Loan Defaulter Exploratory Data Analysis (EDA) — Assignment

Objective The goal of this project is to analyze banking loan data to identify patterns of loandefaults. Through this exercise, you will learn how to:

- Perform data cleaning and exploratory data analysis (EDA).
- Apply banking domain KPIs like Debt-to-Income ratio (DTI), Loan-to-Value ratio (LTV), and EMI-to-Income ratio.
- Interpret the data not just statistically, but from a financial risk management perspective.

Step 1 : Data Understanding & Cleaning

Tasks

- Load the dataset. Check shape, data types, and summary statistics.
- Handle missing values (income, loan amount, credit history).
- Detect and treat outliers (very high income, loan amounts, EMIs).

✓ Domain Questions:

- In banking, what does a missing credit history imply?
- Should extremely high loan amounts always be treated as errors,
- or could they represent high net-worth customers?

Step 2: Univariate Analysis

Tasks

- Distribution of target variable (Default vs Non-default).
- Loan amount distribution.
- Applicant income distribution.
- Age distribution of applicants.
- Credit history distribution.

Domain Questions:

- Why is the Credit Score / Credit History one of the strongest indicators of risk?
- What does a skewed income distribution mean for banks when assessing risk?

Step 3: Bivariate Analysis

Tasks

- Default rate across loan types (personal, home, vehicle, education).
- Default by employment type (salaried, self-employed, business owner).
- Loan-to-Income ratio vs default rate. Credit history vs default probability.
- Correlation heatmap for numerical variables.

Domain KPIs to Calculate:

- Debt-to-Income Ratio (DTI): Total EMI / Applicant Income.
- Loan-to-Value Ratio (LTV): Loan Amount / Property Value.
- EMI-to-Income Ratio: EMI / Income.

Domain Questions:

- At what DTI ratio do defaults become more frequent?
- Does a higher LTV always mean more risk for the bank?

Step 4: Multivariate Analysis

Tasks

- Combine features: Income + Loan Amount + Tenure vs Default.
- Cluster customers (K-Means or Hierarchical) to segment defaulters.
- Apply feature importance (Decision Tree / Random Forest) to find key predictors.

✓ Domain Questions:

- Which combinations of features (e.g., low income + high loan amount) create the highest risk?
- How can banks use clustering to define customer risk buckets?

Step 5: Risk Profiling

Tasks

- Segment customers into Low-Risk, Medium-Risk, and High-Risk based on KPIs and observed patterns.
- Identify early warning signals: e.g., high DTI + poor credit history.

✓ Domain Questions:

- What real-time monitoring could a bank implement to flag “about-to-default” customers?
- How would you explain these findings to a non-technical risk manager?

Step 6: Domain-Level Insights

Demographic Analysis

- Default by age group, gender, marital status, dependents.

Financial Analysis

- Income vs default rate.
- DTI ratio buckets (e.g., <20%, 20–40%, >40%) vs default.
- LTV ratio buckets vs default.

Credit Behavior Analysis

- Past defaults or delinquency history.
- Distribution of credit utilization.
- Number of open credit lines vs default.

Loan Characteristics

- Loan type (secured vs unsecured). Loan tenure (short vs long-term).
- EMI-to-Income ratio vs default.

Regional / Behavioral Analysis

- Rural vs Urban default rates.
- Regional loan performance.
- Patterns of delayed payments.

Conclusion & Business Takeaways

- Top 3–5 risk drivers of loan default.
- How banks can improve risk assessment using these insights.
- Recommendations for credit policy (e.g., limit loans with LTV > 80% unless strong credit history).

This assignment will include code cells (following your reference style) + visualizations (Seaborn/Matplotlib) + domain commentary so students get a project-like experience.