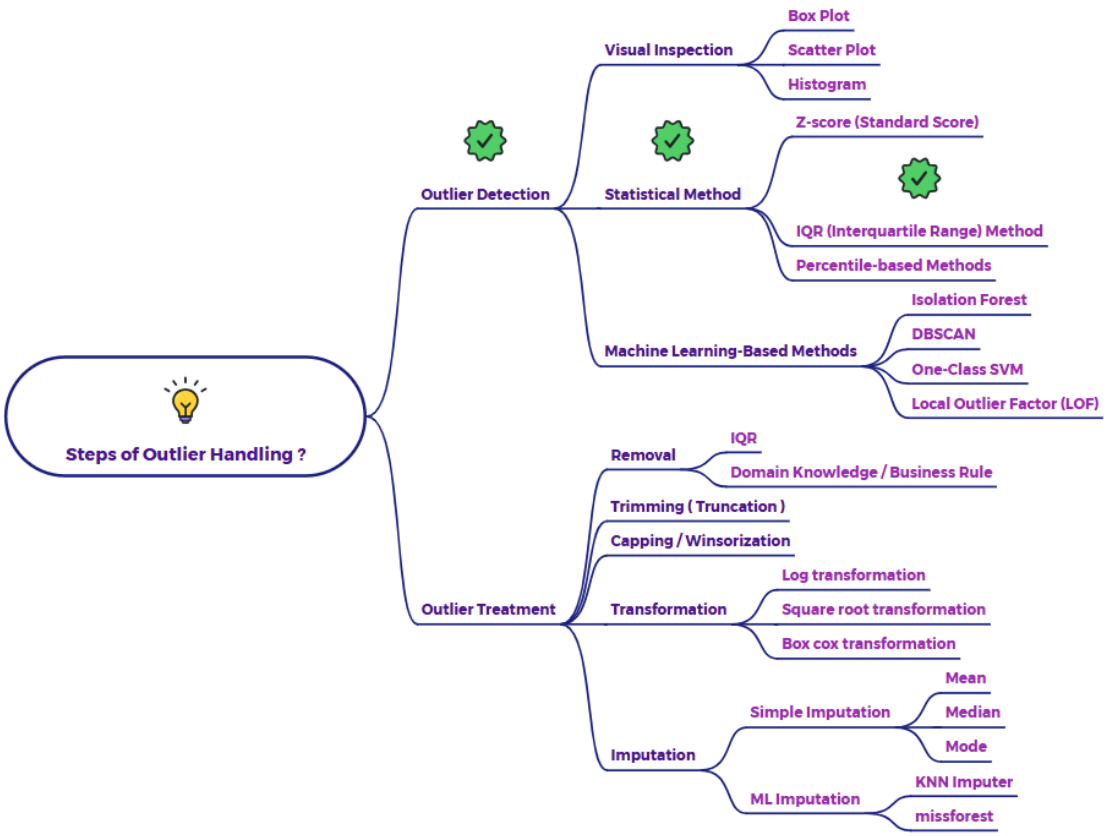
Explain Outlier Detection through Statistical method - IQR



Outlier Detection: Statistical Method - IQR (Interquartile Range) Method

The IQR method is a robust statistical technique used to identify outliers. It's less sensitive to extreme values than methods that rely on the mean and standard deviation.

How to Calculate and Identify Outliers using the IQR Method

1. **Calculate the First Quartile (Q1):** Q1 is the 25th percentile of the data. It separates the lowest 25% of the data from the rest.
2. **Calculate the Third Quartile (Q3):** Q3 is the 75th percentile of the data. It separates the highest 25% of the data from the rest.
3. **Calculate the Interquartile Range (IQR):** The IQR is the difference between Q3 and Q1:

IQR = Q3 - Q1

4. **Define the Outlier Boundaries:** Outliers are defined as data points that fall below the lower boundary or above the upper boundary:
    - Lower Boundary = Q1 - 1.5 * IQR
    - Upper Boundary = Q3 + 1.5 * IQR

The 1.5 multiplier is a common choice, but it can be adjusted (e.g., to 1 or 2) to make the method more or less sensitive to outliers.

5. **Identify Outliers:** Any data point that falls outside these boundaries is considered a potential outlier.

Example (with an outlier)

Let's consider a dataset representing the daily number of customer support tickets received by a company:

[12, 15, 10, 14, 13, 16, 11, 18, 12, 10, 100, 13, 14, 12, 11, 15, 14, 12]

In this dataset, most values are between 10 and 18. However, the value 100 is significantly higher, indicating an unusual spike in support tickets. Let's see how the IQR method identifies it.

1. **Calculate Q1 and Q3:**
    - First, sort the data: [10, 10, 11, 11, 12, 12, 12, 12, 13, 13, 14, 14, 14, 15, 15, 16, 18, 100]
    - Q1 (25th percentile) = 11
    - Q3 (75th percentile) = 14.5
2. **Calculate the IQR:** IQR = Q3 - Q1 = 14.5 - 11 = 3.5
3. **Define the Outlier Boundaries:**
    - Lower Boundary = Q1 - 1.5 * IQR = 11 - 1.5 * 3.5 = 5.75
    - Upper Boundary = Q3 + 1.5 * IQR = 14.5 + 1.5 * 3.5 = 19.75
4. **Identify Outliers:**
    - In our dataset, the value 100 is greater than the upper boundary of 19.75. Therefore, 100 is identified as an outlier.

Benefits of Using the IQR Method for Outlier Detection

- **Robustness:** The IQR method is less sensitive to extreme values than methods based on the mean and standard deviation. This makes it a good choice when dealing with data that may already contain outliers.
- **Easy to Understand and Implement:** The calculations involved are relatively simple.
- **Works Well with Skewed Data:** The IQR method can be effective even when the data is not normally distributed.

Limitations

- **Univariate Only:** The IQR method is applied to individual variables and does not consider relationships between variables.
- **May Miss Some Outliers:** In some cases, particularly with complex or unusual distributions, the IQR method might not identify all outliers.