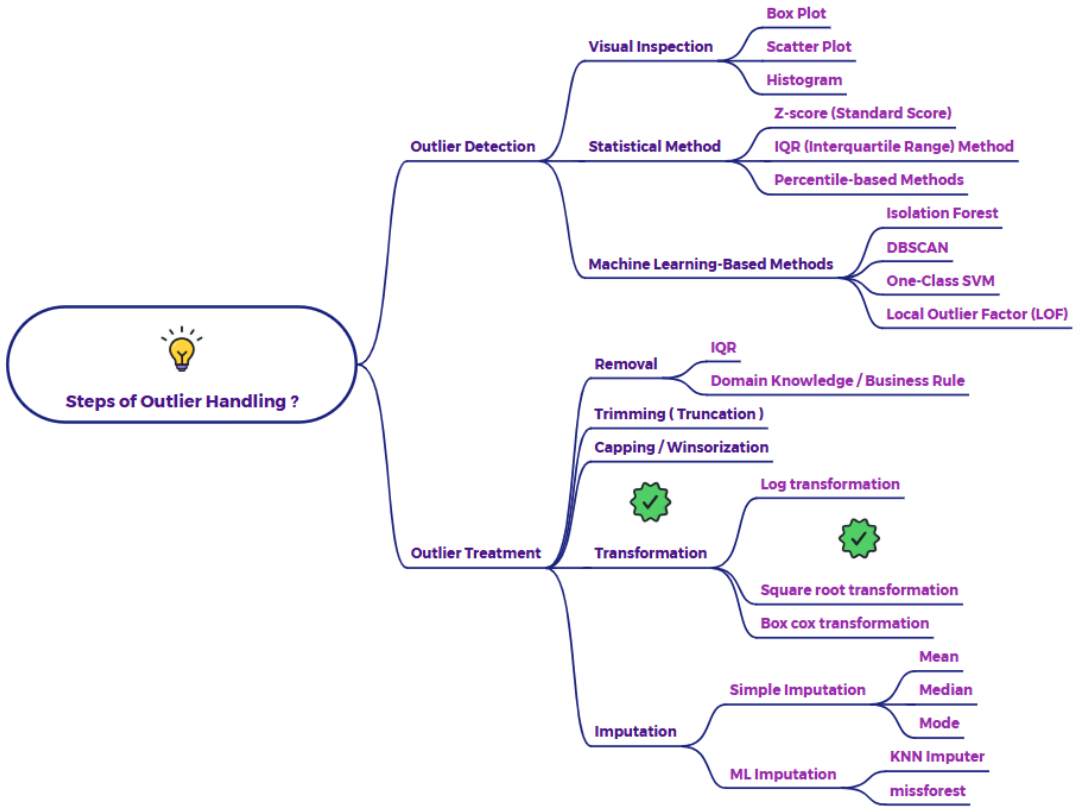
Explain Outlier treatment through transformation (Square root transformation)



Outlier Treatment: Transformation - Square Root Transformation

Square root transformation is a technique that can be applied to data to reduce the effect of outliers and make the distribution more normal. It involves applying the square root function to each data point.

How it Works:

Similar to the log transformation, the square root transformation compresses the scale of the data, especially for larger values. However, it's less drastic than the log transformation. It pulls in the larger values towards the center, thus reducing the impact of outliers and making the data less skewed.

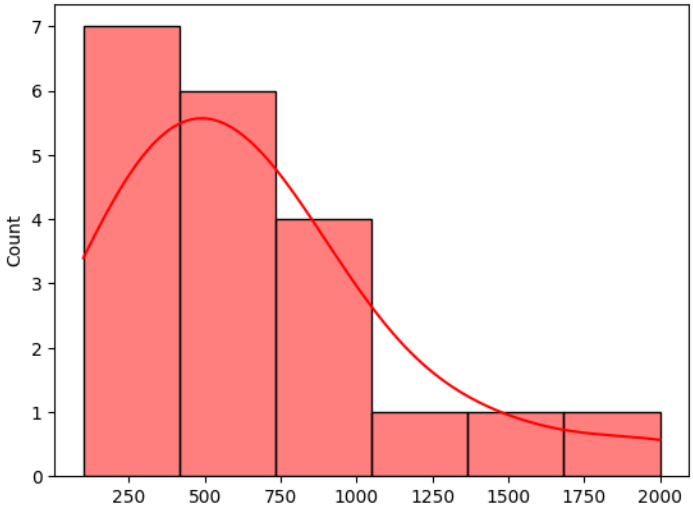
When to Use Square Root Transformation for Outlier Handling:

- **Right-Skewed Data:** Like log transformation, it is most effective when dealing with data that is right-skewed.
- **Positive Data:** The square root function is only defined for non-negative values. So, it can only be applied to variables where all values are zero or positive.
- **Example:**

Let's consider a dataset of the number of website visits per day:

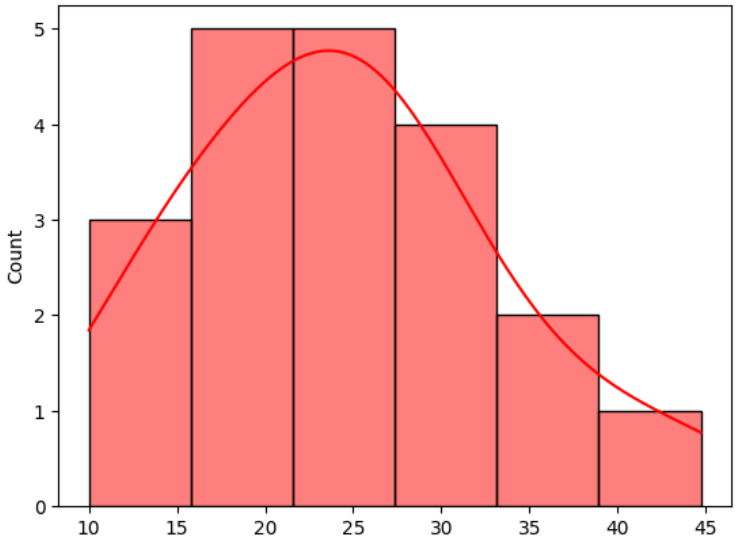
Visits = [100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 750, 800, 900, 1000, 1200, 1500, 2000]

This data is likely to be right-skewed, with a few days having very high visits. The values 1200, 1500, and 2000 are potential outliers, lets visualize the distribution:



After square-root transformation the dataset will look like:

Square root transformed visits data: [10 , 12.24 ,14.14 ,15.81, 17.32,18.70,20 , 21.21, 22.36 , 23.45 , 24.49 , 25.49 , 26.45 , 27.38 , 28.28 , 30 , 31.62 , 34.64 , 38.72 , 44.72] and the visual representation as :



As you can observe, the transformed data has a smaller range, and the large values are closer to the other values compared to the original data.

Benefits:

- Reduces the effect of outliers.
- Can make skewed data more normally distributed.
- Simpler to interpret than log transformation.

Cautions:

- Only applicable to non-negative data.
- Less effective than log transformation in dealing with very extreme outliers.