Assignment No. 02

Aim:
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.

Prerequisites:
1. Prior knowledge of Python programming.
2. Google Colab / Python IDE

Objectives: To learn the concept of how to display summary statistics for each feature available in the dataset. Implement a dataset into a dataframe. Implement the following operations: Importing the libraries

1. Importing Libraries
2. Importing the Dataset
3. Scan all variables for missing values and inconsistencies
4. Scan all numeric variables for outliers
5. Apply data transformations on at least one of the variables

Theory:

1. Importing Libraries

" Numpy: Numpy Python library is used for including any type of mathematical operation in the code. It is the fundamental package for scientific calculation in Python. It also supports to add large, multidimensional arrays and matrices."

"Pandas: The Pandas library, which is one of the most famous Python libraries and used for importing and managing the datasets. It is an open-source data manipulation and analysis library. "

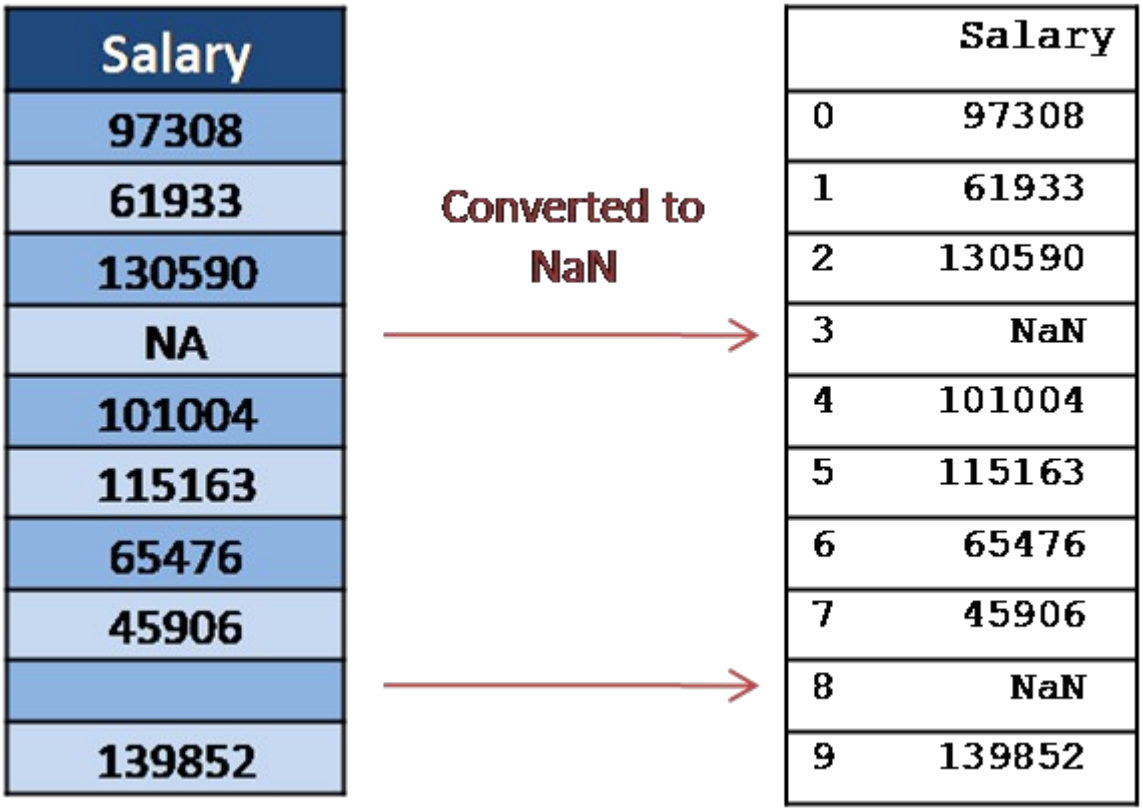
2. Importing the Dataset

dataset = pd.read_csv('data2.csv')

3. Scan all variables for missing values and inconsistencies

There are two types of missing values in every dataset:

1. **Visible errors**: blank cells, special symbols like **NA** (Not Available), **NaN** (Not a Number), etc.
2. **Obscure errors**: non-corrupt but **invalid values**. For example, a negative salary or a number for a name.



In Pandas, we have two functions for marking missing values:

- `isnull()`: mark all NaN values in the dataset as True
- `notnull()`: mark all NaN values in the dataset as False.

Now that we have marked all missing values in our dataset as NaN, we need to decide how do we wish to handle them. The most elementary strategy is to remove all rows that contain missing values or, in extreme cases, entire columns that contain missing values.

inplace=True causes all changes to happen in the same data frame rather than returning a new one.

To drop columns, we need to set axis = 1.

We can also use the how parameter.

- **how = 'any'**: at least one value must be null.
- how = 'all': all values must be null.

Removing rows is a good option when missing values are rare. But this is not always practical. We need to replace these NaNs with intelligent guesses.

There are many options to pick from when replacing a missing value:

- A single pre-decided constant value, such as 0.
- Taking value from another randomly selected sample.
- Mean, median, or mode for the column.
- Interpolate value using a predictive model.

We will use `fillna()` to replace missing values in the **'Salary'** column with 0.

Replacing NaNs with the value from the previous row or the next row: This is a common approach when filling missing values in image data. We use method = 'pad' for taking values from the previous row.

We use `method = 'bfill'` for taking values from **the next row**.

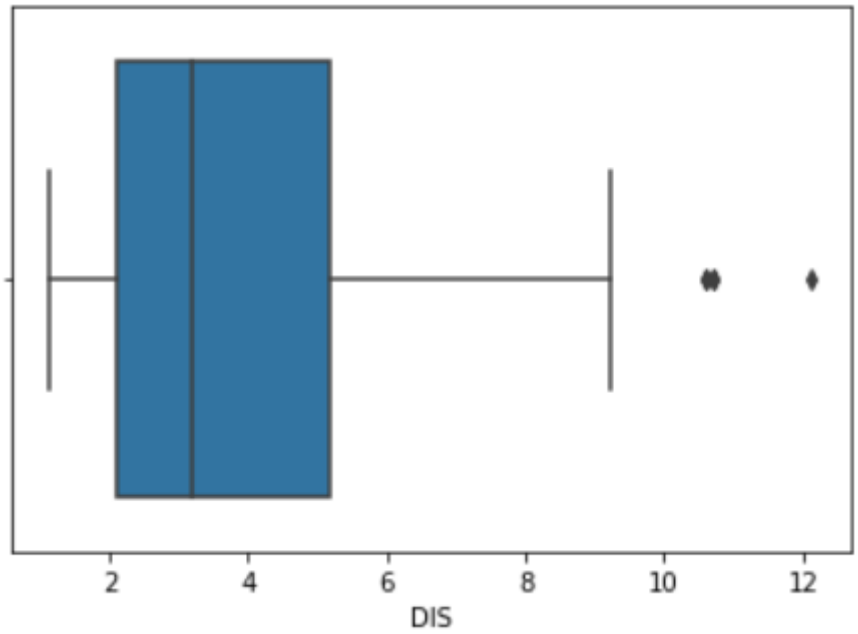
The replace method is a more generic form of the fillna method. Here, we specify both the value to be replaced and the replacement value.

4. Scan all numeric variables for outliers

An Outlier is a data-item/object that deviates significantly from the rest of the (so-called normal)objects. They can be caused by measurement or execution errors. The analysis for outlier detection is referred to as outlier mining. There are many ways to detect the outliers, and the removal process is the data frame same as removing a data item from the panda's data frame. Outliers can be detected using **visualization, implementing mathematical formulas** on the dataset, or using the statistical approach. All of these are discussed below.

1. Visualization

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.



2. Z-score

Z- Score is also called a standard score. This value/score helps to understand that 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.

Zscore = (data_point -mean) / std. deviation

3. IQR (Inter Quartile Range)

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

Trimming: It excludes the outlier values from our analysis. By applying this technique our data becomes thin when there are more outliers present in the dataset. Its main advantage is its **fastest** nature.

Capping: In this technique, we cap our outliers data and make the limit i.e, above a particular value or less than that value, all the values will be considered as outliers, and the number of outliers in the dataset gives that capping number.

5. Apply data transformations on at least one of the variables
Suppose we implement our machine learning model on such datasets. In that case, features with tremendous values dominate those with small values, and the machine learning model treats those with small values as if they don't exist (their influence on the data is not be accounted for). To ensure this is not the case, we need to scale our features on the same range, i.e., within the interval of -3 and 3.

Standard Scaler	$\frac{x_i - \text{mean}(\boldsymbol{x})}{\text{stdev}(\boldsymbol{x})}$
MinMax Scaler	$\frac{x_i - \min(\boldsymbol{x})}{\max(\boldsymbol{x}) - \min(\boldsymbol{x})}$
Robust Scaler	$\frac{x_i - Q_1(\boldsymbol{x})}{Q_3(\boldsymbol{x}) - Q_1(\boldsymbol{x})}$

Input:
Structured Dataset: product purchase
Dataset File: data2.csv

Output:
1. Importing Libraries
2. mporting the Dataset
3. Scan all variables for missing values and inconsistencies
4. Scan all numeric variables for outliers
5. Apply data transformations on at least one of the variables

Conclusion:
Hence, we have studied conversion of dataset into a dataframe and different data preprocessing, formating and normalization techniques.