wrangle_report

January 16, 2023

0.1 Reporting: wragle_report

In this report,I will briefly describe my wrangling efforts.

First, I imported all the necessary packages/libraries (pandas, json, request) and then I also downloaded and imported all the three required datasets.I dowloaded the 'twitter-archive-enhanced (1).csv' dataset that was already provided and I programmatically downloaded the 'image-predictions.tsv' using the request library. However, with the last dataset I used the json file provided because I was having some challenges using the twitter api.I thus read the json file line by line into a pandas dataframe, specifically the 'tweet_id', 'favorite count'(i.e the 'likes') and 'retweet_count'.

In the Assessing Data section, I assesed all the datasets both visually and programmatically. In the visual assessment I took a glance at all the datasets in an attempt to identify some quality and tidiness issues pertaining to the respective dataframes.With the programmatic assessment I used some pandas fuctions also in an attempt to identify some quality and tidiness issues pertaining to the respective dataframes. During and after the assessment, I managed to address/ identify about not less than eight quality issues and about two tidiness issues.

In the Data Cleaning section, I tried to clean(as much as possible) the Quality and Tidiness issues identified in the Assessing Data section. I addressed these issues using the Define-Code-Test framework. I defined how to go about the issues, wrote the codes to address them and also tested the codes to confirm whether or not the issues had been addressed.I managed to clean them as much as possible as I can.

Finally, I joined/concatenated the three dataframes together and stored the cleaned master DataFrame in a CSV file with the main one named twitter_archive_master.csv.