PICT, PUNE

Assignment No. 2

Title - Data Wrangling 2

Problem statement

Create an 'Academic performance' dataset of students and perform following operations using python.

1) Scan all variables for missing values inconsistiness If there are missing values and/or inconsistencies use any of the suitable techniques to deal with it.

2) Scan all numeric variables for outliers, If there are outliers use any of the suitable technique to deal with it

3) Apply the 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 nonlinear relation into linear one, or to decrease the skewness and convert the distribution into a normal distribution.

Learning Objectives

1) To understand and implement various data processing and outliers handling techniques.

2) Understand and implement various data transformation techniques to change scalar or convert distribution to a normal distribution.

PICT, PUNE

Learning outcomes

By performing this assignment, we will be able to

1) Various data preprocessing and outliers handling techniques.

3) various data transformation techniques to change scale or convert the distribution to normal distribution.

Software Requirements

1) python 3.10 (with pandas, numpy, matplotlib, seaborn scipy).

2) jupyter Notebook

Hardware Requirement

1) Intel is 10th generation

2) 8GB RAM.

Theory

Data Wrangling

- Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis.

- This process typically includes manually converting and mapping data from one raw form into another format to allow for more convenient consumption and organisation of the data

package used

1) pandas

- It is a software library in python for data manipulation and analysis.

PICT, PUNE

2) Numpy

It is a library in python used to add support for large collection multidimensional arrays and matrices along with large collection of high level mathematical function to operate on the arrays.

3) scipy-stats

It is a library used to perform advanced statistical operations on dataset.

4) Matplotlib

- It is a library used for plotting simple graphs and also for data visualization to detect outliers, data transformation visualization etc.

5) seaborn

- It is a library used to plot detailed and visually good looking graphs for more advanced data visualization to detect outliers and check it data transformation happned correctly.

Data transformations

skewness is defined as the measure of assymetry of the probability distribution of the real valued random variable (feature/attributes)

the skewness outside the range of -1 to 1 are highly skewed and those in the range of -0.5 to 0.5 are symmetrical or normal, remaining are moderatly skewed.

PICT, PUNE

- to convert the distribution to normal distribution means we need to change the skewness and take it in the range -0.5 to 0.5
- some of the data transformation techniques are.
  1) logarithmic transformation
  2) Reciprocal transformation
  3) square root transformation
  4) Exponential transformation
  5) Box-cox transformation.
- The data transformation techniques can only be applied on positive values.

Detect and eliminate outliers.
- outliers are the data objects that deviate significantly from rest of the data objects and behaves in different manner.
- They are detected on the basis of the kind of dataset attributes use are dealing with.
- some of the ways to detect outliers are.
  1) Using Boxplot or scatterplot (Data visualization)
  2) Using Inter-quartile range.
  3) Using z score
  4) Using standard deviation

Functions used
1) df.mean()
- returns the mean of each numeric attribute

2) df.max()
- for calculating the maximum numeric value

PICT, PUNE

3) df.quantile
- returns the values at the given quantile over requested axis.

4) df.skew
- returns the skewness of the values of numeric attributes dataframe df.

5) seaborn.kdeplt(df[col])
- plots a KDE plot for column 'col'

6) nplog(df[col])
returns the dataframe array object of logarithmic values.

7) scipy.stats.boxcox(df[col])
- returns the transformed values of each numeric record of column 'col' of dataframe 'df' so that the transformed values from a normal distribution after box cox transformation is applied.

Observations
1) out of the two plots to observe the outliers, boxplot is preferred for univariate while scatterplot is preffered bivariate / multivariate outliers.
2) out of all the data transformation techniques, logarithmic and reciprocal transformation gives most in accurate results.

PICT, PUNE

3) The data transformations techniques can only be applied to positive data points.

Conclusion

Hence, after performing this assignment, we learnt various processes and techniques to handle outliers and apply data transformations to convert distribution to normal distribution, we successfully implemented the techniques on provided dataset.