Sourav Kotkar
TE-3 (N3)
31378

Performance date: 11/01/2022
Submission date: 17/01/2022

Page No.:
Date:

DSBDAL

Group-A : Assignment-1

* Title :-

Data Wrangling I

* Problem Statement :-

Perform the following operations using Python on any open-source dataset (e.g. data.csv).

1. Import all the required Python Libraries.
2. Locate an open-source data from the web. Provide a clear description of the data and its source.
3. Load the dataset into pandas' data frame.
4. Data Preprocessing: check for missing values in the data using pandas isnull(), describe() function to get some initial statistics. Provide variable descriptions. Types of variables etc. Check the dimensions of the Data frame.
5. Data Formatting and Data Normalization: Summarize the types of variables by checking the data types of the variables in the dataset. If variables are not in the correct data type, apply proper type conversions.
6. Turn categorical variables into quantitative variables in Python.

* Learning Objectives :-

1. To load dataset in Pandas' dataframe.
2. To understand data preprocessing steps.
3. To learn data formatting & normalization.

* Dataset Used :-

Melbourne Housing Market

* H/W & S/W Req :- 8GB RAM laptop, Windows 10, Jupyter notebook, Python 3.9

* Learning Outcomes :-

Students are able to

1. Import & implement pandas library in Python.
2. Implement different data preprocessing steps on a given dataset.
3. Apply data formatting & data normalization on the given dataset.

* 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 & 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 the actual analysis of the data.

The process of data wrangling may include further munging, data visualization, data aggregation, training a statistical model, as well as many other potential uses.

Data wrangling typically follows a set of general steps which begin with extracting the data in a raw form from the data source, "munging" the raw data or parsing the data into predefined data structures, and finally depositing the result content into a data sink for storage and future use.

⇒ Data Formatting :-

- Data formatting is the process of transforming data into a common format, which helps users to perform comparisons. An example of not formatted data is the following: the same entity is referred in the same column with different values, such as New York and NY.
- We should also make sure that every column is assigned to the correct data type.
- Categorical data should have all the same formatting style, such as lower case.

⇒ Data Normalization :-

Normalization is a technique often applied as a part of data preparation for machine learning. The goal of normalization is to change the values of numeric columns in the dataset to use a common scale, without distorting differences in the range of values or losing information. Normalization is also required for some algorithms to model the data correctly.

⇒ Pandas :-

Pandas is a python library for data analysis. It is a powerful and flexible quantitative analysis tool. Pandas is built on top of two core python libraries - matplotlib for data visualization and Numpy for mathematical operations. Pandas acts as a wrapper over these libraries, allowing you to access many of matplotlib's and Numpy's methods with less code. It has functions for analyzing, cleaning, exploring & manipulating data.

⇒ Functions used :-

1) read_csv() - It is used to read a comma-seperated values (csv) file into DataFrame. It also supports optionally iterating or breaking of the file into chunks.

2) head() - It returns the first n rows for the object based on position.

3) tail() - It returns the last n rows for the object based on position.

4) describe() - It gives descriptive statistics which include those that summarize the central tendency, dispersion and shape of a dataset's distribution, excluding NaN values.

5) info() - It prints information about a DataFrame including the index dtype and columns, non-null values and memory usage.

6) size - It returns the no. of rows times no. of columns.

7) shape - It returns a tuple representing the dimensionality of the dataframe.

8) isnull() - It returns a boolean same-sized object indicating if the values are NA. NA values, such as None or numpy.NaN gets mapped to True values. Everything else gets mapped to False values.

9) dtypes - It returns a series with the data type of each column.

10) astype() - It is used to cast a pandas object to a specified datatype.

11) unique() - It returns unique values in a column.

12) get_dummies() - It is used to convert categorical variable into dummy/indicator variables.

* Conclusion :-

Through this assignment, we have successfully studied & implemented data preprocessing, data formatting & data normalization using various functions of Pandas library on a given dataset in Python.

* Observations :-

- The Melbourne Housing Market contained null values for `Car', `BuildingArea', `YearBuilt' and `CouncilArea' columns. To deal with the null values, we have either removed the rows or replaced the null values with mean of the column.
- For `Price', `Bathroom', `Bedroom2' and `PropertyCount' columns, the data was stored in float which we converted into integer datatype.
- The `Price' column contained very large values, so it has been normalized in the range of 0 to 1.
- `Type' and `Method' columns have been converted from categorical variables to quantitative variables.