Wrangle act report

Author: Iloezumma Ifeanyi

Date: September 5, 2022

The report will give detailed steps on how data was gathered, assesses and cleaned to complete the wrangle and analyze project in the Udacity Data Analysis Nanodegree.

1.0 Python Libraries

The first step of the wrangle process was to import all the required python libraries that will be needed to complete the project. Amongst them are pandas, numpy, matplotlib, seaborn, json, requests, tweepy, Image, BytesIO, create_engine

2.0 Data Gathering

There are three dataset to be used for the project;

- The WeRateDogs Twitter archive data – this can be directly downloaded from the URL (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv) provided in the documentation of the project. It is then uploaded as twitter_archive_enhanced.csv to the project workspace and read using the pandas read_csv function to a data frame **rate_df**.
- Next is the tweet image prediction data – this will be downloaded programmatically using the requests library from the url (https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv) provided in the documentation of the project. It is then read using the pandas read_csv function into a data frame **image_df**.
- Lastly, is the retweet count and favorite count – this has been provided as a tweet_json.txt file by Udacity to be used for the project. It was read using the pandas read_json function into a data frame tweepy_df.

3.0 Assessing data

After gathering the datasets for the project, they were assessed visually one after the other by displaying the data frame and checking for quality and tidiness issues. They were also assessed programmatically using pandas function like info(), describe(), isnull(), notnull() etc.

The following quality and tidiness issues was identified;

Quality issues

rate_df data frame

1. Data type of timestamps column is object
2. Keep original ratings (not retweet)
3. Delete columns that will not be needed for analysis
4. The character case for dog names are not consistent
5. Missing values in the name and dog stage columns denoted as 'None'
6. Null value in the name column denoted as 'none'

image_df data frame

7. Some columns have non descriptive column names

8. Duplicate images

tweepy_df data frame

9. Keep original ratings (not retweet)
10. Delete columns not used for analysis
11. Column name ID is not consistent with other data frame

Tidiness issues

1. Dog stage columns: doggo, floofer, pupper, and puppo can be merged into one column in rate_df data frame
2. Merge all the data frame into a master data frame

4.0 Cleaning the data

The following cleaning activities will be performed on the datasets

rate_df data frame

1. Change the data type of timestamp column to datetime64
2. Filter out retweets and retain only original tweets
3. Drop columns that are not be needed for analysis
4. Change the character case for dog names to capitalize each word
5. Replace rows with 'none' as missing values NaN in the name columns
6. Filter out null values in the name column

image_df data frame

7. Rename non descriptive column names
8. Remove duplicated images

tweepy_df data frame

9. Remove retweets
10. Drop columns that are not needed for analysis
11. Rename the column name ID to tweet_id

Tidiness issues

12. Merge the doggo, floofer, pupper, and puppo columns into one column called dog_stage in rate_df
13. Merge all the data frame into a master dataframe called twitter_archive_master

5.0 Storing the data

The merged dataset twitter_archive_master was saved in the project workspace using the pandas to_csv function. It was also saved in a twitter.db database using the create_engine function.

Note: See the file wrangle_act.ipynb in the project workspace to see the relevant codes to achieve the wrangling process.