WeRateDogs (@dog_rates) – A Twitter Archive Analysis

Date:10/8/2020

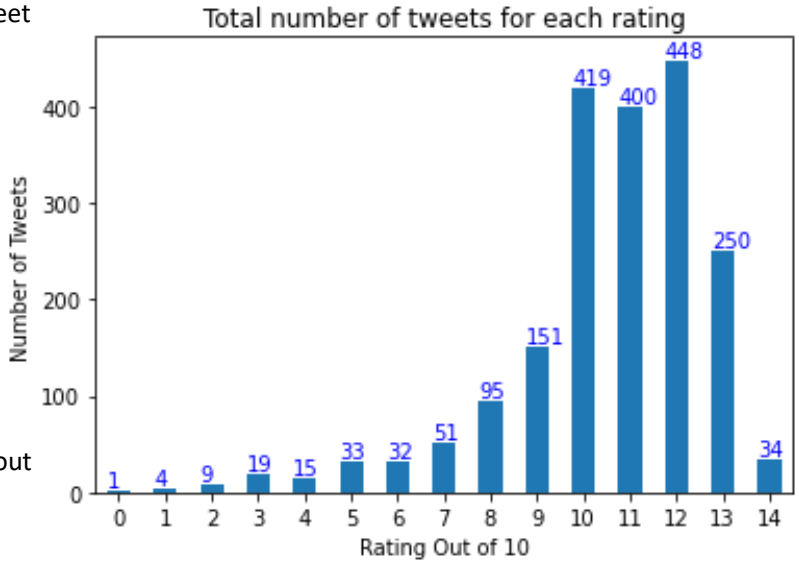
Written by: Miles Murphy

WeRateDogs may be one of the most valuable resources of modern times. The amount of joy that this Twitter Account gathers, enhances, and redistributes to the rest of the world is remarkable. Their humor may be extremely dog-centric, even 'dog'matic at times, but they fully embrace the mission of uniquely reviewing and rating every dog which comes their way.

As a Udacity Data Analysis student, I was tasked with the responsibility of taking the WeRateDogs Twitter archive, merging it with two other gathered datasets (one from a machine learning course and one from the Twitter API), and then thoroughly cleaning the combined dataset for at least 2 tidiness and 8 quality cleaning issues. Personally, I was not satisfied with those minimums and proceeded to clean further.

To initiate this analysis these three datasets were first uploaded into a Jupyter Notebook for analysis to be conducted in a pandas dataframe. These flexible dataframes allow for extensive data storage and analysis in many ways and the notebook provides convenient, instantaneous results of each sequential step. However, it does take a little bit of work to understand how Jupyter Notebooks work with GitHub. I felt that it was important to force myself to keep using every skill which this course and my previous Nanodegree taught, so I stepped outside of the convenient Udacity-based project workspace and created my own local repository, linked it to GitHub, and began my analysis there.

After a few data tidiness steps and 20 data quality steps, the original WeRateDogs twitter archive was much more concise and ready to provide some interesting insights into its rating tendencies and other details. It was condensed from 2356 tweets which featured images down to 1961. Though, before any analysis is of these 1961 tweets is presented, the extensive data cleaning steps taken to reduce the amount of irrelevant or incorrect (and not easily fixable) tweet data should be briefly commented on. Particular attention was paid to remove tweets which were retweets, replies, did not have dog images, or had ratings which were not specific to a dog(s)

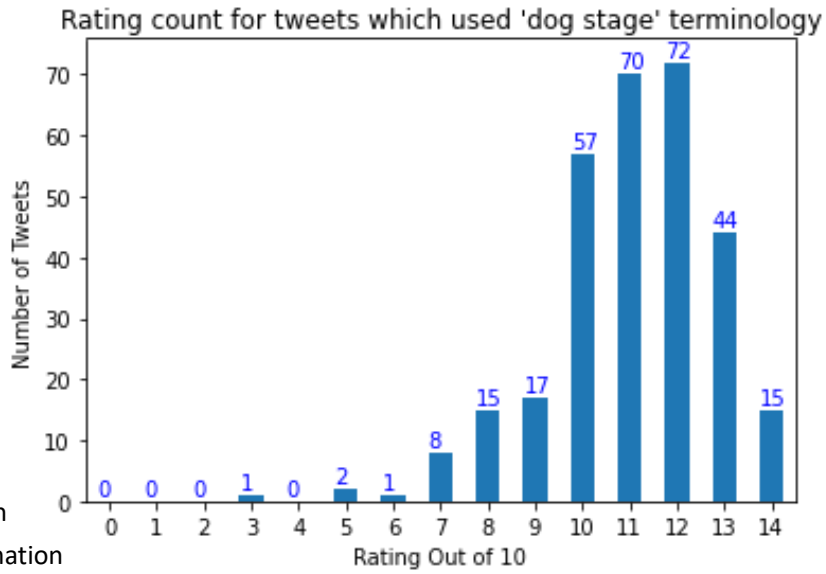


Now, let us look at some insights which were drawn from the remaining tweets. First, the average rating for all these good dogs is 10.54 out of 10 and the rating count's distribution/range can be seen in the adjacent bar graph.

Next, a few other insights related to the clean dataset were generated.

- How many tweets were on the lower end or upper end of the rating scale?
    - Of the 1961 tweets, 1180 or 60.17% of the tweets have ratings lower than 5/10 or greater than 10/10.
- How often did the image prediction pick a dog on its first attempt to match the image?
    - It picked a dog, though maybe not the correct dog, with an average confidence of 0.594071
- How many tweets featured ‘dog stage’ terminology and what percentage of the total number of tweets?
    - 302 of the 1961 tweets or 15.40% of the total featured dog stage terminology
- What is the most popular dog stage term for those tweets which utilized the terminology?
    - Pupper is the most popular dog term with 201 individual uses (209 total), followed by doggo with 62 individual uses (72 total)

Finally, the information you have all been waiting for. It was clear, from a visual analysis of the tweets containing dog stage terminology, that tweets which used ‘dog stage’ terminology generally received ratings of 10 or higher, as seen in the following bar graph.



As interesting as some of these insights may be, they, and the visualizations represent, are only the tip of the iceberg. There are several additional cleaning steps which could further refine the data. For example, removing tweets which do not have image prediction information matching a dog would further slim the dataset, though it may also remove some legitimate dog photos which are misidentified. Additional analysis can also be performed on the current state of the clean dataset. For instance, a key factor which is not included in this brief analysis is an examination as to whether the ratings of tweets were higher with or without the ‘dog stage’ terminology.

Though, regardless of what nature of data wrangling took place surrounding the twitter archive, we can assuredly state that WeRateDogs features some information packed and humors tweets which tend to keep both their followers and aspiring data analysts entertained for hours on end.