PICT, PUNE

Assignment - 2

Name : Mehul Oswal DOC: 20/1/22

Class : TE - 4 DOS: 20/1/22

Roll : 31494

TITLE : Data Wrangling, Part - II

PROBLEM STATEMENT :

Create an 'Academic performance' dataset of students and perform the following operations on Python

1. Scan all variables for missing values and inconsistencies. If there are any 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 technique 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 reason : 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 and convert the distribution into a normal distribution.

PICT, PUNE

LEARNING OBJECTIVE

To learn and understand data wrangling in Pandas
To deal with missing values / inconsistencies
To deal with outliers in the dataset
To learn and perform data transformation methods.

LEARNING OUTCOMES

Students will be able to
- Perform handling of outliers in the dataset
- Perform data transformation for better understanding of variable.
- Detect and remove data inconsistencies.

THEORY

Data preprocessing is mainly to check the data quality. It can be checked by:

1) Accuracy
2) completeness
3) Consistency
4) Timeliness.

Binning:

1) Smoothing by bin mean method: here, the values in the bins are replaced by mean value of bin.

PICT, PUNE

2) By bin median : replacing by bin median values

3) By bin boundary : using minimum and maximum values of the bin values are taken and the values are replaced by closed boundary value.

Outliers :

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

- Mean is accurate measure to describe data when we do not have outliers present.
- Median is used if outliers are present in data
- Mode is used if there is outlier and $\geq 1/2$ of data is same..

Mean is the only measure of central tendency that is affected by outliers which inturn impacts standard deviation.

Some techniques to detect outliers –

→ Boxplot
→ Z-score
→ Inter Quartile Range.

PICT, PUNE

Some techniques to treat the outliers

→ Trimming / Removing the outlier :
Although not a good practice.

→ Quartile based flooring or capping :
at a certain value above go percentile value or floored at a value below 10 percentile.

→ Mean / Median imputation
As mean is highly influenced by outliers, advised to replace outliers with median value.

Normalisation is a technique with the goal to change the values of numeric columns to a common scale without distorting differences in the range of values or losing information.

z - score is a variation of scaling that represents the number of standard deviation away from mean. Ensures your feature distribution has mean = 0 and std dev = 1. Useful when there are few outliers but not so extreme that you need clipping

Another normalization method is the Min-Max scaling
All features are transformed into range [0, 1].

PICT, PUNE

Analysis

- Data preprocessing is very crucial part as if data is not of standard quality
- We performed data cleansing, integration, transformation data reduction to ensure the quality of dataset.
- We apply the technique of IQR to detect outliers

  $IQR = Q_3 - Q_1$

  $UB = Q_3 + 1.5 * IQR$

  $LB = Q_1 - 1.5 * IQR$
- We scaled the data to the desired range.

CONCLUSION

Understood the concept of data processing in terms of way to do it and achieved good quality of dataset.