wrangle_report

February 21, 2023

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.

I was given the twitter-archive-enhanced.csv file which contains data for the tweeter user @weRateDogs tweet archive. I loaded the csv file into a dataframe using the pandas method `read_csv()`. However, this data isn't sufficient for the analysis required, so, I had to gather data from other sources.

The stages of wrangling done in the project can be identified as 1. Gather 2. Access 3. Clean

Gather: I gathered data from 2 other sources. A. From a provided url, using the `Request` library: This data contains the `image_description` data. To gather this data, the content from the url was downloaded and written into a .tsv file. B. From an API call: The twitter API was used here. I needed additional information such as the retweet_count and the favorite_count which can only be fetched from the API. This data was fetched from the API and saved in a file called `tweet_json.txt`. At the end of the gathering phase, I have 3 different dataframes that represented my data from all 3 sources, `df_twitter_archive`, `df_image_prediction` and `df_tweet_data`.

Access: At the accessing stage, I went through all of these 3 dataframes to identify issues with their content that. The issues identified can be categorized into the `Quality Issues` and `Tidyness Issues`. The assessment was done both visually and programatically. The issues are identified using pandas method such as `value_counts()`, `describe()`, `sort_values()` on all the dataframes I have. The following issues are identified #### Quality issues ##### df_twitter_archive table:

1.Missing data on `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, drop these columns as they have very little non null values.

2.`expanded_urls`has incomplete data of 2297 out of 2367

3.We only want original tweets, not retweeted data. This table has retweeted data i.e data with `retweeted_status_id` not null.

4.`timestamp` column is of type object instead of type datetime

df_image_prediction table: 5.The prediction columns `p1`, `p2`, `p3` values are separated by underscore instead of space. Eg `Old_English_sheepdog` should be `Old English sheepdog`

6.Inconsistent case for `p1`, `p2` and `p3` columns.

7.Missing records 2075 instead of 2356

df_tweet_data table: 8.We do not need tweets from August 1, 2017. Data from this date should be excluded from analysis

Tidiness issues 1.Four different columns for dog type on df_twitter_archive table (`doggo`, `floofer`, `pupper`, `puppo`) should be colapsed into one

2.The `favorite_count` and `retweet_count` column of the `df_tweet_data` should be added to the grand table. The other columns already exists on the table

Clean: Cleaning is the last phase of the wrangling process where all the identified issues are rectified. This was donw using methods like `drop()` to drop columns not required, `fillna()` to fill missing values amongst others. During cleaning, I got a grand master dataframe referred to as `df_rate_dog_tweets`, that has all the cleaned data in just one file. The clean data is saved to csv file using the `to_csv()` method afterwatrds.