SUPERIOR UNIVERSITY LAHORE

GOLD CAMPUS

NAME: MUHAMMAD AWAIS **(065)**

PROGRAM: ARTIFICIAL INTELLIGENCE

SEMESTER: 3rd SEMESTER

SECTION: BSAI-3A

SUBJECT: ARTIFICIAL INTELLIGENCE

Submission Title: ASSIGNMENT (DOCUMENT)

Submitted To: SIR ANSAR

Assignment 1 — Exploratory Data Analysis (EDA)

Dataset: Loan.csv

Introduction

In this assignment, we perform **Exploratory Data Analysis (EDA)** on the Loan.csv dataset using an Object-Oriented Programming (OOP) approach.
The goal is to:

- Load and understand the dataset
- Clean and preprocess data
- Perform statistical analysis
- Visualize relationships and distributions

Structure Overview

The complete workflow is divided into **five classes**, each responsible for one specific task of the EDA pipeline:

Class Name	Responsibility
DataLoader	Load and explore dataset
DataPreprocessor	Handle missing values, encode categorical, and scale numerical data
DataCleaner	Remove duplicate records
DataAnalyzer	Generate summary statistics and feature overview
DataVisualizer	Visualize dataset through graphs and charts

Class Details

DataLoader

This class is responsible for loading the dataset and showing its initial overview.

Functions:

- load_data():
  - Loads the CSV file
  - Displays dataset shape
  - Shows first (head()) and last (tail()) five rows
  - Displays data information (info())

Purpose:
Helps understand data size, types, and structure.

DataPreprocessor

Handles missing values, categorical encoding, and data scaling.

Techniques Used:

- **Label Encoding:** Converts categorical text data into numeric form.
- **Mean & Mode Imputation:** Fills missing values.
- **Standard Scaling:** Standardizes numerical data to mean=0 and std=1.

Purpose:
Makes data ready for analysis and modeling.

DataCleaner

Focuses on removing duplicate entries.

Function:

- clean(): Detects and removes duplicate rows.

Purpose:
Ensures data quality and consistency.

DataAnalyzer

Generates insights and basic understanding of features.

Functions:

- summary_statistics(): Displays mean, median, min, max, and count.

- feature_understanding(): Lists columns with their data types.

Purpose:
Helps understand data characteristics numerically and structurally.

DataVisualizer

Creates visual representations to better understand trends and patterns in the data.

Visuals Generated:

1. **Histograms:**
   Shows distribution of numerical variables.
2. **Countplots (for all categorical columns):**
   Displays category frequencies.
3. **Boxplots:**
   Identifies outliers in numerical features.
4. **Correlation Heatmap:**
   Shows correlation between numerical variables.
5. **Pairplot:**
   Displays pairwise relationships between features.
6. **Special Plot — Loan Status Count:**
   Displays distribution of loan approval (Loan_Status).

Purpose:
To visually identify relationships, trends, and anomalies in the data.

Workflow Execution

Step	Operation	Class Used
1	Load dataset and display info	DataLoader
2	Handle missing data and scale values	DataPreprocessor
3	Remove duplicate records	DataCleaner
4	Generate summary and understand columns	DataAnalyzer
5	Visualize dataset with multiple plots	DataVisualizer

Technologies Used

Tool	Purpose
Python	Programming language
Pandas	Data manipulation
Matplotlib / Seaborn	Visualization
Scikit-learn	Encoding & scaling
OOP Concepts	Code organization