ASSIGNMENT No. 2

Aim : Data Wrangling II.

Create an "Academic performance" dataset of students & perform the following using python.

1. Scan all variables for missing values & inconsistencies. If there are missing values &/or inconsistencies, use any of the suitable techniques to deal with them.

2. Scan all numeric variables for outliers. If there are outliers, use any of the suitable techiques to deal with them.

3. Apply data transformations on atleast one of the variables. The purpose of this transformation should be one of the following reasons: to change the scale for better understanding of the variable, to convert a non-linear relation into a linear one, or to decrease the skewness & convert the distribution into a normal distribution.

4. Reason and document your approach properly.

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

→ P.T.O.

Prerequisite : 1. Basic of programming.
2. Concept of Data preprocessing, Data formatting, Data Normalization & Data cleaning.

Theory: Detailed Explaination of Exploratory data analysis using Iris Dataset.
For complete code please visit :
https://github.com/Naidu-Bharyal Exploratory Data-Analysis-on Iris-Dataset.

CSV file / Dataset - Academic performace

· Required libraries

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt.
import seaborn as sns.
import math
```

· Functions used -
Create dataset by using roll.no, marks, name & grade.

→ P.T.O.

- df = pd.dataframe({"rollno": rollno, "Name": name, "Marks": marks, "grade": grade})
- df.info ()
- df.describe ()
- df.datatypes
- df.columns , df.read_csv ("academic perf.csv")
- df.isnull ().sum()
- first_outlier
- second_outlier
- df.loc[15] = first_outlier
- df.loc[16] = second_outlier
- sns.countplot ()
- sns.boxplot ()

→ P.T.O.

· df = df.drop ( )

scaling the marks column: import minmax scaler from sklearn] preprocessing, scaler = minmax scaler ()
df[['marks']] = scaler.fit-transfer (df[['marks']]).

<u>Conclusion</u> : Hence we have throughly studied the/ 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. Provide a clear description of data & its source.
3. Load the Dataset into pandas data frame.
4. <u>Data preprocessing</u> : check for missing values in the data using pandas isnull (), describe () function to get some initial statistics. check the dimension of the data frame.
5. <u>Data formatting & Data Normalization</u> : Summarize the type of variables by checking the datatype of the variables in the dataset. If variables aren't in correct datatype, apply proper type conversion.
6. Turn categorial variables into quantitative variables in python. In addition to the codes & o/p, explain every operation that you do in the above steps & explain everything that you do to import/read/ scrape the dataset.

→ P.T.O.

Q1. What is exploratory data analysis?

→ 1. Exploratory data analysis is a task of analyzing the data using simple tools from 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 correlations based on which we can either go back to the data preprocessing step or move on to modeling.

3. Once exploratory data analysis is complete & insights are drawn, its features can be used for supervised & unsupervised machine learning modeling

2 Importance of EDA.

→ Many Data scientists will be in a hurry to get to the machine learning stage, Some either entirely skip exploratory process or do a very minimal job. This is a mistake with may many implications, including generating inaccurate models, generating accurate models but on the wrong data, not creating the right types of variables in data preparation, & using resources inefficiently because of realizing only after generating models that perhaps the data

→ P.T.O.

is skewed, or has outliers, or has too many missing values or finding that some values are inconsistent.

Which parameters used to create "Academic performance" dataset.

"roll.no", "name", "marks", "grade".

Which library is used for scaling the marks column.

import minmaxscaler from sklearn preprocessing
scaler = minmaxscaler()

df[['marks']] = scaler.fit_transform(df[['marks']])