Assignment no:1 (DSBDAL)

Page No.
Date

Name :- Om khedkar
Roll no :- 31325

Batch :- L-3
Date :- 19-01-2022

* Problem statement :-

Perform the following operations using Python on any open source dataset.

1. Import all required Python libraries.
2. Locate an open source data from the web
3. Load the Dataset into Pandas dataframe.
4. Data preprocessing : Checks for missing values in data using isnull(), describe() function to get some initial statistics. Types of variables & check the dimensions of data frame.
5. Data formatting & Data normalization : Summarize the types of variables by checking the data types of variables in the data set.
6. Turn categorical varicible into quantitative variables in Python.

* Learning objective :-

1. To learn & understand data wrangling using Pandas.
2. To perform data preprocessing, formatting & normalization.
3. To perform encoding on categorical values.

31325

Page No.
Date

* Learning outcomes :-
students will be able to
- perform basic data preprocessing, data formatting & data normalization.
- perform encoding for conversion.

* S/w & H/w requirements :-
windows 10 o.s. (64 bit)
Jupyter notebook.

* Theory :-
while working with tabular data stored in excel sheet or in a dataframe Pandas is the best tool which helps to explore & process data
In pandas a dataset is called dataframe. Pandas supports integration with many file formats (csv, excel, sql, json). Importing data from each of these data source is provides by function with prefix read_*.
when we want to select a single column of Pandas dataframe we use column name ~~is~~ label in [].

The describe() method gives quick overview of numerical data in dataframe. The aggregative statistic can be calculated for multiple columns at same time using describe() method.

31325

Page No.
Date

Pandas represents missing data with a special float value NaN. isnull() method can be used to find fields with missing values.
df.shape() returns a tuple of the shape of underlying data.
df.size() returns number of elements in the underlying data.
df.dtypes is used to find datatypes of variables in the dataframe.
If variable in a dataframe is not in the correct data type, it can be converted to specified datatype using df.astype(dtype).

Need of data wrangling :-
i> To make row data usable
ii> Quality of data ensured.
iii> Supports timely decision making
iv> Noisy, flawed, missing, data are cleaned.

* Methodology :-

1. Importing required libraries.
2. Loading dataset.
3. Data preprocessing (handling missing values)
4. Data formatting & normalization
5. categorical to quantitative variable conversion.

31325

Page No.
Date

* conclusion :-

Through this assignment we have successfully performed preprocessing, formatting & normalization of data using Pandas on a data set.