PICT, PUNE

31118

DSBDA Lab

Date: 30-Jan-22

Assignment-02- Data wrangling-II

Title: Data Wrangling

Problem statement:

Create an academic performance dataset of students & perform the following operations using python:

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

2) Scan all numeric variable for outliers. If there are use any of the suitable techniques to deal with them.

3) Apply data transformations on at least 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.

Reason & document your approach properly.

Objective:

i) find null values & handle them.

ii) Scan data for outliers & handle them.

iii) Apply data transformation on data.

Outcome:

Student will be able to handle null value, outliers & transform data.

PICT, PUNE

S/w & H/w requirements.

operating system: windows-10-Home, 8GB RAM, 64-bit
Programming language: Python 3.8
Programming Tool: Jupyter Notebook & Pandas.

Theory:

Data Wrangling / data muging:
The activity of taking input data frame its original state to a format where we can perform meaningful analysis on it is called data wrangling.

Missing Values:
In order to check whether our dataset contains missing values, we use function isnull()
Syntax: df.isnull()
will return if cell of dataset is NaN or not.

Technique to handle missing values.
1) Drop missing values
- dropna() function used
- syntax: df.dropna(axis=0) - delete row.
df.dropna(axis=1) - delete column.

2) Replace missing values with a value.
- A good strategy when dealing with missing values involves their replacement with another values
- for numerical value replace with mean/media

- for categorical values replace with most frequent value.
- function used findna().

3) keep missing value as it is

Outliers:

1) An outlier is a data point that differs significantly from other observations.

2) Visualizing outliers:

- A way to visualize we the outlier is the boxplot
- Observations shown outside the whiskers are outliers.

Handle Outliers:

1) Interquartile Range (IQR) method
2) z-score method

IQR: Method:

- Data points that falls outside of 1.5 times of an interquartile range above the 3rd quartile (Q3) & below the 1st quartile (Q1) are outliers.

Transformation:

Standardization:

- It doesn't have any fixed min./max value. Here, the values of all the columns are scaled in such a way that they all have mean = 0 & standard deviation = 1. This scaling technique

PICT, PUNE

is prefered if outliers present in the dataset.

Conclusion:

Through this assignment we performed data wrangling such as handling missing value outliers & transform data by standardization.