# 📘 Day 8 – Outliers & Standardization

---

## 🔷 1. Outliers & IQR Rule

We already learned outliers are **extreme values**.

Day 8 formalizes the method to detect them using **quartiles (Q1, Q3)**.

- **Q1** = 25th percentile
- **Q2 (Median)** = 50th percentile
- **Q3** = 75th percentile
- **IQR (Interquartile Range)** = Q3 − Q1

👉 Outlier thresholds:

- **Mild outliers:**

  $$Lower\ bound = Q1 - 1.5 \times IQR, Upper\ bound = Q3 + 1.5 \times IQR$$
  $$\text{Lower bound} = Q1 - 1.5 \times IQR, \quad \text{Upper bound} = Q3 + 1.5 \times IQR$$

- **Strong outliers:**

  $$Lower\ bound = Q1 - 3 \times IQR, Upper\ bound = Q3 + 3 \times IQR$$
  $$\text{Lower bound} = Q1 - 3 \times IQR, \quad \text{Upper bound} = Q3 + 3 \times IQR$$

### Example:

- Q1 = 10k, Q2 = 1 lakh, Q3 = 5 lakh
- IQR = 5L − 10k = 4.9L
- Upper bound = 5L + 1.5×4.9L ≈ 12.35L

👉 Anyone earning **≥12.35L per month** = outlier.

📌 In **Python boxplots**, the default threshold = **1.5×IQR** (mild outliers).

---

## 🔷 2. How to Deal with Outliers

Outliers impact **mean** but not **median**.

Options:

1. **Drop them (not recommended)**
    - If only 1–2% of data are outliers → dropping might be fine.
    - BUT: you also lose related info (other columns).
2. **Replace with Median**
    - Since median is robust, we can replace outliers with the **50th percentile value**.
3. **Cap with Q1 & Q3 (Winsorization)**
    - If value > Q3 → set = Q3.
    - If value < Q1 → set = Q1.
    - This reduces outlier effect but keeps all rows.

👉 Example:

If income = 100L, Q3 = 5L → cap it at 5L.

---

## 🔷 3. Why Scale Data?

In datasets, features have different ranges.

- Age ≈ 20–80
- Income ≈ 10k – 1 crore

⚠️ Problem:

Some ML models use **distance-based calculations** (like k-NN, SVM, clustering).

If features are not scaled → large-value features dominate.

Example:

Distance between (20, 50,000) and (30, 20,000):

$$(30-20)2+(20000-50000)2\sqrt{(30-20)^2+(20000-50000)^2}$$

→ Income dominates, age ignored.

---

## 🔷 4. Standardization (Z-Score Scaling)

👉 Formula:

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

- Mean becomes 0
- Standard deviation becomes 1
- Result = **standard normal distribution**

### Properties:

- After standardization:

  $$\mu = 0, \sigma = 1\mu = 0, \quad \sigma = 1$$

- Variance = 1
- Works well when data is normally distributed.

👉 Example:

If student marks = 70, mean = 60, SD = 5:

$$z = 70 - 605 = 2z = \frac{70-60}{5} = 2$$

= “score is 2 SDs above average”.

---

## 🔷 5. Normalization (Min–Max Scaling)

👉 Formula:

$$x\prime = x - min(x)max(x) - min(x)x' = \frac{x-\min(x)}{\max(x)-\min(x)}$$

- Scales values to [0,1] range.
- Useful when we want all features between 0–1 (e.g., neural networks).

### Example:

Data = {20, 30, 50}, min = 20, max = 50

- For 30:

$$30-2050-20=1030=0.33\frac{30-20}{50-20}=\frac{10}{30}=0.33$$

---

## 🔷 6. Empirical Rule (for normal data)

- 68% of data → within 1 SD of mean
- 95% → within 2 SD
- 99.7% → within 3 SD

👉 Outliers beyond ±3 SD = strong outliers.

---

## ✅ Summary

- **Outliers**: detected using IQR rule (1.5×IQR = mild, 3×IQR = strong).
- **Handling outliers**: Drop, Replace with median, or Winsorize.
- **Scaling**: ensures features are comparable in models.
- **Standardization**: (z-score) → mean=0, SD=1.
- **Normalization**: (min-max) → scale between 0–1.
- **Empirical Rule**: 68-95-99.7 rule for normal distribution.

---

## 📝 Practice Problems

1. Dataset incomes: {10k, 20k, 30k, 40k, 12L}.
   - Find Q1, Q3, IQR.
   - Detect if 12L is an outlier (mild or strong).
2. Marks: {20, 30, 40, 50, 60}, mean = 40, SD = 10.
   - Standardize each mark (z-score).
3. Heights: {150, 160, 170, 180}, min=150, max=180.
   - Normalize values into [0,1].

---