DATA WRANGLING REPORT

BY KARTHIKRAJ
25 AUGUST 2018

The data wrangling process for wrangle and analyze project is the main task that gives well defined structured dataset that gives many solutions to the problems for business purpose. This process is a tedious and took more time . The data wrangling process is divided into 3 steps:

1. Gathering.
2. Assessing.
3. Cleaning.

The gathering of twitter dataset was done from three different source to make the worthy report about the dataset. First i gathered the twitter archive dataset(twitter-archive-enhanced.csv) that was provided directly by udacity and i downloaded manually. Second, i gathered tweet image predictions programmatically using Requests library and the following url 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv' . Image prediction dataset which is downloaded as tsv file used to extract the image of the dog through neural network. Third , i gathered the dataset using twitter API. It is tough comparing to other gathering it tooks lot of time to understand . i viewed lots of website and examples codes to understand the functioning of the process. My mentor also helped me to do this. Using the tweet_id from the twitter archive dataset each tweet json dataset is stored using the dataframe with specific required column in a tweet_json.txt.

Assessing of the three dataset requires more attention to find the dirty and tidy data.i found the basic dirty data is found using the info() method which tells us about the number of null values in the dataset . sample() and head() method used to view random dataset to find the major issue in the dataset. Many column data types are not properly assigned for the values in the dataset so they are assessed. Assessing becomes iterative process when the dog stages are grouped in a single in the cleaning process. it show more than 80% data as null values . After assessing process is over , i copied the dataset into separate dataframe to clean the issues assessed in assessing process.

Cleaning the dataset is the final step in data wrangling that makes the data clean and understandable by three steps define , respective code to clean and test the dataset whether it is clean properly . drop() are used to remove the unwanted rows and columns. replace() to change the misspelled words . astype() to change the datatype of the values. Assessing and cleaning steps are used as iterative steps to make the clean dataset which is good for analyzing.

Final three cleaned dataset are merged into a new dataset(twitter_archive_master.csv) using inner join and merge() method which has clean data that good worth for analyze and visualization.