Rupesh Dharme
31124

DATE PAGE

Assignment 02

DOP 10/01/2022 DOS 20/01/2022

Title : Data wrangling II

Problem statement:

Create academic performance dataset of students and perform following operations using python

1. scan all variables for missing values and inconsistencies. If there are missing values/ inconsistenc use any of suitable techniques.

2. Scan all neumeric variables for outliers If there are, use any technique to deal with them.

3. Apply data transformation on at least one of variables with any of following purpose:

to change scale for better understanding.
to convert non-linear to linear relation or
to decrease skewness and convert to normal distrubution.

learning Objectives:

To learn and understand data wrangling.
To deal with missing values/inconsistencies
To deal with outliers in dataset
To learn and perform data transformations

learning outcomes:

student will be able to

Rupesh Dharme
31124

DATE PAGE

- perform handling of outliers
- perform data transformations for better understanding.

H/W S/W req:

Windows 10, 64 bits, 8 GB RAM, 512 GB SSD intel i5, vscode, jupyter notebook.

Theory:

An outlier is an observation in given dataset that lies far from rest of the observations. It is a larger or smaller than remaining values. It may occur due to variability in data / experimen or human error. They may indicate heavy skewness

- mean is accurate measure to present data when we do not have outliers.
- median is used when outliers are present.
- mode is only measure of central tendency that is used with outliers when more than half of data is same.

Some techniques to detect outliers

1. Boxplot
2. Z-score
3. Inter Quartile Range

Some techniques to trim outliers:

1. Trimming 2. Quantile based flooring or capping
3. mean / median imputation.

Rupesh Dharme
31124

DATE PAGE

As mean is highly influenced by outliers, advised to replace outliers with median value.

Normalizat^n is a technique with the goal to change the values of numeric columns to a 'common scale without distorting differences in the ranges of values or losing information

Z-score is a variation of scaling that represents the numbers of standard deviations away from mean Ensures your feature distribution has mean = 0 and std dev = 1. Useful when there are few outliers butnot so extreme that you need clipping.

Another normalization method is the Min-Max scaling. All features are transformed into the range[0, 1] meaning minimum corresponds to 0 and maximum to 1.

Analysis:-

i) The dataset has a shape of (1000, 8)

ii) There are null values in 'math score', 'reading score' 'writing score'

iii) 'Math score' column is given in string data type so we type cast it into Int64.

iv) By plotting box plot, we come to know that there are outliers

$IQR = Q_3 - Q_1$

Upper bound $= Q_3 + 1.5 * IQR$

lower bound $= Q_1 - 1.5 * IQR$

Rupesh Dharme
31124

DATE PAGE

vi) We drop the rows having outliers.

vii) we apply one hot encoding on categorical columns to ensure there is liner relationship.

Conclusion :-

We have successfully implemented data analysis dataset.