wrangle_report

October 6, 2022

1 Wrangle_Report

by Francis Wobulu Wesa

1.1 Indtoduction

Thi project helped me put in practice what i had learned in my Data Wrangling class.The dataset used is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs. This report decribes my wrangling efforts.

Tasks Undertaken - Gathering Data - Accessing Data - cleaning Data

1.1.1 Gathering Data.

This step involved obtaining three dataset from diffrent sources and loading them into a pandas data frame for use - **Twitter archive file**:- twitter_archive_enhanced.csv was provided by udacity, i donwnlaoded it then uploaded it manually ono the jupiter notebook working space whereby i was able load and read it using pandas - **Tweet image predictions**:-This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information - **Twitter API & JSON**:- I was able to download the tweetjson.txt file programmatically using the Request libray and the url provided and was able to extract impotart data, retweet count and favorite count bases on tweet_id.

1.1.2 Accessing Data

Once allthe data was obtained and load into tables using pandas dataframe tools, it was time for assesment.i used two metthods to access the data. - **Visual Accessment**:- I visually assesed the data by loading the data frames in jupyter notebook and looking through the data for any tidy and cleanliness issues. - **Programmatical Accessment**:- i used diffrent methods available like;- info(), value_counts(), duplicated() and many more to root out the tidyness issues and dirt present in my datasets.
Cleanind Data After accessing my data both visually and programmatically, i listed doen the cleanliness and tidyness issues down so that my cleaning process can be guided and made much simpler since i know the issues to be tackled. These were the cleaning issues tackled.

1.Only intrested in original tweets thus have to remove retweets that contain @RT

2.Missing values in `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id`, `retweeted_status_user_id`,`retweeted_status_timestamp`

3.Timestamp is an object data type, it should be datetime.

4.on twitter archive in the name column, None appears instead of NaN value

5.Text of the tweet is not visible, it can be used to give additional info

6.Drop the rating_denominator column and only use the rating_numerator column out of 10

7.We have dulpicates in image prediction under jpg_url

8.Inconsitent data in p1, p2, p3 columns some are in uppercase while some are in lower case

These were the tidiness issues tackled

1.keep only one prediction of dog breed with its confidence level

2.the dog stage is one variable and hence should form single column. But this variable is spread across 4 columns - doggo, floofer, pupper, puppo

3.`twitter archive`,`image_prediction` and `tweet_json` all the data belongs to one table because they are all characteristics of the tweets

All this were done in three steps. 1. Define 2. Code 3. Test

i had to define the issue i was cleaning up, the code written to finally carry out the cleaning task and lastly i had to test to make sure the data was clean.

1.1.3 Analyzing and Visualization

lastly i had to carry out a few analysis on the clean data and create a couple of visualization to support my analysis.

1.1.4 Conclusion

Data warngling is a key step in the data analysis process. one should be familiar with the necessary tools and procedures used to gatther, accesses and clean data for efficient analysis.

```
In [ ]:
```