wrangle_report

September 8, 2022

0.1 Reporting: wragle_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.

This are the steps I took in Wrangling my data:

STEP 1: Gathering The data was gathered from three different sources. The first dataset was a csv file that was gathered by downloading on the Udacity page i.e twitter-archive-enhanced.csv. The second dataset was downloaded programmatically using the url that was given. I imported the requests module and used the get function to download the dataset i.e image-predictions.tsv. The third dataset was collected from Twitter API. I queried Twitter for their Access token, Access token secret, consumer secret and consumer key which allowed me to access information from their database. I accessed the information that I needed using the tweet_id in the twitter-archive-enhanced.csv then I stored it in a txt file i.e tweet_json.txt.

STEP 2 : Accessing I used pd.read_csv to create a dataframe for the three dataset(archive, image and json_tweets). After that, I accessed them manually(by calling out their names e.g archive, image and json_tweets). By accessing them manually, I noticed a lot of things like the presence of null values, the use of none and so on. Then I programmatically accessed them using pandas info() , describe() , value_counts(), head(), tail(), duplicated() and so on. While using this programmatically assessment, I noticed some data quality and tidiness issues which I jotted down below the assessment step.

STEP 3: Cleaning Before starting the cleaning step, I made a copy of each data frame which made it easier for me to clean. The cleaning step was divided into cleaning tidiness issues first and data quality issues next. It was divided into Define-Code-Test stages. I first noted down the issues, then defined it, wrote the code to do as the definition says and tested the code if it worked. The cleaning step was achieved using pandas merge(), melt() , replace(), astype(), dropna(), drop(), drop_duplicates() and so on. After cleaning, I saved the clean dataset i.e df_all to a csv file i.e twitter-archive-master.csv.