wrangle_report

August 2, 2022

0.1 Reporting: wrangle_report

- Create a **300-600 word written report** called "wrangle_report.pdf" or "wrangle_report.html" that briefly describes your wrangling efforts. This is to be framed as an internal document.

0.2 WRANGLING EFFORTS

This wrangling process involved gathering,assessing, cleaning and exploring three datasets. These include

- 'twitter-archive-enhanced.csv'
- 'image-predictions.tsv'
- 'tweet-json.txt'

0.2.1 GATHERING DATA

- For this aspect of wrangling, different methods were required to gather each dataset.

twitter-archive-enhanced.csv

  - I acquired the twitter archive dataset by directly downloading it then read it into a pandas dataframe after importing the pandas module.

image-predictions.tsv

  - The image predictions dataset needed to be gotten from the internet. To do this, I used Requests library to download it using a url by importing the necessary modules (requests and os). After this, I read it into a dataframe.

tweet-json.txt

  - This required me getting it from twitter using their API but I couldn't get the developer account so I directly downloaded it and read it into a dataframe.

0.2.2 ASSESSING DATA

- For this, I was to detect at least eight quality issues and two tidiness issues using both visual and programmatic assessment. Below are the Quality and Tidiness issues I detected from the three datasets while paying attention to specific key points such as original ratings in the data i.e no retweets.

0.2.3 Quality issues

Twitter archive data

1. Wrong datatypes of some data
2. Unneccesary data
3. Source column is duplicated in both tweet json and twitter archive file
4. Name column contains non-valid names
5. Source column as a link not as a category
6. Retweets in the dataframe

Image predictions

7. Some columns contain inconsistent data

Json data

8. Id column is not similar to the id columns in other dataframes

0.2.4 Tidiness issues

1. Merge datasets
2. Expanded urls have multiple variables

0.2.5 CLEANING DATA

- This involved cleaning the data according to the rules of tidy data. Prior to cleaning, I made a copy of the original data. I used the , Define, Code, Test method throughout the cleaning process to define what problem I was tackling, write code to solve it and finally check the results. I was guided by the quality and tidiness issues gotten earlier during this process.
- Following this process, I merged the cleaned dataframes into one master dataframe and stored it in a csv file called 'twitter_archive_master.csv'

0.2.6 ANALYSIS AND VISUALIZATION

- After cleaning and storing the data, I proceeded to analyse and visualize the data to gather at least three insights and one visualization