PICT, PUNE

Assignment - 1

Name : Menul Oswal DOC : 6/01/22

Class : TE4 DOS : 11/01/22

Roll : 31444

TITLE : Data Wrangling, Part-1

PROBLEM STATEMENT :

Perform the following operations using python on any open source data set (eg. data.csv)

1) Import all the required python libraries.
2) Locate an open source data from web.
3) Load data set into pandas data frame.
4) Data pre-processing : check for missing values in the data using pandas describe().
5) Data formatting and Data Normalization : Summarize the type of variable by checking the data types of variable in data set.
6) Turn categorical variables into quantitative variables in python.

LEARNING OBJECTIVE

One should be able to make raw data useful by applying scientific data processing libraries.

PICT, PUNE

LEARNING OUTCOME

Understand the importance of data wrangling and role of scientific data processing libraries

THEORY

Data Wrangling :

Data wrangling sometimes referred to as data munging, is the process of transforming and mapping data from one raw data into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics.

The goal of data wrangling is to assure quality and useful data. Data analysis typically spend the majority of their time in the process of data wrangling compared to actual analysis of data.

Pandas :

It is an opensource pythons library. It provides ready to use high-performance data structures and data analysis tools. It runs on top of Numpy. It has a higher level interface and provides streamlined, alignment of tabular data and powerful time-series functionality.

PICT, PUNE

Pandas Dataframe :

Dataframe is the most important and widely used data structure and is a standard way to store data It has data aligned in rows and columns like SQL table.

pandas.DataFrame (data, index, columns, dtype, copy)

Creating a DataFrame :

```
df = pd.DataFrame ({
        "state" : ['AP', 'MH']
        "capital" : ['Hyderabad', 'Mumbai']
        })
```

Methods and functions

1) pandas.read_csv()

This method is used to read a comma seperated values file in DataFrame.

2) data_frame.head (limit)

By default returns first five rows of dataframe.

3) data_frame.shape

Returns a tuple representing the dimensions.

PICT, PUNE

4) value_counts()

Returns counts for each unique value in the column you selected

5) .isnull().sum()

Returns count of null values in column.

6) data_frame.dtypes

This returns the dtypes in the DataFrame.

7) pandas.map()

Used to map values from two series having one column same

8) data_frame.astype()

9) df['col'].unique()

Analysis:

Pandas is an very efficient scientific data processing library by using which a person can scrap out useful data from raw data for processing and using the data in many applications.

PICT, PUNE

Conclusion :

Data Wrangling is one of the most important technique to turn raw data into useful asset where pandas performs an very important role of converting the so called raw data into productive data sets that can be utilized later for different purposes.