Act Report

February 17, 2018

Amrut Deshpande

Analyzing and Visualizing WeRateDogs data after Wrangling process

Introduction

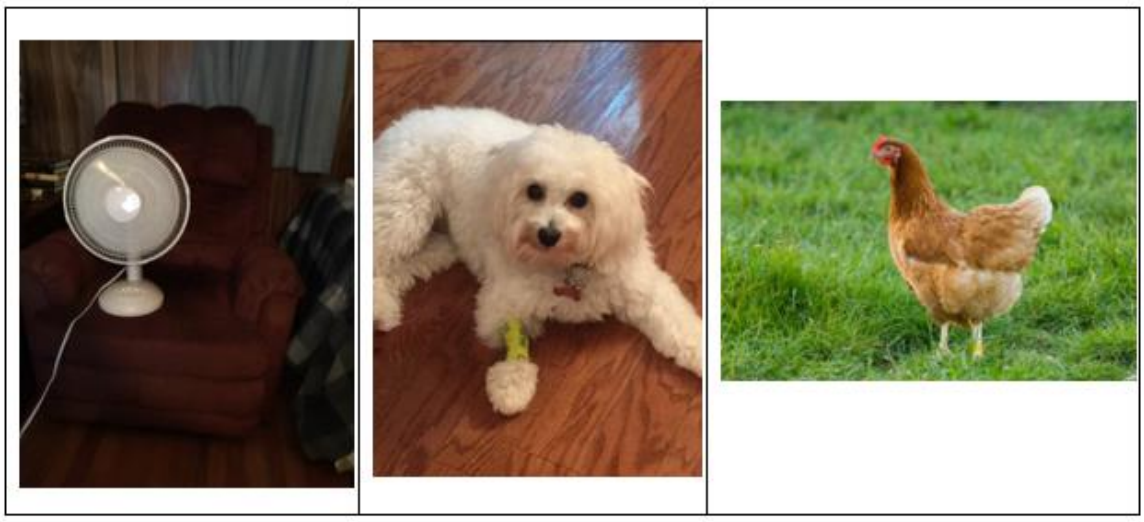
WeRateDogs https://t.co/N7sNNHAEXS is an open source twitter handle for professional dog ratings. This twitter account has received international coverage for its popularity and was created on November 2015, by college student Matt Nelson. Dogs are the most preferred pets and there are plenty of dog lovers worldwide who love to see their adorable pets get good ratings.

The ratings given by Matt Nelson on user-submitted pups almost ranks greater than 10 over a base denominator of 10(Yes, that's quite strange right!!) along with his witty, unique captions that just relates to dog's cuteness and yes thereby has gained over 5.6 Million followers. WeRateDogs shared their twitter archive with Udacity. This archive contained basic tweet information like tweetID, text, source etc. Additional gathering querying twitter's API was done to gather interesting data such as favourite and retweet counts for extensive analysis. After obtaining master dataset via cleaning, I began my analysis section by computing descriptive statistics.

Continuing with my project I started noting points and firstly, I noticed an outlier with a very large rating of 1776. This built a curiosity and I began extracting the details of this tweet. Turns out this dog is Atticus, he is quite simply America's most preferred dog. This tweet was posted on July 4, 2016 which is its Independence Day. The dog breed classifier couldn't predict the breed as Atticus was dressed up way too cool with the national flag, cool shades and other props. Here's the image of Atticus which I could download from the analysis.

I also investigated the data for most favourite dog. This is Stephan, he has received the most favourite count and highest retweet count. Dog's breed classifier correctly predicted this breed as chihuahua/Corgi mix with a prediction confidence of 0.51.

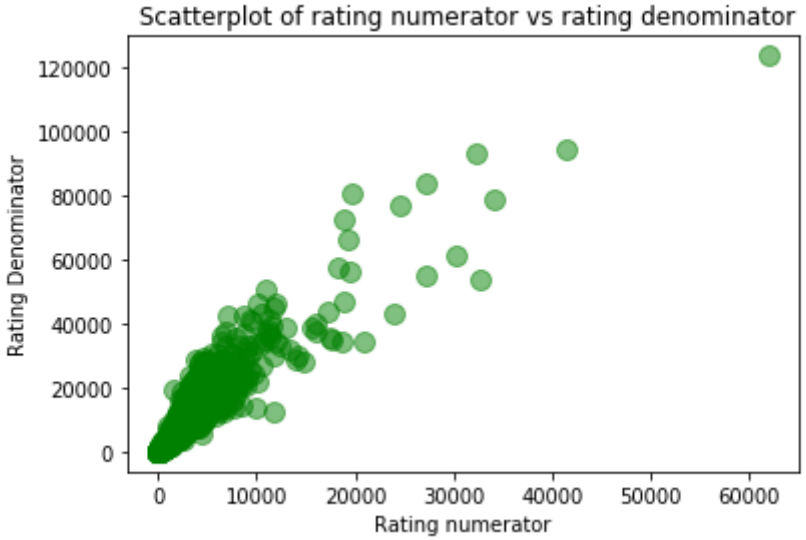
I also found out the dogs which got the lowest numerator rating of 1 and I have added the image below. Mainly the images do not show the actual picture of the dog and so the rating. Whereas the middle image is unluckiest as it is cute to look. It may be an outlier and the tweet did not have right hashtag or the twitter did not have any followers.



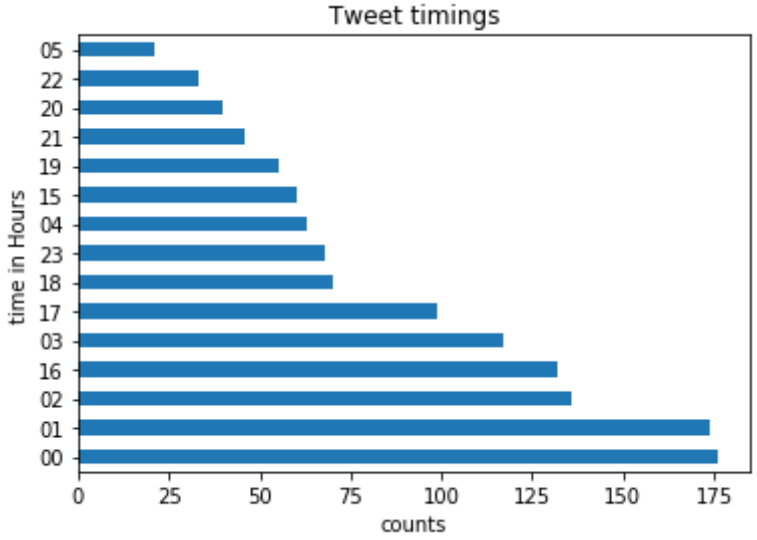
Statistical analysis

	rating_numerator	rating_denominator	img_num	p1_conf	p2_conf	p3_conf	favorite_count	retweet_count
count	1300.000000	1300.000000	1300.000000	1300.000000	1.300000e+03	1.300000e+03	1300.000000	1300.000000
mean	12.843077	10.545385	1.186923	0.587045	1.371542e-01	6.144363e-02	8350.230000	2560.027692
std	51.127955	7.871481	0.540562	0.273533	1.018995e-01	5.200750e-02	11512.741927	4084.210096
min	1.000000	2.000000	1.000000	0.044333	1.011300e-08	1.740170e-10	80.000000	13.000000
25%	10.000000	10.000000	1.000000	0.354718	5.440723e-02	1.649338e-02	1736.000000	593.000000
50%	11.000000	10.000000	1.000000	0.579762	1.203825e-01	4.961540e-02	3875.000000	1284.000000
75%	12.000000	10.000000	1.000000	0.836836	1.987905e-01	9.470035e-02	10345.250000	3043.250000
max	1776.000000	170.000000	4.000000	1.000000	4.676780e-01	2.710420e-01	123834.000000	61860.000000

I conducted statistical analysis on favorite and retweet count variables to investigate their relationship. The analysis confirmed a strong positive correlation infering that as favorite count increases, the retweet count increases too. This makes sense as people who retweet a tweet basically does that if they find it interesting and they like it. Here is a scatterplot of their relationship
Scatterplot confirms the strong correlation found between these two variables.



Further I began analysing the most popular time hours for people to tweet and I developed a histogram of most popular time to tweet by extracting hours. Poeple tend to tweet the most at 12am, 1am and 2am. This result was quite astonishing as I expected it to be during late evenings. Looks like internet dominates one's time.



1.5 Conclusion

This project aided me achieve a new skill of Data Wrangling which is a very important need of data science domain. I feel confident of wrangling massive datasets and generate a master dataset ready for inferential statistics and analysis.

My major findings are:

1. ‘Stephan’ as the most favourite dog with highest favourite count and retweet count because of its cool outlook and American connection
2. Few dogs also get very low rating and those should be because of tweet reach.
3. Favourite count and retweet count data are strongly positively correlated.
4. Midnight to 2a.m is the most popular time to tweet about dogs, which is quite surprising.