

# EDA-Intro

## Importance of Exploratory Data Analysis (EDA)

### What is EDA?

- **Definition:** Exploratory Data Analysis (EDA) is an approach to analyzing data sets to summarize their main characteristics, often using visual methods. It is a critical first step in the data analysis process, enabling analysts to understand the data's structure, detect anomalies, test hypotheses, and check assumptions.

### Why is EDA Important?

- **Data Understanding:** Gain insights into the data's distribution, trends, and relationships.
- **Quality Assessment:** Identify and address data quality issues like missing values, outliers, and inconsistencies.
- **Hypothesis Generation:** Formulate hypotheses for further statistical analysis or modeling.
- **Informing Modeling:** Guide the selection of appropriate models and techniques based on data characteristics.

### Example Scenario

- **Business Scenario:** A retail company wants to understand its sales performance.
    - **Without EDA:** The company might make uninformed decisions based on incomplete understanding.
    - **With EDA:** By performing EDA, the company can identify sales trends, peak periods, underperforming products, and customer preferences, leading to informed strategies.
-

# Introduction to Key Concepts

## Data Analysis vs. Data Analytics

- **Data Analysis:**
  - Focuses on processing and performing statistical analysis on existing data.
  - Primarily descriptive and diagnostic.
  - Example: Calculating the average sales for the past year.
- **Data Analytics:**
  - Encompasses a broader range of techniques to extract insights from data.
  - Includes predictive and prescriptive analytics.
  - Example: Using machine learning to forecast future sales trends.

## Types of Data

- **Quantitative Data:**
  - **Discrete:** Countable values (e.g., number of customers).
  - **Continuous:** Measurable values (e.g., revenue, temperature).
- **Qualitative Data:**
  - **Nominal:** Categories without a natural order (e.g., product categories).
  - **Ordinal:** Categories with a natural order (e.g., customer satisfaction ratings).
- **Time-Series Data:** Data collected at specific time intervals (e.g., monthly sales data).
- **Cross-Sectional Data:** Data collected at a single point in time across different subjects (e.g., survey responses).
- **Panel Data:** Combines cross-sectional and time-series data (e.g., tracking sales across regions over multiple months).