00. Handling Outliers

Methods:

1. Z-Score
2. IQR
3. Modified Z-Score

1. Z-Score

- It measures how far away the data in terms of a mean

Formula:

$Z = X-\mu / \sigma$

If |Z| is greater than 3, meaning it is an outlier

Use case:

- When there is unusual selling price or high transaction amount and more
- It follows **Gaussian distribution**, recommended to use when data is **normally distributed**

2. IQR

- It detect the outlier by analyzing the data distribution between **Q1 (25 percentile)** and **Q3 (75 percentile)**

Formula:

IQR = Q3 - Q1

Use case:

- Uses to identify unusual selling price or very big amount of transaction amount
- If data is heavly skewed or tailed, it is best to use to detect outliers

3. Modified Z-Score

- It usually used for small dataset with messed with many outliers
- It considered robust because it use**s median** instead of **mean**

Formula:

$M = 0.6745 * (X\text{-median}) / MAD$

$MAD = \text{median}(| Xi - \text{median}(X) |$

if if `|M| > 3.5` it is outlier else no

Use case:

- For small dataset with extreme outliers