wrangle_report

July 1, 2022

0.0.1 DATA WRANGLING REPORT

Compiled by Victor Adegboyega Adedeji

Goal: The aim of this project is to fully understand data wrangling process, and also wrangle data beloging to @weratedogs. the process required to effective analyse these datapoints includes 1. Gathering 2. Accessing 3. Cleaning 4. Analysing

Project Processes

Gathering Data

1. twitter_archive_enhanced.csv: document was provided by Udacity and it contains a sample data of 2356 rows.

2. image_prediction.tsv: this file was hosted on Udacity server which i programatically downloaded and clean. There were 2075 records (url = 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv' response = requests.get(url) response)

3. tweet-json.txt: This Twitter scraped file was provided by Udacity and it contains 2327 sample of twitter data (df_list = [] with open('tweet-json.txt','r', encoding='utf-8') as file: for line in file.readlines(): lines = json.loads(line) tweet_handle = lines['id'] total_retweet = lines['retweet_count'] total_favorite = lines['favorite_count'] df_list.append({'tweet_handle':tweet_handle,'total_retweet':total_retweet,'total_favorite':total_favorite}) df_tweet_data = pd.DataFrame(df_list, columns=['tweet_handle','total_retweet','total_favorite'])

df2=df_tweet_data df2.head(10))
Various Python librabry was employed in gathering these data from different sources available

Assessing Data After import these tables, the data was assessed Visually and Programmatically

Visually: assessing data by scrolling through and also downloading the csv and tsv file to manually assess it

Programmatically: Functions like .head(1) .info(), .value_counts(), shape, .describe(),Programmatically: Functions like .head(1) .info(), .value_counts(), shape, .describe(), .isnull(), .head(), .tail(), .isnull(), .head(), .tail()

Cleaning Data his part of the data wrangling process was divided into three parts: Define, Code and Test.

These three steps were each on the issues stated in the assess section.

to properly clean, the following issue was raised and properlly taken care of

Quality issues

1. Inconsistent column title (tweet_handle in place of tweet_id)
2. Wrong data type for tweet ID in df1, df2, df3. The correct data type is String
3. wrong data type in df1 timestamp column
4. tweet id with values in retweet column needs to be removed
5. Null value in the expanded url which needs to be dropped
6. Dog names with small letter are not dogs, needs to be dropped
7. Inconsistent letter case for data in P1, P2, P3 some are lower case, while others are uppercase
8. Confidence level can be converted to percentage fore easy analysis
9. remove column retweeted_status_id, retweeted_status_timestamp,retweeted_status_user_id. contains a lot of null cell

Tidiness issues

1. The dog stages needs to be merged into a single column
2. Merge df1, df2, df3 into a single file

Storing the Data After completing the gathering, assessing and cleaning process, I saved the merged data in a csv file named twitter_archive_master.csv

Conclusion This project has so much increase my skill level and expose me to dealing with difficult i might encounter working with big data

```
In [ ]:
```