**Different ways to detect outliers,**

Detecting outliers is crucial in data analysis to identify data points that deviate significantly from the rest of the dataset. Here are several common methods to detect outliers:

**Ways to remove outliers:**

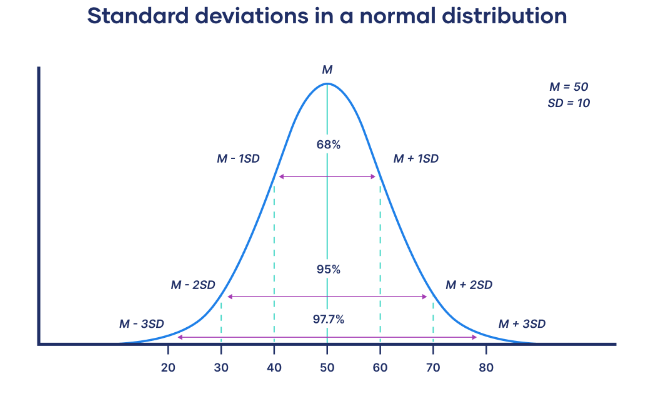
1. **Trimming**
2. **Capping**

**Different ways to detect outliers**

1. **Z-Score: or Standard Deviation Method**:
2. **Box Plot Method** **or** **Interquartile Range (IQR) Method**
3. **Percentile/Winsorization:**

1) **Standard Deviation Method**: Outliers can be identified by looking for data points that lie beyond a certain number of standard deviations from the mean. Typically, data points that are more than 3 standard deviations away from the mean are considered outliers.

2) **Z-Score:** Z-score measures how many standard deviations a data point is from the mean. Data points with a Z-score greater than a certain threshold (e.g., 3 or -3) are considered outliers.



**3) Box Plot Method (Tukey's Method)** **Interquartile Range (IQR) Method:**

Box plots visually display the distribution of data and identify outliers as points that fall outside the whiskers of the box plot. Whiskers typically extend to 1.5 times the interquartile range (IQR) from the first and third quartiles.

IQR=*Q*3−*Q*1

This method defines outliers as data points that fall

- below Q1 - 1.5 \* IQR or

- above Q3 + 1.5 \* IQR,

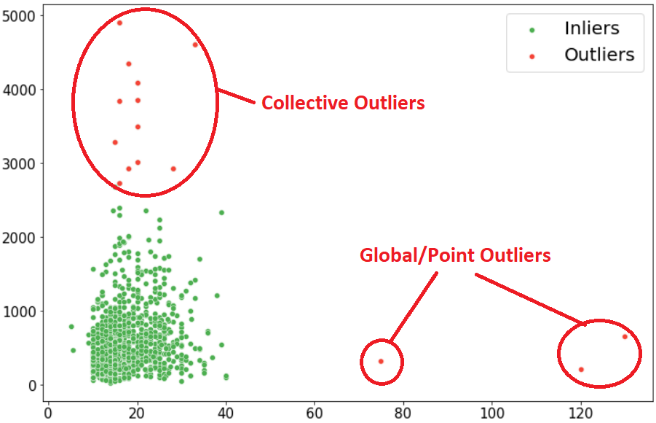
where Q1 and Q3 are the first and third quartiles, respectively.

**4) Percentile/Winsorization:**

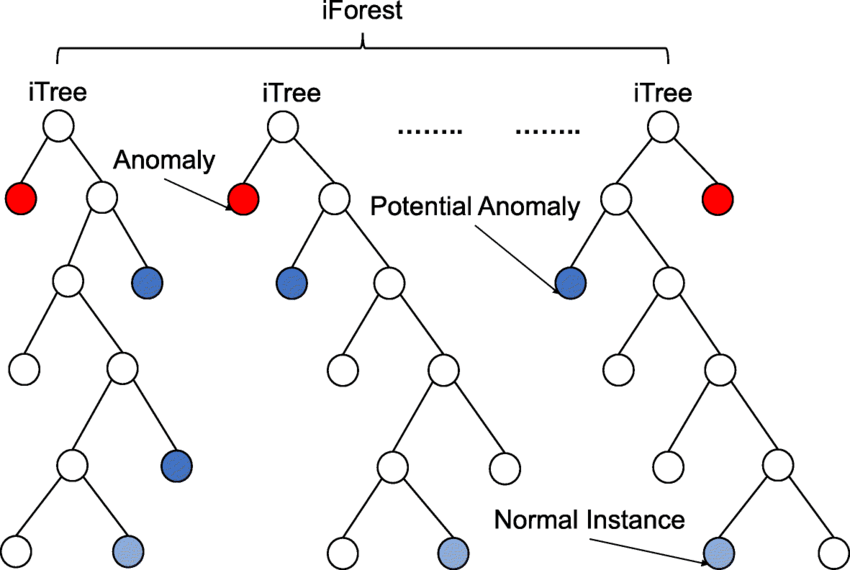
Winsorization is a method used to reduce the influence of outliers in a dataset by replacing extreme values with less extreme ones. In Winsorization, instead of removing outliers outright, you adjust them to a specified percentile value.

Here's how you can perform Winsorization using percentiles:

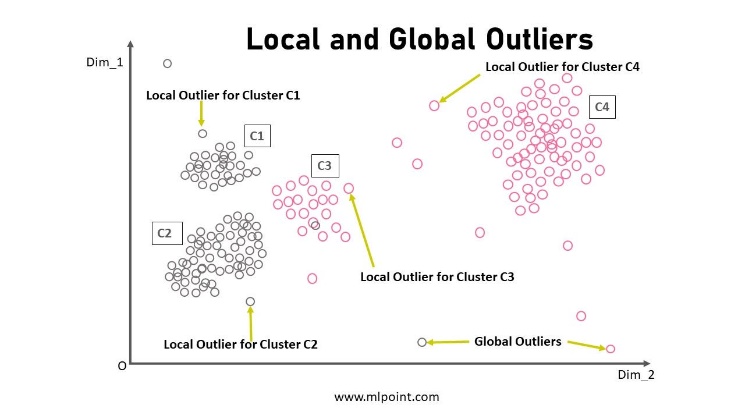
1. **Calculate Percentiles**: Determine the desired percentiles of the data. Typically, you'll calculate the lower and upper percentiles, such as the 5th and 95th percentiles.
2. **Identify Outliers**: Identify data points that fall below the lower percentile or above the upper percentile.
3. **Winsorize Outliers**: Replace outliers with values corresponding to the lower or upper percentile. For example, if an outlier falls below the 5th percentile, you would replace it with the value at the 5th percentile. Similarly, if it falls above the 95th percentile, you would replace it with the value at the 95th percentile.
4. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise)**: It's a clustering algorithm that identifies outliers as points that are in low-density regions.



7. Isolation Forest: This is an unsupervised learning algorithm that identifies anomalies by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of that feature.



8. Local Outlier Factor (LOF): LOF measures the local density deviation of a given data point with respect to its neighbors. Points with significantly lower density than their neighbors are considered outliers.



9. Robust Random Cut Forest (RRCF): It's an algorithm for detecting outliers in streaming data. It builds a tree-based ensemble model to detect anomalies.

10. Histogram-based Outlier Detection: This method involves creating a histogram of the data and identifying bins with significantly lower frequencies as potential outliers.

Each method has its advantages and limitations, and the choice of method depends on factors such as the nature of the data, the presence of noise, and the specific goals of the analysis. It's often advisable to use multiple methods for outlier detection to gain a comprehensive understanding of the data.