DSBDAL

Roll no. 31126

Assignment - 2

Title: Data wrangling II

Problem statement:

Create an "Academic Performance" dataset of students and perform the following operations using python

1. Scan all variables for missing values & inconsistencies. If there are missing values / inconsistencies, use any of the technique to deal with them.
2. Scan all numeric variables for outliers. If they outlies, use any suitable technique to deal with them.
3. Apply data transformations on atleast one of the following reasons
   to change the scale for better understanding of the variable,
   to convert non-linear relation into a linear one or to decrease skewness and convert to normal distribution

learning Objectives:

- To learn & understand data wrangling in pandas.
- To deal with missing values / incosistencies
- To deal with outliers in 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.

H/W & S/W Requirements:

Windows 10 64 bit, 8GB RAM, 256GB SSD

VS code, Python 3.8

Theory:

An outlier is an observation in a given dataset that lies far from 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 is present in dataset.
- Mode is used if there is outlier & $\geq \frac{1}{2}$ of data is same Mean is the only measure of central tendency that is affected by outliers which in turn impacts standard deviation.

* Some techniques to detect outliers :-

→ Boxplot

→ Z-score

→ Inter quantile Range

* Some techniques to treat the outliers :-

→ Trimming / Removing the outlier :

Although not a good practice

→ Quantile based flooring or ~~capping~~ capping :

at a certain value above 90 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.

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

Z-score is a variation of scaling that represents the number of standard deviations away from mean Ensures your feature distribution has mean = 0 & std 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 the range [0,1] meaning minimum corresponds to 0 & maximum to 1.

Analysis:

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

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

iii) 'Math sore' column is given in string data type so we type cast it into int 64.

iv) By plotting box plot, we come to know that these are outliers in every numeric column.

v) We apply the technique of interquantile range to detect outliers.

$IQR = Q_3 - Q_1$

upper bound $= Q_3 + 1.5 \times IQR$

lower bound $= Q_1 - 1.5 \times IQR$.

vi) We drop the rows having outliers

vii) We apply one hot encoding on categorical columns to ensure there is linear relationships

CONCLUSION:

We have successfully implemented data wrangling on dataset.