Experiment No. 2

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.

Data Wrangling in Python

Data Wrangling is the process of gathering, collecting, and transforming Raw data into another format for better understanding, decision-making, accessing, and analysis in less time. Data Wrangling is also known as Data Munging.
Importance Of Data Wrangling
Data Wrangling is a very important step. The below example will explain its importance as : Books selling Website want to show top-selling books of different domains, according to user preference. For example, a new user searches for motivational books, then they want to show those motivational books which sell the most or have a high rating, etc.

Data wrangling in python deals with the below functionalities:
1. Data exploration: In this process, the data is studied, analyzed and understood by visualizing representations of data.
2. Dealing with missing values: Most of the datasets having a vast amount of data contain missing values of NaN, they are needed to be taken care of by replacing them with mean, mode, the most frequent value of the column or simply by dropping the row having a NaN value.
3. Reshaping data: In this process, data is manipulated according to the requirements, where new data can be added or pre-existing data can be modified.
4. Filtering data: Some times datasets are comprised of unwanted rows or columns which are required to be removed or filtered
5. Other: After dealing with the raw dataset with the above functionalities we get an efficient dataset as per our requirements and then it can be used for a required purpose like data analyzing, machine learning, data visualization, model training etc.
Below is an example which implements the above functionalities on a raw dataset:
Data exploration, here we assign the data, and then we visualize the data in a tabular Format.

Outlier detection is usually performed in the Exploratory Data Analysis stage of the Data Science Project Management process, and our decision to deal with them decides how well or bad the model performs for the business problem at hand. The model, and hence, the entire workflow, is greatly affected by the presence of outliers.

Outlier Detection Methods

1. Statistical Methods

Simply starting with visual analysis of the Univariate data by using Boxplots, Scatter plots, Whisker plots, etc., can help in finding the extreme values in the data. Assuming a normal distribution, calculate the z-score, which means the standard deviation (σ) times the data point is from the sample's mean. Because we know from the Empirical Rule, which says that 68% of the data falls within one standard deviation, 95% percent within two standard deviations, and 99.7% within three standard deviations from the mean, we can identify data points that are more than three times the standard deviation, as outliers. Another way would be to use InterQuartile Range (IQR) as a criterion and treating outliers outside the range of 1.5 times from the first or the third quartile.

2. Proximity Methods

Proximity-based methods deploy clustering techniques to identify the clusters in the data and find out the centroid of each cluster. They assume that an object is an outlier if the nearest neighbors of the object are far away in feature space; that is, the proximity of the object to its neighbors significantly deviates from the proximity of most of the other objects to their neighbors in the same data set. The usual approach is as follows – Fix a threshold and evaluate the distance of each data point from the cluster centroid and then remove the outlier data points and go ahead with the modeling.

Proximity-based methods are classified into two types: Distance-based methods judge a data point based on the distance(s) to its neighbors. Density-based determines the degree of outlines of each data instance based on its local density. DBScan, k-means, and hierarchical clustering techniques are examples of density-based outlier detection methods.

3. Projection Methods

Projection methods utilize techniques such as the PCA to model the data into a lower-dimensional subspace using linear correlations. Post that, the distance of each data point to a plane that fits the sub-space is calculated. This distance can be used then to find the outliers. Projection methods are simple and easy to apply and can highlight irrelevant values.

The PCA-based method approaches a problem by analyzing available features to determine what constitutes a "normal" class. The module then applies distance metrics to identify cases that represent anomalies.

Conclusion: Hence we have thoroughly studied how to perform the following operations using Python on created dataset (e.g. data.csv / Dictionary)