EDA - exploratory Data Analysis

Exploratory Data Analysis (EDA) is the process of analyzing, visualizing, and understanding a dataset before applying any machine learning models. It helps in identifying patterns, detecting outliers, handling missing values, and discovering relationships between variables.

Understanding the Dataset

Before performing any analysis, we need to understand the dataset’s attributes and their types.

1. Understanding Attributes and Their Types

Binary Attributes – Attributes with only two possible values (e.g., Yes/No, Male/Female).

Categorical Attributes – Attributes with discrete values belonging to categories (e.g., Color: Red, Blue, Green).

Continuous Attributes – Attributes with numerical values within a range (e.g., Age, Salary).

Dependent & Independent Variables –

Independent Variables (X) – Features used as inputs.

Dependent Variable (Y) – Target variable (output to predict).

2. Understanding Data Using Statistical Methods

Mean (Average) – Measures central tendency.

Standard Deviation (std) – Measures data dispersion.

Variance – Measures how spread out data points are.

Correlation Analysis – Determines the relationship between numerical features.

Major Steps of EDA

1. Understanding the Dataset

Identify attributes and attribute types (binary, categorical, continuous).

Check units of numerical attributes (e.g., dollars, meters, years).

View a few rows of the dataset using .head().

Study metadata (if available) to understand attribute meanings.

2. Data Preparation

Handle missing values by either filling (.fillna()) or dropping (.dropna()).

Handle null values by estimating or imputing missing data.

Detect and remove duplicate values if present.

Feature Selection – Keep only relevant features.

Feature Extraction – Derive new meaningful features for models.

3. Handling Outliers

Detect Outliers using statistical and visualization techniques.

Statistical Measures like Interquartile Range (IQR) and Z-score to find extreme values.

Visualization Methods:

Box Plot – Identifies outliers in numerical data.

Histogram – Checks for skewness and unusual values.

Handling Outliers – Drop or transform extreme values based on impact.

4. Data Visualization

Visualizing data helps extract meaningful insights.

Univariate Plotting (Single Variable)

Histogram – Shows the distribution of numerical data.

Box Plot – Identifies outliers and spread of the data.

Count Plot & Bar Plot – Used for categorical data analysis.

Bivariate Plotting (Two Variables)

Scatter Plot – Shows the relationship between two numerical variables.

Line Plot – Used for time-series data.

Heatmap – Displays correlations between numerical variables.

5. Feature Engineering

Create new features by combining existing ones.

Feature selection – Remove irrelevant or redundant features.

Binarization & Categorization – Convert continuous variables into categories.

Normalize Numerical Features – Scale values to a standard range (e.g., Min-Max scaling, Z-score).