wrangle_report

December 13, 2020

1 Wrangle Report

1.0.1 By Reham Metwally Maree

2 Introduction :

The purpose of this project is to put in practice what I learned in data wrangling data. The dataset that is wrangled is the tweet archive of Twitter user @dog_rates , also known as 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.

3 Project details :

My tasks in this project are as follows:
1- Gathering data
2- Assessing data
3- Cleaning data

4 Gathering data

The data for this project consist on three different dataset that were obtained as following:

4.0.1 Twitter archive file:

The twitter_archive_enhanced.csv was provided by Udacity and downloaded manually.

4.0.2 Twitter API & JSON:

By using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

4.0.3 The tweet image predictions:

Which breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information.

5 Assessing data

Once the three tables were obtained I assessed the data as following:

5.0.1 Visually:

I used two tools:
1- by printing the three entire dataframes separate in Jupyter Notebook.
2- by checking the csv files in Excel.

5.0.2 Programmatically:

by using different methods (info, value_counts,sample, duplicated, groupby, ...).

5.0.3 Then I separated the issues encountered in quality issues and tidiness issues.

6 Cleaning data

This part of the data wrangling was divided in three parts:
1- Define
2- Code
3-Test

- Copies of the original pieces of data are made prior to cleaning.

- All issues identified in the assess phase are successfully cleaned using Python and pandas.

- A tidy master dataset with all pieces of gathered data is created.

7 Storing data

I Save master dataset to a CSV file.

8 Analysis & Visualization

I made 3 insights and a visulization about the dataset after storing the datasets.

9 Conclusion :

- I have used Python programming language and some of its packages.
- There are several advantages of this tool (as compared to Excel) that is used by many data scientists.

- For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.
- It is strong in dealing with big data (much better than Excel).
- It can deal with a large variety of data (unstructured data like JSON (Tweets) or also structured data from ERP/SQL databases.
- Handling, assessing, cleaning and visualizing of data is possible programmatically using code.