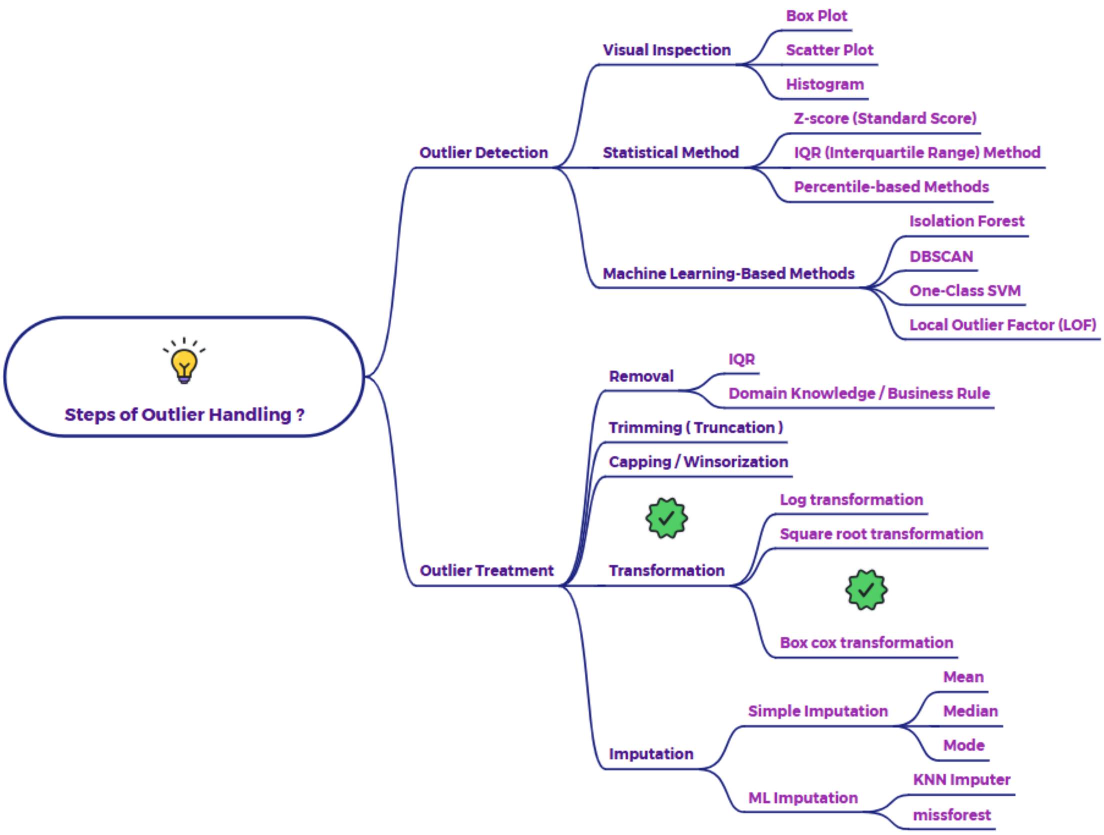
Explain Outlier treatment through transformation (Box cox transformation)



Outlier Treatment: Transformation - Box-Cox Transformation

The Box-Cox transformation is a powerful technique used to transform non-normally distributed data into a more normal distribution. While its primary goal isn't solely outlier handling, it effectively reduces the impact of outliers by compressing the range of extreme values.

How it Works:

The Box-Cox transformation is defined by the following formula:

$$\begin{cases} y = \dfrac{x^{\lambda} - 1}{\lambda} \text{ where } \lambda \neq 0 \\ y = \ln x \text{ where } \lambda = 0 \end{cases}$$

Where:

- (x) is the original data value.
- (y) is the transformed data value.
- (λ) (lambda) is the transformation parameter that is determined from the data.

The Box-Cox transformation essentially finds the best possible value of λ that makes the data as close to a normal distribution as possible. It automates the process of finding an appropriate power transformation.

When to Use Box-Cox Transformation for Outlier Handling:

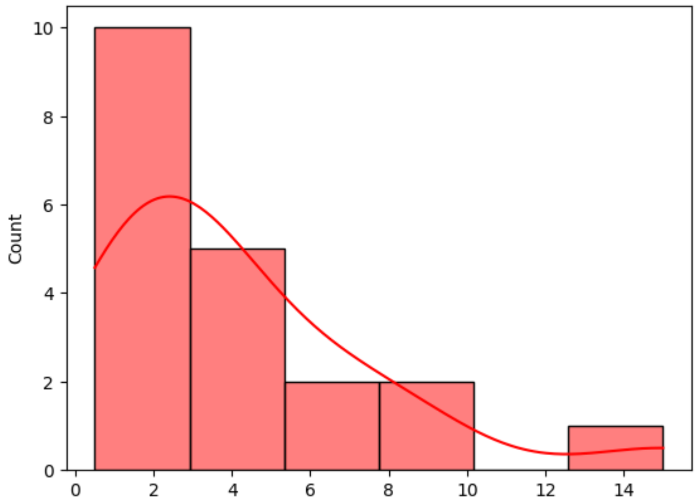
- **Non-Normal Data:** When your data is significantly skewed, and you want to make it more symmetric.
- **Positive Data:** The Box-Cox transformation requires the data to be strictly positive.
- **Stabilizing Variance:** When the variance of the data is not constant across different levels of the independent variable (heteroscedasticity).

Example:

Let's consider a dataset of the time it takes for a website to load (in seconds):

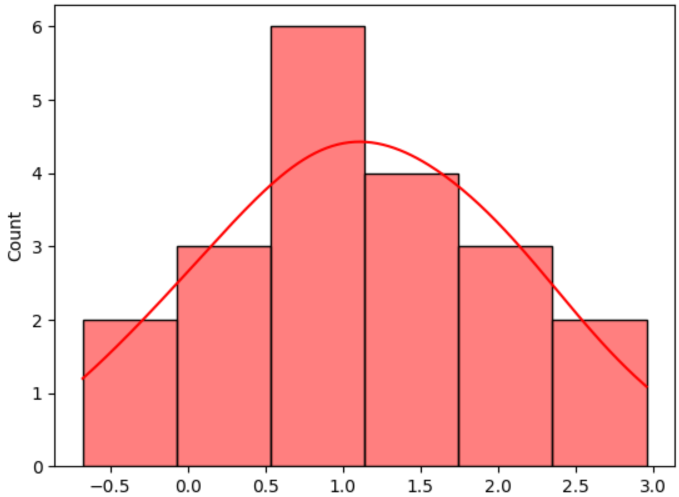
loading_times = [0.5, 0.8, 1.0, 1.2, 1.5, 1.8, 2.0, 2.2, 2.5, 2.8, 3.0, 3.5, 4.0, 4.5, 5.0, 6.0, 7.0, 8.0, 9.0, 15.0]

This data is likely to be right-skewed, with a few long loading times (e.g., 15 seconds). Let's visualize the distribution:



Box cox transformed data looks like below:

Box-Cox transformed loading times: [-0.67, -0.22, 0 , 0.18 , 0.41 , 0.59 , 0.70 , 0.80 , 0.94 , 1.06 , 1.13 , 1.30 , 1.45 , 1.57 , 1.69 , 1.89 , 2.07 , 2.22 , 2.36 , 2.95] and the visualization looks like below :



The transformation will reduce the impact of the large values, effectively handling the potential outlier.

Benefits:

- Effectively handles non-normality and reduces skewness.
- Often stabilizes variance.
- Reduces the impact of outliers.
- Provides a family of transformations, allowing for flexibility.

Cautions:

- Requires strictly positive data.
- The transformed data can be harder to interpret in its original units.
- The optimal λ needs to be estimated, which adds complexity.