Group A

Assignment No: 2

Title of Assignment: Data Wrangling. II

Create an " Academic Performance" data of student and Perform the following operations using python.

1. Scan all variables for missing values and inconsistencies. If there are missing values and for inconsistencies, use any of the suitable techniques to deal with them.
2. Scan all numeric variables for outeriers. If there are outliers, use any of the suitable techniques to deal with them.
3. Apply data transformations on at least one of the variable. the purpose of this transformations should be one of the following reasons: to change the scale for better understanding of the variable. to convert a non-linear rotation into a linear one, or to decrease the skewness and convert the distribution into a normal distribution.

Reasons and document your approach properly.

Objective of the Assignment: students should be able to perform the data wrangling operation using python on any open source dataset.

Prerequisites

1. Basic of python programming.
2. concept of Data proprocessing, Data formating. 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 student performance.
- The feature of the dataset are: Math-score, Reading Score, writing score, Placement-score, club-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 depends on placement score.

To fill the values in the dataset. The RANDBETWEEN is used Returns a random integer number between the number 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.

step1: open microsoft Excel and click on save As. select other Formats.

step2: Enter the name of dataset and save the dataset ar type csv (MS-DOS)

step3: Enter the name of features as column header. fill the data by using RANDOMBETWEEN function.

Scroll down the cursor for 30 rows to create 30 instances, Repeat this for the features, Reading-score, writing-score, placement-score.

Step4: In 20% data, fill the impurities, The range of maths score is [60.80] updating a few instances values below 60 or above 80

2. Identification and Handling of Null values.

Missing Data and occur when no information is provided for one or more items or for whole unit. missing Data is very big problem in real-life scenarios. missing Data can also refer to as NA values in pandas. In dataframe sometimes many dataset. simply. arrive with missing data.

In pandas missing data is represented by two value.

1. None: None is a python singleton object that is often used for missing data in python code.

2. NaN: NaN (an acronym for Not a Number) is a special floating point value recognised by all system that use the Standard IEE floating point representation

useful function for detecting, removing and replacing null values in Pandas ~~function~~: Dataframe:

- isnull()
- notnull()
- dropna()
- fillna()
- replace()

1. checking for missing values using isnull() and notnull()

- Checking for missing values using isnull()

In order to check null values in pandas Dataframe, isnull() function is used. This function return dataframe of Boolean values which are True or for NaN values.

- 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 which are false for NaN values.

2. Filling missing values using dropna(), fillna(), replace()

In order to fill null values in dataset, fillna(), replace() functions are used. This function replace NaN values with some values of their own. All these function help in filling null values in dataset of dataframe.

- filling null values in dataset.

To fill null values in dataset use inplace=true

```
m-v = df['math score'].mean()
df['math score'].fillna(value=m-v, inplace=True)
df
```

Deleting null values using dropna() method

In order to drop null value from a dataframe, dropna() function is used. this function drop rows/column of dataset with Null values in different ways.

1. Dropping rows with at least 1 null value.
2. Dropping rows if all values in that rows are missing
3. Dropping columns with at least null value
4. Dropping Rows with at least null value csv file.

8. Indentification and Handling outliers.

8.1 Identification of outliers.

one of the most importance step or part of data preprocessing is detecting and treating the outliers as they can negatively affect the stastistical analysis and training process of machine learning.

1. what are outliers

we all have heard of the idiom 'odd one out'' which means something unusual in comparison to the order in a group.

2. why do they occur?

An outlier may occur due to the variability in the data, or due to experimental error/ human error

3. what do they affect?

In statistics, we have three measures of central tendancy namely Mean, Median and Mode. They help us describe the data.

4. Detecting outliers:

If our dataset is small, we can detect the outlier by just looking at the ~~atleast~~ dataset. But what if we have a huge dataset. how do we identify the outliers then? we need to use visualization and mathematical techniqu Below are some of the techniques of detecting outliers.

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

4.1 Detecting outliers using Boxplot.

It captures the summary of the data efficeatively with only a simple box and wishes.

32) Handing of outliers:

For removing the outliers, 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.

- Trimming / removing the outlier
- Quantile based flooring and capping
- mean / median imputation

4. Data Transformation for the purpose of:

Data transformation is the process of converting raw data into a format or structure that would be more suitable for model building and also data discovery in general.

The process of data transformation can also be reffered to as extract/transform/load (EFL). The extraction phase involves identifying and pulling data from the various source systems that create data and then moving the data to a single repository. The data transformation involves steps that are

- smoothing: It is process that is used to remove noise from the dataset using some algorithms. It allows for highlighting important features present in the dataset.

- Aggregation: Data collection or aggregation is the method of storing and presenting data in a summary format. The data may be obtained from multiple data sources to integrate these data sources into a data analysis description.

- Generalization: It converts low-level data attributes to high-level data attributes using concept hierarchy

- Normalization: Data normalization involves converting all data variables into a given range.

- min-Max normalization: This transforms the original data linearly.
- Normalization by decimal scaling: It normalizes the values of an attribute by changing the position of their decimal points.

- Attribute of feature construction.

New attributer constructed from the given ones: where new attributes are created& applied to assist the mining process from the given set of attributes.

a. To change the scale for better understanding (Attribute or feature construction).
Here the club_join_Date is transfered to duration

1

b. To decrease the skewness and convert distribution into normal distribution (Normalization by decimal Scalin

Data skewness: It is asymmetry in a statical distribution in which the curve appears distored, or skewe either to the left or the right. skewness can be quantified to define the extent to which a distribution differers from a normal distribution.

Normal Distribution: In a normal distribution, the graph appear as a classical symmetrical "bell-shaped curve"

A positively skewed distribution: means that the extradat results are larger. This skews the data in the it that brings the mean (averages) up. The mean will be larger that the median in Positively skewed distribution.

- A negatively skewed distribution means the opposite. That the extrems data results are smaller. This means that the mean is brought dam. & the median is larger than the mean in a negatively skewed distribution.

- Reducing skewness: A data transformation may be used to revise skewness. A distribution that is symmetric or nearly so is often easier to handle and interpreted than a skewed distribution.

- Conclusion:- In this way we have explored the functionsof the python library for data identifying & handling the outliers. Data Transformation techniques are explored with the purpose of creating the new variable and reducing the skewness from dataset.