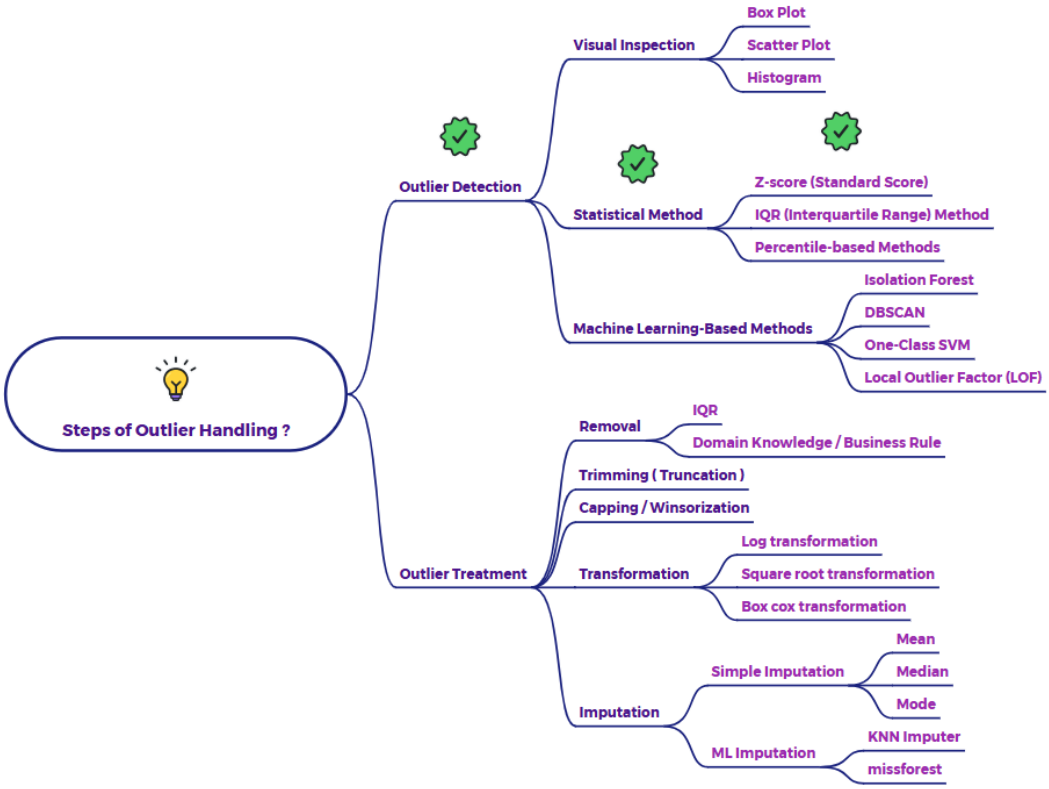
Explain Outlier Detection through Statistical method [Z-score (Standard Score)]



Outlier Detection: Statistical Method - Z-score (Standard Score)

The Z-score is a statistical measure that quantifies how far a data point deviates from the mean of the dataset, measured in terms of standard deviations. It's a very common and effective way to detect outliers, especially when the data is approximately normally distributed.

How to Calculate and Interpret Z-scores

1. Calculate the Mean (μ): Calculate the average of all the data points in the dataset.
2. Calculate the Standard Deviation (σ): Calculate the standard deviation, which measures the amount of variation or dispersion of the data points from the mean.
3. Calculate the Z-score for Each Data Point (x): For each data point, use the following formula:

Z = (x - μ) / σ

4. Identify Outliers: Data points with Z-scores that exceed a certain threshold are considered potential outliers. A common threshold is:
    - **|Z| > 3 (i.e., Z-score greater than 3 or less than -3)**

This means that these data points are more than 3 standard deviations away from the mean. However, the threshold can be adjusted depending on the specific application and the desired sensitivity to outliers.

Example

Let's consider a dataset of the heights (in inches) of adult males in a specific community:

[68, 70, 72, 69, 71, 73, 67, 74, 70, 65, 75, 72, 71, 70, 78]

In this dataset, most heights are between 65 and 75 inches. However, the value 78 is somewhat higher. Let's see how the Z-score method identifies it.

1. Calculate the Mean (μ): μ = (68 + 70 + 72 + 69 + 71 + 73 + 67 + 74 + 70 + 65 + 75 + 72 + 71 + 70 + 78) / 15 = 70.4
2. Calculate the Standard Deviation (σ): σ ≈ 3.15
3. Calculate the Z-score for Each Data Point: For example, for the height 78: Z = (78 - 70.4) / 3.15 ≈ 2.41

Here are the Z-scores for all the data points: [-0.76, -0.13, 0.51, -0.44, 0.19, 0.83, -1.08, 1.14, -0.13, -1.71, 1.46, 0.51, 0.19, -0.13, 2.41]

4. Identify Outliers: Using the threshold of |Z| > 3, there are no outliers. If we use a threshold of |Z| > 2, then 78, with a Z-score of 2.41, would be considered a potential outlier.

Benefits of Using Z-score for Outlier Detection

- Simple and Easy to Understand: The Z-score is a straightforward concept and calculation.
- Standardized Measure: It provides a standardized measure of how far a data point is from the mean, making it easier to compare outliers across different datasets.

- Works Well with Normal Distributions: It's particularly effective when the data is approximately normally distributed.

Limitations

- Sensitive to Non-Normal Data: If the data is significantly skewed or has a non-normal distribution, the Z-score method may not be as reliable.
- Affected by Outliers: The mean and standard deviation, which are used to calculate Z-scores, are themselves affected by outliers. This can sometimes mask the presence of extreme outliers.
- Univariate Only: The Z-score method is applied to individual variables (univariate) and does not consider relationships between variables.