PICT, PUNE

Assignment 01

Title - Data Wrangling

problem statement

perform the following operations using python or any open source data set (e.g. data.csv)

1) import all the required libraries
2) locate an open source data from web.
3) load data set into pandas data frame.
4) Data processing: check for missing values in the data using pandas insull(), describe() etc.
5) Data formatting and data Normalization: summariz the types of variables by checking the data types. (i.e. character, numeric, integer, logical)
the variables in the data set, IF the variable used are not in the correct data type, apply further type conversion.
6) Turn categorical variables into quantative variables in python in addition to the codes and outputs explain every operation that you do int the above steps and explain everything that you do to import /read/ scrap the data set.

learning objectives

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

PICT, PUNE

Learning Outcomes

understood the importance of data wrangling and role of scientific data processing libraries.

H/W & S/W Requirement

8 GB RAM
Windows 10
Jupyter Notebook
python 3.0

Theory

Data Wrangling.

Data wrangling sometimes reffered to as data mungling 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 variety of downstream purpose such as analysis.

- The goal of the data wrangling is to assure quality and usefull data.
- Data Analysis typically spend the majority of their time in the process of data wrangling compared to the actual analysis of data.
- the process of data wrangling may include further munging, data visualization, data aggregation, training and statistical model as well as many other uses.
- Data wrangling typically follows a set of general steps which begin with extracting the data in raw form the data source, munging the raw data as forcing the data into classified data structure.

methods and functions used.

(i) pandas.read.csv()

- This methods is used to read a comma separated values file into Data frame.

- Also supports operationally iterating or breaking of the file into chanels.

2. Data Formatting

- Data formatting is the process of transforming data into common format, which helps users to perform comparisons.
- An example of not formatted data is the following the same entity is reffered in the same column with different values such as Newyork & NY.
- We should also make sure that every column is assigned to the correct data type.

Data Normalization

Normalization is a technique often applied as a part of data preparation of me the goal of normalization is to change the values of numeric columns in the dataset to a common scale. without distorting difference in the ranges values.

- Normalization is also required for some algorithms to model the data correctly.

PICT, PUNE

2) data frame.head(limit)
returns the first 5 rows of the data frame, to override the default, you may insert value between the braces to change the number of rows.
Ex. head(10) will return 10 rows.

3) data frame.shape
returns a tuple representing the dimensions.
for ex. an output (100,14) represents 100 rows & 14 columns.

4) data frame.describe()
provides descriptive statistics that summarize the central tendency, dispersion and shape.

5) value count
this method returns count for each unique value in the column you selected.

6) is NULL.
returns data frame with value 'True' where it finds NULL value and False where it encounters any value.

7) df['col'] unique()
returns the unique values in column 'col'

PICT, PUNE

8) isnull().sum()
returns count of null values in the column.

9) data frame.dtypes
this attribute returns the dtype in the data frame it returns the series with the data type of each column

10) pandas.map()
it is used to map values from two series having one column same, for mapping two series, the last column of the first series should be same as index column of the second series, also the values should be unique.

packages/ module/ libraries used.

1) pandas :
- pandas is software library written for the python programming language for data manipulation and analysis.
- In particular it offers data structure and operations for manipulating numerical tables and times series it is a free software library released under the three clause BSD license.
- The name is derived from the term 'panel data' an ecomometics terms for data set that includes observations are multiple time period for the same individuals.
- It allows importing various file formats such as CSV, JSON, SQL datasets & queries.

PICT, PUNE

2) Numpy

- Numpy is a python library adding support for large, multi dimensional arrays and matrices along with large collections of high level mathematical functions to operate on these arrays.
- The ancestor of numpy, numeric was originally created by several other developers.
- IN 2005, Travious oliphet created numpy by incorporating features of comparing num arrays into numeric with extensive modifications.
- Numpy is open-source software and has many contributors it is a numfocus fiscally sponsored project.

conclusion

Data wrangling is one of the most important technique to turn raw data into usefull asset where pandas perform on very important role of converting the raw data into productive data set that can be utilized for different purpose, Hence we perfoimed different operation using these libraries.

observations.

- the melbourne lashing market contained all the null values for 'car', Bulding Area, 'yearbill', 'councilarea', columns for dealing with it.
- we have removed the rows and added it with mean values.
- Type & method columns have been converted from categorical variable to quantative variable.