act_report

January 24, 2018

1 Data Analysis and Visualization Report - 'WeRateDogs' twitter dataset

by L

1.1 Introduction

'WeRateDogs' is a twitter account which rates dogs. It has a huge fan following and people from all over the world send images. The rating used by 'WeRateDogs' is peculiar. Most of the times, the rating denominator is out of 10 (ofcourse, this alos has exceptions). The rating numerator is almost always greater than 10! 'WeRateDogs' also have good international coverage and their tweets are most of the times very humerous!

A typical tweet by the 'WeRateDogs' twitter account is as given below:

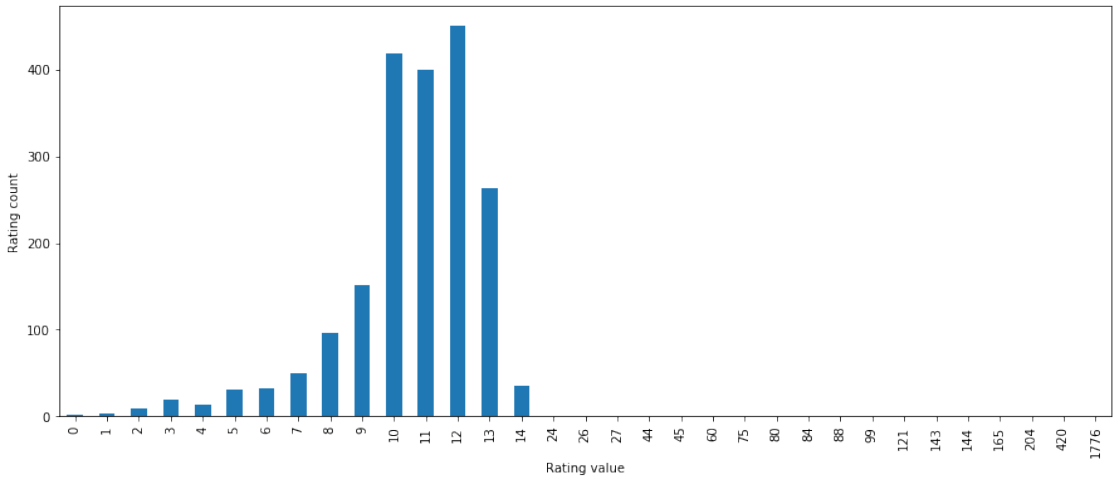
This report will document the analysis and visualizations performed on the wrangled and cleaned 'WeRateDogs' final dataset.

1.2 Analysis and Visualizations

1.2.1 Numerator Rating

The rating numerator is what sets 'WeRateDogs' aside. The denominator is almost always 10. Hence, let us look at the numerator. Plotting a bar chart of the numerator ratings:

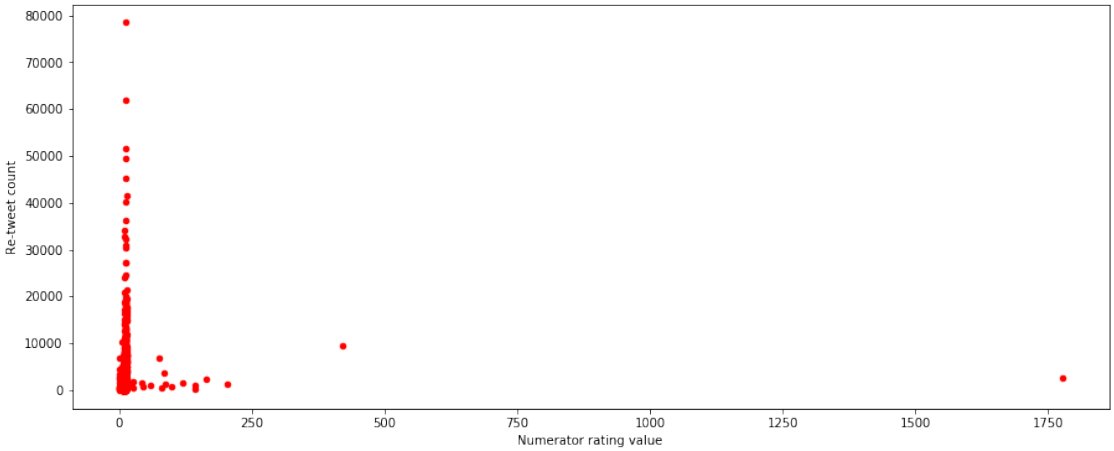
Out[6]: Text(0,0.5,'Rating count')



The above plot shows that, a dog image is most likely to receive a numerator rating of 10, 11, 12 or 13 with 12 being the most common rating value. There are quite a few ratings below 10 which are most likely not dog images. Also, there are some very highly rated images as can be seen from the plot as well as the value counts. it would be interesting to see the number of re-tweets for these images.

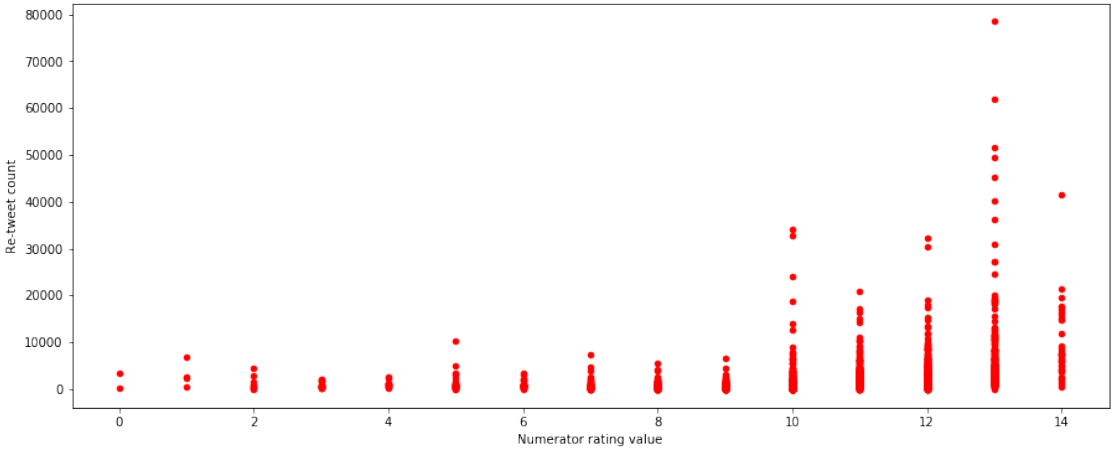
1.2.2 Numerator Rating vs Re-tweet Counts

Out[7]: `Text(0,0.5,'Re-tweet count')`



From the above plot, it is evident that the number of retweets does not correspond to rating value. Narrowing out the x-axis for more visual inspection:

Out[8]: `Text(0,0.5,'Re-tweet count')`



```
Tweet counts for each numerator rating:
                   tweet_id
                     count
rating_numerator
0                        2
1                        4
2                        9
3                       19
4                       14
5                       31
6                       32
7                       50
8                       96
9                      151
10                     419
11                     399
12                     451
13                     263
14                      36
24                       1
26                       1
27                       1
44                       1
45                       1
60                       1
75                       1
80                       1
84                       1
88                       1
99                       1
121                      1
143                      1
144                      1
165                      1
204                      1
420                      1
1776                     1
```

Most tweet ratings seems to be in the range 10 - 13.
From the plot above, most retweets seems to be for dogs from tweets which has a numerator rating of 13. This also shows that most humourous tweets have ratings of 13. There are only very few exceptions with higher rating greater than 13.

```
The maximum retweets for a tweet is:  78466
```

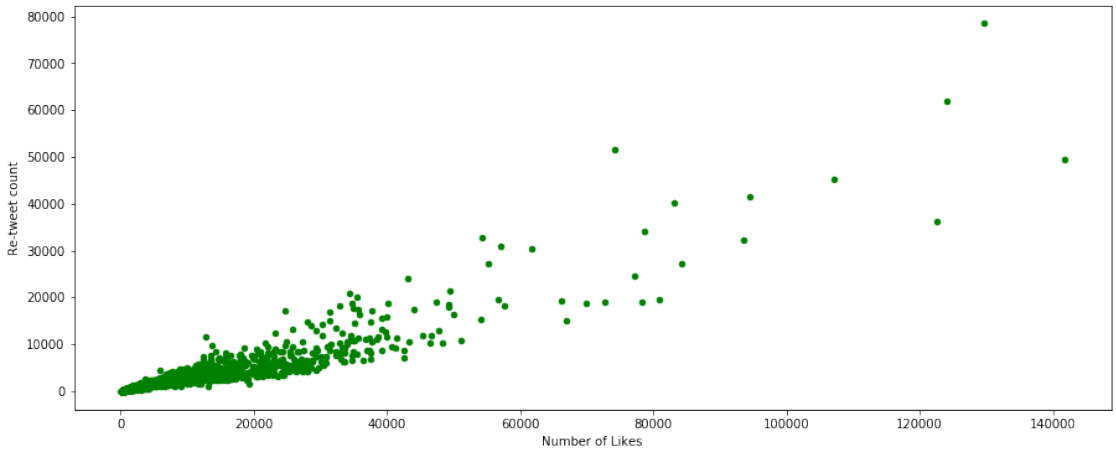
1.2.3 The most re-tweeted dog image is as given below with 78466 re-tweets!

This tweet is of a dog in a swimming pool. The rating for this tweet is only 13/10 but the text is so humourous that I believe that is the reason for the highest number of re-tweets for this tweet. One another reason could be that this tweet has a video of a dog swimming in the pool and then realising it can actually stand without swimming (and not an image).
The text for this tweet is - ***"Here's a doggo realizing you can stand in a pool. 13/10 enlightened af (vid by Tina Conrad)"***

I also found that the number of retweets mentioned in the tweet when I accessed this picture is only 78463, 3 re-tweets less than what it was when I extracted the data (78466). This doesn't seem to be an extraction error. Most probably this may be due to tweet deletion or other causes as twitter data is dynamic and subject to changes.

1.3 Number of Likes vs Re-tweet count

```
Out[13]: Text(0,0.5,'Re-tweet count')
```



The number of likes for a tweet is highly proportional to the number of retweets. This means a person is likely to retweet a tweet if he 'likes' it.

1.3.1 Tweet image with most likes

```
The maximum likes for a tweet:  141665
```

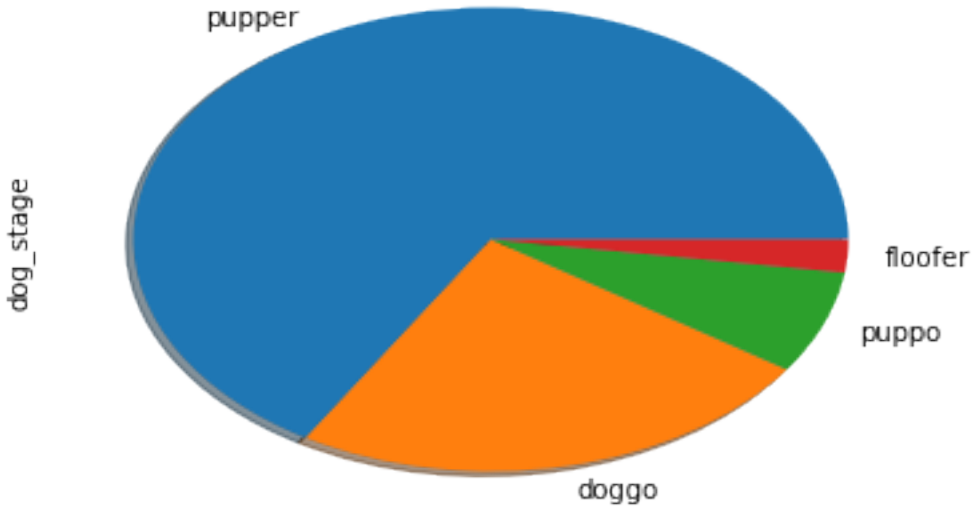
The above image has the most likes on weRateDogs with 141665 likes!!

1.4 Dog Stages

```
The various dog stage counts are:
 pupper     203
doggo       74
puppo       22
```

```
floofer      7
Name: dog_stage, dtype: int64
```

```
Out[17]: <matplotlib.axes._subplots.AxesSubplot at 0x7f52d36267f0>
```



From the plot above, most common dog stage in the weRateDogs tweets (from available data) is 'pupper.

1.5 Conclusion:

WeRateDogs twitter account rates dog images with a humourous text. It has a huge fan following around the world.
Since the data for thse visualizations were extracted from twitter, it required a lot of wrangling effort to geta clean dataset that could be used for the above analysis and visualization.

1.6 Future Work

A lof of more wrangling needs to be done. Some of the interesting things that I intent to do in the next version are as follows:
1. Analyze the dog breeds and find some statistics like breed with most likes, most retweets, most rating, most common breed etc.
2. Bar plots for numerator ratings for each breed

1.7 References:

1. http://pandas.pydata.org/pandas-docs/version/0.15.0/visualization.html