EDA Summary Report

1. Introduction

The purpose of this report is to perform exploratory data analysis (EDA) on a customer credit dataset. The goal is to identify trends, detect data quality issues, and extract useful insights that could support predictive modeling or business decision-making.

2. Dataset Overview

This section summarizes the dataset, including the number of records, key variables, and data types. It also highlights any anomalies, duplicates, or inconsistencies observed during the initial review.

- Number of records: 10,000

- Key variables: Customer_ID (Unique ID), Age, Gender, Income, Credit_Score, Loan_Status

- Data types: Categorical (Gender, Loan_Status), Numerical (Age, Income, Credit_Score)

3. Missing Data Analysis

Identifying and addressing missing data is critical to ensuring model accuracy. This section outlines missing values in the dataset, the approach taken to handle them, and justifications for the chosen method.

- Variables with missing values: Income (2.1%), Credit_Score (0.7%)

- Missing data treatment: Median imputation was applied for numerical values. No records were deleted.

4. Key Findings and Risk Indicators

This section identifies trends and patterns that may indicate risk factors for delinquency. Feature relationships and statistical correlations are explored to uncover insights relevant to predictive modeling.

- Correlations observed between key variables: Credit_Score positively correlates with Income (0.62); negative correlation between Age and Loan_Status (-0.31)

- Unexpected anomalies: A few records have unusually high income values (> ₹10,00,000) needing verification.

5. AI & GenAI Usage

Generative AI tools were used to summarize the dataset, impute missing data, and detect patterns. This section documents AI-generated insights and the prompts used to obtain results.

Example AI prompts used:

- 'Summarize key patterns in the dataset and identify anomalies.'

- 'Suggest an imputation strategy for missing income values based on industry best practices.'

6. Conclusion & Next Steps

The EDA revealed key relationships and quality issues in the dataset. Missing values were minimal and resolved using median imputation. Certain variables show strong correlation, which may be useful for modeling. Next steps include feature engineering, scaling of variables, and model prototyping using logistic regression and tree-based classifiers.