Wrangle Report

Author: Omar Abdelaziz

Date: October 1, 2020

The report describes the efforts of the Data Wrangling process in WeRateDogs project which consists of three steps:

1. Gathering.
2. Assessing.
3. Cleaning.

Note: This is a project of the Udacity's Data Analysis Professional Nanodegree.

Gathering

The project consists of three datasets from three different sources. The first dataset contains information about tweets gathered from twitter archive from WeRateDogs account and contains the basic information of tweets. Although it's a large dataset, it was missing some information such retweet counts and favorite counts. The second dataset contains image predictions of the dog breed coming from a neural network on some of the tweets of that twitter archive. The third and the last dataset was file contains the data queried from Twitter's API using a Python Library called Tweepy to obtain further information about the tweets that stored in the archive file using TweetID.

The first two datasets are downloaded programmatically using a Python library called Requests and the files stored in a folder inside the project directory called 'WeRateDogs/resources' .

The third dataset is created after querying through Twitter API using Tweepy, the process was as follows, I created a list of TweetIDs stored in the twitter archive and for each TweetID, I managed to get the status's json object using get_status() method (of the tweepy.API class) and stored each tweet in a file called tweet_json.txt. Then, stored the file in the following directory 'WeRateDogs/exports' and through the program, I opened and read each line of the file as Python Dictionary and stored the Tweets in a Python List. Finally, the third dataset was created and filtered to columns of interest such tweetID, favorite count and retweet count and saved the file in the exports' directory.

The process of creating directories and save files in these directories was all done programmatically.

Assessing

The process of Assessing the data consists of two approaches, the first one is to assess the data visually, just scrolling through the data, and the second one is to assess the data programmatically using Pandas Methods such .iloc(), .info(), .duplicated().sum(), .unique(), .value_counts() or slicing the DataFrame with a Boolean series to check specific conditions.

There were two criteria that we are interested in, quality and tidiness. For each table and at the end of Assessing Process, I write down the problems of the datasets then apply the cleaning process.

Quality refers to: issues related to the content of the data.

Tidiness refers to: issues related to the structure of the data.

Cleaning

In the final step of data wrangling process, I fix the problems I found in the Assessing Process by creating user-defined functions or using Pandas built-in functions such as: .astype(), .apply() .

Conclusion

Putting it all together, now I have a cleansed mastered version of the three datasets in the directory ‘WeRateDogs/exports’ named with twitter_archive_master.csv . Finally, I can draw analysis and conclusions.