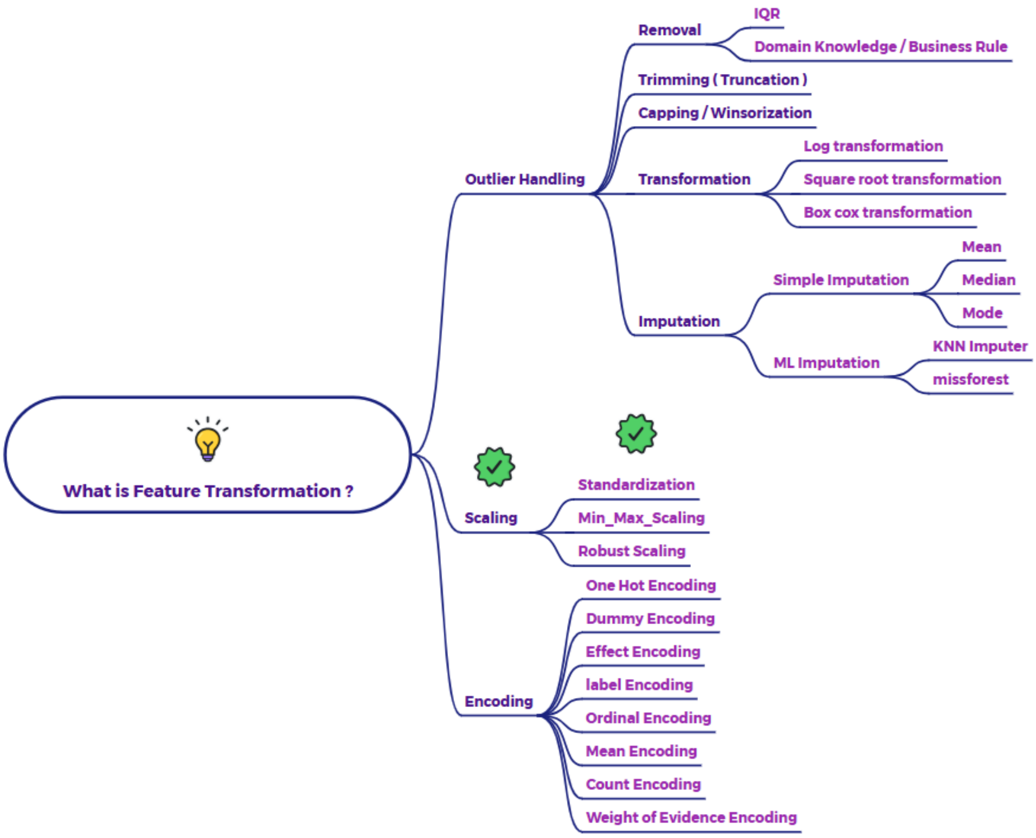
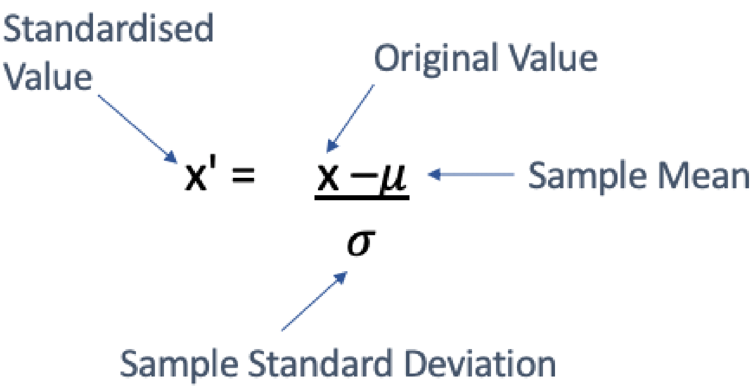
Explain Standardization for numerical variables



Standardization (Z-score Scaling)

1. **Explanation of Standardization**
    - Standardization is a scaling technique that transforms numerical data to have a mean of 0 and a standard deviation of 1. It involves rescaling the values by subtracting the mean and dividing by the standard deviation.
2. **How to Calculate Standardization**



The formula for standardization is $x' = (x-\mu)/\sigma$

Where:

- x : Original value
- μ : Mean of the variable
- σ : Standard deviation of the variable

Example:

Let's say we have the following data for a variable "Age": 25, 30, 35, 40, 45

Calculate the mean (μ): (25 + 30 + 35 + 40 + 45) / 5 = 35

Calculate the standard deviation (σ): The standard deviation is approximately 7.07.

Standardize each value:

- For 25: (25 - 35) / 7.07 ≈ -1.41
- For 30: (30 - 35) / 7.07 ≈ -0.71
- For 35: (35 - 35) / 7.07 = 0
- For 40: (40 - 35) / 7.07 ≈ 0.71
- For 45: (45 - 35) / 7.07 ≈ 1.41

So, the standardized "Age" values are: -1.41, -0.71, 0, 0.71, 1.41

3. When to Use Standardization

- When your data has a Gaussian (normal) distribution, or when the algorithm you're using assumes a Gaussian distribution.
- When you don't have specific knowledge about the distribution of your data.
- Standardization is generally preferred for algorithms that are sensitive to the scale of the data, such as:
  - Principal Component Analysis (PCA)
  - Linear Regression
  - Logistic Regression
  - Support Vector Machines (SVM)
  - Neural Networks

4. Strengths and Weaknesses of Standardization

- **Strengths:**
  - Not sensitive to outliers.
  - Transforms data to a standard scale, making it easier to compare variables.
  - Can improve the performance of many machine learning algorithms.
- **Weaknesses:**
  - Assumes data is normally distributed, which may not always be the case.
  - The exact shape of the original distribution is not preserved.