Report Data Wrangling

Project Data wrangling

WeRateDogs

By

Shahad Alotaibi
Email:shahadalotaibi9@gmail.com

Data Analysis Nanodegree Program

2020-2021

Project Overview

Real-world data rarely comes clean. Using Python and its libraries, you will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. You will document your wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries) and/or SQL.

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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.
WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

The following packages (libraries) need to be installed. You can install these packages via conda or pip. Please revisit our Anaconda tutorial earlier in the Nanodegree program for package installation instructions.

- pandas
- NumPy
- requests
- tweepy
- json

Twitter archive file : twitter-archive-enhanced-2.csv

Enhanced Twitter Archive
The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which I used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, I have filtered for tweets with ratings only (there are 2356).

The tweet image predictions, i.e.This file (image_predictions.tsv)

Image Predictions File
One more cool thing: I ran every image in the WeRateDogs Twitter archive through a neural network that can classify breeds of dogs*. The results: a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

The last process was dealt with after the process of taking the backup copy of all the three files and working on cleaning the data with the backup copy, and then creating a third file that combines the three files together to get one file that contains all the information without disintegration and access to the meaning of this data Between photos and tweets
This is the information that the file contains after merging the three files together using Tweet ID

Merage All Dataset into one file

In [124]:
```
File_copy.describe()
```

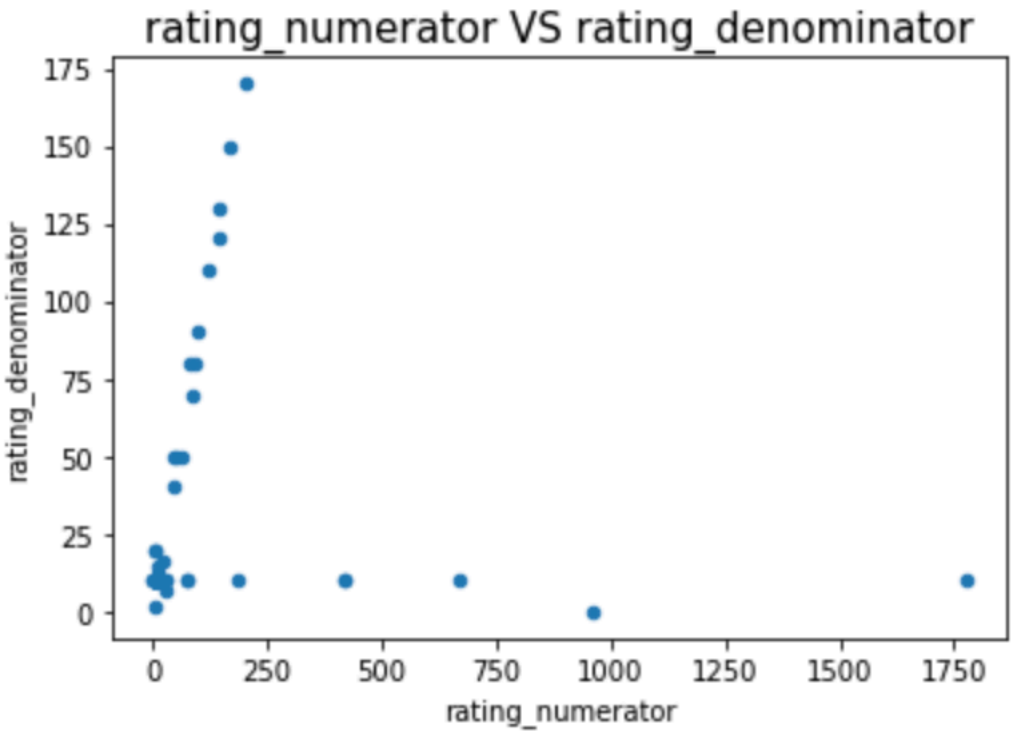
Out[124]:

	tweet_id	in_reply_to_status_id	in_reply_to_user_id	retweeted_status_id	retweeted_status_user_id	rating_numerator	rating_denominator	favorite
count	2.356000e+03	7.800000e+01	7.800000e+01	1.810000e+02	1.810000e+02	2356.000000	2356.000000	2354.00000
mean	7.427716e+17	7.455079e+17	2.014171e+16	7.720400e+17	1.241698e+16	13.126486	10.455433	8080.96856
std	6.856705e+16	7.582492e+16	1.252797e+17	6.236928e+16	9.599254e+16	45.876648	6.745237	11814.77133
min	6.660209e+17	6.658147e+17	1.185634e+07	6.661041e+17	7.832140e+05	0.000000	0.000000	0.00000
25%	6.783989e+17	6.757419e+17	3.086374e+08	7.186315e+17	4.196984e+09	10.000000	10.000000	1415.00000
50%	7.196279e+17	7.038708e+17	4.196984e+09	7.804657e+17	4.196984e+09	11.000000	10.000000	3603.50000
75%	7.993373e+17	8.257804e+17	4.196984e+09	8.203146e+17	4.196984e+09	12.000000	10.000000	10122.25000
max	8.924206e+17	8.862664e+17	8.405479e+17	8.874740e+17	7.874618e+17	1776.000000	170.000000	132810.00000

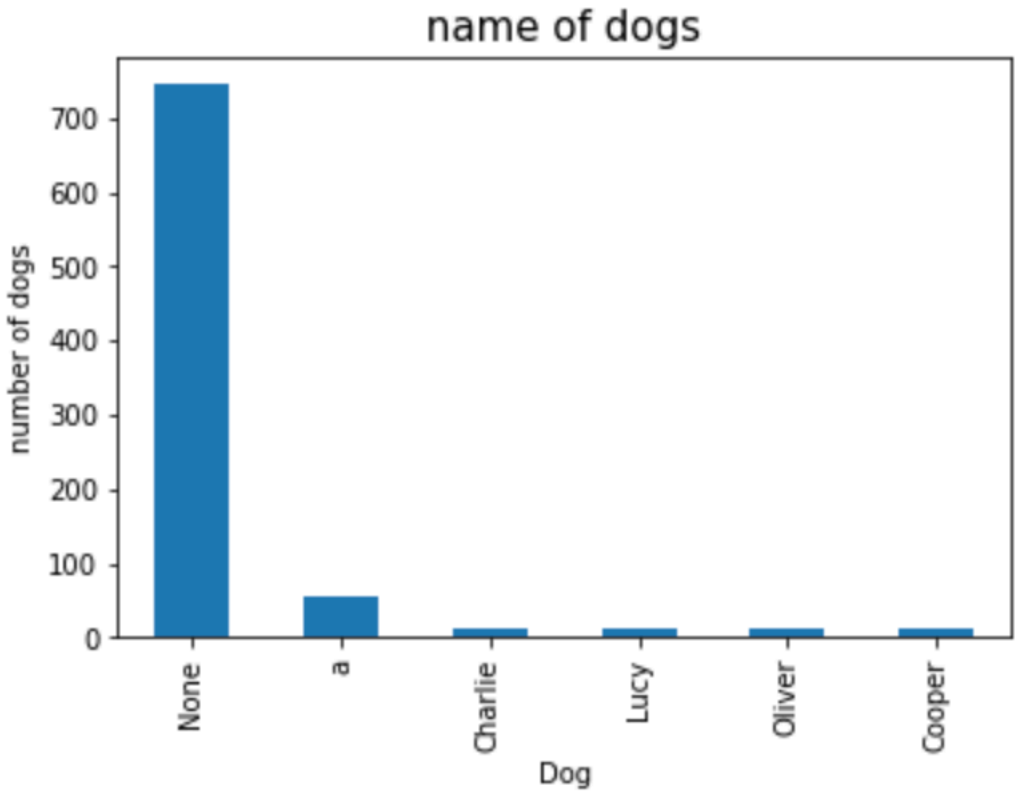
Out[124]:

ed_status_id	retweeted_status_user_id	rating_numerator	rating_denominator	favorites	retweets	img_num	p1_conf	p2_conf	p3_conf
.810000e+02	1.810000e+02	2356.000000	2356.000000	2354.000000	2354.000000	2075.000000	2075.000000	2.075000e+03	2.075000e+03
.720400e+17	1.241698e+16	13.126486	10.455433	8080.968564	3164.797366	1.203855	0.594548	1.345886e-01	6.032417e-02
.236928e+16	9.599254e+16	45.876648	6.745237	11814.771334	5284.770364	0.561875	0.271174	1.006657e-01	5.090593e-02
.661041e+17	7.832140e+05	0.000000	0.000000	0.000000	0.000000	1.000000	0.044333	1.011300e-08	1.740170e-10
.186315e+17	4.196984e+09	10.000000	10.000000	1415.000000	624.500000	1.000000	0.364412	5.388625e-02	1.622240e-02
.804657e+17	4.196984e+09	11.000000	10.000000	3603.500000	1473.500000	1.000000	0.588230	1.181810e-01	4.944380e-02
.203146e+17	4.196984e+09	12.000000	10.000000	10122.250000	3652.000000	1.000000	0.843855	1.955655e-01	9.180755e-02
.874740e+17	7.874618e+17	1776.000000	170.000000	132810.000000	79515.000000	4.000000	1.000000	4.880140e-01	2.734190e-01

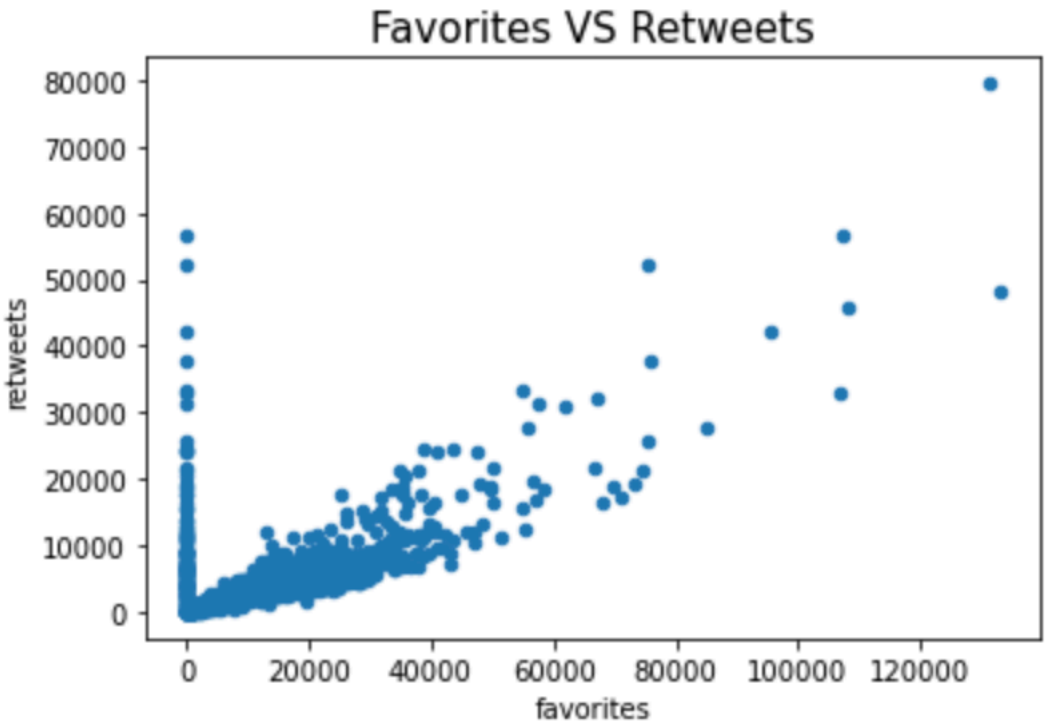
analyzing and visualizing of data



We see here rating_numerator that a ratio is close and in a sequence other than a rating_denominator ratio, it is far apart and there is a difference in its degrees



Here it becomes clear to us the types of animals of dogs and the percentage of their presence



Here it becomes clear to us the difference between liking and the number of retweets, the difference between them and the extent of their use

Summary

Download the data to be analyzed The files that must be viewed and made commensurate with each other, because they serve one account, and they are three files that were previously explained, namely (Twitter archive file: download this file manually by clicking the following link: twitter_archive_enhanced.csv The tweet image predictions, i.e.This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv Twitter JSON)

View the data and know its characteristics through the use of many Python programs in pandas Find and delete duplicate data Query for data that does not contain value Delete data that does not contain a value Describe data that contains individual values and make use of them Know the quality of the image data used in the image data Delete the columns that do not lead to a cognitive value in analyzing the data or that are not useful in the analysis.

Rather, it is considered an obstacle in the analysis. If there are many columns, we must focus on the columns that lead to a result in their values and be used.