the WeRateDogs Project

An analysis of data wrangled from twitter

Project Details

Your tasks in this project are as follows:

- Data wrangling, which consists of:
- Gathering data (downloadable file in the Resources tab in the left most panel of your classroom and linked in step 1 below).
- Assessing data
- Cleaning data
- Storing, analyzing, and visualizing your wrangled data
- Reporting on 1) your data wrangling efforts and 2) your data analyses and visualizations

Gathering data

the gathering data that is first and important step in data wrangling .

1-getting data from file 'twitter-archive-enhanced.csv' using pandas.

2-download file image_predictions.tsv programmatically using requests.

3-read the json file (tweet_json.txt)

WeRatejDogs gave Udacity exclusive accssto their Twitter archive for this project in the form of a csv file . this archive contains basic tweet data(tweetID , timestamp, text) they stood on Aug 1,2017 . Each tweet image was run through a convolutional neural network with the purpose of analyzing the images to correctly identify the dog breeds . the convolutional neural network with the purpose of analyzing the images to correctly identify the dog breeds . the convolutional neural network predictions were programmatically downloaded using the requests python

library as a tsv file .And finally ,using the tweet IDs from the WeRateDogs archive I queried the Twitter API for each tweet's JSON data using the python 's Tweepy library I stored each tweet's entire set of JSON data , whichIwould later use to analyze the tweet's retweet and favorite counts

Assessing data

Quality

1- Delete columns that won't be used for analysis

2-unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone

3-Delete unreasonable rate rows at the column that has the value of (rating_numerator / rating_denominator) 'rate'

4-Rating_denominator should have 10 .

5-Remove outlier rating

6-Remove all un-original tweets (retweets).

7-Change datatype of tweet_id column to a string and Change datatype timestamp column to data time.

8-Change missing values in 'name' from 'None' to NaN (dog stages already covered).

Tidiness

1-Dog "stage" variable in four columns: doggo, floofer, pupper, puppo

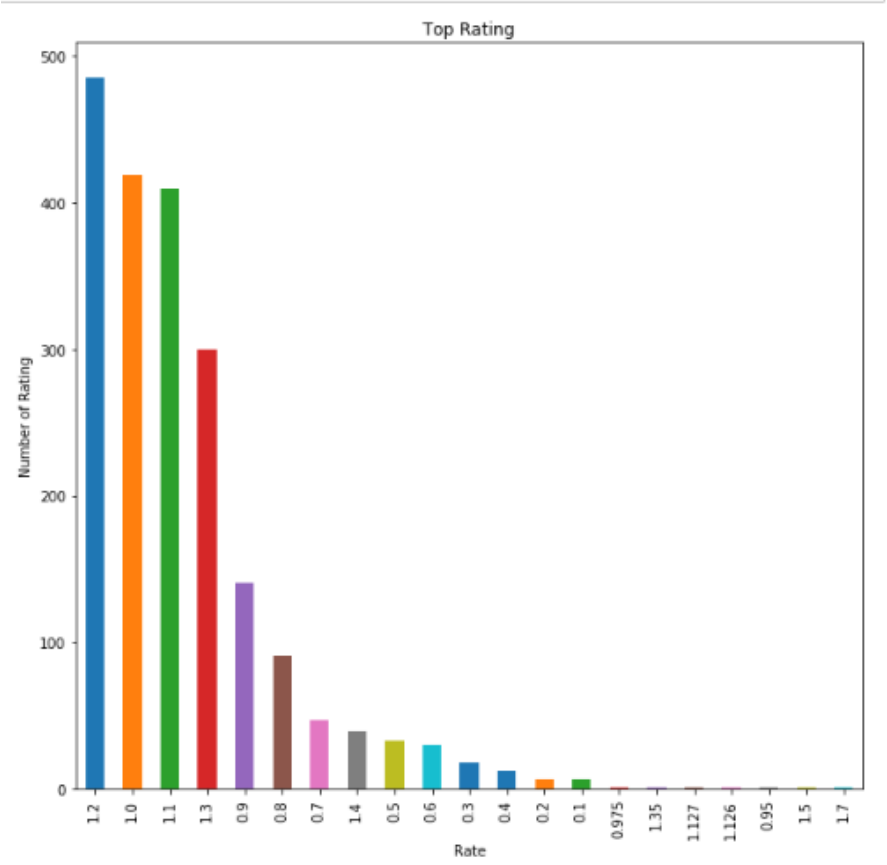
2-Join 'json_tweets' and 'arc_df'

Cleaning Data

You should take copy for each file to clean it .

used basic python function like duplicates, drop ,sort , value_counts and others to comply with above mentioned point. I struggle with few issues and had to spend a lot of time to get my understand .

Visualizing



The most rate is 1.2

Storing Visualizing Data for this Project

Store the clean DataFrame(s) in a CSV file with the main one named

twitter_archive_master.csv . If additional files exist because multiple tables are

required for tidiness, name these files appropriately. Additionally, you may store the cleaned data