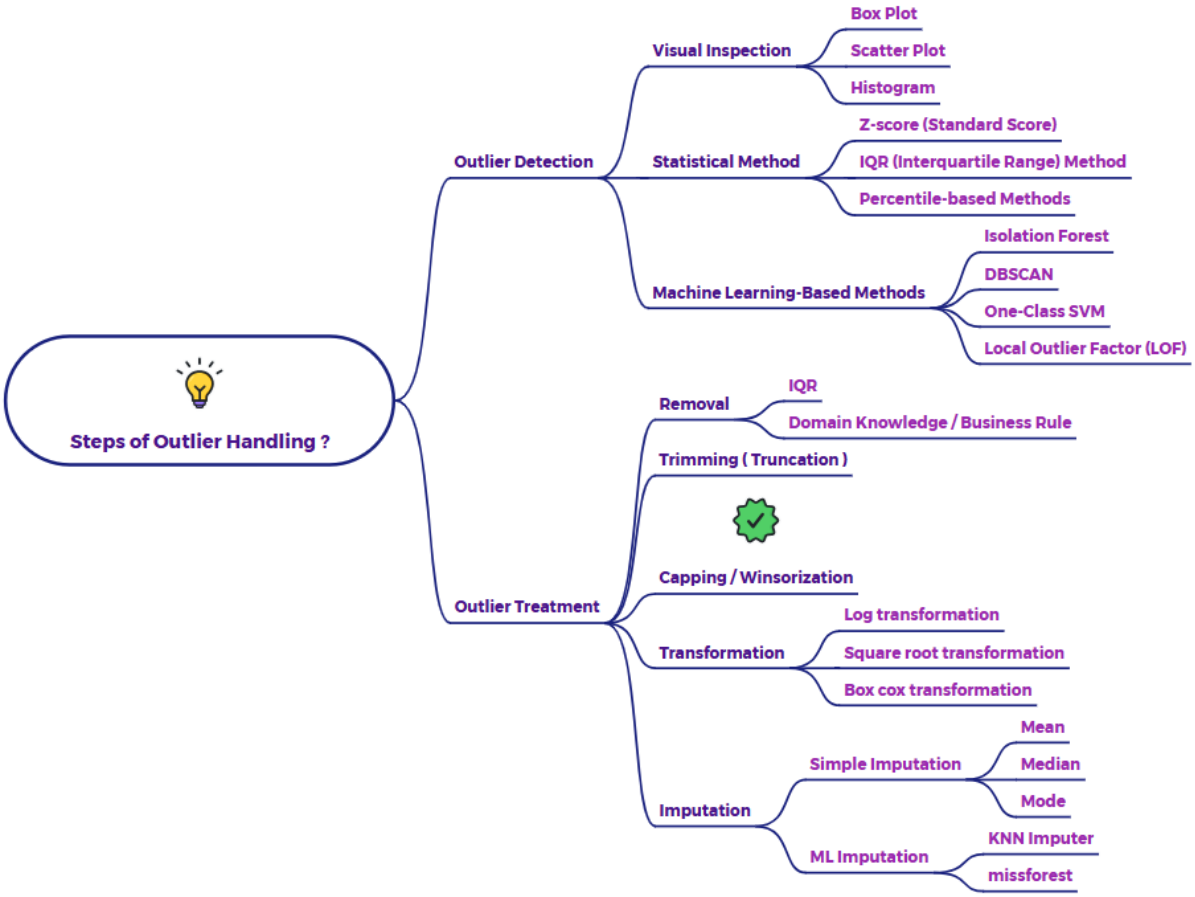
## Explain Outlier treatment through Capping (Winsorization)




### Outlier Treatment: Capping / Winsorization

Capping, also known as Winsorization, is a method of handling outliers where extreme values are replaced with the nearest plausible values, rather than being removed entirely. In other words, outliers are "capped" or "brought in" to a specified range.

### Process:

1. **Define the Capping Percentiles:** Determine the lower and upper percentiles to use for capping (e.g., the 1st and 99th percentiles, or the 5th and 95th percentiles).
2. **Identify Capping Limits:** Calculate the data values corresponding to the chosen percentiles. These values will serve as the "caps."

3. **Cap the Outliers:**
    - Any data point below the lower percentile's value is replaced with the lower percentile's value.
    - Any data point above the upper percentile's value is replaced with the upper percentile's value.
    - Values within the defined percentile range remain unchanged.

**Example:**

Let's consider a dataset of daily website traffic for a month:

[100, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 500, 600, 700, 800, 900]

Here, the values 500, 600, 700, 800, and 900 are significant outliers. Let's cap the data at the 5th and 95th percentiles.

1. **Define the Capping Percentiles:** We'll use the 5th and 95th percentiles.
2. **Identify Capping Limits:**
    - The 5th percentile is 100.
    - The 95th percentile is 350.
3. **Cap the Outliers:**
    - Any value below 100 is replaced with 100. (There are no values below 100 in this dataset.)
    - Any value above 350 is replaced with 350.

The capped dataset becomes:

[100, 120, 130, 140, 150, 160, 170, 180, 190, 200, 210, 220, 230, 240, 250, 260, 270, 280, 290, 300, 310, 320, 330, 340, 350, 350, 350, 350, 350, 350]

**When to Use Capping/Winsorization:**

- When you want to reduce the influence of outliers without discarding data points entirely.
- When you suspect that the outliers might contain valuable information but are distorting your analysis due to their extreme values.

- When you want to make your data more robust to outliers for use in statistical modeling or machine learning.

**Benefits:**

- Retains more data than removal (trimming).
- Reduces the impact of outliers.

**Cautions:**

- Can still distort the distribution of the data, though less so than leaving outliers as they are.
- The choice of capping percentiles can be subjective and influence the results.