Handling Outliers – Z-Score Method

1. What is the Z-Score Method?

- The **Z-score** measures how many **standard deviations** a data point is from the **mean**.
- Formula:

$$Z = \frac{x - \mu}{\sigma}$$

where:

- x = individual data point
- μ\mu = mean of the feature
- σ\sigma = standard deviation

- Typically, if **Z > 3** or **Z < -3**, the point is considered an **outlier** (threshold can vary).

2. Z-Score-Based Outlier Treatment

- After calculating Z-scores:
  - Remove rows where Z-score exceeds the threshold.
  - This helps **reduce noise** and improve model performance.

- **Important**: Always apply **feature scaling/normalization** before using Z-scores if data isn't already standardized.

3. Capping

- Instead of removing outliers, you can **cap (clip)** values at a certain threshold.
- ***Example:***
  df['column'] = np.where(df['column'] > upper_limit, upper_limit, df['column'])
- **Capping preserves row count**, which is useful when you want to **keep all data points** but reduce the impact of extreme values.

➢ Key Takeaway:

Z-score is a **simple and effective** method for detecting and treating outliers in normally distributed data. Depending on the use case, you can choose to **remove or cap** them to improve model quality.