12/03/2021

ACT-REPORT

On Analyzing and Visualizing WeRateDogs Twitter Account

By: Ahmed Mohamed Soliman

Introduction:

Real-world data rarely comes clean. The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This project works through the data wrangling process, focusing on the gathering, assessing and cleaning of data. There are visualization and observation from the analysis provided as well.

-The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Gather:

This project gathered data from the following sources:

• The WeRateDogs Twitter archive. The 'twitter-archive-enhanced.csv' file was provided to Udacity Students. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 6000+ of their tweets as they stood on August 1, 2017.

• The Tweet image prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity Students.

• tweet_json.txt to gather each tweet's retweet count and favourite ("like") count at minimum and any additional data I find interesting ,because i found alot of trouble talking to twitter support about having thier API so i handeld the project with the tweet_json file.

Assessing Data:

Once the data was gathered, I began to assess the data on both quality and tidiness issues:

There are four main issue in quality dimensions:

1. Completeness: Missing data

2. Validity: Does the data make sense

3. Accuracy: Inaccurate data

4. Consistency: Standardization

And There are three main requirements for tidiness:

1. Each variable forms a column

2. Each observation forms a row

3. Each type of observation unit forms a table

Clean:

Cleaning data is tedious and often iterative. Just when data analyst believe they found all quality and tidiness issue, they often found additional issue arises. The cleaning process involves three steps:

1. Define: Determine exactly what needs to be clean and how.

2. Code: Programmatically clean the code

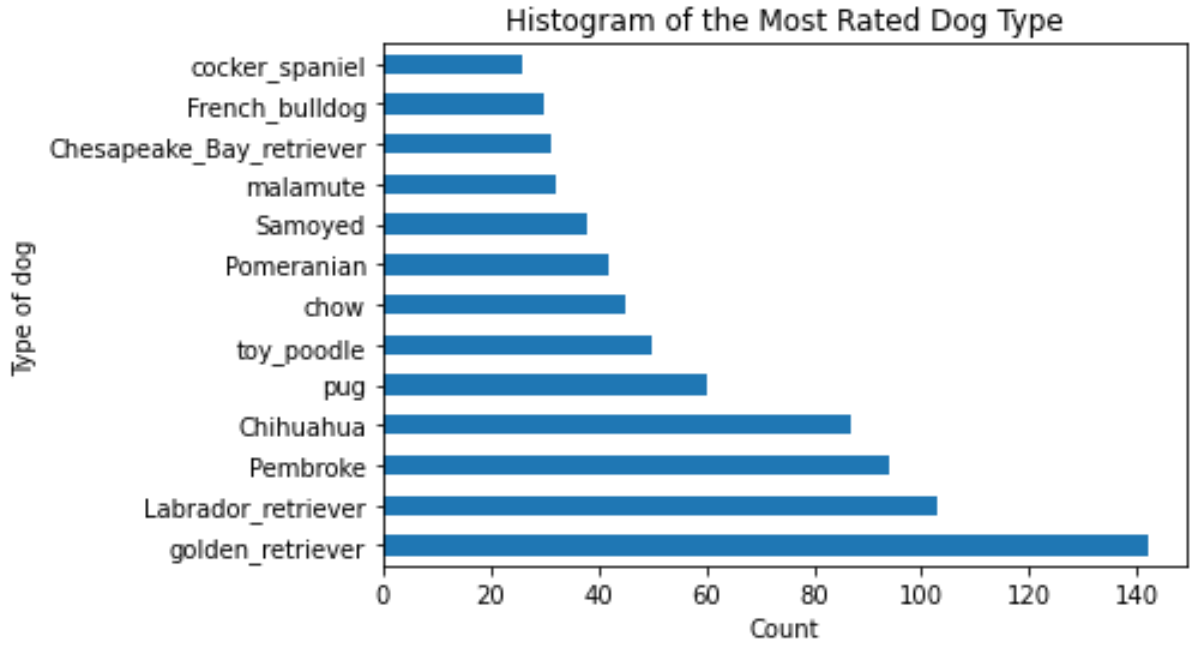
3. Test: Evaluate the code to ensure the data set was cleaned properly.

Analysis and Visualization:

There are serval analysis, which I have done and those are in following:

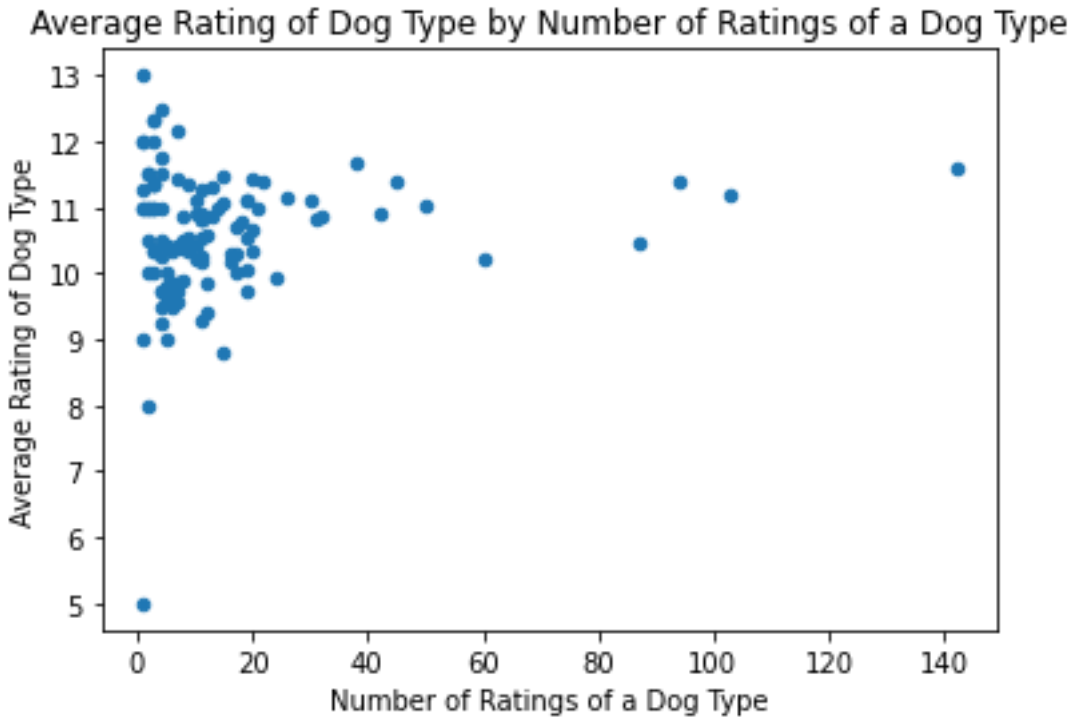
1- Most common dog type

WeRateDogs has over 6000+ tweets. I was able to analyzed around 1500+ tweets. The most rated dog was golden retriever with more than 140 ratings according to the following graph.



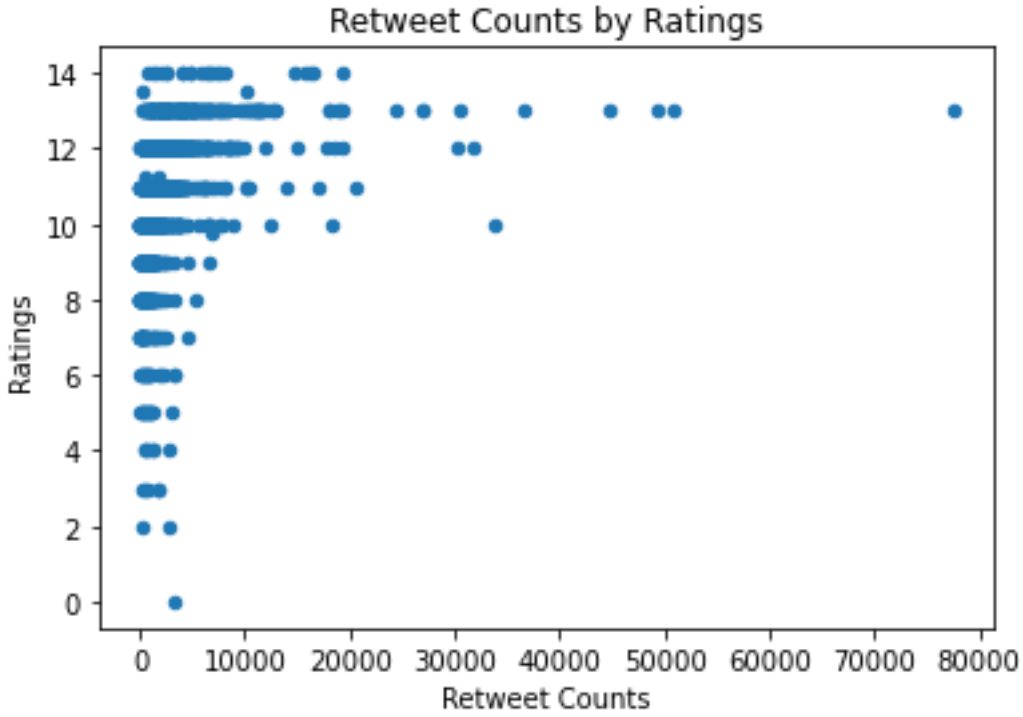
2- Lowest average rating among dog types

According to the graph The blue dot on the very right represents the approximately 140 ratings of golden retriever. On average these ratings are high 11.59. This really shows that the golden retriever is very popular.



3- Retweet counts

Amazingly many tweets have been retweeted more than 1000 times some more than even 50000 times. There is not a clear relationship between the ratings and the retweets.



Conclusion:

The write up offers a straight look at the data wrangling process. There is so much more that can be done with this data set