UDACITY PROJECT II

WRANGLE REPORT

Introduction

The data Wrangling and Analysis project involves data sourcing, wrangling, analysis and visualization of tweets and postings from the we rate dogs page on the tweeter social media platform.

The project involves gathering data from 3 different sources; a given csv file, downloaded via url and web scrapping via a tweeter API. The data is wrangled and cleaned to obtain good insight.

Libraries Used

The below python libraries were used for the project

- Numpy
- Pandas
- requests
- os
- urllib
- IPython.

Data Source

To better manage my documents, a folder was created programmatically to receive and organize all data set downloaded for the project.

Data Set 1*:* The Udacity page provided a direct download access to the "twitter-archive-enhanced.csv" file which was read into a python data frame "df_archive"

Data Set 2: The second data set was scrapped off a webpage via a given URL ('https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'). This data is a tsv file with a pipe (|) delimiter and it's a prediction of the dog kind from its image. The data was downloaded and read into "df_image" data frame

Data Set 3: Due to my inability to obtain a developer account on twitter for the API, I utilized the "tweet_json.txt" file provided by Udaclity and read the content line by line into "df_tweet" data frame.

Assessment

For each data set, an in-depth assessment is carried out by visualizing the data set in Microsoft excel.

Metadata information, descriptive statistical and duplicates are also checked for programmically in python using the info(), describe() and value_counts() methods.

A total of ten (10) issues were noted and documented after the assessment.

This includes eight (8) quality issues listed below

- Invalid datatype '"timestamp", "retweeted_status_timestamp
- Invalid denominator rating – "rating_denominator"
- Invalid dog names – "names"
- Null represented as None – "doggo", "floofer", "pupper" & "puppo" columns
- Data contains retweets which must be removed
- Empty columns - "geo", "coordinates" and "contributors"
- Abbreviated header name - "lang" and "id"
- Missing data for some columns

The two(2) tidiness issues.

- "doggo", "floofer", "pupper" & "puppo" columns must be merged as one
- "in_reply_to_status_id", "in_reply_to_user_id" and "source" repeated in 2 data sets.

Cleaning

Various NumPy and Pandas methods were employed to deal with the above issues A summary is provided below

- Invalid datatype - The to_datetime() method is used to convert data type from object to datetime
- Invalid denominator rating – records are evaluated and dropped using the drop() method
- Invalid dog names – records are identified and corrected using the replace() method
- Null represented as None - records are corrected using the replace() method with np.nan
- Data contains retweets - records are identified and dropped from the data frame
- Empty columns are identified and dropped using the drop() method
- Abbreviated header name – record is identified and corrected using the rename() method
- Missing data for some columns – Columns with more than 50 percent empty records are dropped using the dropna() method. This correction is however carried out at the final stage of the data cleansing to avoid dropping columns that are meant to be merged.

- "doggo", "floofer", "pupper" & "puppo" – String concatenation is used to merge the content of these columns
- "in_reply_to_status_id", "in_reply_to_user_id" and "source" repeated – the drop() method is used to drop these columns from the archive -enhanced data set

Merging & Storing the data

After the above cleaning steps have been implemented, the resulting data sets are merge together and save to csv as "twitter_archive_master.csv" using the to_csv() method.