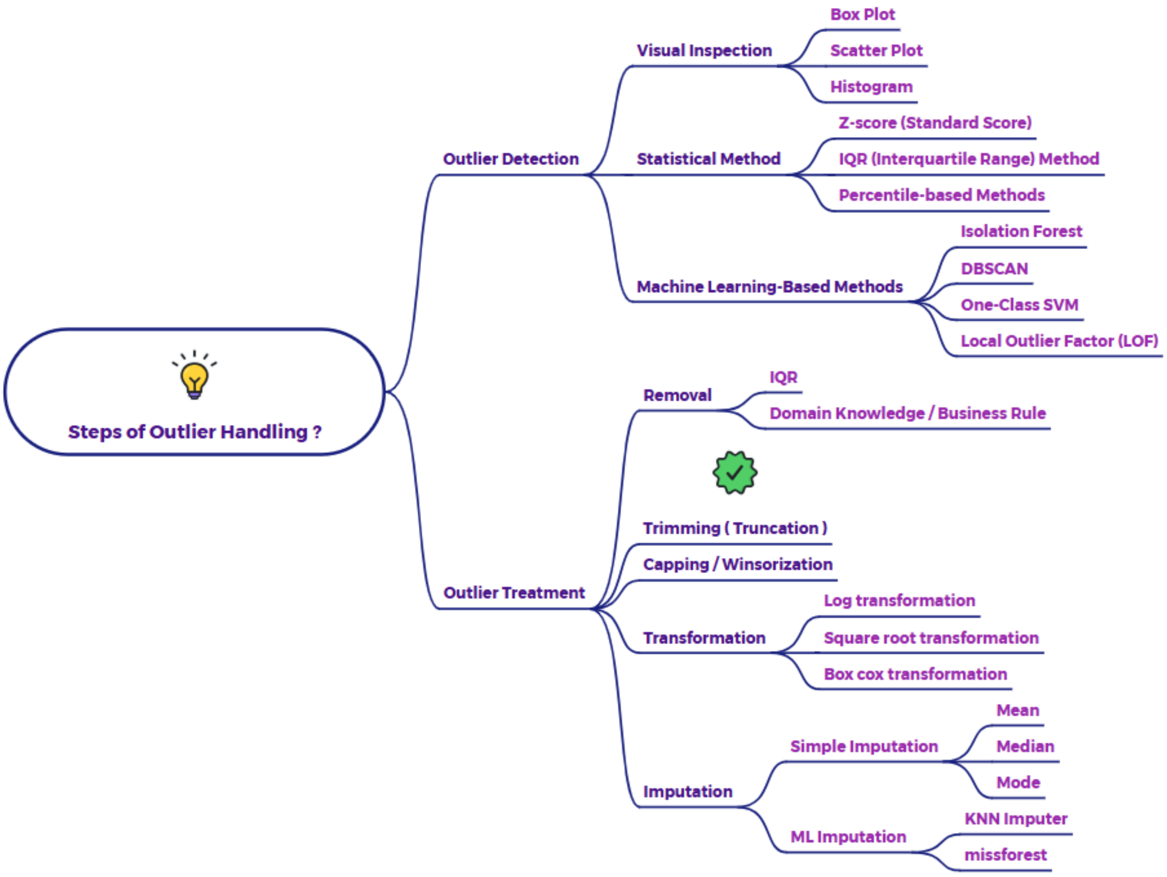
Explain Outlier treatment through Trimming (Truncation)



Outlier Treatment: Trimming (Truncation)

Trimming, also known as truncation, is a straightforward method for handling outliers. It involves removing a specific portion of the data from either or both ends of the distribution. Essentially, you "cut off" the extreme values.

Process:

1. **Define the Trimming Percentage:** Determine the percentage of data you want to remove from each tail of the distribution (e.g., 5% from the lower tail and 5% from the upper tail).
2. **Calculate Percentiles:** Calculate the corresponding percentiles for the chosen percentages. For example, if you're trimming 5%, you'd calculate the 5th and 95th percentiles.

3. **Remove Data Points:** Remove any data points that fall below the lower percentile or above the upper percentile.

Example:

Suppose we have a dataset of employee salaries (in thousands of dollars):

[40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120, 200]

In this case, 200 is a significant outlier. Let's trim the top 5% of the data.

1. **Define the Trimming Percentage:** We'll trim 5% from the upper tail.
2. **Calculate Percentiles:**
    - Since there are 16 data points, 5% of 16 is 0.8. We'll round it up to 1.
    - To trim 5% from the top, we need to find the value at the (100-5)th percentile, i.e., 95th percentile.
    - The 95th percentile lies between the 15th and 16th value. In this case it is 120.
3. **Remove Data Points:** We remove any data points greater than 120.

The trimmed dataset becomes:

[40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 100, 110, 120]

When to Use Trimming:

- When you want a simple and quick way to remove outliers.
- When you believe that the outliers are not crucial to your analysis and might be distorting your results.
- When you have a large enough dataset that removing a small percentage of data won't significantly reduce your sample size.

Cautions:

- Trimming can lead to a loss of information, as you are discarding data points.
- It's essential to choose the trimming percentage carefully. Trimming too much data can lead to underrepresentation of the underlying distribution.
- It assumes that the outliers are at the extreme ends of the distribution.