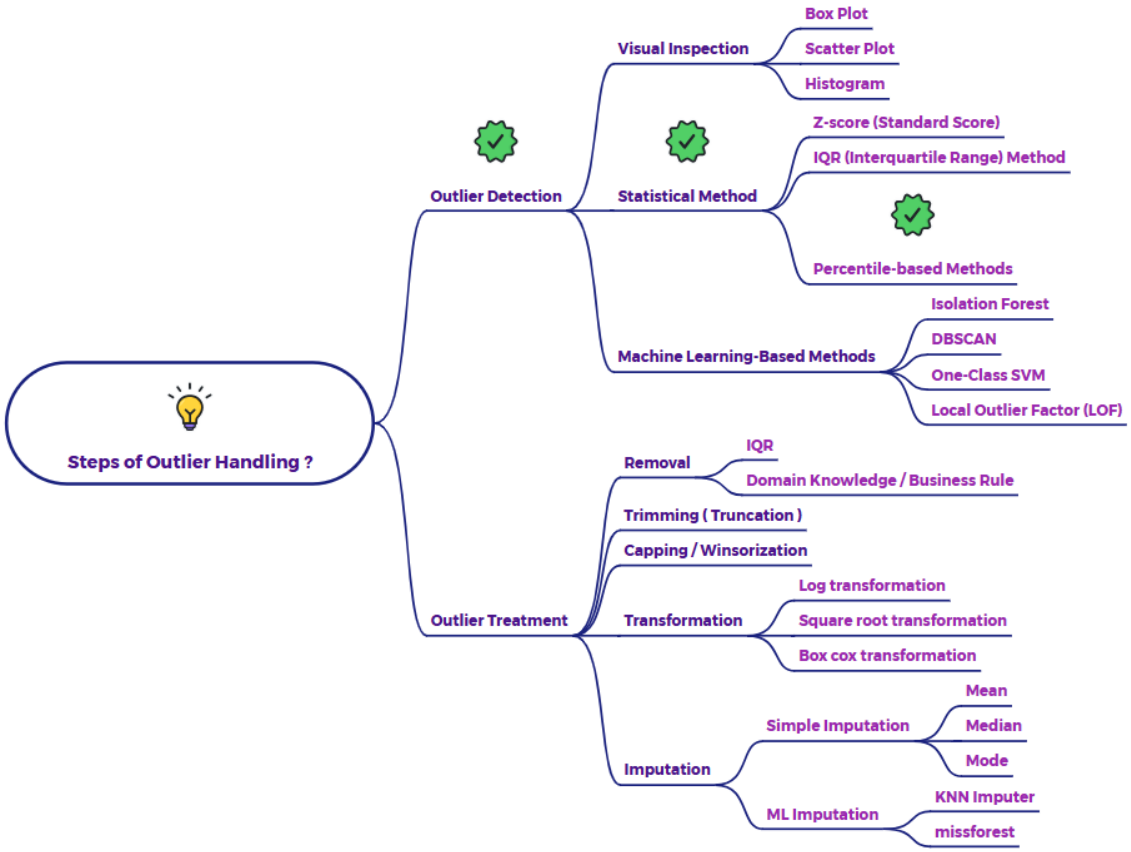
Explain Outlier Detection through Statistical method (Percentile based method)



Outlier Detection: Statistical Method - Percentile-Based Method

Percentile-based methods identify potential outliers by setting thresholds at specific percentiles of the data distribution. Data points that fall below a chosen lower percentile or above a chosen upper percentile are flagged as outliers.

How to Use Percentile-Based Outlier Detection

1. **Choose Lower and Upper Percentiles:** Select the percentiles to define the outlier boundaries. Common choices include:
    - 1st and 99th percentiles
    - 5th and 95th percentiles
    - 2.5th and 97.5th percentiles

2. **Calculate Percentile Values:** Determine the data values that correspond to the selected lower and upper percentiles.
3. **Define Outlier Boundaries:**
    - Lower Boundary = Value at the Lower Percentile
    - Upper Boundary = Value at the Upper Percentile
4. **Identify Outliers:** Any data point that is *strictly* below the Lower Boundary or *strictly* above the Upper Boundary is considered a potential outlier.

Example

Let's use a dataset of customer ages:

[22, 25, 28, 30, 32, 35, 40, 45, 50, 60, 65, 70, 80, 95, 120]

1. **Choose Percentiles:** Let's use the 5th and 95th percentiles.
2. **Calculate Percentile Values:**
    - First, sort the data: [22, 25, 28, 30, 32, 35, 40, 45, 50, 60, 65, 70, 80, 95, 120]
    - With 15 data points:
        - 5th percentile: 15 * 0.05 = 0.75. We typically round *up* to the nearest integer index (1 in a 0-based index). The value at index 1 is 25.
        - 95th percentile: 15 * 0.95 = 14.25. We round *down* to the nearest integer index (14 in a 0-based index). The value at index 14 is 120.
3. **Define Outlier Boundaries:**
    - Lower Boundary = 25
    - Upper Boundary = 120
4. **Identify Outliers:**
    - In our dataset:
        - 22 is *below* the Lower Boundary of 25, so it's an outlier.

    - 120 is *equal to* the Upper Boundary of 120, so it's *not* an outlier.

Benefits of Percentile-Based Outlier Detection

- **Robustness:** Percentiles are not greatly influenced by extreme values.
- **Simplicity:** The method is easy to understand and implement.
- **Flexibility:** The choice of percentiles can be adjusted.

Limitations

- **Data Size:** Less reliable with very small datasets.
- **Univariate:** Applied to individual variables.
- **Choice of Percentiles:** Can be somewhat subjective.