# Data Wrangling Project Report

*By Nan Lin*

*Date: Feb 18, 2018*

*[550] words*

In this project, my goal is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualization. WeRateDogs is a very popular Twitter account with over 5.68M followers and has received more than 130K likes. WeRateDogs gained its popularity by rating people's uploaded image of their dogs and gave a humorous comment about the dog. Their rating system is based on fraction with the denominator fixed at 10 and the numerator is always greater than 10 because "They are good dogs Brent".

The data are from three different sources:

- Enhanced Twitter archive in the form of a CSV file. This file contains basic tweet information such as tweet ID, timestamp, text, etc.
- Additional JSON data extracted via the Twitter API using the Python's Tweepy library based on the tweet IDs from above archive.
- The neural network predictions results downloaded programmatically using the Requests library. This table presents the classification results of each image alongside each tweet ID, URL, and the confidence level.

After iteratively assessing and cleaning data from these three sources, I conducted exploratory data analysis on the cleaned dataset. Firstly, I programmatically assess the statistic of the different column:

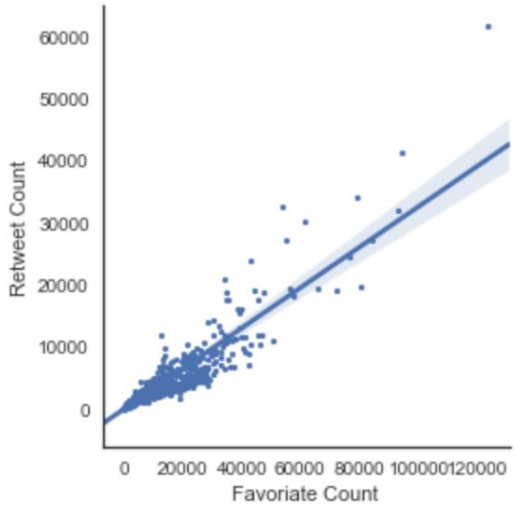
| | rating_denominator | numerator | img_num | p1_conf | p2_conf | p3_conf | favorite_count | retweet_count |
|---|---|---|---|---|---|---|---|---|
| **count** | 1289.0 | 1289.000000 | 1289.000000 | 1289.000000 | 1.289000e+03 | 1.289000e+03 | 1289.000000 | 1289.000000 |
| **mean** | 10.0 | 12.181164 | 1.187742 | 0.586774 | 1.376677e-01 | 6.155534e-02 | 8387.331265 | 2564.332040 |
| **std** | 0.0 | 50.517051 | 0.542142 | 0.273681 | 1.020348e-01 | 5.214194e-02 | 11556.047600 | 4094.384441 |
| **min** | 10.0 | 1.000000 | 1.000000 | 0.044333 | 1.011300e-08 | 1.740170e-10 | 80.000000 | 13.000000 |
| **25%** | 10.0 | 10.000000 | 1.000000 | 0.354674 | 5.478680e-02 | 1.640380e-02 | 1723.000000 | 590.000000 |
| **50%** | 10.0 | 11.000000 | 1.000000 | 0.578120 | 1.208530e-01 | 4.990060e-02 | 3875.000000 | 1277.000000 |
| **75%** | 10.0 | 12.000000 | 1.000000 | 0.836572 | 1.995440e-01 | 9.475920e-02 | 10437.000000 | 3051.000000 |
| **max** | 10.0 | 1776.000000 | 4.000000 | 1.000000 | 4.676780e-01 | 2.710420e-01 | 123484.000000 | 61590.000000 |

There are some interesting observation I gained from this chart:

- Mean rating for a dog image is 12.843/10 with first quartile in 10/10 and third quartile in 12/10
- Mean favorite count is 8387 with maximum value of 123484; The median favorite count is 3875
- Mean retweet count is 2564 with maximum value of 61590; The median retweet count is 1277

- The neural network algorithm has the highest confident in the first iteration with mean confident of 0.5868

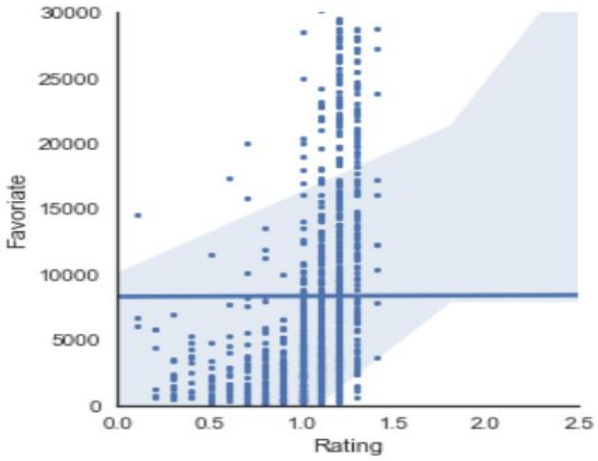
Based on the statistical analysis, We have a clear overview of this dataset. Secondly, I would dive deeper into the relationship between retweet count and favoriate count and additionally, how the ratings(calculated as rating numerator/rating denominator) changed over time.




From this visual, we can see a strong linear correlation between favorite and retweet count, which makes a lot of sense because people are more likely to retweet when they like this tweet. Retweet and favorite both present the popularity of a tweet. Furthur, in our dataset, the tweet with the most favorite also get the highest retweet count. It's a very funny video. You can check it out through the link below:

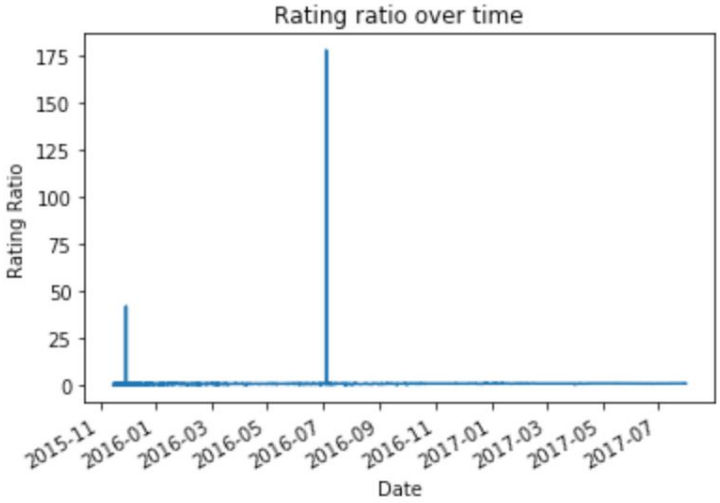

https://twitter.com/dog_rates/status/807106840509214720?lang=en

Next, I tried to answer the question that whether higher-rated dogs get more favorites in WeRateDogs by plotting a linear relationship between favorite count and rating ratio:

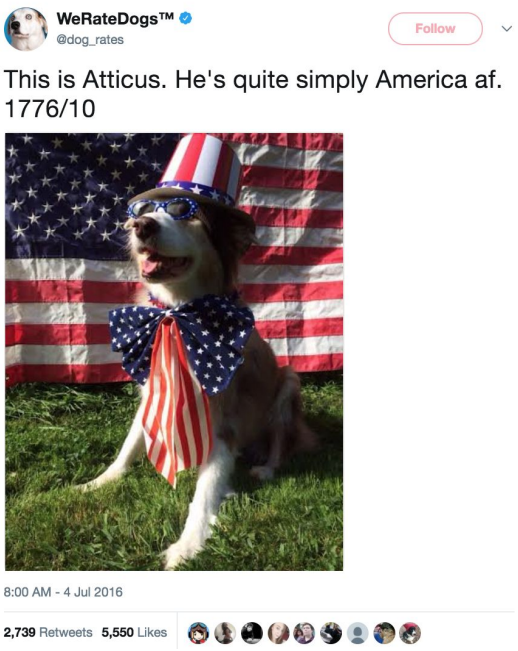



It appears no linear relationship between a dog's popularity and its rating ratio. Further analysis is needed to draw a conclusion about the relationship between these two variables.

Last but not least, I visualize the rating ratio over tweet times:

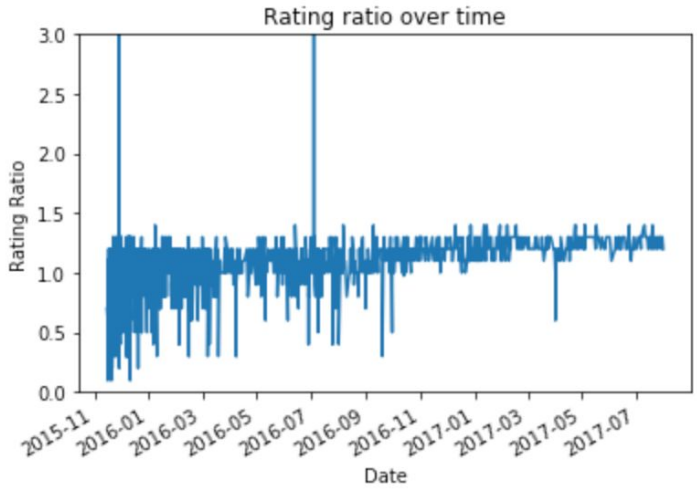



Basically, the rating ratio is within the range of 0 to 5 with two outliers. The maximum value of the rating is 177.6. The second maximum value is 42.0.
If we take a look at the images in these two tweets, we can find very interesting facts:





1776 is the year when America declared its independence from Britain.
On the right is Snoop Dogg, a famous singer.
Thus these two abnormally high rating was given for comic effect. Now let's limit the y-axis to see the trend in detail:




Based on this visual, we see that a few dogs have scores of close to zero. There are lower scores were given in the relatively earlier time. A majority of scores falls between 1.0 to 1.5. Over time, the scores tend to increase.