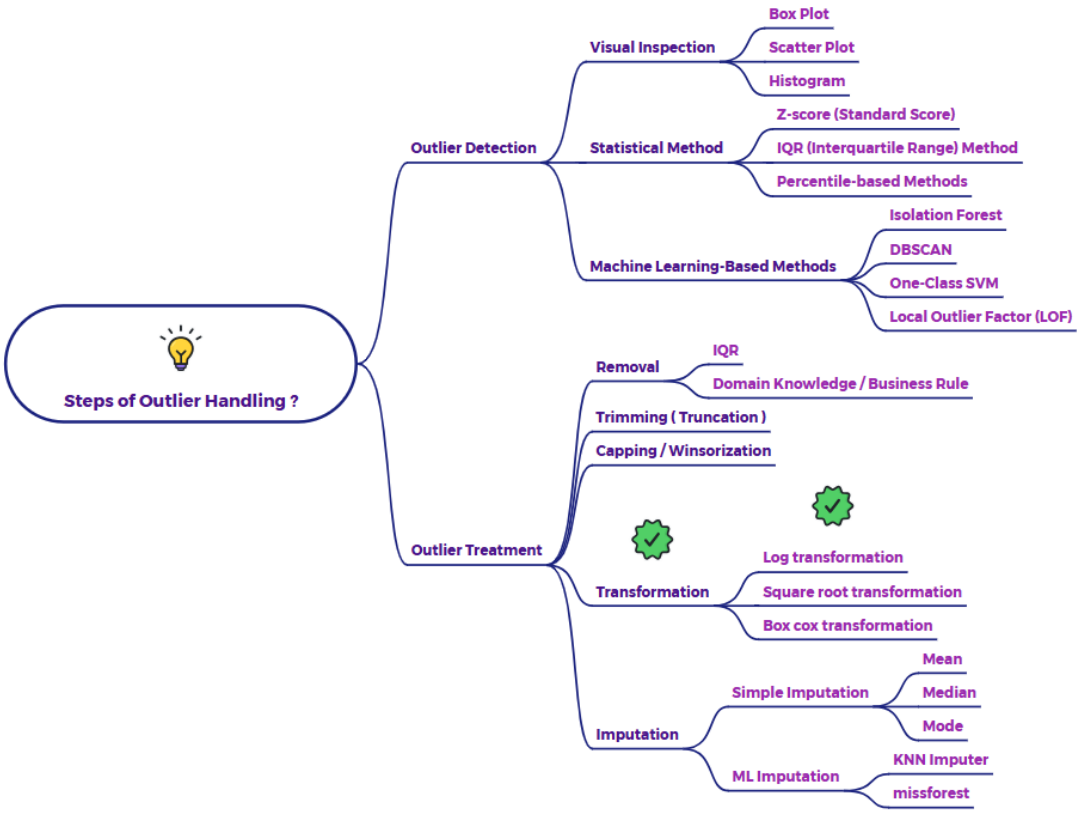
Explain Outlier treatment through transformation - Log transformation



Outlier Treatment: Transformation - Log Transformation

Log transformation is a mathematical technique that can be applied to data to make it more normally distributed and to reduce the impact of outliers. It involves applying the logarithmic function to each data point.

How it Works:

The log function compresses the scale of the data, particularly for large values. This compression has the effect of pulling in extreme values (outliers) towards the center of the distribution, making the data less skewed.

When to Use Log Transformation for Outlier Handling:

- **Right-Skewed Data:** Log transformation is most effective when dealing with data that is right-skewed. In a right-skewed distribution, the tail is longer on the right side, and outliers are more likely to be large values.

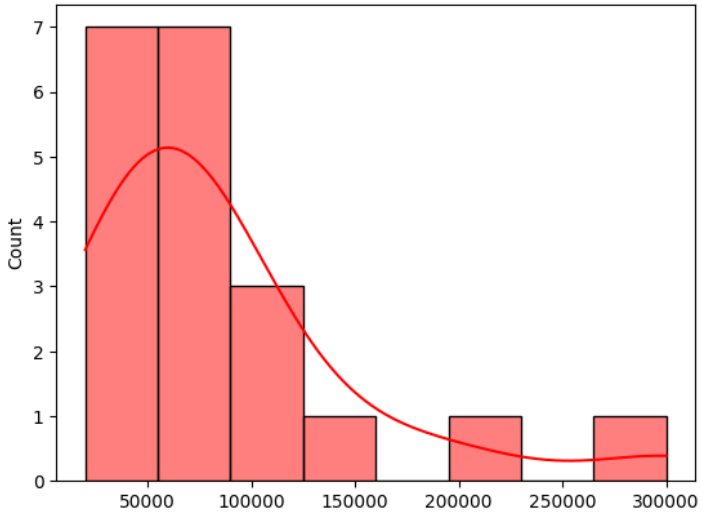
- **Positive Data:** The log function is only defined for positive values. Therefore, log transformation can only be applied to variables where all values are positive.

Example:

Let's say we have a dataset of customer income (in dollars):

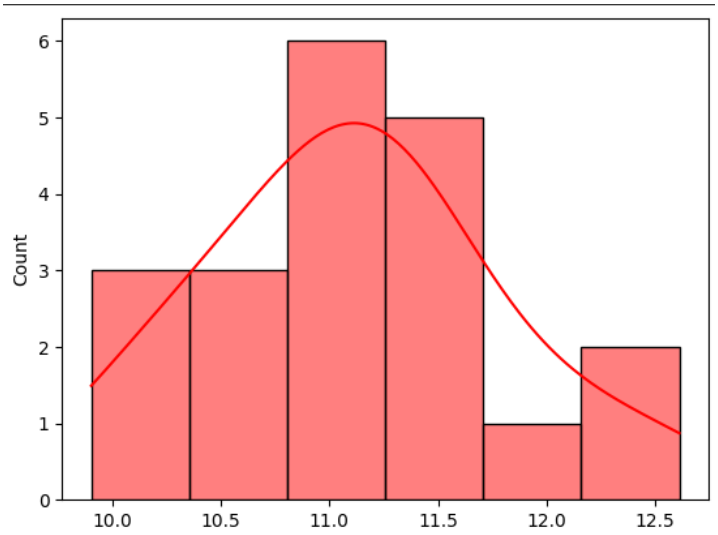
income = [20000, 25000, 30000, 35000, 40000, 45000, 50000, 55000, 60000, 65000, 70000, 75000, 80000, 85000, 90000, 100000, 120000, 150000, 200000, 300000]

This data is likely to be right-skewed, with a few high-income earners. The values 200000 and 300000 are potential outliers. Check the histogram below:



After we apply log transformation, the dataset will look like:

Log-transformed income data: [9.90348755 , 10.1266311 , 10.30895266 ,10.46310334,10.59663473, 10.71441777 , 10.81977828 ,10.91508846 , 11.00209984 , 11.08214255, 11.15625052 ,11.22524339,11.28978191, 11.35040654, 11.40756495 ,11.51292546 ,11.69524702 ,11.91839057 ,12.20607265 ,12.61153775]



As you can see, the difference between the smaller values in the log-transformed data is larger than the difference between the larger values. This is how log transformation reduces the impact of outliers.

Benefits:

- Reduces the effect of outliers.
- Can make skewed data more normally distributed.
- Can stabilize variance.

Cautions:

- Only applicable to positive data.
- The transformed data is more difficult to interpret in its original units.
- If the data contains values close to zero, log(x) approaches negative infinity. In such cases, you might need to add a small constant to all values before applying the log transformation (e.g., log(x + 1)).