KARNAVATI UNIVERSITY

Data Science

By:
Dr. Shikha Deep

OUTLIERS

- **Outliers** are values that are **very different** from the rest of the data.
- They may indicate variability, errors, or something important.

Ex: (1) In a dataset of ages [22, 24, 23, 25, 21, 100], the value 100 is an outlier.

(2) fraud detection, equipment failure

Types of Outliers/Anomalies :

(1) Point Anomalies (Global Outliers): A single data point that is very different from the rest of the data. (By Box Plot and Z-score).

(2) Contextual Anomalies (Conditional Outliers) : A data point that is normal in some contexts but anomalous in others. (By Time Series)

(3) Collective Anomalies (Group Outliers) : A group of data points that together show an anomaly, even if individual points seem normal.

Methods to Detect Outliers

(1) IQR (Inter quartile range)

(2) Z-score

(3) Visulization Methods – Histogram, Scatter Plot

(4) Isolation Forest (ML-Based)

(5) DBSCAN (Clustering based)

(1) IQR (Inter-quartile Range) : Visulization

Code 1.

```
import pandas as pd
import matplotlib.pyplot as plt

data = [3, 3, 7, 8, 8, 10, 11, 12, 15, 18, 40]  # 40 is an outlier
df = pd.DataFrame(data, columns=['Years'])

# Box Plot
plt.boxplot(df['Years'])
plt.title('Box Plot for Teaching Years')
plt.show()
```

Code 2. Numerical representation

```
import numpy as np
data = [10, 12, 14, 15, 15, 16, 18, 19, 20, 29, 100]
data_sorted = sorted(data)

# Calculate Q1, Q3 and IQR
Q1 = np.percentile(data_sorted, 25)
Q3 = np.percentile(data_sorted, 75)
IQR = Q3 - Q1
lower_bound = Q1 - 1.5 * IQR
upper_bound = Q3 + 1.5 * IQR
```

```
# Detect outliers
outliers_iqr = [x for x in data if x < lower_bound or x > upper_bound]

print("Q1 =", Q1)
print("Q3 =", Q3)
print("IQR =", IQR)
print("Lower Bound =", lower_bound)
print("Upper Bound =", upper_bound)
print("Outliers using IQR method:", outliers_iqr)
```

(2) Z-Score

The Z-score (or standard score) tells you how many standard deviations a data point is from the mean.

Formula:

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

Where:

x = data point

μ = mean of the data

σ = standard deviation

Z-Score Outlier Rule:

- If $|Z| > 3$, the point is considered an outlier.
- That means the value is more than 3 standard deviations away from the mean.


Thanks!