Box Plot

Problem:
Suppose we are given following numerical data:

[5, -2, 20, 13, 4, -20, 15, 8, 10, 12, 6, 40, 45, 1]

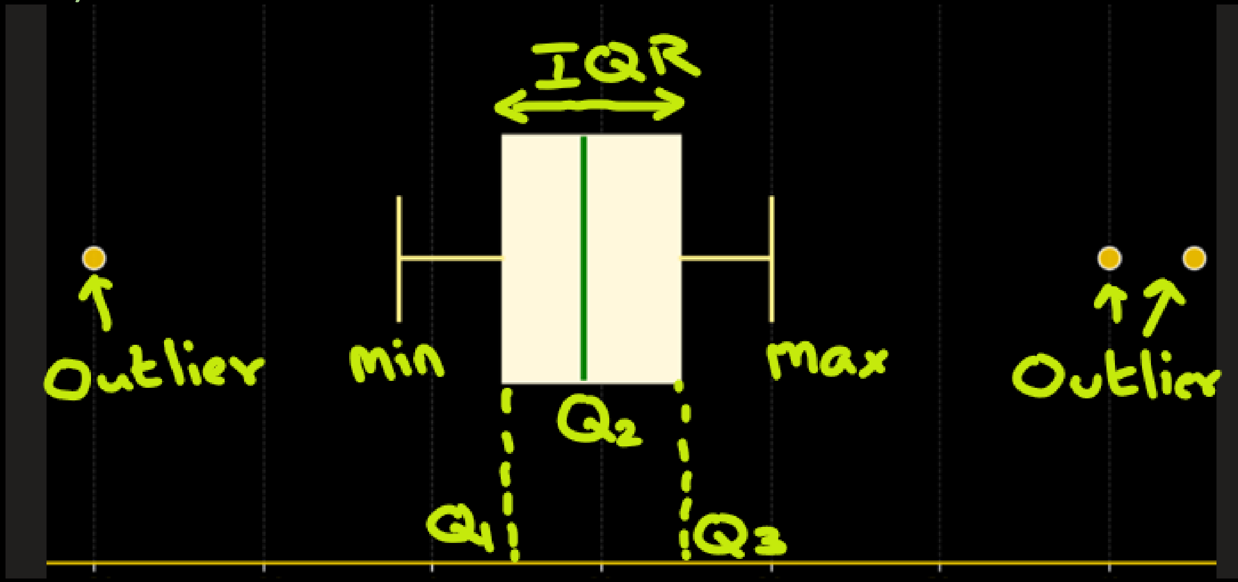
Which single graph can show the data’s **spread, center, outliers, and symmetry**—all at once ?

A. Box plot
B. Scatter plot
C. Histogram

Box Plot

A **boxplot** (a.k.a. **box-and-whisker plot**) is a **graphical representation of data distribution** based on **five key summary** statistics:

1. **Minimum** whisker– smallest non-outlier value
2. **Q1 (First Quartile)** – 25% of data lies below this (25th percentile)
3. **Q2 (Median)** – middle value (50th percentile)
4. **Q3 (Third Quartile)** – 75% of data lies below this (75th percentile)
5. **Maximum** whisker– largest non-outlier value



The box plot show following:

- The **box** shows the **Interquartile Range (IQR)**
  IQR = Q3 - Q1 (the middle 50% of data).
- The **line inside** is the **median** — the center of your data.
- The **whiskers** extend to min and max values (excluding outliers).
- **Dots or stars outside whiskers** indicate **outliers**.

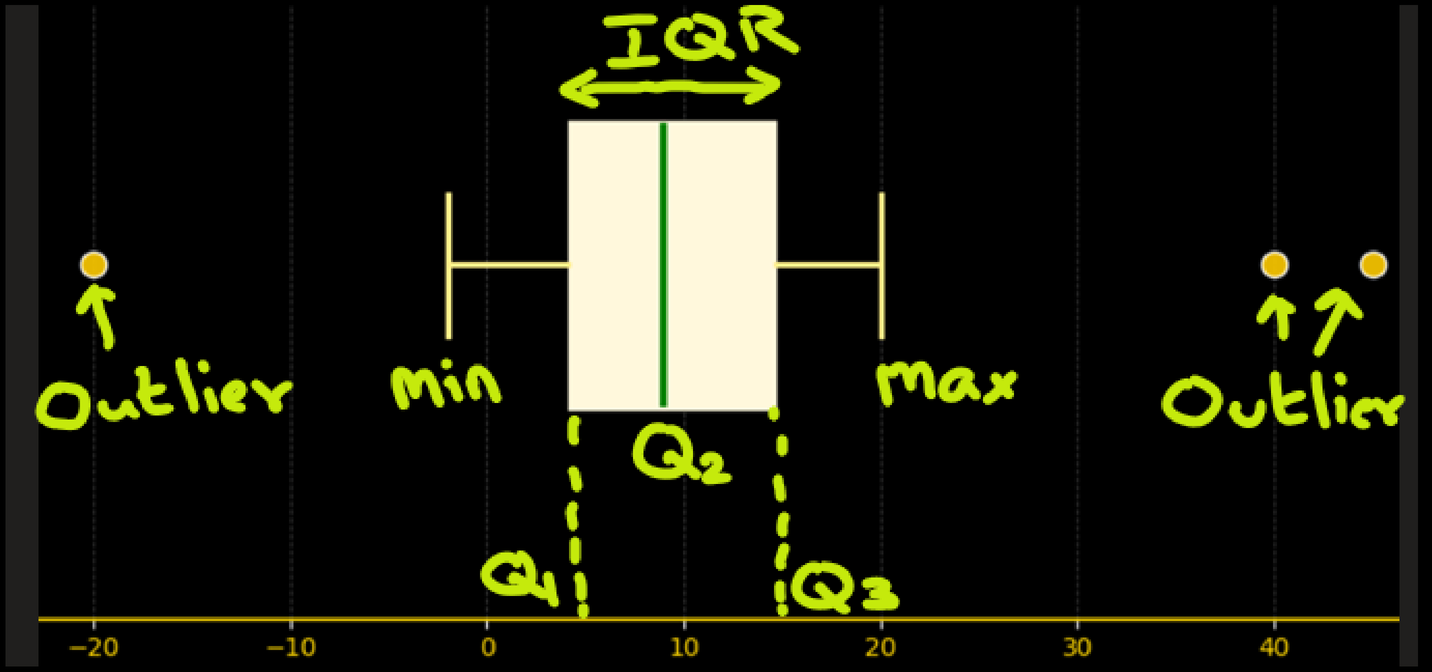
Box Plot

Example: Let’s say we have temperature readings:

[5, -2, 20, 13, 4, -20, 15, 8, 10, 12, 6, 40, 45, 1]

I want to see the 5 key summaries: **Minimum, Q1, Q2, Q3, Maximum**

Ans:
Let’s sort the data first:
-20, -2, 1, 4, 5, 6, 8, 10, 12, 13, 15, 20, 40, 45



From numerical computation we see that,
Minimum whisker= -2
Q1: 4.25
Q2: (Median) 9.00
Q3: 14.50
Maximum whisker: +20

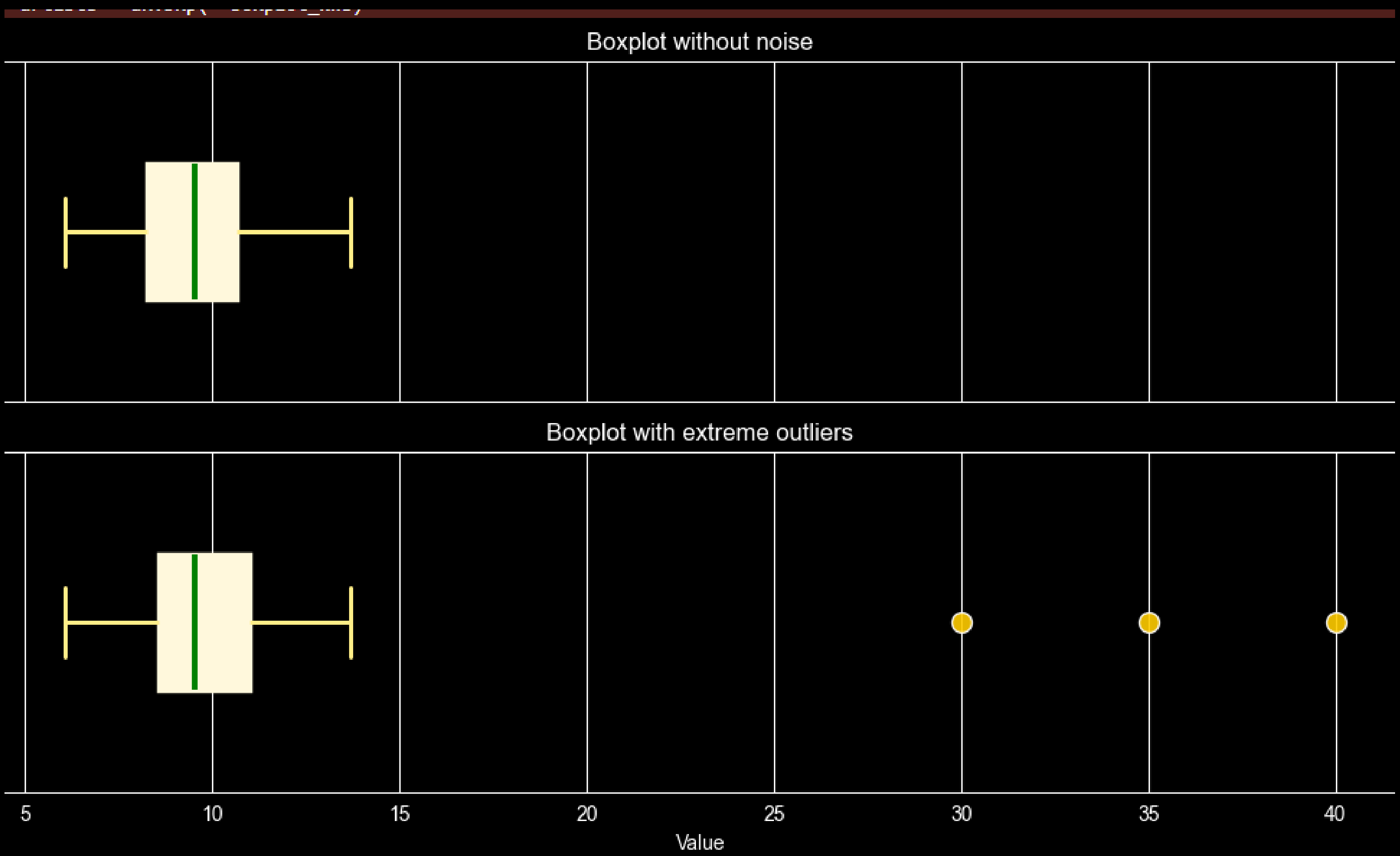
Outliers: [-20, 40, 45]
IQR = Q3 – Q1
= 14.5 – 4.25
= 10.25 (the middle 50% of the data lies within a range of 10.25 units)

Box Plot: Robust To Noise

Example: Let’s say we receive some noise to temperature readings:
[10.99, 9.72, 11.3, 13.05, 9.53, 9.53, 13.16, 11.53, 9.06, 11.09, 9.07, 9.07, 10.48, 6.17, 6.55, 8.88, 7.97, 10.63, 8.18, 7.18, 12.93, 9.55, 10.14, 7.15, 8.91, 10.22, 7.7, 10.75, 8.8, 9.42, 8.8, 13.7, 9.97, 7.88, 11.65, 7.56, 10.42, 6.08, 7.34, 10.39, 11.48, 10.34, 9.77, 9.4, 7.04, 8.56, 9.08, 12.11, 10.69, 6.47, **30**, **35**, **40**]
The noise is shown in red underline. Let’s plot boxplot for both data without and with noise

We see that

- The median (center line) barely changes.
- The box (IQR) stays almost the same.
- Outliers are shown separately as points: showing that the boxplot is robust to extreme values.



Boxplot without noise

Boxplot with extreme outliers

Box Plot: Comparing Groups

Box plots are used to compare distribution of data among groups

Setosa

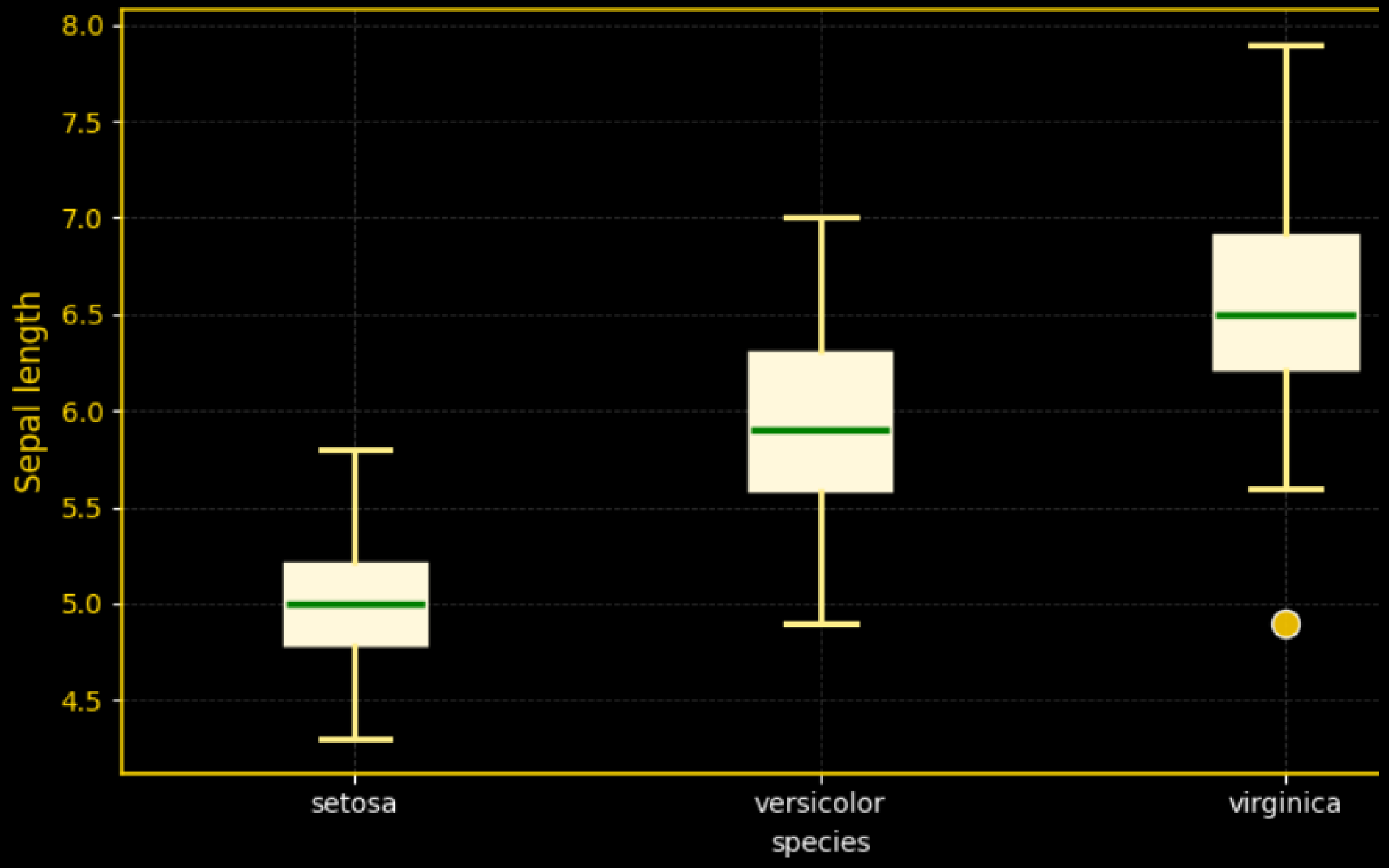
- Has the smallest sepal lengths overall.
- The median is around 5.0 cm.
- The spread (IQR) is tight → values are consistent.
- No outliers

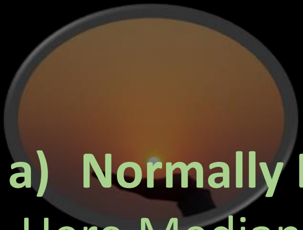
Versicolor

- Intermediate sepal length (median around 5.9–6.0 cm).
- Slightly wider IQR → moderate variation.
- No outliers

Virginica

- Has the largest sepal lengths, median around 6.5–6.6 cm.
- Slightly more spread, indicating more variation within the species.
- Whiskers extend higher → presence of some longer sepals.
- Has an outlier

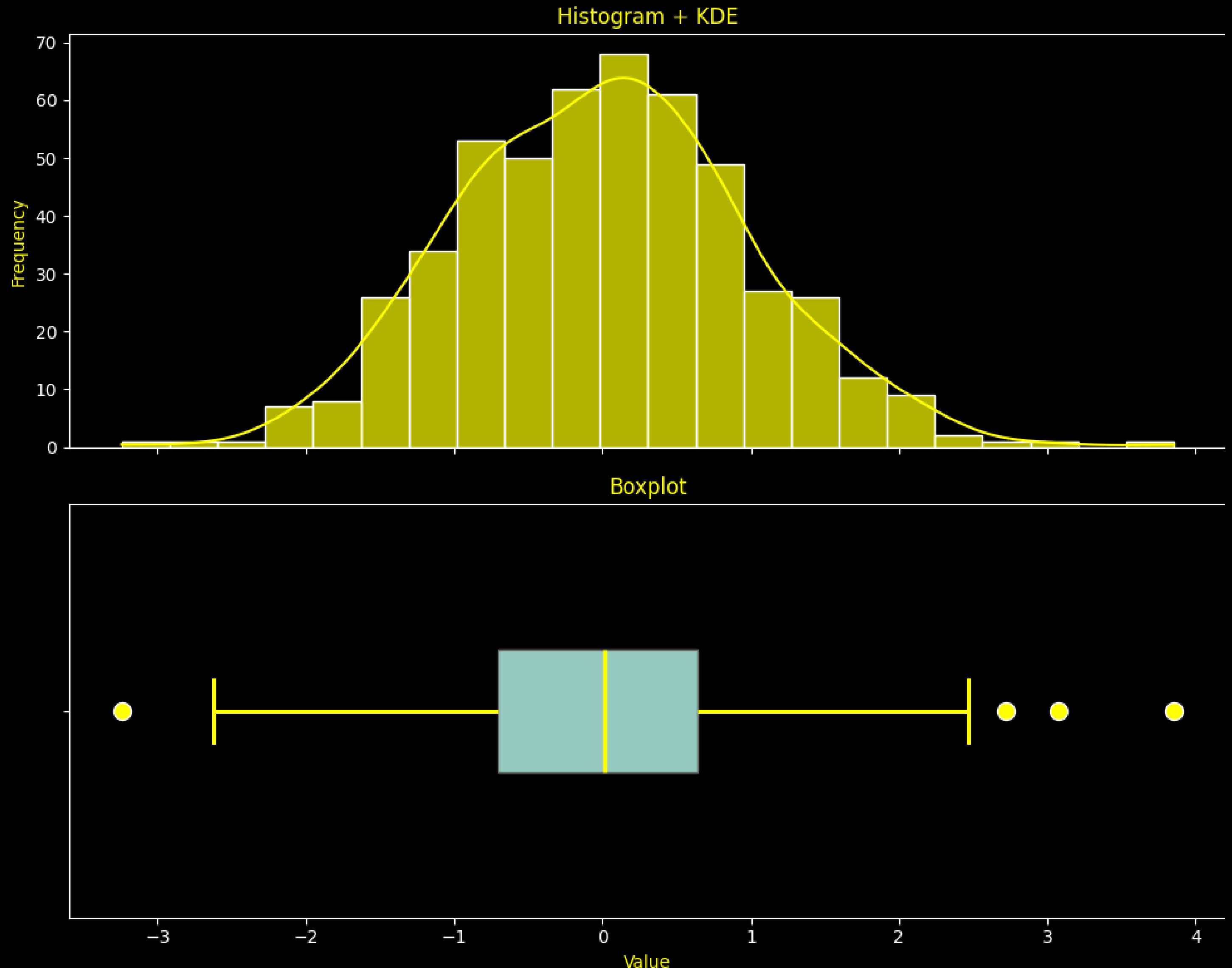




Box Plots shows Skewness of the data

a) Normally Distributed :

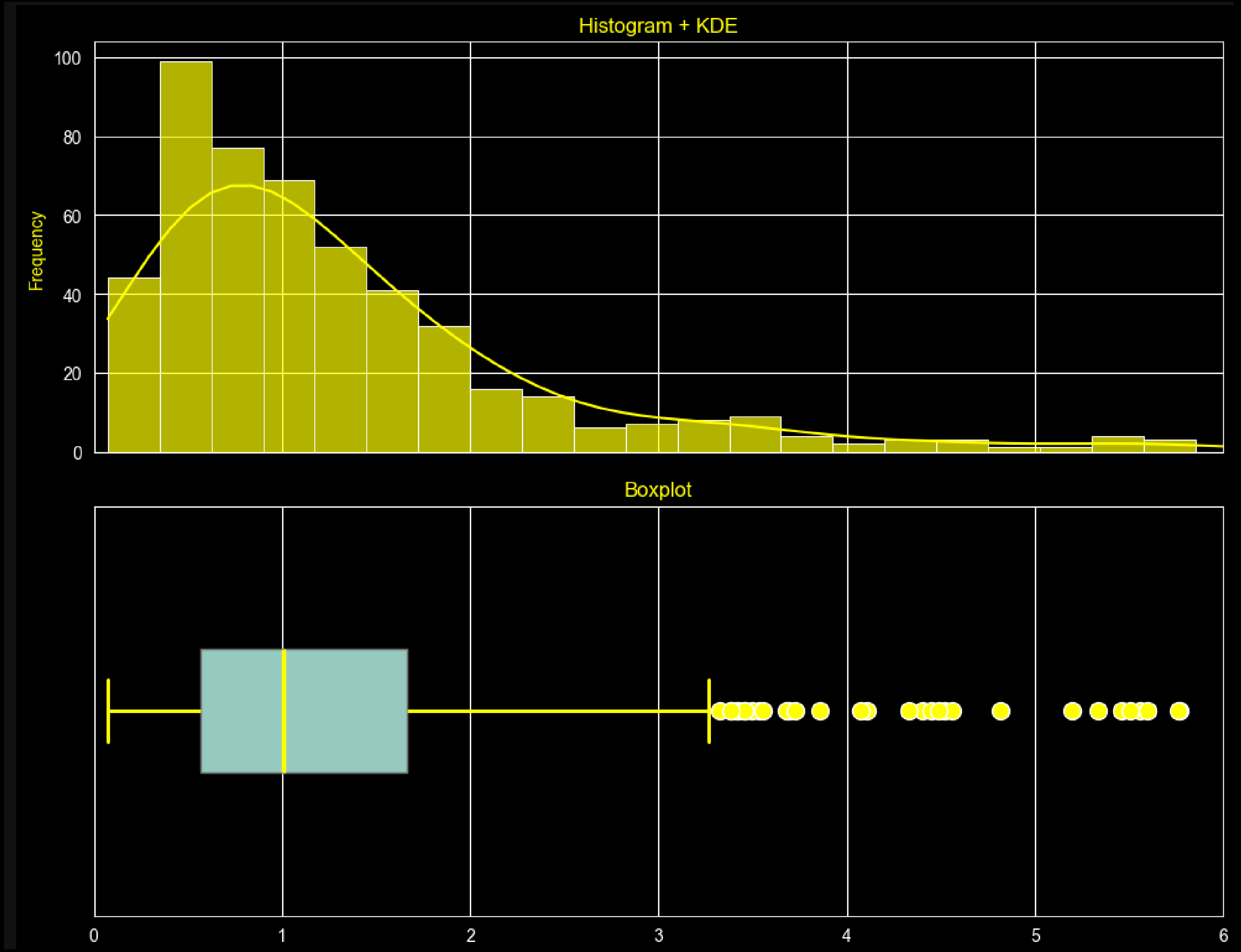
Here Median is at the **center** of the Box and the **whiskers** are almost the **same on both the ends**



Box Plots shows Skewness of the data

b) Positive Skew (Right Skew):

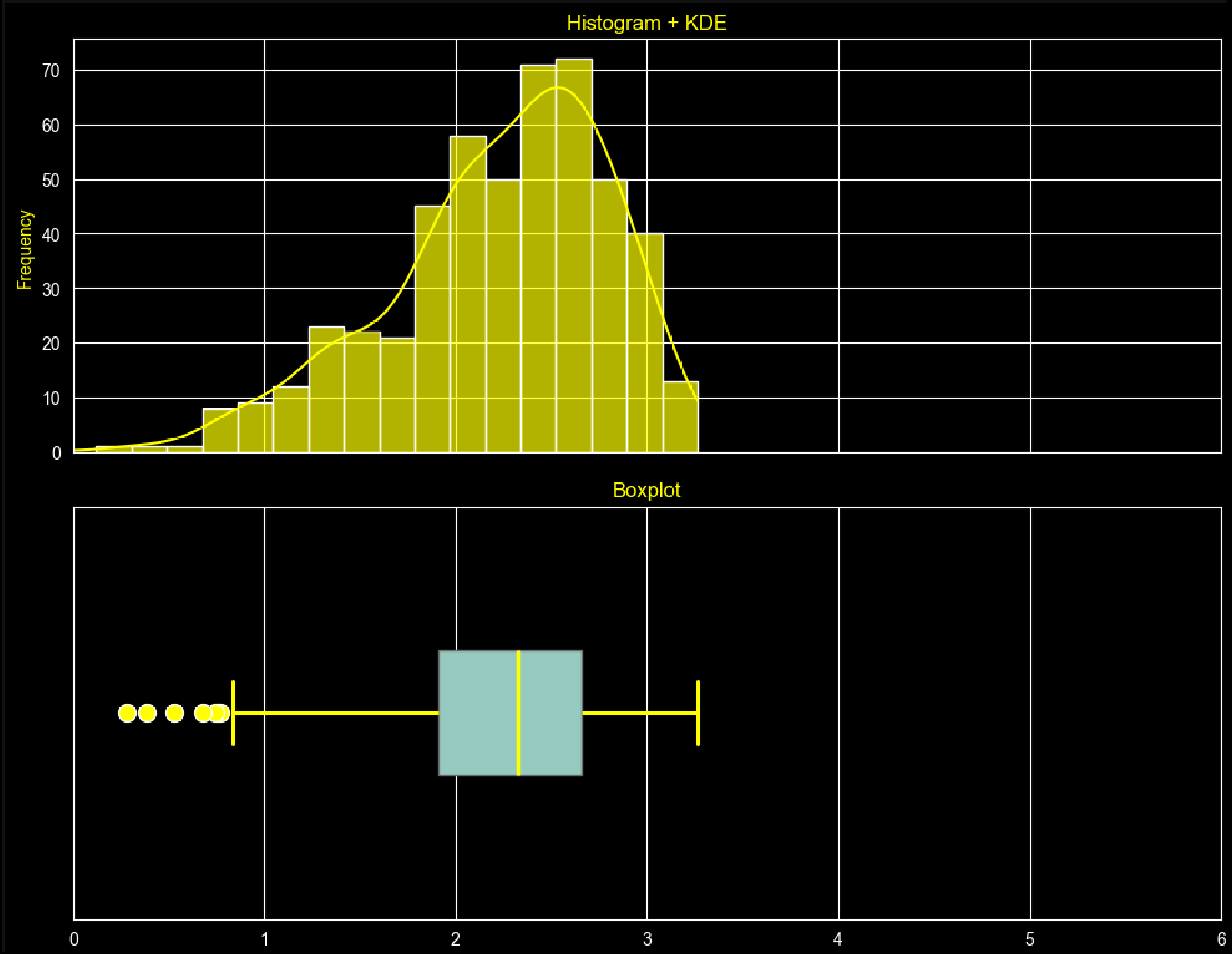
Here the Median(Q2) lies **closer to the First Quartile(Q1)** and the **whisker at the lower end is shorter**



Box Plots shows Skewness of the data

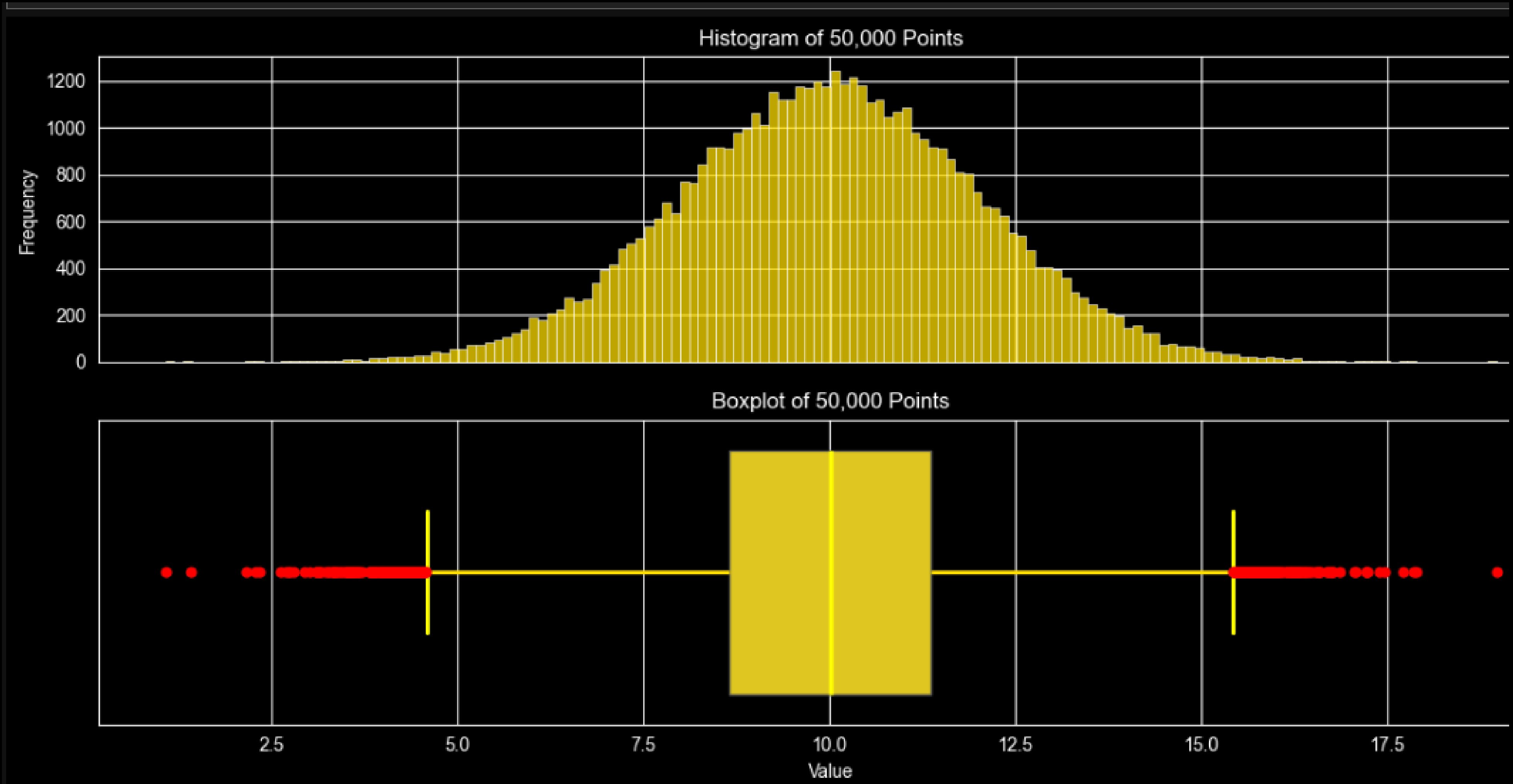
c) Negative Skew (Left Skew):

Here the Median(Q2) lies **closer to the Third Quartile(Q3)** and the **whisker at the upper end is shorter.**



Box Plots: Great with Large Datasets for Summary

- Histogram needs many bins to make sense and can look cluttered with 50k points.
- Boxplot summarizes the entire dataset in a clean, compact view with median, IQR, and outliers.
- Boxplot still looks clean while the histogram may need many bins and can look messy



What Makes Boxplots Special ?

Outlier Detection – Instantly flags unusual data points that could indicate errors or interesting phenomena.

Robust to Noise – Uses medians and quartiles, so it’s **less affected by extreme values** than mean-based charts.

Group Comparison – Easily compare distributions between multiple categories side by side.

Skewness – Helps to identify skewness in data, if any.

Great with Large Datasets for Summary – While histograms need many bins, boxplots scale beautifully even with huge datasets. It summarizes thousands of data points in one small plot.

Hfdslfds

fksdljfsD|]

fkdslfa