Day 23 – Exploratory Data Analysis (EDA)

Exploratory Data Analysis (EDA) is the process of examining, visualizing, and preparing raw data before applying any machine learning model or statistical technique.

EDA helps to:

- Understand the structure, trends, and patterns in the data
- Detect missing values, outliers, and anomalies
- Discover relationships between variables
- Guide feature engineering and model selection

---

1. What is Raw Data?

Raw data is the initial form of data collected from various sources like surveys, sensors, logs, web scraping, etc.

It often contains:

- Missing values
- Inconsistent formats
- Duplicates
- Irrelevant features
- Mixed data types

---

2. Data Cleaning

Data cleaning involves correcting or removing inaccurate, corrupted, or incomplete data.

Common tasks include:

- Handling missing values
- Fixing inconsistent data types
- Removing duplicate rows
- Filling in or dropping nulls
- Renaming columns for clarity

Cleaning ensures the dataset is ready for analysis and modeling.

---

3. Variable Identification

Identifying the role of each column helps shape the direction of analysis and modeling.

Type	Description	Examples
Independent (IV)	Input features	Age, Salary, Gender
Dependent (DV)	Output/target variable	Purchased, Outcome

Example:
If you're predicting whether someone will buy a product, the input features like `Age`, `Income`, etc., are IVs and the `WillBuy` column is the DV.

4. Univariate Analysis

This type of analysis focuses on a single variable at a time.

For **numerical variables**:

- Histogram
- Boxplot
- KDE (Kernel Density Estimation)

For **categorical variables**:

- Count plot
- Pie chart

Univariate analysis helps understand the distribution and frequency of values.

5. Bivariate Analysis

Bivariate analysis studies the relationship between two variables.

Common comparisons:

- Numerical vs Numerical → Scatter plot, Correlation
- Categorical vs Numerical → Boxplot, Barplot
- Categorical vs Categorical → Grouped barplot, Crosstab

6. Correlation

Correlation measures the strength and direction of a linear relationship between two numeric variables.

Correlation Value	Meaning
1	Perfect positive
0	No correlation
-1	Perfect negative

Important for feature selection, multicollinearity checks, and relationship understanding.

7. Outlier Detection

Outliers are unusual data points that differ significantly from others.

Common detection methods:

- Boxplot
- Z-score
- IQR method

Outliers can:

- Skew statistical summaries
- Mislead machine learning models

8. Feature Engineering

Feature engineering involves transforming or creating new features to improve model performance.

Key techniques:

- Label encoding / One-hot encoding
- Creating dummy variables
- Binning numeric data (e.g. age groups)
- Generating interaction terms

This step is crucial for preparing categorical and continuous variables for ML models.

9. EDA Process Overview

Step-by-step EDA approach:

1. Load the dataset
2. Explore dataset dimensions and types
3. Handle missing values and duplicates
4. Identify variable roles (IV/DV)
5. Conduct univariate and bivariate analysis
6. Visualize relationships and distributions
7. Create new features if needed
8. Finalize clean dataset for modeling

Summary

EDA is a **critical pre-modeling step** in any data science project. It provides insights into the structure, quality, and characteristics of the dataset, which directly impacts model accuracy and interpretability.

Libraries commonly used in EDA:

- `pandas` – data handling
- `numpy` – numerical operations
- `matplotlib` & `seaborn` – visualizations
- `scipy` – statistical tests

Introduction to Machine Learning (Post-EDA)

Once you've completed data cleaning and exploratory data analysis (EDA), the next step is to apply machine learning models to draw predictions and patterns from your data.

Supervised Learning

Supervised learning uses **labeled data** (i.e., input features + target output) to train a model.

There are two major types:

1. Regression

Regression is used when the **target variable (dependent variable) is continuous** (numeric).

Common Regression Algorithms:

1. Simple Linear Regression
2. Multiple Linear Regression
3. Gradient Descent Variants:
    - Batch Gradient Descent (BGD)
    - Stochastic Gradient Descent (SGD)
4. Lasso & Ridge (L1 & L2 Regularization)
5. K-Nearest Neighbor Regressor (KNN)
6. Support Vector Regressor (SVR)
7. Decision Tree Regressor (DTR)
8. XGBoost Regressor (XGB)
9. LightGBM Regressor (LGB)
10. Artificial Neural Network (ANN) Regressor
11. Time Series Forecasting
12. Random Forest Regressor

Use regression when your output is something like **price, score, distance, revenue**, etc.

Examples:

- Predicting house price
- Forecasting stock price
- Estimating salary based on experience

---

2. Classification

Classification is used when the **target variable is categorical** (labels or classes).

Common Classification Algorithms:

1. Logistic Regression
2. Support Vector Machine (SVM)
3. K-Nearest Neighbor Classifier (KNN)
4. Naive Bayes (Bayesian Theorem)
5. Decision Tree Classifier
6. Random Forest Classifier
7. Ensemble Learning Techniques
8. XGBoost Classifier
9. LightGBM Classifier
10. Artificial Neural Network (ANN) Classifier

Use classification when your output is like **yes/no, spam/ham, pass/fail, disease/no disease**, etc.

Examples:

- Spam vs Not Spam
- Disease vs No Disease
- Will Buy vs Won’t Buy

Regression vs Classification

Feature	Regression	Classification
Target Variable Type	Continuous (numeric)	Categorical (labels/classes)
Output Example	Price, Score, Temp	Yes/No, Spam/Not, Category
Evaluation Metric	MAE, MSE, RMSE, R^2 Score	Accuracy, F1 Score, AUC
Sample Algorithm	Linear Regression	Logistic Regression

Unsupervised Learning – Clustering

While regression and classification are types of **supervised learning**, clustering belongs to a different category called **unsupervised learning**.

In unsupervised learning, the dataset has **no labeled output**. The goal is to discover **hidden patterns or natural groupings** in the data.

What is Clustering?

Clustering is a technique used to **group similar data points together** based on their features, similarity, without knowing the target variable in advance.

Common Clustering Techniques:

1. Principal Component Analysis (PCA) – Dimensionality reduction & grouping
2. K-Means Clustering
3. Hierarchical Clustering
4. DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

Use clustering when you're doing:

- Customer segmentation
- Anomaly detection
- Pattern discovery without predefined categories

Real-World Examples:

- Grouping customers by purchasing behavior
- Organizing news articles by topics
- Segmenting users by browsing activity
- Detecting fraudulent transactions

Clustering is especially useful when you want to **explore structure** in the data or **segment** it for targeted analysis before applying supervised techniques.

Summary Table

Task Type	Target Variable	Algorithms Example
Regression	Continuous	Linear Regression, XGBoost, ANN
Classification	Categorical	Logistic Regression, SVM, RF, ANN

Task Type	Target Variable	Algorithms Example
Clustering	Not available	K-Means, PCA, DBSCAN