Basic Techniques To Handle Outliers

1) Trimming: (Remove Outliers)

Temperature reading:

- 40, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 56, 89

1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5

Basic Techniques To Handle Outliers

2) Drop rows with extreme values:
Works well if outliers are **data errors**
(e.g., negative height, impossible values).

	Age	Income	Position
	23	40	HR
	62	89	HR
	53	89	HR
Drop this row -->	-12	23	IT
	62	89	HR

However, it’s essential to **carefully consider the percentage to be trimmed** to avoid removing too much data

Basic Techniques To Handle Outliers

3) Capping: (a.k.a. Winsorization)

- 40, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 56, 89

1, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 5, 5

Replace extreme values with nearest acceptable boundary.

Example: Cap at 5th and 95th percentile.

Best for preserving dataset size and avoiding bias.

Basic Techniques To Handle Outliers

Capping prevents the complete removal of outliers and instead modifies their values to **align them with the nearby observations**.

This approach helps **control the impact of outliers** while **retaining their presence in the dataset**

Basic Techniques To Handle Outliers

4) Imputation:

Replace outliers with a more “typical” value (mean, median, mode).

Temp: - 40, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 56, 89

Temp: M, 1, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, M, M

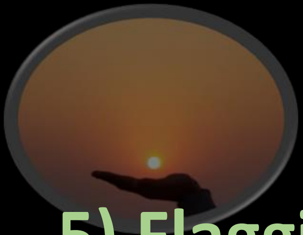
Here M is median or mean value

Shoe Size: 21, 2, 1, 2, 2, 2, 2, 2, 3, 4, 4, 5, 26, 89

Shoe Size: 2, 1, 2, 2, 2, 2, 2, 2, 3, 5, 4, 4, 2, 2

Here 2 is the mode.

Best for **small datasets** where every record matters.



Basic Techniques To Handle Outliers

5) Flagging Outliers:

Create a new binary column.

Example: is_outlier = 1 if outlier else 0

Age	Income	Position	Is_outlier
23	40	HR	0
29	89	HR	0
51	89	HR	0
98	23	IT	1
45	89	HR	0

Keeps information about outliers without deleting them.

Summary

Question	**Action**
Is it a data error?	Remove
Is it rare but valid?	Keep
Is it hurting model?	Transform / cap
Is it the target case?	Definitely keep
Is model sensitive?	Handle carefully
Is model robust?	Maybe ignore

Takeaway

1) Outliers should be detected during EDA,

2) handled during preprocessing, and

3) reconsidered during modeling — never removed blindly.

EXTRA

Fhdsklf
Fjdsklf
Fjskldf