act_report

February 16, 2018

1 Analyzing and Visualizing WeRateDogs data after Wrangling process

1.1 By, Mithun Muralidhar

1.2 Feb 2, 2018

1.3 Introduction

WeRateDogs https://t.co/N7sNNHAEXS is a open source for professional dog ratings. This twitter account was created on November 2015, by college student Matt Nelson and has received international coverage for it's popularity. Ofcourse there are plenty of dog lovers worldwide and who not loves 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 favorite and retweet counts for extensive analysis.

After obtaining master dataset via cleaning, I began my analysis section by computing descriptive statistics.

With this information, 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 af... This tweet was posted on July 4, 2016 and we can recall this date when America declared it's Independence from Britain on July 4, 1776. The dog breed classifier couldn't predict the breed as Atticus was dressed up way too cool!!!

	rating_numerator	rating_denominator	favorite_count	retweet_count	p1_conf	p2_conf	p3_conf
count	1300.000000	1300.000000	1300.000000	1300.000000	1300.000000	1.300000e+03	1.300000e+03
mean	12.843077	10.545385	8348.503846	2561.196923	0.587045	1.371542e-01	6.144363e-02
std	51.127955	7.871481	11507.761964	4085.916003	0.273533	1.018995e-01	5.200750e-02
min	1.000000	2.000000	80.000000	14.000000	0.044333	1.011300e-08	1.740170e-10
25%	10.000000	10.000000	1734.750000	593.750000	0.354718	5.440723e-02	1.649338e-02
50%	11.000000	10.000000	3876.500000	1285.000000	0.579762	1.203825e-01	4.961540e-02
75%	12.000000	10.000000	10355.000000	3043.750000	0.836836	1.987905e-01	9.470035e-02
max	1776.000000	170.000000	123867.000000	61882.000000	1.000000	4.676780e-01	2.710420e-01

Descriptive statistics

title

Here is Atticus 

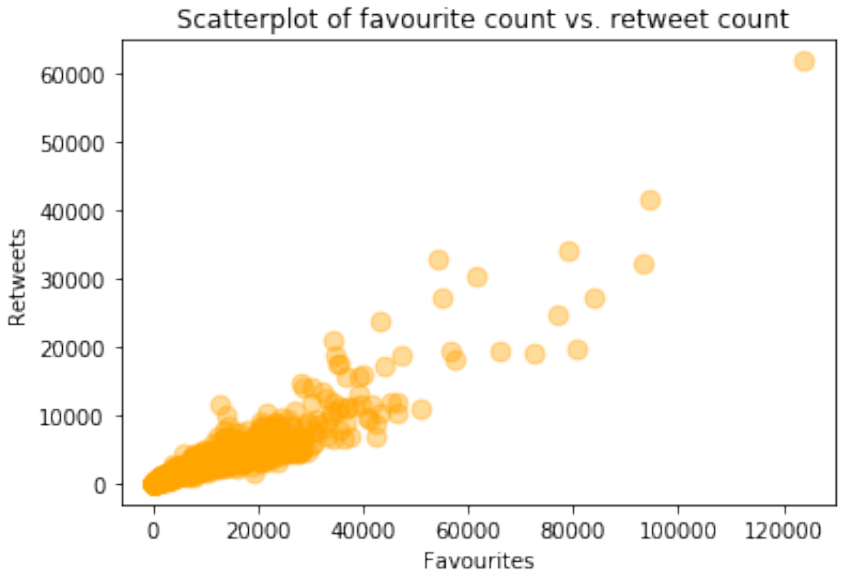
Next I investigated the data for most favorite dog. This is Stephan, he has received the most favorite count and also highest retweet count. Dog's breed classifier correctly predicted this breed as chihuahua/Corgi mix with a prediction confidence of 0.51.

1.4 Statistical analysis

I conducted statistical analysis on favorite and retweet count variables to investigate their relationship. The analysis confirmed a strong positive correlation of 0.92 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 corrleation found between these two variables.

Further I began analyzing the most popular month and time for people to tweet.I developed a histogram of most popular month and most popular time to tweet by extracting month and hour



scatterplot

data from the date and time columns respectively. Below is the histogram

As evident from the graph, it seems December is the most popular month to tweet followed by November and January. I guess this is the time of the year with most holidays like Thanksgiving(November), Christmas(December) where people generally tend to spend most of their time with families and home, giving more reasons to tweet about their dogs.

Here is the histogram of most popular time to tweet

Poeple tend to tweet the most at 12am, 1am and 2am. This result was quite surprising for me, as I expected it to be during late evenings. Looks like internet dominates one's time.

1.5 Conclusion

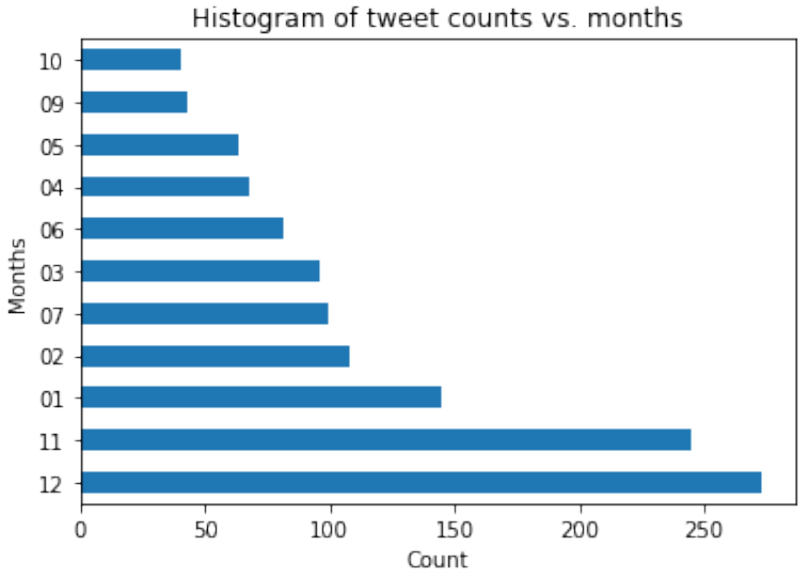
This project helped me achieve a new level of skill-set. I feel confident of wrangling massive datasets and generate a master dataset ready for analysis. I believe this is the core skill-set of any given professional data analyst.Some of the results obtained are

1)'Stephan' as the most favorite dog with highest favorite count and retweet count as well.

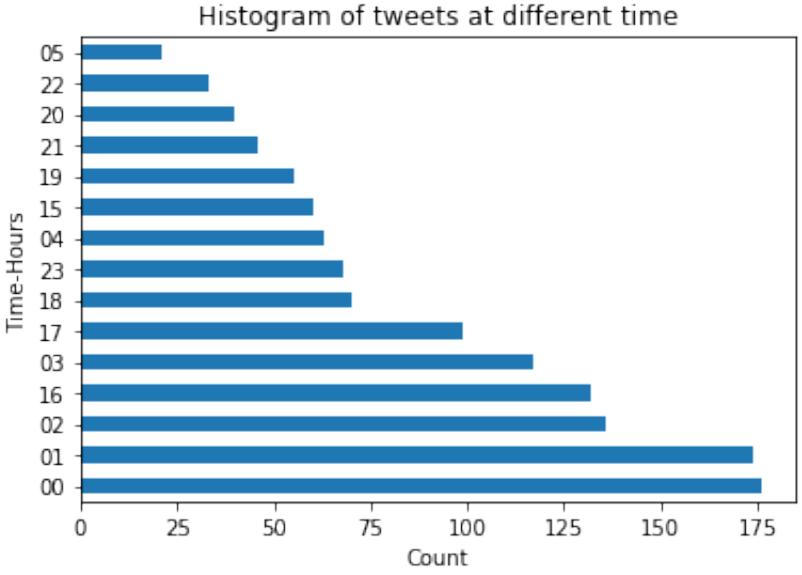
2)Favorite and retweet count data are strongly positively correlated

3)December is the most popular month to tweet followed by November and January (Winter season)

4)12am is the most popular time to tweet about dogs followed by 1am and 2am respectively(Late nights)


Histogram of tweet counts vs. months
Months
10
09
05
04
06
03
07
02
01
11
12
0
50
100
150
200
250
Count

title


Histogram of tweets at different time
Time-Hours
05
22
20
21
19
15
04
23
18
17
03
16
02
01
00
0
25
50
75
100
125
150
175
Count

title