

EDA-Scatter Plot Analysis

Scatter Plot Analysis for Exploratory Data Analysis (EDA)

A scatter plot is a graphical representation of the relationship between two continuous variables. It's a fundamental tool in Exploratory Data Analysis (EDA) to understand the correlation, distribution, and patterns in data.

Key Concepts:

1. **X-axis:** The independent variable, also known as the predictor or feature.
2. **Y-axis:** The dependent variable, also known as the response or target variable.
3. **Scatter Plot:** A graphical representation of the relationship between X and Y variables.
4. **Correlation Coefficient** (e.g., Pearson's r): Measures the strength and direction of the linear relationship between X and Y.

Types of Scatter Plots:

1. **Positive Correlation:** As X increases, Y also tends to increase.
2. **Negative Correlation:** As X increases, Y tends to decrease.
3. **No Correlation:** No apparent pattern or relationship between X and Y.

Example:

Suppose we're analyzing the relationship between a student's study time (X) and their exam score (Y). We create a scatter plot using Python with the following code:

```
import matplotlib.pyplot as plt

# Sample data
study_time = [2, 4, 6, 8, 10]
exam_score = [60, 70, 80, 90, 100]

plt.scatter(study_time, exam_score)
plt.xlabel('Study Time (hours)')
plt.ylabel('Exam Score')
plt.title('Study Time vs. Exam Score')

# Show the plot
plt.show()
```

Interpretation:

In this example, we observe a positive correlation between study time and exam score. As students spend more hours studying (X-axis), their exam scores tend to increase (Y-axis). This suggests that there is a strong linear relationship between these two variables.

However, if the scatter plot showed no apparent pattern or relationship, it would indicate that there is no correlation between study time and exam score.

Common Applications of Scatter Plots in EDA:

1. **Identifying correlations:** To understand how different variables relate to each other.
2. **Visualizing distributions:** To examine the shape and spread of data across multiple variables.
3. **Detecting outliers:** To identify unusual patterns or anomalies in the data.

By analyzing scatter plots, you can gain valuable insights into your data, identify relationships between variables, and inform future analyses or modeling efforts.

Curated by Brajesh Kumar