What are Outlier in data science ?

In data science, outliers are data points that significantly deviate from the general pattern or distribution of the dataset. They are values that lie far away from the majority of the data.

Characteristics of Outliers

- Extreme Values: Outliers are unusually high or low values compared to the rest of the data.
- Rarity: They occur infrequently.
- Deviation: They deviate substantially from the expected norm.

Examples of Outliers

1. **Age:** In a dataset of primary school students' ages, an age of 45 would be an outlier. Most students would be between 5 and 12 years old, while 45 is far outside that range.

2. **Income:** In a dataset of salaries for entry-level software engineers, a salary of $500,000 per year would be an outlier. Most entry-level salaries might range from $60,000 to $100,000.
3. **Test Scores:** In a class of 100 students, if 99 students score between 70 and 95 on a test, and one student scores 20, that score of 20 is an outlier.
4. **Height:** In a dataset of women's heights, a height of 7 feet would be an outlier. While some women are tall, 7 feet is exceptionally rare.
5. **Sales Data:** For a small retail store, daily sales might typically range from $500 to $2,000. If, on one particular day, sales reach $10,000, that could be considered an outlier. This could be due to a special promotion, a one-time large order, or an error in recording the data.

Causes of Outliers

Outliers can arise for various reasons, including:

- Measurement or Recording Errors: Mistakes in data collection or entry.
- Data Corruption: Errors during data transmission or storage.
- Genuine Extreme Values: Rare but legitimate values in the distribution.
- Sampling Problems: Non-representative samples.

Impact of Outliers

Outliers can have a significant impact on data analysis and machine learning:

- Distorted Statistics: They can skew the mean and variance of a variable.
- Biased Models: They can disproportionately influence the training of machine learning models.
- Misleading Visualizations: They can make it difficult to visualize the true pattern of the data.