Box Plot

A **box plot** (also known as a **box-and-whisker plot**) is a statistical visualization that helps summarize the **distribution of numerical data** and detect **outliers**. It is widely used in **Exploratory Data Analysis (EDA)** to understand data spread and variability.

1. Why Use a Box Plot?

✅ Summarizes large datasets concisely.

✅ Identifies **median, quartiles, and outliers**.

✅ Helps detect **skewness** and **spread of data**.

✅ Compares distributions across multiple categories.

2. Understanding Box Plot Components

A **box plot** consists of:

(a) Median (Q2 - 50th Percentile)

- The **middle value** of the dataset (when sorted).
- **Divides data into two halves**.
- Represented by a horizontal line inside the box.

(b) Quartiles (Q1 & Q3) and Interquartile Range (IQR)

- **Q1 (25th Percentile):** Middle value of the lower half of data.
- **Q3 (75th Percentile):** Middle value of the upper half of data.
- **IQR (Interquartile Range):**

  IQR = Q3 - Q1

  - Represents the **middle 50% of the data**.
  - Helps detect **outliers**.

(c) Whiskers (Data Spread Limits)

- **Lower Whisker:** Extends to the smallest value within $Q1 - 1.5 \times IQR$
- **Upper Whisker:** Extends to the largest value within $Q3 + 1.5 \times IQR$
- **Values beyond whiskers** are considered **outliers**.

(d) Outliers (Extreme Data Points)

- **Values outside** the range of whiskers are plotted as **dots**.
- Indicates **anomalies or rare values** in data.

---

3. Box Plot Interpretation

(a) Symmetric Distribution (No Skewness)

- Median is **centered** inside the box.
- Whiskers are **equal in length**.
  📌
  Example: Normally distributed data.

(b) Right-Skewed Distribution (Positive Skewness)

- **Median closer to Q1**.
- **The right whisker is longer**.
  📌
  Example: Salaries (few very high values).

(c) Left-Skewed Distribution (Negative Skewness)

- **Median closer to Q3**.
- **The left whisker is longer**.
  📌
  Example: Exam scores (most students score high).

(d) Presence of Outliers

- Outliers appear as **individual points** outside whiskers.
  📌

Example: Errors in data entry, rare events in finance.

---

4. Creating a Box Plot in Python

```
import matplotlib.pyplot as plt
import seaborn as sns

# Sample dataset
data = [10, 12, 15, 18, 20, 22, 25, 27, 30, 35, 40, 100]  # 100 is an outlier

# Create Box Plot
plt.figure(figsize=(6,4))
sns.boxplot(data=data, color="skyblue")
plt.title("Box Plot Example")
plt.show()
```

📌 **Observations:**

✅ **Whiskers show data spread.**

✅ **Box shows central 50% data.**

✅ **Outlier (100) appears as a separate dot.**

---

5. Comparing Multiple Distributions

Box plots can compare **distributions across different categories**.

```
import seaborn as sns
import pandas as pd

# Sample Data
df = pd.DataFrame({
    "Category": ["A"]*10 + ["B"]*10 + ["C"]*10,
    "Value": [10, 12, 15, 18, 20, 22, 25, 27, 30, 35,
              15, 18, 22, 24, 28, 30, 32, 35, 38, 40,
              12, 16, 19, 21, 23, 26, 29, 31, 33, 37]
})

# Create Box Plot
```

```
plt.figure(figsize=(8,5))
sns.boxplot(x="Category", y="Value", data=df, palette="coolwarm")
plt.title("Box Plot by Category")
plt.show()
```

📌 **Observations:**

✅ **Box heights show data spread per category**.

✅ **Comparing medians helps understand group differences**.

---

6. Key Takeaways

✅ **Box plots summarize distributions using five key statistics (min, Q1, median, Q3, max).**

✅ **IQR helps detect outliers.**

✅ **Skewness affects whisker length.**

✅ **Useful for comparing multiple groups.**
