wrangle_report

June 27, 2022

1 WRANGLE REPORT

Data wrangling alludes to the entire cycle that involves how you get data and how you make it investigation ready.It begins with Gathering , Assessing lastly cleaning the data.In this project,I will wrangle data from a twitter handle called ***@WeRateDogs*** with the assistance of twitter's API .WeRateDogs is a twitter page that rates canines as per their appelling apperance.

1.0.1 Presentation

This Wrangle Report is a piece of a Data Analyst Nanodegree presented by Udacity. The venture means to assemble information from Twitter API and Udacity given tweet data, to make examination about the tweets and the anticipated canine's variety.

Data Wrangling follows,

`Data Gathering, Data Assessing, Data Cleaning`

1.**Data Gathering**-I have assembled the documents twitter_archive_enhanced.csv and image_predictions.tsv, which are given by Udacity utilizing the solicitations bundle,

The twitter_archive_enhanced.csv record contains essential tweet information (tweet ID, timestamp, message, and so on) for 2356 of their tweets as they remained on August 1, 2017. As I want additional data from the WeRateDogs client, I have accumulated information from Twitter API utilizing the tweepy bundle and put away it as text_json.txt for the previously mentioned period (questioning by tweet_id present in twitter_archive_enhanced.csv)

I have separated a few highlights like retweet_count, favorite_count from the json and made a dataframe called tweets.

The accumulated information are stacked into three distinct DataFrame,

****twiter_archive*** : Loaded information from twitter_archive_enhanced.csv ***image_prediction*** :loaded information from image_predictions.tsv tweet
tweet: loaded information from data.csv

2.**Data Assessment** :The two kinds of Data Assessment performed, Visual appraisal: Each piece of assembled information is shown in the Jupyter Notebook. Once showed, information are furthermore surveyed in an outside application (google sheets) Programmatic appraisal: pandas' capabilities/techniques are utilized to evaluate the information.

Strategy :Visual and Programmatic Issue Type : Quality and Tidiness

3.**Data Cleaning** First Step: I have duplicated all the three DataFrames utilizing .copy() technique,

```
df1 = Twitter_archive.copy()
df2= image_prediction.copy()
df3= tweet.copy()
```

Further Steps:

I have dropped the Uninterested information segments that I won't use in my examination. For Erroneous information types, I have changed over it utilizing .astype() technique for tweet_id and pd.to_datetime() strategy for timestamp

For the amazing rating_numerator and rating_denominator, I have done message scratching for relating perception. From the text, it was seen that a few perceptions had wrong evaluating and transformed it automatically by setting 10 as the denominator for every denominator value. Also, a few other high qualities is valid and are because of twitter given evaluations to gathering of dogs.

The dog's names issue was addressed by supplanting the qualities beginning with lower case by making every initial character of the names capital.

I have likewise combined 4 sections (doggo, pupper, puppo, and floofer) into one, which I have packaged and named as dog_stage. Where, some have more than one phase.

At long last, I have addressed the cleanliness issues by blending every one of the three information outlines by tweet_id and put away it in ace DataFrame.

Ultimately, I have reported and tackle 8 quality issues and 2 cleanliness issues. However, this expert data isn't thoroughly liberated from issues, as Data Wrangling is an iterative cycle.

I have put away the wrangled data in twitter_archive_master.csv record with a minored number of issues, and prepared for a Data Analysis.

In [ ]: