Title of the Assignment 2: Data Wrangling, II

Create an “Academic performance” dataset of students and perform the following operations usingPython.

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.

Objective of the Assignment: Students should be able to perform thedata wrangling operation usingPython on any open source dataset

---

Prerequisite:

1. Basic of Python Programming

2. Concept of Data Preprocessing, Data Formatting , Data Normalization and Data Cleaning.

Contents for Theory:

1. Creation of Dataset using Microsoft Excel.

2. Identification and Handling of Null Values

3. Identification and Handling of Outliers

4. Data Transformation for the purpose of :

a. To change the scale for better understanding

b. To decrease the skewness and convert distribution into normal distribution

---

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: Math_Score, Reading_Score, Writing_Score, Placement_ScoreClub_Join_Date .
- Number of Instances: 30
- The response variable is: Placement_Offer_Count .
- Range of Values: Math_Score [60-80], Reading_Score[75-,95], ,Writing_Score [60,80], Placement_Score[75-100], Club_Join_Date [2018-2021].

- The response variable is the number of placement offers facilitated to particular students, which is largely depend on Placement_Score 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 integerRANDBETWEEN 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.s

2.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

3. Identification and Handling of Outliers

3.1 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.

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.

Why do they occur?

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

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.

3.2 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

- Boxplots
- Scatterplots
- Z-score
- Inter Quantile Range(IQR)

3.2.1 Detecting outliers using Boxplot:

It captures the summary of the data effectively and efficiently with only a simple box and whiskers. Boxplot summarizes sample data using 25th, 50th, and 75th percentiles. One can just get insights(quartiles, median, and outliers) into the dataset by just looking at its boxplot.

3.2.2 Detecting outliers using Scatterplot:

It is used when you have paired numerical data, or when your dependent variable has multiple values for each reading independent variable, or when trying to determine the relationship between the two variables. In the process of utilizing the scatter plot, one can also use it for outlier detection.

To plot the scatter plot one requires two variables that are somehow related to each other. So here Placement score and Placement count features are used.

3.2.3 Detecting outliers using Z-Score:

Z-Score is also called a standard score. This value/score helps to understand how far is the data point

from the mean. And after setting up a threshold value one can utilize z score values of data points to

define the outliers.

3.2.4 Detecting outliers using Inter Quantile Range(IQR):

IQR (Inter Quartile Range) Inter Quartile Range approach to finding the outliers is the most commonly used and most trusted approach used in the research field.

IQR = Quartile3 – Quartile1

To define the outlier base value is defined above and below datasets normal range namely Upper and Lower bounds, define the upper and the lower bound (1.5*IQR value is considered) :

upper = Q3 +1.5*IQR

lower = Q1 – 1.5*IQR

In the above formula as according to statistics, the 0.5 scale-up of IQR (new_IQR

= IQR + 0.5*IQR) is taken.

5. Handling of Outliers:

For removing the outlier, one must follow the same process of removing an entry from the dataset using its exact position in the dataset because in all the above methods of detecting the outliers end result is the list of all those data items that satisfy the outlier definition according to the method used.

Below are some of the methods of treating the outliers

- Trimming/removing the outlier
- Quantile based flooring and capping
- Mean/Median imputation

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.