wrangle_report

September 8, 2022

0.1 Reporting:

0.2 Wrangle Report

This project is about gathering, assesing and analyzing a data from WeRateDogs Twiiter account The data we gathering includes a csv file data from WeRateDogs twiiter archive which is exlusively sent to Udacity for this project. The second one is an image predictions file which contains neural network classifier on breeds of dogs. And finally a data gathered through twitter api. To analyse the data given it has to be assed and clened to answer our questions we need. let's see details below.

0.2.1 Gathering The Data

Enhanced Twitter Archive The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets, but not everything. So the enhanced of this twitter archive contains filtered tweets which contains ratings only. This data contains totaly 2356. Thsi data is given us by Udacity and accessed by downloading the file.

Image Predictions File This table data contains an image predictions file which a neural network that can classify breeds of dogs ran on the WeRateDogs twitter archive and gives this result. This data is accessed using the request library function from the udacity server.

Data via the Twitter Api This data is gathered using the twitter api using pythons tweepy library on the WeRateDogs twitter accound gathering all required data based on the availbale twitter ids on the enhanced twitter archive. After gathering all required datas we stored them on tweet_json.txt file. Then reading each tweet's json data and we have created a dataframe for each tweet ids.

0.2.2 Assesnig The Data

On this step we assesd the data and found some quality issue and tiddness issue some of the are the following.

```
Multiple null values are non-null on some columns
timestamp column datatypes should be datetime but it's string type.
Invalid rating_denominator values. some denominator values looks like very unreal.
Invalid rating_numerator which is less than 10.
Duplicated jpg_urls which means there are duplicated image prediction there.
```

```
p1, p2, and p3 should be changed to catagoral datatype
Renaming columns to more appropriate names.
Since df_tweet and df_archive have the smae tweet_id we can merge both dataframe in to one dataf
and etc...
```

This are some of the issues we assed on the gathered data.

0.2.3 Cleaning The Data

Based on the raised quality and tiddness issue we have cleaned the given data on the copy of each dataframes. By deleting unecessary columns duplicated data and also merging or melting some columns in to one. after cleaning each table we have merged the two tables which is Enhanced Twitter archive data table and the table we gathered via twitter api.

So finally after cleaning all the issues raised we have saved in to csv files.