wrangle_report

September 15, 2022

0.1 Wragle Report - WeRateDogs

0.1.1 Synopsis

Along the process of **Data Wrangling** of tweet data from the Twitter user @dog_rates also known as WeRateDogs, I have found several issues in the **Quality** and **Structure** of data collected by different means.

I have analyzed, cleaned and combined all the data into a new DataFrame. And stored it in twitter_archive_master.csv file. > **Note** : The version of file is from the First Iteration of Wrangling. This is not totally free of issues.

The Wrangling steps involved in this project are gathering of data, assessing its quality and structure, and cleaning it before producing meaningful insights and analysis. The main steps involved are three: - Gathering - Assessing - Cleaning

0.1.2 1. Gathering

In this process, I gathered data from three different media: - Downloading a csv file(twitter_archive_enhanced) manually and reading it into a dataframe. - Downloading a tsv file(image_predictions) programmatically using the requests library and reading the contents into a dataframe. - The third process required using the txt file(tweet-json) provided by Udacity. The content of the file in json format was read line by line into a dataframe using the necessary json methods.

The twitter_archive_enhanced.csv file contains basic tweet data (tweet ID, timestamp, text, etc.) for 2356 of their tweets as they stood on August 1, 2017.Further information was extracted from the image_predition url which was gotten using request method and lastly, I extracted data from tweet-json file for the tweet_id present in "twitter_archive_enhanced.csv" and stored it as "tweet.txt"

The gathered data are loaded into three different DataFrame,

- TAE : Loaded data from twitter_archive_enhanced.csv
- image : Loaded data from image_predictions.tsv
- tweet : Loaded data from tweet-json file

0.1.3 2. Assessing

In this process, each piece of gathered data from the gathering process was assessed to detect quality and tidiness issues. The assessment was done in two ways: - Visual Assessment: Here, each piece of gathered data was displayed in the jupyter notebook and external application(sublime)

to detect any of the issues mentioned above. - Programmatic Assessment: Here, each piece of gathered data was assessed using some pandas methods in order to detect any of the issues stated above.

Quality Issues

twitter_archive_enhanced **(TAE)**

1. Timestamp column should be changed datetime
2. Source column should be cleaned to make it more straightforward
3. Tweets without image urls should be dropped, keeping only the original tweets with images
4. Retweets (tweets with retweeted_status_id populated) should be dropped
5. Some names are not in tweet and some entries are incorrect, any name not in proper case should be NaN as unavailable
6. "None" string values in dataframe should be NaN to indicate the values that are not available

image_predictions (image)

7. Remove underscores and proper case image descriptions

tweet_json (tweet)

8. Remove duplicate data

Tidiness issues

1. Doggo, floofer, pupper, and puppo columns in twitter_archive should be replaced with single column and value via melt (then checked manually and cleaned or dropped in case of multiple values)
2. Merging the three data sets into one. You have split this into two tidiness issue when actually it is one tidiness issue about merging the three data sets.

0.1.4 3. Cleaning

First step: A copy all the three DataFrames using .copy() method,

- TAE_clean = TAE.copy()
- image_clean = image.copy()
- tweet_clean = tweet.copy()

After a copy each of the datasets was made, the issues documented while assessing were properly cleaned using **define-code-test framework** , after which the cleaned datasets were merged into a master dataset('twitter_archive_master.csv').