Outliers in Data

What is an Outlier?

An outlier is a data point that is significantly different from the rest of the data.

It lies far away from other values and can distort statistical results such as mean, standard deviation, or regression.

Example:

If most students scored between 60-80, but one scored 10 or 100, those could be outliers.

Example: Scores: 65, 68, 70, 72, 75, 78, 10 (Outlier)

Causes of Outliers:

1. Human error (e.g., typing mistake)
2. Measurement error (e.g., faulty sensor)
3. Natural variation (e.g., rare event)
4. Fraud or noise

Why Outliers Matter?

- Can skew results (mean, standard deviation)
- Can stretch graphs in visualizations
- Can confuse machine learning models or cause overfitting

How to Detect Outliers?

1. Visualization Techniques:
   - Box Plot
   - Scatter Plot
   - Histogram

2. Statistical Methods:
   - Z-score: If $|Z| > 3$ Outlier
   - IQR Method:

     Outlier < Q1 - 1.5×IQR or Outlier > Q3 + 1.5×IQR

Outliers in Data

How to Handle Outliers?

- Remove: Delete outliers if they're due to error
- Cap/Impute: Replace with max/min or average values
- Transform: Use log/sqrt to reduce impact
- Use robust models: Models like Decision Trees handle outliers well

Summary:

What: Extreme value far from rest of data

Detection: Z-score, IQR, Visual plots

Impact: Skews stats, affects model performance

Handling: Remove, transform, or use robust methods