Wrangle_Report

By Precious Okon

Udacity Student

A written report and it briefly describing my wrangling efforts. According to udacity this report is to be framed as an internal document.

Project Goal

To wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The data is gotten from @WeRateDogs. I have effectively gathered the necessary data, assessed it both visually and programmatically, thoroughly cleaned the dataset, merged the different dataset together and used it for Analysis.

Project Steps:

Gathering Data

I made use of three Dataset for this project which include;

- Twitter archive Dataset which was provided by udacity as twitter-archive-enhanced.csv. I manually downloaded it then I uploaded it into my Jupyter Notebook Workspace. I used pandas library pd to read the file into a dataframe and I put it into a variable name 'df1'

- Tweet Image Prediction Dataset which I was to download it programmatically. I imported the request library and I used the get library function to get the contents of the dataset from its URL. Using the Python with open function. I open a tsv file and I saved the contents of the data into the tsv file name image-prediction.tsv and I now read it into a dataframe with variable named 'df2'

- Tweet Json File which was to be gotten from twitter API. I applied for a twitter developer accounts, I put in the necessary information needed and i also stated all the reasons why I needed the Twitter developer account and what I wanted to use the twitter API for. Unfortunately I wasn't approved and I had no other option than to use the second method provided by Udacity. Udacity provided the twitter_api.py which is the Twitter API code to gather some of the required data for the project. I Read the code and comments, I understood how the code works, then copy I copied and pasted it into my notebook. Udacity also gave the tweet_json.txt: which is the resulting data from twitter_api.py. I can proceeded to download the tweet_json.txt file, i uploaded on my jupyter notebook workspace. I use the with open function to pen a list [] and append the json file into the list, I read the list into a pandas dataframe with the relevant colums 'tweet_id', 'retweet_count' and 'favorite_count' I then saved it into a csv file as tweet-json.csv. I read it into a variable name 'df3'

Accessing Data

After gathering the three dataset and read it into a variables names df1, df2, df3. I then proceeded to access the dataset visually and programmatically

- Visually: I read the three datasets individually into jupyter notebook and I scrolled through each columns and rows

- Programmatically: I accessed the data set programmatically by using different functions such as .shape-to know the numbers of rows and columns for each dataset, .dtypes-to know the data type of each columns in a dataset, .info()-to know the information on each dataset, .isnull()-to check for null values, .duplicated()-to check for duplicate values, .nunique()-to check for unique values, .describe()-to check statistical information of each dataset.
  I then listed out the 9 Quality issues and 2 Tidiness issue I found.

Cleaning Data

I proceeded to cleaning the dataset, I made a copy of the three dataset from df1, df2, df3 to df1_clean, df2_clean and df3_clean respectively and I divided my cleaning effort into three which is define, code and test. Below are the three cleaning efforts I made

- I combined the different Dog Stages column into a single column which I named stage

- I dropped the rows that contain tweet for retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp and I also dropped the columm as well because they are not needed for our analysis

- I dropped the columns in_reply_to_status_id and in_reply_to_user_id because it contains a whole lot of missing values as well and needs to be dropped

- I also dropped the expanded_urls column because it is not needed in our analysis

- I changed Timestamp column from int to datetime

- I changed Tweet id column from int to string

- I changed Column 'p1', 'p2', 'p3' from its inconsistent value format to lower case

- I also changed {Tweet ids column in image_prediction dataset(df2) and and tweet_json dataset(df3) from int to string

- I changed the column name in tweet-json file from 'id' to 'tweet_id'

- Finally I merged the three dataset into one from their tweet_id columns

Storing Data

After gathering, accessing and clean the dataset, I merged the dataset into one and saved it into a csv file name twitter-archive-master.csv and I read it as master.

Conclusion

This project took me so long to complete, I was so exhausted but I eventually overcome and I was able to finally complete it, I did a lot of research and learning while working on this project and it has helped me to learn about more packages and appraise my skills on pandas functions and coding.

Finally I successfully wrangle a dataset and got information’s from it and I ended up visualizing on the dataset