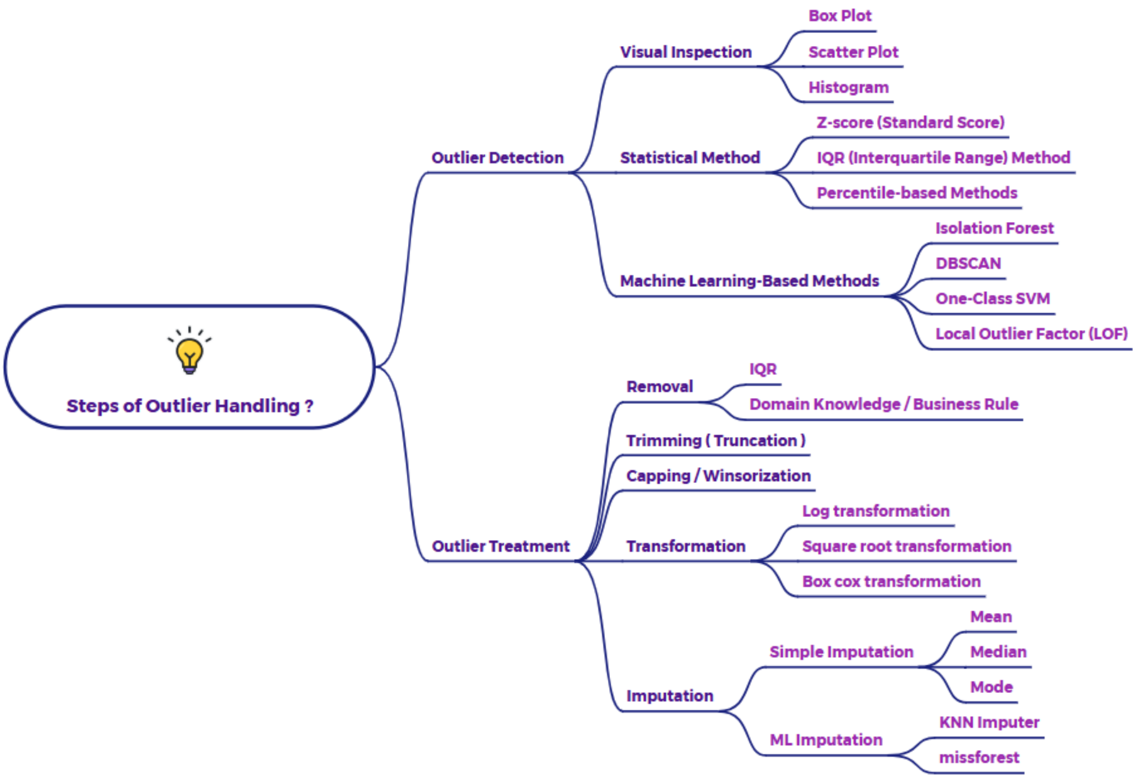
Different stages in Outlier handling?



Outlier handling is a two-stage process:

1. Outlier Detection

- This stage involves identifying data points that deviate significantly from the norm. The image presents three categories of methods:
- **Visual Inspection:**
    - **Box Plot:** Displays the distribution of data and can highlight potential outliers as points outside the whiskers.
    - **Scatter Plot:** Useful for identifying outliers in bivariate data, where outliers appear far from the general pattern.
    - **Histogram:** Shows the frequency of data values and can reveal outliers as isolated bars at the tails of the distribution.

- **Statistical Methods:**
  - **Z-score (Standard Score):** Measures how many standard deviations a data point is from the mean; values with high Z-scores are considered outliers.
  - **IQR (Interquartile Range) Method:** Defines outliers as values falling below Q1 - 1.5 * IQR or above Q3 + 1.5 * IQR.
  - **Percentile-based Methods:** Identify outliers as values below a low percentile (e.g., 5th) or above a high percentile (e.g., 95th).
- **Machine Learning-Based Methods:**
  - **Isolation Forest:** Isolates outliers by randomly partitioning the data space; outliers require fewer partitions to be isolated.
  - **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):** Clusters data points based on density and labels points in low-density regions as outliers.
  - **One-Class SVM (Support Vector Machine):** Learns a boundary around normal data and identifies points outside this boundary as outliers.
  - **Local Outlier Factor (LOF):** Compares the local density of a data point to that of its neighbors; outliers have a significantly lower local density.

2. Outlier Treatment

- Once outliers are detected, this stage involves deciding how to handle them. The image presents several treatment methods:
- **Removal:**
  - **IQR:** Remove data points identified as outliers using the IQR method.
  - **Domain Knowledge / Business Rule:** Remove outliers based on specific knowledge about the data or established rules.
  - **Trimming (Truncation):** Remove a fixed percentage of extreme values from both ends of the distribution.

- **Transformation:**
    - **Log Transformation:** Compresses the scale of the data and can reduce the impact of right-skewed outliers.
    - **Square Root Transformation:** Similar to log transformation but less extreme in its effect.
    - **Box-Cox Transformation:** A family of transformations that can make the data more closely fit a normal distribution.
- **Imputation:**
    - **Simple Imputation:**
        - **Mean:** Replace outlier values with the mean of the remaining data.
        - **Median:** Replace outlier values with the median of the remaining data.
        - **Mode:** Replace outlier values with the mode of the remaining data (for categorical data).
    - **ML Imputation:**
        - **KNN Imputer:** Imputes missing values using the K-nearest neighbors algorithm.
        - **MissForest:** Imputes missing values using a random forest algorithm.

N.B: Machine learning based Outlier detection will be handled as a separate section called Anomaly detection.