EDA-Data Distribution

Here is a summary of data distribution and how it can be used in Exploratory Data Analysis (EDA) with an example:

What is Data Distribution?

Data distribution refers to the way the values in a dataset are spread out. It describes the shape, center, and dispersion of the data.

Key Concepts:

1. **Mean**: The average value of the data.
2. **Median**: The middle value of the data when it's sorted in ascending or descending order.
3. **Mode**: The most frequently occurring value in the data.
4. **Standard Deviation (SD)**: A measure of the spread or dispersion of the data from its mean.

Types of Data Distribution:

1. **Symmetric**: The data is evenly distributed on both sides of the mean, with a bell-shaped curve.
2. **Skewed**: The data is not symmetric, with most values clustering around one side of the mean.

Example:

Suppose we have a dataset of exam scores for 100 students:

Student	Score
1	80
2	70
3	90
...	...
100	60

Data Distribution:

- **Mean**: 75 (average score)
- **Median**: 75 (middle value when sorted)
- **Mode**: 80 (most frequent score, which is 10 students got a score of 80)

The data distribution for this example would be symmetric, with most scores clustering around the mean and median.

Using Data Distribution in EDA:

Data distribution can help us:

1. Identify outliers or anomalies
2. Determine if the data is normally distributed (e.g., for statistical tests)
3. Visualize the spread of the data using histograms or box plots

In summary, understanding data distribution is essential in EDA to gain insights into the nature and behavior of the data.

Curated by Brajesh Kumar