Name: Divyanshu Bansal
Sap id:500087339
Roll no:R2142201836

How to find outliers/ Anomalies in the data which can be exploited

- **Visualization:** We can better comprehend the data and outliers by using a variety of visualisations. The amount of variables you're analysing will determine the sort of plot you choose. The most well-liked visualisation techniques for identifying outliers in data include the following:

  I. Histogram
  II. Box Plot
  III. Scatter Plot

HISTOGRAM:
We may view the distribution of the data using a histogram. Some of the statistical methods used to find outliers need data to have a normal distribution. The z-score technique shouldn't be utilised to identify outliers if the data does not have a normal distribution.

BOX PLOT:
We may locate the single-variable outliers or outliers using a box plot. Box plots are helpful because they display the data's minimum and highest values as well as the median and interquartile range.

SCATTER PLOT:
Using a Scatter plot, it is possible to review multivariate outliers or the outliers that exist in two or more variables.

- Statistical Methods:

I. Z- Score

One popular technique for rating anomalies in one-dimensional data is the Z-score. The method below may be used to get the Z-score for each data point if you know the mean and standard deviation of the data:

$$Z = \frac{(x - \mu)}{\sigma}$$

Where,

Z= Z- Score
X= Individual data point
μ= Mean of the data
σ= Standard deviation of the data

The Z-score calculates the number of standard deviations that separate each data point from the mean. A high absolute Z-score value denotes a data point that deviates considerably from the mean and may be regarded as an oddity or outlier.
A threshold for the Z-score can be used to specify which data points are anomalous. Z-scores larger than 2 or 3, which indicate that data values falling outside of 2 or 3 standard deviations of the mean are identified as potentially anomalous, are often employed thresholds.
Additionally, Z-score assumes that the data is normally distributed, so it may not be appropriate for datasets with non-normal distributions.

II. Interquartile Range (IQR)

Statistics that split a dataset into four equal portions are known as quantiles. 25% of the data are below the first quartile (Q1), which is represented by Q1. Seventy-five percent of the data falls below the third quartile (Q3), which is the 75th percentile. The median (Q2), which splits the data into two equal portions, makes up the second quartile.

The IQR is defined as the range between the first quartile (Q1) and the third quartile (Q3). Mathematically, IQR = Q3 - Q1.

Outliers are detected via fences, which are thresholds. They are often described as a multiple of the IQR. In order to determine the lower fence, Q1 is subtracted 1.5 times the IQR, and in order to determine the higher fence, Q3 is added 1.5 times the IQR. The "normal" range of the data is determined by these fences.

Any observation that lies outside the upper or lower fences is regarded as a probable outlier or anomaly. These observations deviate considerably from the remainder of the data and could call for more research.
An observation X is considered an outlier if:
$X < Q1 - 1.5 * IQR$ (lower fence)
$X > Q3 + 1.5 * IQR$ (upper fence)

The method can detect outliers that deviate significantly from the central distribution of the data, while being less affected by extreme values compared to methods based on the mean and standard deviation.