PICT, PUNE

Aarkin Karnik
31140

Assignment - 1

- Title : Data Wrangling - I

- Problem Statement:
Perform the foll. operations using Python on any open-source dataset.

1) Import all required libraries

2) Locate an open-source data from the web.

3) Load the data set into panda's dataframe.

4) Data Preprocessing:
· Check for missing values in the data. Use pandas describe() function to get some initial statistics.

5) Data fromatting & Data Normalization
Summarize the types of variables by checking the data types of the variables in the dataset.

6) Turn categorical values into Quantative variables

- Objectives:
1) To learn & understand data wrangling using Pandas
2) To perform data preprocessing, formatting & normalization.
3) To perform one hot encoding on categorical variable

31140

- Outcomes:

students will be able to

1) Perform basic data preprocessing, data formatting & data normalization.

2) Perform encoding for conversion

- S/W & H/W Requirements:

windows -10, 64 bit, 4 GB RAM, 512 GB SSD, Intel i3-10H, Python 3.8

- Theory:

When working with tabular data, such as data stored in spreadsheets or databases, pandas is the right tool. Pandas helps to explore, clean & process data.

In pandas, a data table is called Dataframe Pandas support integration with many file formats (csv, excel, sql, json). Importing data from each of these data sources is provided by function with prefix read_*.

Similarly, to_* methods are used to store data.

When selecting a single column of pandas dataframe, the result is a pandas series. To select the column, use column label in [].

Pandas represents missing data with special float value NaN. series.isna() & series.notna() can be used to filter rows. .dropna() is used to drop rows with missing values. .fillna() is used to fill rows with missing values.

df.shape returns a tuple of shape of underlying data.
df.size returns number of elements in underlying data.
df.astype(dtype) converts / casts the type of an object to specified datatype.

- Analysis / Methods:

The given dataset contained 13580 rows X 21 columns with missing values in some columns that was filled with default '0'. Some columns that didn't satisfy dtype were type casted to appropriate dtype. One of the categorical variables 'Type' was converted to numerical variables by use of get_dummies(). The end results were printed on console & the dataframe was saved in file.

- Conclusion:

Successfully performed the mentioned operations on the given dataset.