Ehsan Jafari-Shirazi

Wrangle and Analyze Data Project –
Wrangling Twitter Archived Dataset: WeRateDogs

The dataset used in this project is an archived dataset in Twitter known as WeRateDogs. WeRateDogs is a very popular Twitter account which has received international media coverage. The account is used for rating dogs (based on the shared picture) along with a humorous comment about the dog. The fun part is the way the ratings are assigned, as the numerators are almost always greater than 10, versus the denominator is 10 for the most part.

This report is to analyze the wrangled version of the dataset queried from the Twitter account.

I have done the analysis in two parts:

1- Providing a few Insights by looking at some descriptive statistics.
2- Using Visualization within matplotlib in Python to obtain some insights about the trend of Dog Ratings over time.

Part 1: Insights obtained through descriptive statistics:

Using some simple descriptive statistics such as mean, min and max, the following insights were obtained:

- Mean rating numerator is 12.85 with an outlier of 1776.

- Mean rating denominator is 10.54 (median of 10) with an outlier of 170.

- Mean favorite count is 8344.6 and a maximum value of 123490.

- Mean retweet count is 2552.8 and a maximum value of 61590.

- The neural network performed the best on the 1st iteration with a mean prediction of 0.587.

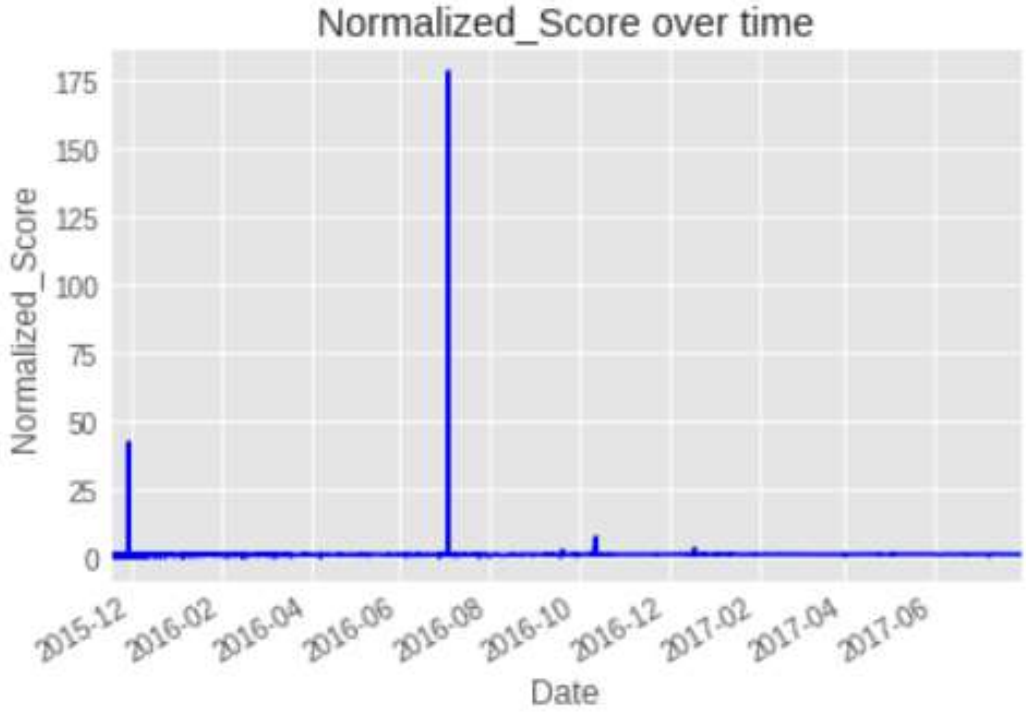
Investigating the rating_numerator outlier, we can find out that it's belonged to a Dog named Atticus. His picture can be seen through the following link:

https://twitter.com/dog_rates/status/749981277374128128/photo/1

Furthermore, by investigating the maximum value for favorite count and retweet count, it can be found that *the same dog named Stephan has both the highest favorite and retweet counts.*

Part 2: Insights obtained through Visualization:

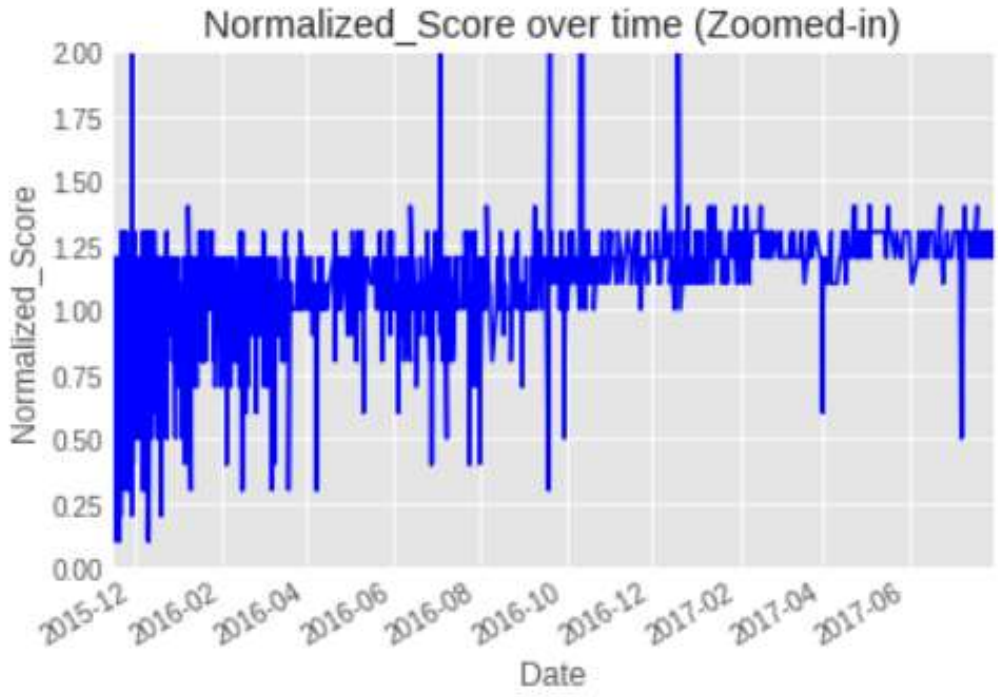
First I made a new variable called "Normalized_Score" that normalizes the rating by dividing the numerator over the denominator, and then the following graph was plotted to illustrate the trend of the Normalized_Score variable over time:



As it can be seen in the graph above, there are a couple of outliers. Specifically, *the largest outlier (Normalized_Score > 175) is easily recognizable*. This was also

identified in part 1 (descriptive statistics). The dog name is Atticus and we could see his picture as well.

To be able to see the trend of the data over time the outliers needs to be masked in the graph and to do so, the graph can be zoomed by limiting the y-axis in the range of 0 to 2:



From the above graph, it can be observed that *the scores (ratings) have been improved over time.* Besides a few outliers, *the Normalized Score mostly is above 1 (i.e. Numerator Rating is greater than 10) for the most recent data.* It can be seen that *there were many scores less than one and close to zero in the beginning, for the time period of late 2015 to early 2016.*