Assignment No - 02 (Group A)

Problem statement:

Data Wrangling II

Create an "Academic performance" dataset of students and perform the following operations using Python.

1. Scan all variables for missing values and inconsistencies. If there are missing values and/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 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 and convert the distribution into a normal distribution.

Reason and document your approach properly

Pre-requisite

      1. Basic of Python Programming

      2. Concept of Data Pre-processing, Data Formatting , Data Normalization and Data Cleaning

Objective

- Students should be able to perform the data wrangling operation using Python on any open source dataset

Software and Hardware requirements:-

1. **Operating system:** Linux- Ubuntu 16.04 to 17.10, or Windows 7 to 10,
2. **RAM-** 2GB RAM (4GB preferable)
3. IDE :-Anaconda Jupiter Notebook / pycharm /Visual Studio

Theory-

1. Creation of Dataset using Microsoft Excel. The dataset is created in "CSV" format.

● The name of dataset is StudentsPerformance

● The features of the dataset are:

Roll Number, First Name, Middle Initial, Last Name, Gender, No of subjects CC WT AI

DSBDA CGPA

● Number of Instances: 30

● The response variable is: CGPA.

● Range of Values:

The response variable is the number of placement offers facilitated to particular students, which is largely depend on CGPA To fill the values in the dataset the RANDBETWEEN is used.

Returns a random integer number between the numbers you specify Syntax : RANDBETWEEN(bottom, top) Bottom The smallest integer and Top The largest integer RANDBETWEEN will return.

For better understanding and visualization, 20% impurities are added into each variable to the dataset.

Identification and Handling of Null Values Missing Data can occur when no information is provided for one or more items or for a whole unit.

Missing Data is a very big problem in real-life scenarios. Missing Data can also refer to as NA (Not Available) values in pandas. In Data Frame sometimes many datasets simply arrive with missing data, either because it exists and was not collected or it never existed.

For Example, suppose different users being surveyed may choose not to share their income, some users may choose not to share the address in this way many datasets went missing.

Checking for missing values using notnull ()

In order to check null values in Pandas Dataframe, notnull () function is used. This function return dataframe of Boolean values which are False for NaN values. Deleting null values using dropna () method

In order to drop null values from a dataframe, dropna () function is used. This function drops

Rows/Columns of datasets with Null values in different ways.

1. Dropping rows with at least 1 null value

2. Dropping rows if all values in that row are missing

3. Dropping columns with at least 1 null value.

4. Dropping Rows with at least 1 null value in CSV file

Identification and Handling of Outliers

Identification of Outliers

One of the most important steps as part of data preprocessing is detecting and treating the outliers as they can negatively affect the statistical analysis and the training process of a machine learning algorithm resulting in lower accuracy.

1. What are Outliers?

We all have heard of the idiom 'odd one out' which means something unusual in comparison to the others in a group. Similarly, an Outlier is an observation in a given dataset that lies far from the rest of the observations. That means an outlier is vastly larger or smaller than the remaining values in the set.

2. Why do they occur?

An outlier may occur due to the variability in the data, or due to experimental error/human error. Theymay indicate an experimental error or heavy skewness in the data(heavy-tailed distribution).

3. What do they affect?

In statistics, we have three measures of central tendency namely Mean, Median,and Mode. They help us describe the data.Mean is the accurate measure to describe the data when we do not have anyoutliers present. Median is used if there is an outlier in the dataset. Mode is used if thereis an outlier AND about ½ or more of the data is the same.'Mean' is the only measure of central tendency that is affected by the outliers which in turn impacts Standard deviation.

4. Detecting Outliers

If our dataset is small, we can detect the outlier by just looking at the dataset. Butwhat if we have a huge dataset, how do we identify the outliers then? We need to use visualization and mathematical techniques. Below are some of the techniques of detecting outliers are Boxplots, Scatterplots , Z-score and Inter Quantile Range(IQR)

Conclusion:

In this way we have explored the functions of the python library for Data Identifying and handling the outliers.

Data Transformations Techniques are explored with the purpose of creating the new variable and reducing the skewness from datasets