Udacity Project - Analysis of Data
Twitter Data Sets
By Andrey Gorelov
2019

Introduction

After all the Twitter data about WeRateDogs has been gathered, accessed and cleaned, it is time to analyse the insights and demonstrate some visualization. The reason we do the analysis at the end, is to guarantee the most efficient and 'clean' outcome after all the pre-work.

Analyse - 4 Insights and 1 Visual

When it comes to the analysis, I have constructed these three questions to investigate:

1. Most favorite Tweet?
2. Highest retweeted tweet?
3. Top 10 most common dog breeds in the whole data ? (Illustrated by the visual as well)
4. Number of Retweets vs Favorites? (Illustrated by the visual as well)

Question 1

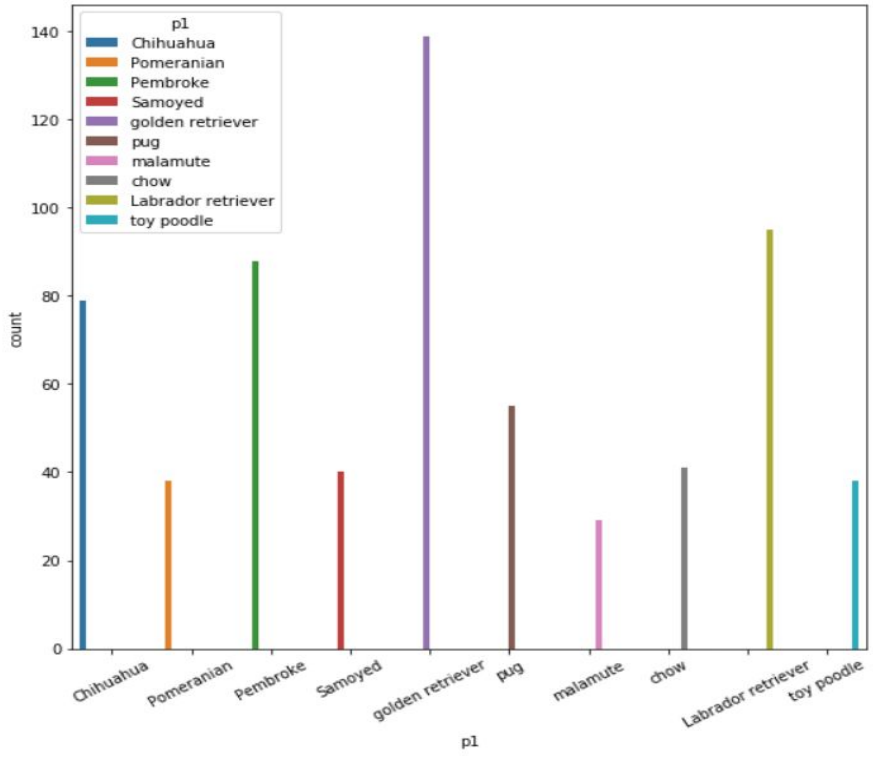
The first analysis made me identify a cute puppo and luckily I also got a chance to take a look at his face. By focusing on the 'favorite' column and putting the order to be descending, I could easily find him. From further code of iloc I managed to get more information about him (as well as this picture.)

Question 2

Similar to the previous question, we run the code that would give us a list of information in the descending order. This way, the highest retweeted tweet belongs to doggo with 79515 retweets. Unfortunately, the picture for that doggo is present but can not be reached.

Question 3

In the next question, I try to find the most common doggo breed in the whole data set. For this I use the command group by to determine the breed, but unfortunately there are too many dog breeds so the graph becomes too populated. Hence, I cut the breed type to top 10 to see the highests breeds. The result shows that the most common breed is Golden Retriever with nearly 140 counts.



Question 4

Last but not least, I try to compare the number of Retweets and Favourites to each other using the regplot. We can see that most of the positive correlation happen in the lower number corner. Some tweets that have been retweeted between 10k and 20k time have 0 favourites. However, when we look at the plot towards the highest numbers, then we can see that favourites and/or retweets both have high amounts.