Insights and visualisation of WeRateDogs twitter data

Souhail YOUSSEFI

Introduction:

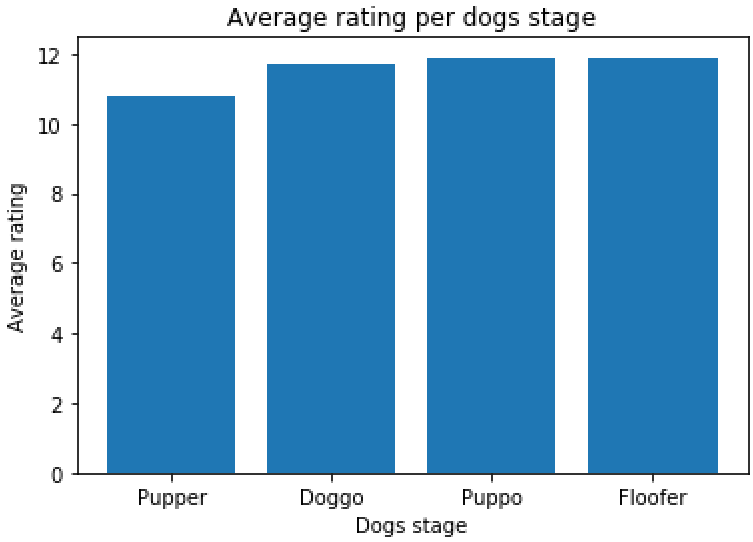
If internet has nowadays unleashed the learning potential of people across the world, it has also managed to create bonding around quirky and fun stuff. The WeRateDogs twitter account is a good example of this second usage of the internet!

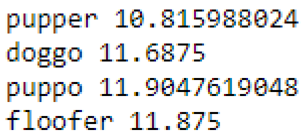
“WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.” *Udacity project definition*

Using different methods, we gathered several tweets from this page, cleaned the data and tried to get some insights from it. This document presents our findings in a Q&A format.

1. Which of the dog’s stage gets the best rating?

In order the answer this question we needed to remove all the rows with dogs_stage as "None" and then it seemed like a bar plot would be the most suitable vizualisation to be able to answer the question and get some insight.



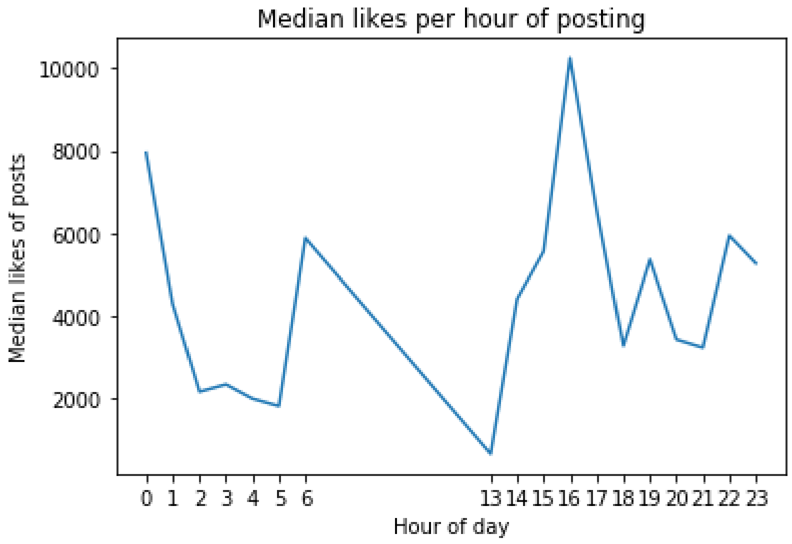


```
pupper 10.815988024
doggo 11.6875
puppo 11.9047619048
floofer 11.875
```

The average ratings for dogs stages are pretty close to each other. The spread of each stage was calculated to check if it is better to use the mean or median. As the spread value is not very high, the mean was used. It appears that the ones classified as "Puppo" achieved the highest average ratings closely followed by "Floofer", then "Doggo" and "Pupper". It is a tight race as all averages lay in a range of 1 point. Finally we should question the average of "Floofer" as it was only calculated on a 3 ratings sample. To draw better conclusions it would be interesting to gather more data about "Floofer" and "Puppo" ratings.

2. Do tweets posted after work hours (in the evening) manage to get more likes?

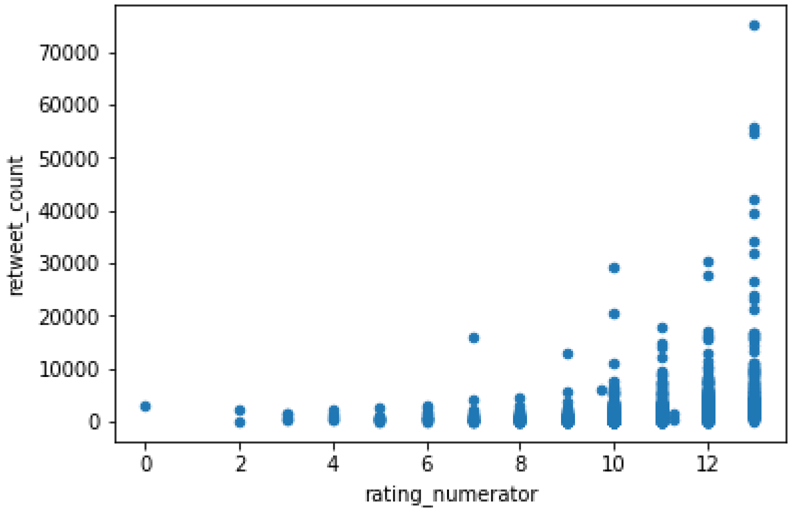
It would be interesting to check if tweets posted after work hours achieve more likes as there is possibly a higher audience. Here, the fact of formatting the timestamp as time date format will be helpful.



It looks like 4pm seems to be the posting time that gets the more median likes (which is towards the end of the work days?). 1pm seems to be the hour of the day with the least number of median likes. We can also notice some peaks that we could link to everyday routine:

- Peak at 6 am (waking up/ breakfast)
- Peak at 7 pm (back from work / dinner time)
- 10pm to midinight (just before bedtime) Of course, we are just seing some correlation here, therefore it does not imply causality. Some other parameters could be involved in the median number of like.

3. Is there a correlation between retweets and the rating?



It looks like there is a correlation between the number of retweets and the rating. It looks like an exponential correlation.