



What are Outliers?

📊 Outliers are the data points that are significantly differ from other data points. This may arise due to the error in the measurement or Wrong data entry.

How are they affecting the Model?

📉 Bias in Training:

The Outliers can Skew the models learning Process. Causing it to focus too much on extreme values rather than learn over all pattern in the data.

🚫 Poor Generalization:

If we train the data with the outliers may perform well in training data and poor performance in the unseen data and leads to the overfitting.

⚠️ Algorithm Sensitivity:

Algorithm like Linear regression and kmeans clustering are highly sensitive to the outliers, Which Drastically reduce the model Performance.

Remove the Outliers:

🔧 As they defecting the model we have some statistical approach to solve this problem, Let discuss about the

1. Interquartile Range (IQR)
2. Z-score

1. Interquartile range:

IQR is the one of the methods to remove the outliers. Here we learn

Steps to Remove the outliers Using the IQR


1. Sort the data
2. Find the Q1 and Q3 (Q1=25th Percentile, Q3=75th Percentile)
3. Calculate the IQR = Q3-Q1
4. Determine the Lower Bound and the Upper bound.

$$\text{Lower Bound} = Q1 - (1.5) * \text{IQR}$$

$$\text{Upper Bound} = Q3 + (1.5) * \text{IQR}$$

5. Identify the outliers. Any point below Lower Bound and above the Upper Bound is considered as the outlier.

Practical Example:

 Dataset = {6,2,5,3,8,1,6,2,5,4,6,3,7,5,6,2,8,9,27}

Step 1: Sort the Data.

Dataset = {1,2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27}

Step 2: Find the Q1 and Q3 (Q1=25th Percentile, Q3=75th Percentile)

Note: For detail understanding of percentile refer this [Medium article](#)

$$Q1 \text{ (25th Percentile)} = (p/100) * (n + 1)$$

$$= (25/100) * (19+1)$$

=5 —> Whole Number So directly take the 5th Index in the Sorted Dataset

$$Q1=3$$

$$Q3 \text{ (75th Percentile)} = (p/100) * (n + 1)$$

$$= (75/100) * (19+1)$$

=15 —> Whole Number So directly take the 15th Index in the Sorted Dataset

$$Q3=7$$

Step 3: Calculate the IQR = Q3-Q1

$$\text{IQR} = 7 - 3 = 4$$

$$\text{IQR} = 4$$

Step 4: Determine the Lower Bound and the Upper bound.

$$\text{Lower Bound} = Q1 - (1.5) * \text{IQR} = 3 - 1.5 * (4) = -3$$

$$\text{Upper Bound} = Q3 + (1.5) * \text{IQR} = 7 + 1.5 * (4) = 13$$

Step 5: Identify the outliers

If a data point falls outside the range [-3, 13], it is considered an outlier, so 27 is the Outlier.

Advantage of the IQR

- ✓ Resistance to the extreme outliers
- ✓ Clear and interpretable boundaries

Disadvantage of the IQR:

⚠ Ignores context of data

The IQR method is purely statistical and does not consider the context or domain knowledge of the data.

Example: In a dataset of salaries (in thousands):

30,35,40,45,50,55,60,65,70,500

However, in the context of salaries, 500 might represent a valid high-income individual (e.g., a CEO), not an error or anomaly. But using the IQR the 500 is the outliers. Hence it not suitable for this Scenario.

⚠ Limited to the Univariate data

The IQR method works only for single variables (univariate data) and cannot handle multivariate data (where outliers depend on relationships between multiple variables).

For example:

Consider a multivariate dataset of height (cm) and weight (kg).

Height in Cm	Weight in Kg
160	60
165	65
170	70
160	120
175	75
180	80
185	85
190	90
195	95
200	100

The IQR method cannot directly handle this data because it involves two variables (height and weight).

A point like (160, 120) might be an outlier because it represents someone unusually heavy for their height, but the IQR method cannot detect this.

Visualize the Outliers with python

A box plot is a powerful visualization tool that provides a clear and concise summary of data distribution. It uses the Interquartile Range (IQR) to highlight the middle 50% of the data, while also identifying outliers as individual points beyond the whiskers. With its ability to display central tendency, spread, and anomalies in one glance, the box plot is an essential tool for data analysis and outlier detection. In this section, we'll explore how to create a box plot using Python to visualize outliers effectively.

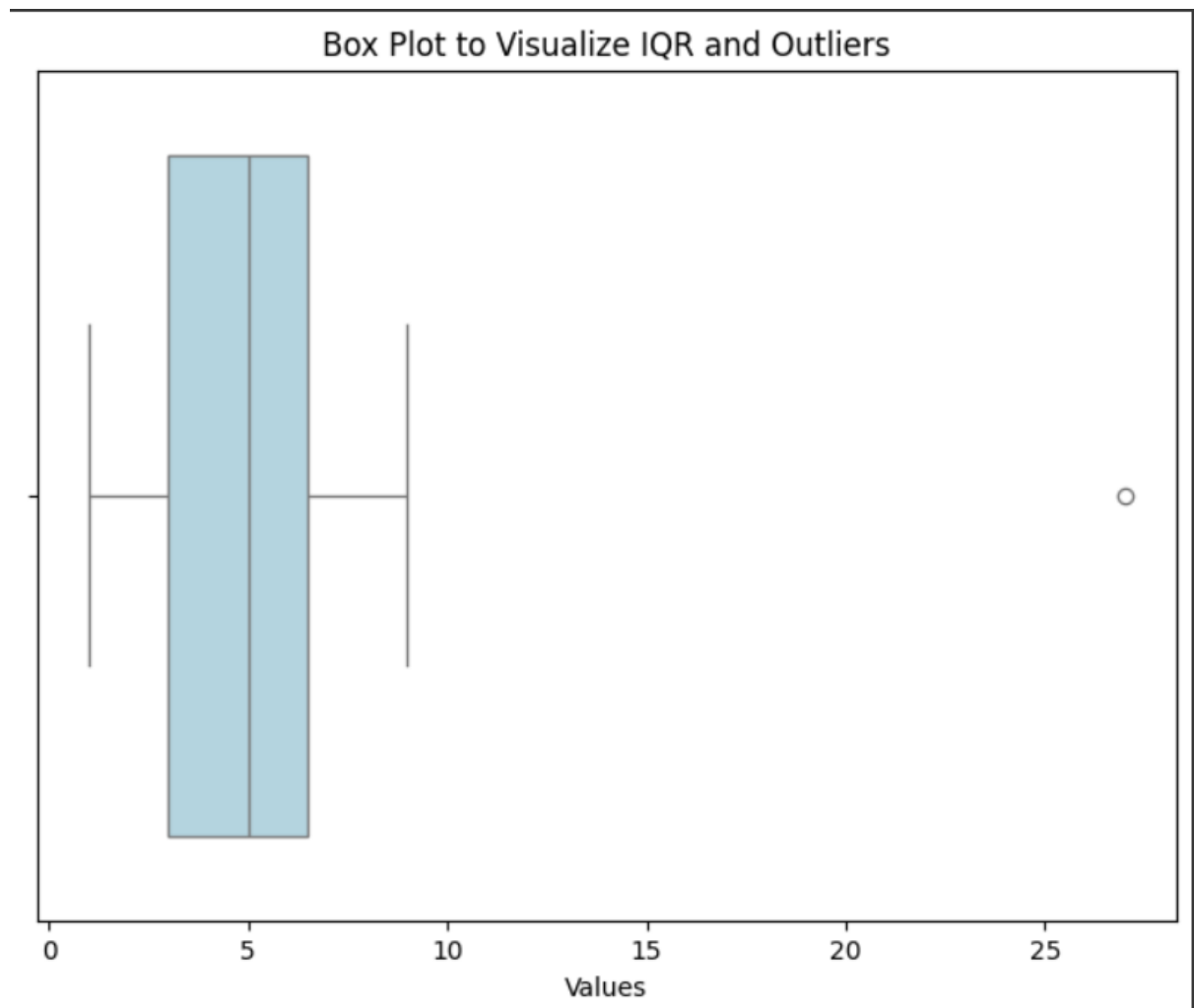
```
import matplotlib.pyplot as plt
import seaborn as sns
# Dataset

data = [1, 2, 2, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 8, 9, 27]

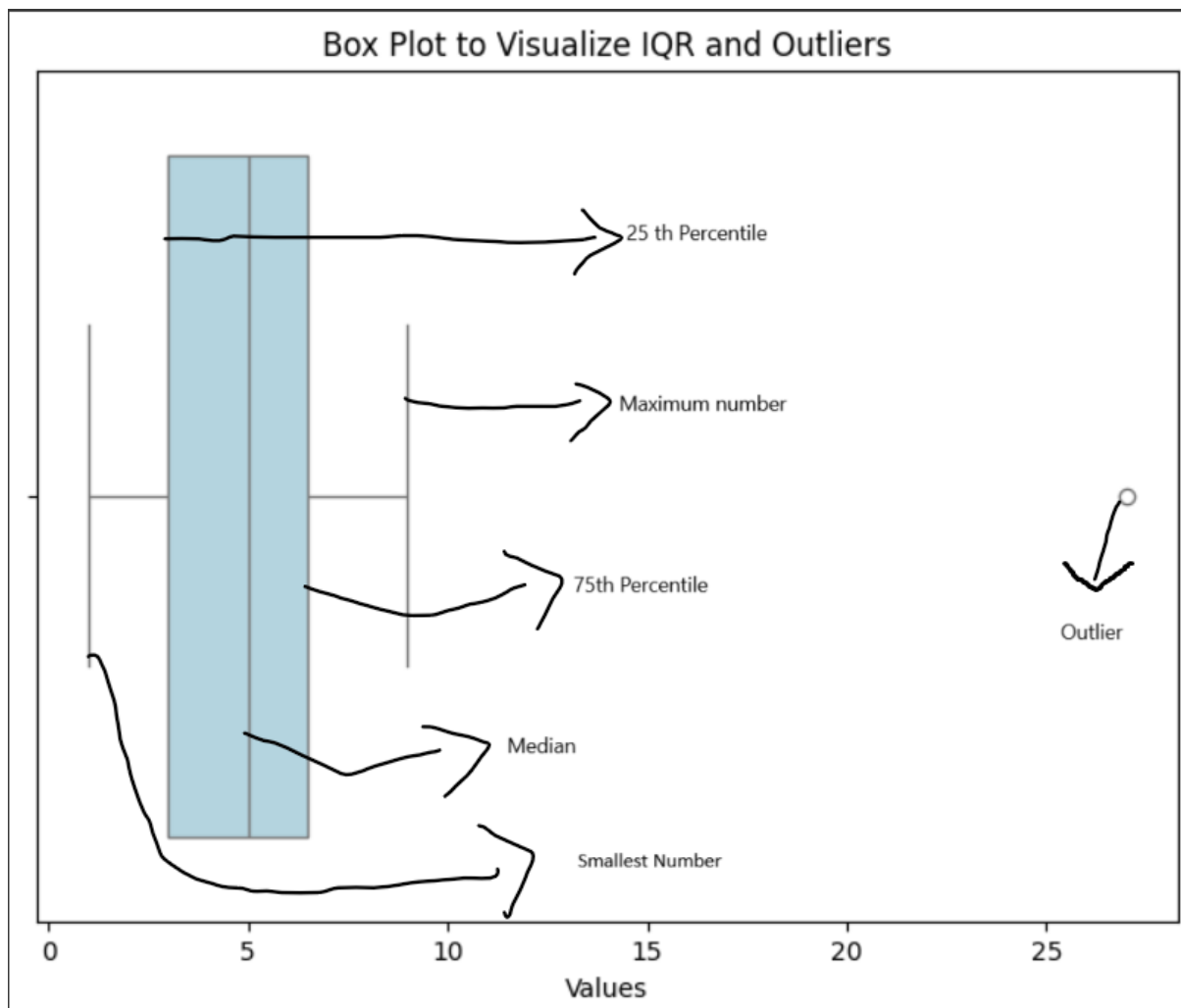
# Create a box plot
plt.figure(figsize=(8, 6))
sns.boxplot(data=data, orient="h", color="lightblue")

plt.title("Box Plot to Visualize IQR and Outliers")
plt.xlabel("Values")
plt.show()
```

Output:



Further Explanation of the output:



Box Plot Details:

- Smallest Number: 1 (end of the lower whisker).
- 25th Percentile (Q1): 3 (lower edge of the box).
- Median (50th Percentile): 5 (line inside the box).
- 75th Percentile (Q3): 7 (upper edge of the box).
- Maximum Number Within Whisker: 9 (end of the upper whisker).
- ! Outlier: 27 (plotted as an individual point beyond the upper whisker)

Conclusion:

The IQR (Interquartile Range) method is a powerful and straightforward tool for identifying outliers in univariate data. By focusing on the middle 50% of the data, it provides a robust and reliable measure that is less affected by extreme values. While it has limitations — such as assuming a symmetrical distribution and being limited to univariate data — its simplicity, clarity, and effectiveness make it a go-to choice for outlier detection in many real-world applications. Whether you're analyzing small datasets or exploring data for the first time, the IQR method is an essential technique to have in your statistical toolkit.

