PROJECT TITLE : PROJECT 2

TITLE :DATA INSIGHTS

STUDENT NAME :DENZEL AKWANY

Introduction

The data wrangling process was successful therefore leading to remarkable insights despite the data being big.

Key insights

The following were main insights from the analysis process.The insights were guided by the following questions:

1. What are the frequent words within the tweets?
2. What are the outliers in the ratings
3. Which tweet ids had the highest volume of likes,time series analysis.

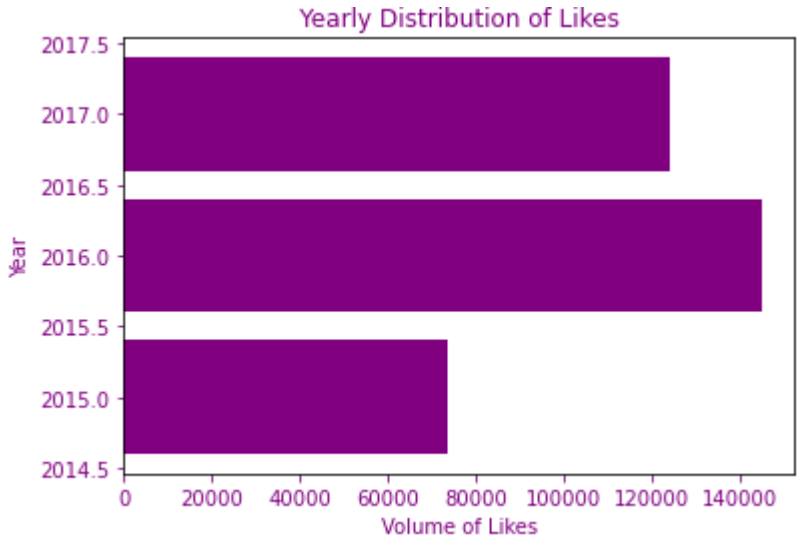
Frequent words

The frequent words from the tweets were identified by using the wordcloud. In the absence os the module , the library can be installed by running **pip install wordcloud.** The wordcloud excluded stop words which are often used in joining of sentences among others.

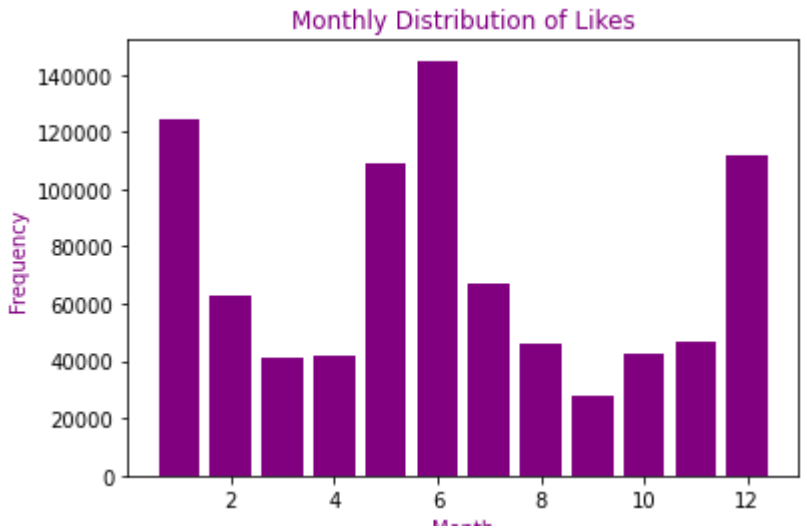
From the image frequent words are such as dog,pet,rates,pup,pupper among others. These are words that can be used in tweet look up and finding people with similar interest in dogs and their ratings.

Time series analysis

Through time series analysis, one is able to visualise trends over time. This is necessary in identifying times that tweets were highly liked,times that were less likes. Knowing the state of the business for the WeRateDogs and its popularity.An upward trend trajectory is always encouraged.

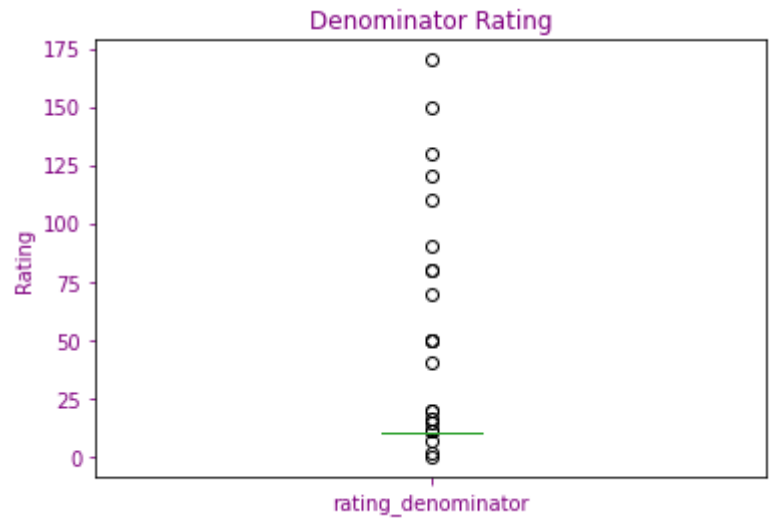


From the time series horizontal bar graph above, the likes doubled in the year 2016 compared to the previous year 2015. However, the likes dropped in 2017.The year 2017 is however,slightly above the average of the three years. This is still good awareness.

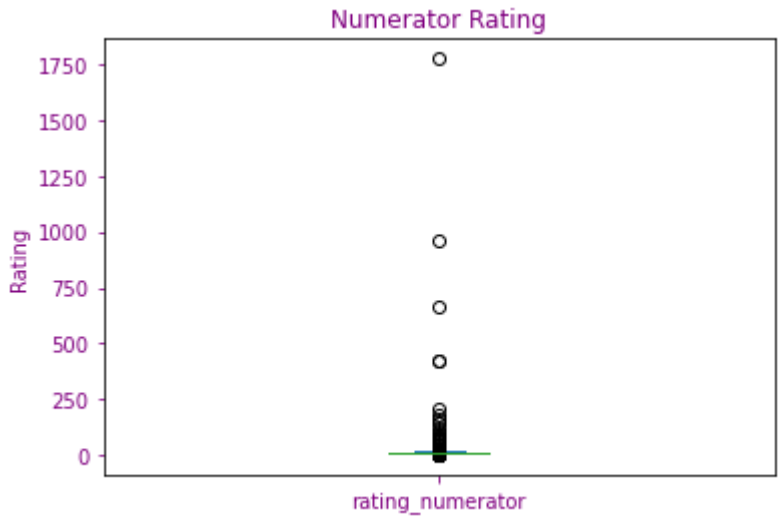


The monthly distribution of likes shows that in the months of January,June and December, there were more likes. This can be associated with seasons such as summer as it warms the Dogs are in parks and also festive periods where individuals are looking for pets for themselves or as presents.

Ratings



The image above shows the rating denominator. Which shows the concentration of ratings as well as outliers. The majority of ratings are within the range of 0 to 25.



The numerator rating gives insights into how people rated the dogs. It is necessary for the site to include a rating limit to avoid such big outliers within the rating. The quality of the final rate is highly influenced by the large numbers.

Volume of likes by tweet id

	tweet_id	retweets
1315	744234799360020481	144891
1931	822872901745569793	124120
1811	807106840509214720	111704
2197	866450705531457537	108921
1275	739238157791694849	107249
...	...	...
1618	782021823840026624	0
2074	841833993020538882	0
1605	780496263422808064	0
1604	780476555013349377	0
1859	814578408554463233	0

Twitter users with the highest volume of likes can be used by the site as dog promoters and their input ,analysis of their tweets,and wordings can be insightful in coming up with promotion videos,words and reaching a wider population. This information is of high value to the marketing team.

Conclusion

The data is large and more insights could be drawn from int. An enrichment of dog purchase or how the rating has influenced purchase or ownership of a particular breed can be highly useful to potential dog customers or sellers. The rating should also be standardised to avoid dispersed values.