wrangle_report

January 16, 2023

0.1 Wrangle report

In this report, I will provide a summary of the gathering steps that I followed to prepare data for analysis in this project.

0.1.1 Step 1 : Gathering

We needed to gather data about the Twitter account "WeRateDogs" that rates dogs in a funny way. To do that, we started from three different resources: a given TSV file, another TSV file that should have been automatically uploaded, and the Twitter API.

A. The given Tsv file 'twitter_archive_enhanced.csv ' We just needed to download the file, upload it to our environment, and read it.

B. Download the file 'image_predictions.tsv' automatically To import this file automatically, we needed to use the special Python library 'requests.' Then, we stored the file.

C. The twitter Api That was the hard part. We first needed to create a Twitter developer account and wait for the approval of our request, which granted us elevated access. Then, we used the Tweepy library to fetch data from the API, and finally stored it as JSON data in a text file then we create from it a datafram and we store it as Tsv file to avoid use the api again (it take at least 30 minutes) .

0.1.2 Step 2 : Assessing

To assess the data I gathered, I visited the Twitter account and watched some videos about the topic to collect information and gain enough background to assess logically. Then, I displayed the data in Excel sheets to visualize it easily. Of course, that was not enough, so I used what we've learned in the classroom and benefited from methods such as info(), describe(), and value_counts() to assess the data automatically. As a result, I finally managed to extract 3 tidiness issues and 8 quality issues.

0.1.3 Step 3 : Cleaning

We started the cleaning process by making copies of the data to keep the original data safe. First, I solved the tidiness issues, and the most important was to merge all three datasets into one because all their data was related. Then, I continued cleaning the quality process using the skills I've

learned in the classroom, like dropping columns and changing column data types. However, I didn't manage to solve some problems due to a lack of data. But in total, I ended up with a ready dataset to be visualized.

In [ ]: