Udacity – Data Wrangling Project

Data Analyst Nano Degree

hind baaqeel

11/28/20

Data Wrangling Act Report

Introduction

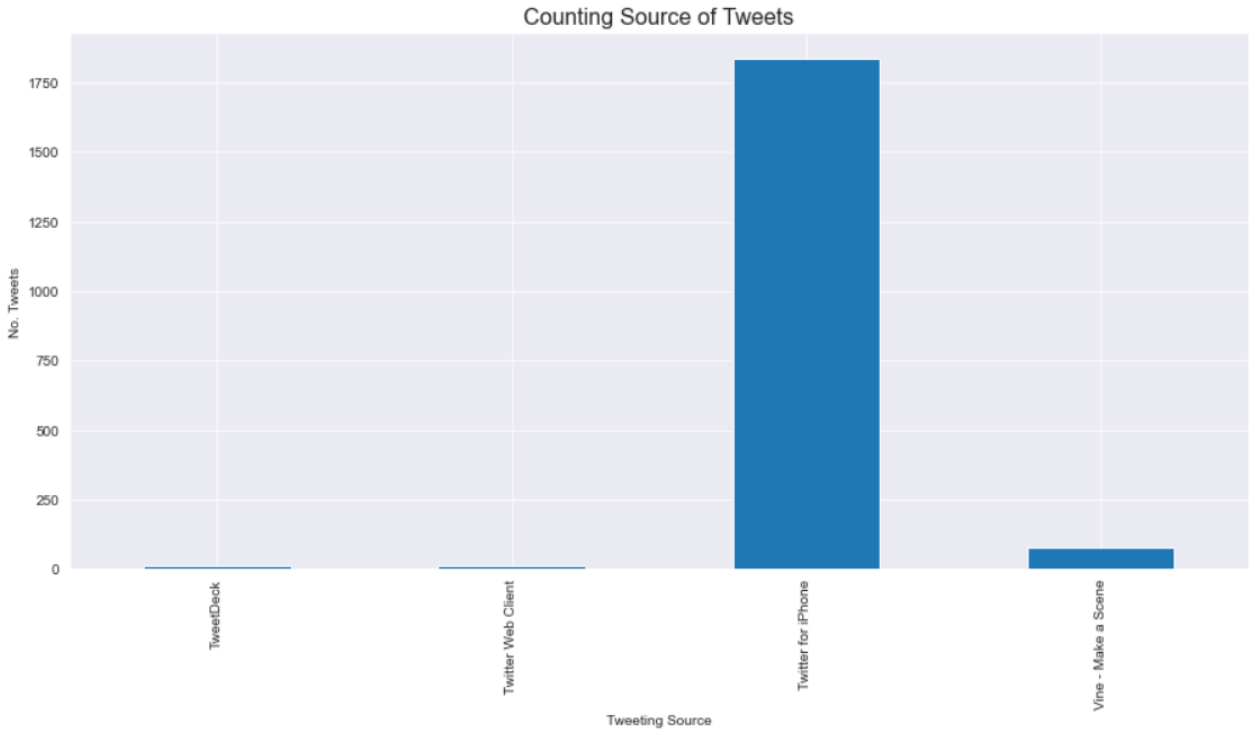
This Project demonstrate the data wrangling process using twitter data on archive called WeRateDogs. In this archive, users are given a dog image and they rate it with their comments. It is worth noting that the setup of the rating system requires that the rating denominator is always higher than 10. In the project, I have used this dataset to practice data wrangling steps through gathering, accessing and cleaning the data inside the archive. In this report, we summarized the data analysis process done after data wrangling.

Exploratory Data Analysis

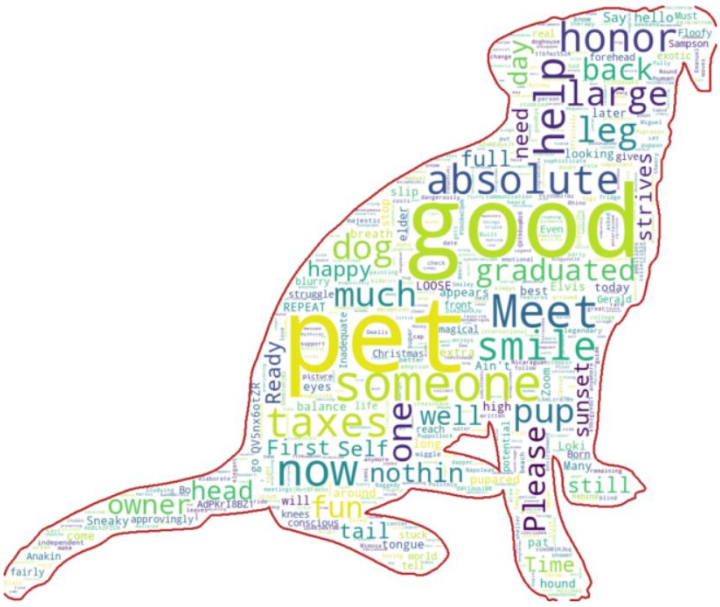
In this report, I aimed to answer some question related to our dataset:

- what is the most used source for rating in WeRateDogs?
- what are the most common words in tweets associated with dogs in doggo stage?
- what is the most common hour of dogs rating tweets?
- What is the stage that got the most average retweet? how about the most favorited one?
- What are the top ten favorited bread? how about the top retweeted?

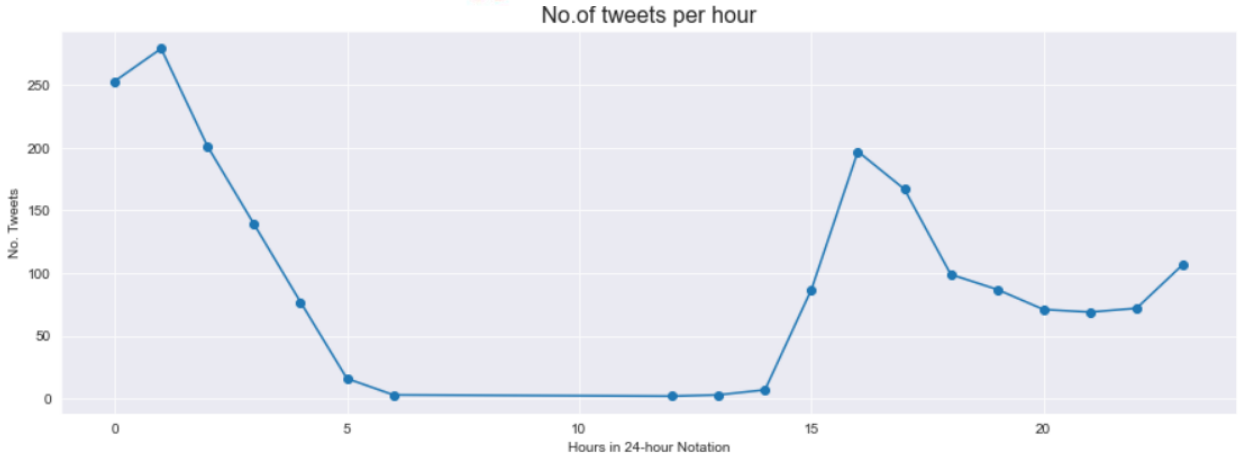
Results:



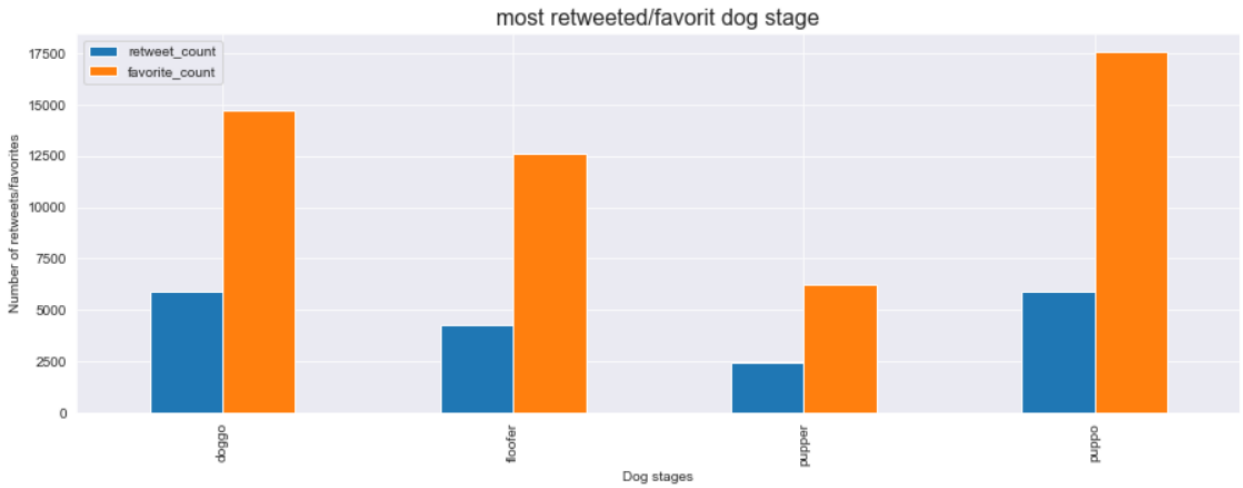
from the figure above, it is clear that twitter for iPhone is the most used application for tweeting in this WeRateDogs followed by Vine



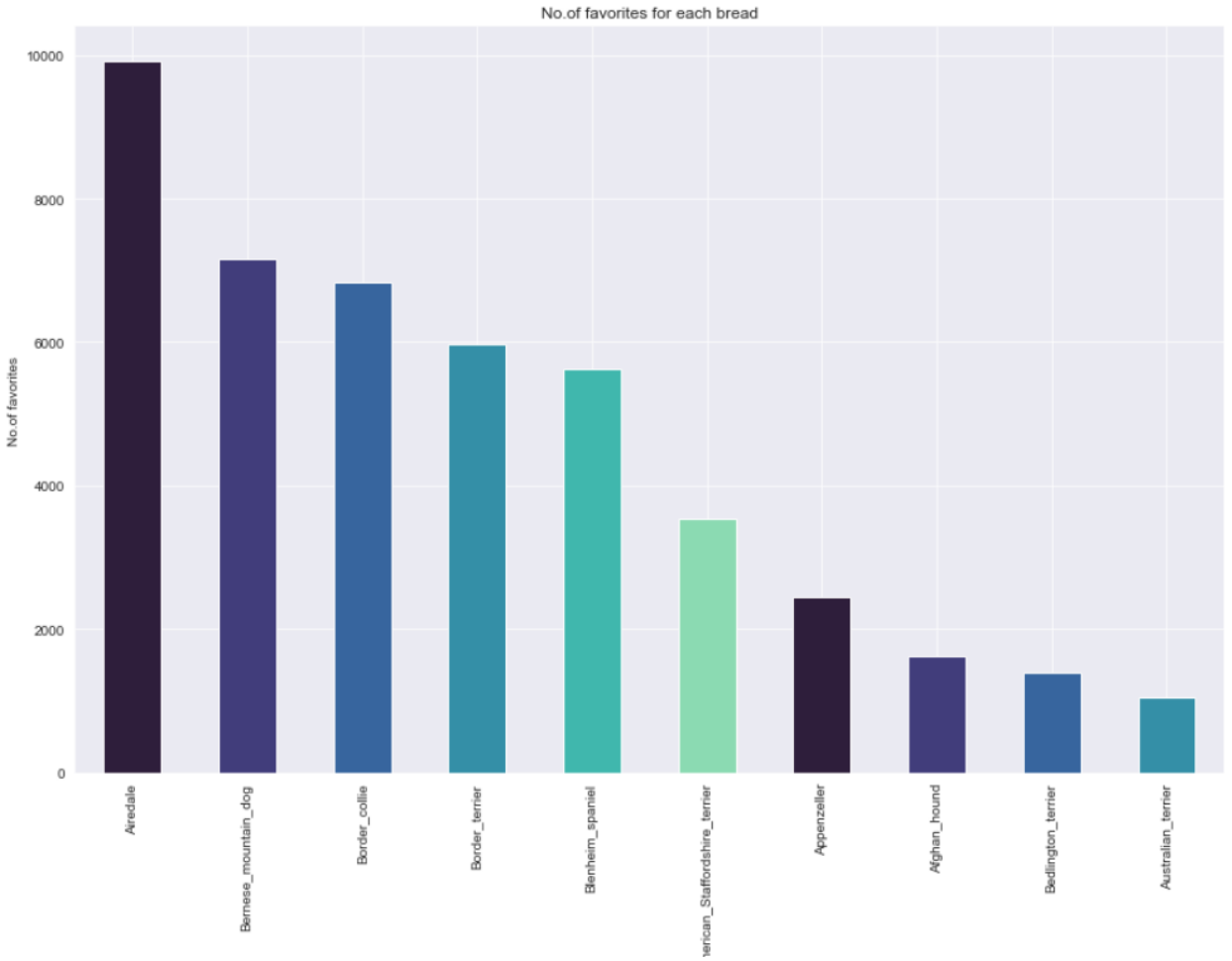
The word cloud generated the most common words associated with the dogs in doggo stage. the words include smile, good, meet, pet and help. which are mostly positive words.



we can see from the above figure that the most common tweeting hour is between 1 am - 2 am



We can see from the above figure that the most retweeted stage is puppo with small difference margin from doggo. In term of favorited tweets, puppo is clearly the highest among the other stages followed by doggo



from the figure, we can see that the top 10 breads in average favorites. On the other hand, the top bread was Airdale followed by Bernese_mountain_dog.

Resources:

https://www.datacamp.com/community/tutorials/wordcloud-python

https://stackoverflow.com/questions/3398852/using-python-remove-html-tags-formatting-from-a-string

https://stackoverflow.com/questions/8209568/how-do-i-draw-a-grid-onto-a-plot-in-python

https://seaborn.pydata.org/tutorial/color_palettes.html

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

https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html