wrangle_report1

January 28, 2023

1 Reporting: Wrangling_Report

2 Data Collection/Gathering:

This project(dataset) is on Data wrangling and I worked on the most popular twitter account(@dog_rates) also known as "WeRateDogs. I used this dataset to carry out data wrangling analysis(Gathering, Assessing and cleaning) and then carried out analysis with visualization to find insights from the data. this dataset rate people's dog with some features and it include three piece of dataset with different features which was gathered, assessed and cleaned,this was then merged together to make a Dataframe and used to carry out some further analysis.

2.1 Twitter_archive_enhanced(Csv Data)

To gather the twitter archive dataset, I directly downloaded the WeRateDogs twitter archive data provided by Udacity's which is in CSV format. I then read the (twitter-archive-enhanced.csv) using Pandas and saving it directly into a DataFrame.

2.2 Tweet image predictions(Tsv Data)

To gather the tweet image predictions, I downloaded the tweet image predictions data from Udacity's servers(url link:("https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv"))using Python's Requests library. I then read the file into a Python Pandas and saved it directly into a DataFrame.

2.3 Twitter API (JSON.txt Data)

To gather the tweet_json.txt Data using the tweepy library to query additional data via the twitter API. I alternatively downloaded the tweet_json.txt data provided by Udacity's. I accessed the json file by reading the text file line by line and used the text to form a dictionaries. I then create a dataframe from tweet of dictionaries and saved it as a dataframe.

2.4 ASSESSING DATA

In this section, I detected and documented Nine (9) quality issues and two (4) tidiness issue. I used both visual assessment and programmatic assessement to assess the data and I used some methods in pandas such as:
.head ()

.tail ()
.describe()
.info ()
.describe ()
.value_counts ()
.duplicate() etc.

2.5 QUALITY ISSUES

df_archive

1.Tweet_id, Retweeted_status_id, Retweeted_status_user_id float datatype should be coverted to appriopriate datatype(Object)

2.Timestamp and Retweete_status_timestamp datatype should be coverted to appriopriate datatype(datetime)

3.Doggo, Floofer, Pupper and Puppo has missing value indicated as "None" instead of "NaN".

4.expanded_url has duplicated value.

5.Name has irrelevant names like 'a', 'not', 'all', 'by', 'the', 'my', 'this', and 'his'.

6.Only the original tweet is needed, drop all retweet columnsi.e in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp are not needed.

df_image

7.The p1,p2 and p3 columns has some lowercase letters.

8.duplicated values in jpg_url.

9.Tweet_id should be converted to object datatype.

df_tweet

10.Id should be rename the appropriate column(Tweet_id) and change to an object datatype.

11.retweet_count should be dropped

2.5.1 Tidiness issues

1.Doggo,Floofer,Pupper and Puppo column should be combined to one column as "dog_stage"

2.the three table should be merged.

2.6 CLEANING DATA

In this section , A copy of each DataFrame was created (df_twitter_archive, df_image_pred, and df_twets_data). All the quality/tidiness issues documented were fixed and were all cleaned using the three-stage model of programmatic data cleaning (Define, Code and Test).

2.7 STORING DATA

In this section,I merged the three peice of cleaned datasets into one using pandas.merge(how=inner,on=tweet_id) library and I saved the file as a csv