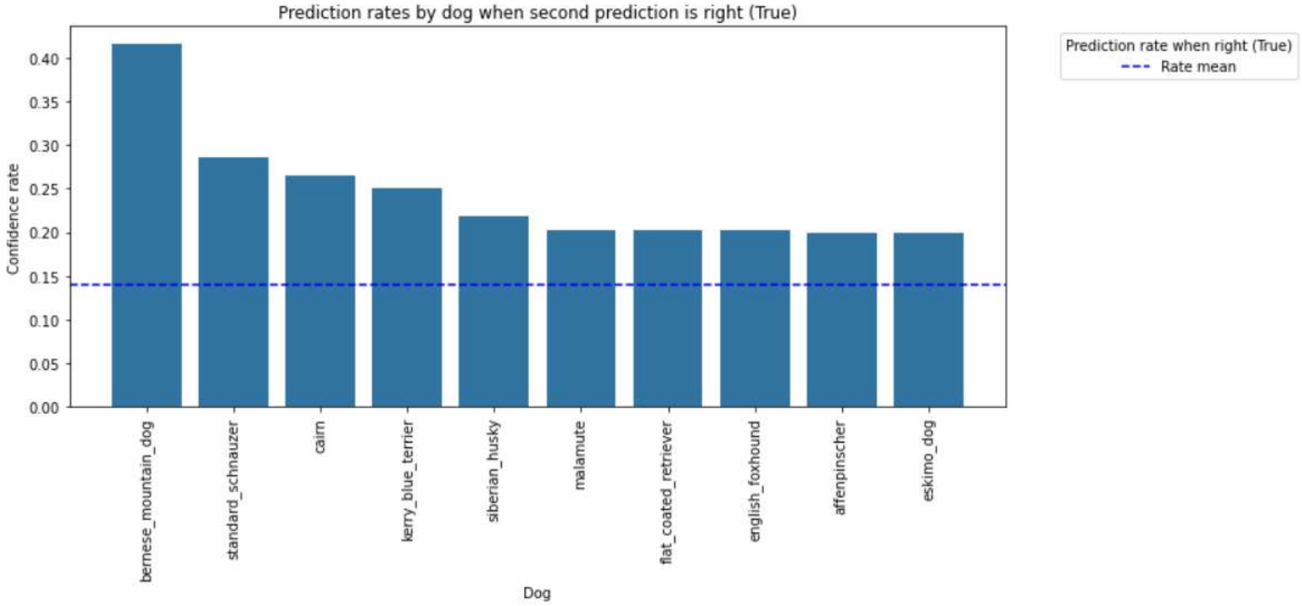
The next thing to see is the ratings by dog category and year. For this a line plot was made:

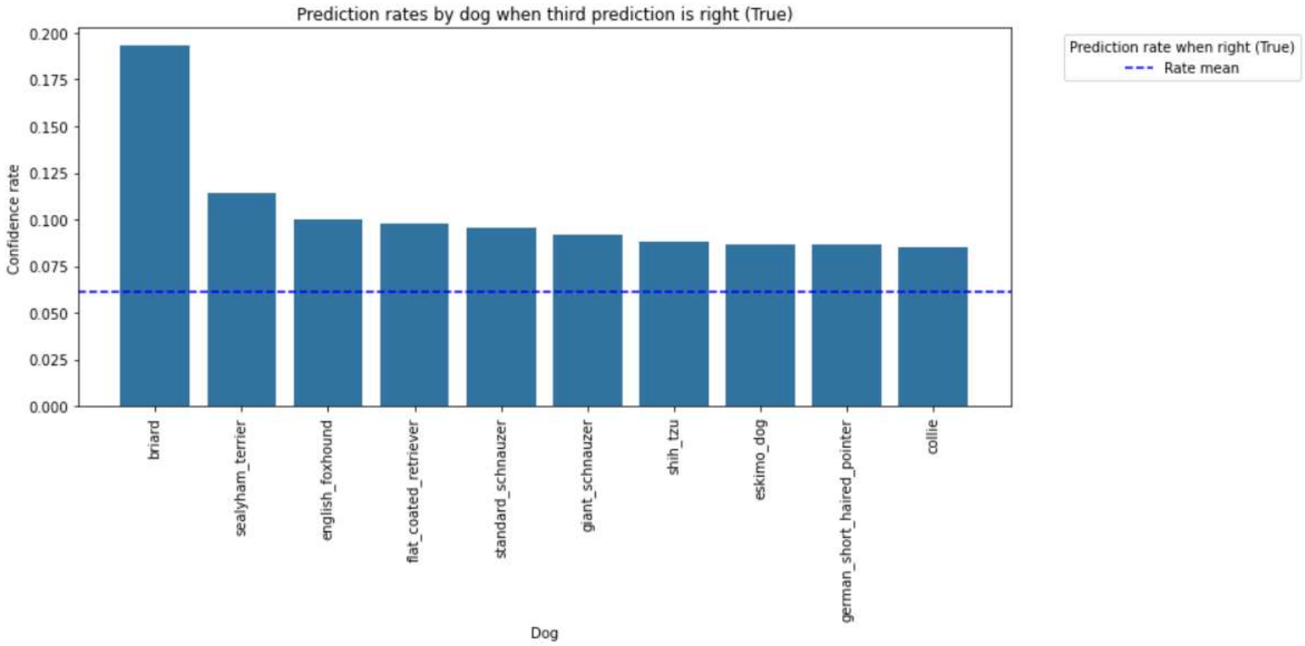


As we can see above, the trend is that the ratings increase year by year and “flooter” is the category with best ratings and “pupper” with the worst ones. In addition, we can see that “pupper” is the only category that was rated in 2015.

Finally, we made a visualization of the prediction rates, when the prediction was right (True) and we compared it with the average for all the prediction1, prediction2 and prediction3 (one visualization for each one). In the bar chart we have visualized just the 10 highest rates and compared them with the corresponding average (blue dashed line):







Comparing the last three visualizations, it is clear that the confidence rate is higher for the first prediction and decreases for the second and third prediction.