DSBDAL.

Divya C. Adak
Roll No. 01
Div. A
TE Comp.

Practical No - 01.

Page No.
Date

Title :-

Data Wrangling, I.

Objectives :-

1. student should be able to perform the data wrangling operation using python on any open source dataset.

Aim :-

Data Wrangling - I perform the following operations using python on any open source dataset (eg. data.csv).

1. Import all the required python libraries.
2. locate an open source data from the web (eg. https://www.google.com) provide clear description of the data and its source.
3. load the dataset into the panda's dataframe
4. Data preprocessing : check for missing value in the data using pandas is null (), describe (), function to get some initial statistics provide variable descriptions. Type of variables, check dimensions of the data frame.
5. Data formatting and Data Normalization : summarize the type of variables by checking the datatypes of variables in the dataset. If variables are not in correct datatype apply proper type conversion

6. Turn Categorical variables into quantitative variables in python.

Requirements :-
1. Basic of python programming.
2. Concept of data preprocessing Data formatting, Data normalization and Data Cleaning.

Theory :-
Data Wrangling in python -
Data Wrangling is the process of gathering, collecting and transforming raw data into another format for better understanding, decision making, accessing and analysis in less time. Data Wrangling is also known as Data mugging.

- Importance of Data Wrangling :

- Data Wrangling is very important step. The below example will explain its importance as -

Book selling website want to show top selling books of different claims, according to user interface.
eg. a new user search for motivational books, then they want to show those motivational books which set all the most

or having a high rating etc.
But, on other their website, there are plenty of raw data from different users. Here, the concept of data mugging or data wrangling is used. As we know Data is not wrangled by system. This process is done by data scientists. So, the data scientist will Wrangle data in such a way that they will sort the motivational books that are sold more or having rating or user buy this book with these package of book etc, on the basis of that, the new user will make choice. This will explain the importance of data wrangling.

- Data Wrangling in Python :-

Data Wrangling is a crucial topic for data and data analysis. Pandas framework of python is used for data wrangling Pandas is an open source library specifically. Data analysis in python deals with the below functionalities.

1. Data exploration -
In this process, the data is studied, analyzed and understood by visualizing representation of data.

2. Dealing with missing values -
Most of the datasets having a vast amount of data contain missing values of NAN, they are needed to care of by replacing them by with mean, mode and most frequent value of the column or simply by dropping the row having a NAN value.

3. Reshaping Data -
In this process, data is manipulated : according the requirements where new data can be added or pre-existing data can be modified.

4. Filtering Data -
Sometimes, dataset are comprised of unwanted rows or columns which are required to be removed or filtered.

5. Other -
After dealing with the raw dataset

Page No.
Date

with the above functionalities, we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization model training etc.

Below is an example which implements the above functionalities on a raw dataset.

- Data Exploration -

Here, we assign the data and then we visualize the data in a tabular format.

- Conclusion :-

Hence, we have throughly studied, how to perform the following operations using python on any open source dataset (eg. data (sv)).

1. Importance all the required libraries.
2. locate an open source data from the web.

(eg. https :// www.kaggle.com)

provide a clear description of the data and its source (ie URL of the website)

3. Load the Dataset into Pandas data frame.

4. Data preprocessing:
Check for missing values in the data using pandas is null (), describe () function to get some initial statistics. Provide variable descriptions, Type of variable etc. Check the dimension of data frame.

5. Data formatting and Data Normalization:
Summarize the type of variables by checking the data types (ie. character, numeric, integer, factor and logical) of the variable in dataset. If variables are not in correct data type, apply proper type conversions.

6. Turn Categorical variables in Python:
In addition, to the codes and o/p, explain every operations that you do in the above steps and explain everything; that you do to import/ read / scrape that the dataset.

Divya C. Adak
Roll No. 01
Div. A
TE Comp.

Practical No - 02.

Title :-

Data Wrangling II.

Aim :-

Data Wrangling II.

Create an "Academic Performance" data set of students and perform the following operations using python.

1. Scan all variables for missing values and inconsistence. If there are missing values and / are inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all the numeric variables for outliers. If there are outliers, use any of the suitable techniques to deal with them.
3. Apply data transformations on an atleast one of the variables. The purpose of this transformation should be one of the following reason :

To change the scale for better understanding of the variable ; to convert non-linear relation into linear one, or decrease the skewness and convert the distribution into normal distribution.

Reason and document approach properly :

- Objectives :-
  student should be able to perform the data wrangling operation using python on any open source dataset.

- Prerequeuisities :-
  1. Basic of python programming.
  2. Concept of data processing, data formatting, data normalization and data cleaning.

- Theory :-
  Detailed explaination of exploratory Data analysis using Iris Dataset.

For complete code please visit :

https : // github . com / Naidu - Bhavya / Exploratory - Data - Analysis - on - Iris - Dataset.

What is exploratory Data Analysis ?

1. Exploratory data analysis is a tasks of analyzing the data using simple tools for statistics, some plotting tools linear Algebra.

2. Exploratory Data Analysis is a crucial step before you jump to machine learning or modeling of your data. By doing this, you can get to know whether the selected features are good enough to model, are all the features required are there any co-orelations based on which ; we can either go back to the data pre-processing step or move on to modeling.

3. Once exploratory data analysis is complete and insight are drawn, its feature can be used for supervised machine learning modeling.

- Importance of EDA :-

Many data scientist will be in a hurry to get all the machine learning stage, some either entirely skip exploratory process or do a very minimal job. This is a mistake with many implications, including generating inaccurate models, generating accurate models but on wrong data, not creating the right type of variables in data preparation and using resource inefficiently because of realizing.

Only after generating models that perhaps the data is skewed, or has outliers, or has too many missing values or finding that some values are inconsistent.

- Conclusion :-

Hence, we have throughly studied how to perform the following operations using python on any open source dataset (eg. data, csv)

1. Import all the required python libraries.
2. locate an open source data from the web (eg. https://www.kaggle.com) provide a clear description of the data and its source (ie URL of the websites)
3. Lord the dataset into pandas data frame.
4. Data processing :

check for missing values in the data using pandas is null (), describe () function to get some initial provide variable description etc. check the dimension of the data frame.

Page No.
Date

Data formatting and Data Normalization: summarize the type of variables by checking the data type (ie character numeric, integer, factor and logical) of the variables are not in correct data type, apply proper type, conversations.

6. Turn categorical variables into quantitative variable in python. In addition, to the codes and outputs explain every operation that you do in above step and explain everything that you do to import / read / scrape the dataset.