# Project Title: Synthetic Lending Operations Dataset for MIS Analysis

## Project Objective

The objective of this project is to create a synthetic dataset that mimics real-world lending operations. The dataset will serve as the foundation for building dashboards, analyzing KPIs, and deriving insights to support decision-making in MIS (Management Information Systems) Analyst roles.

## Problem Scoping

The goal is to simulate data commonly found in lending operations, enabling analysis of loan performance, customer behavior, repayment trends, and risk indicators. The synthetic dataset will help in practicing reporting, dashboarding, and analytical tasks aligned with the responsibilities of an MIS Analyst.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  |  | | --- | --- | --- | | Column | Type | Description | | Customer\_ID | String | Unique customer identifier | | Age | Integer | Age of borrower | | Gender | Categorical | Male / Female | | Income | Numeric | Annual income | | Employment\_Type | Categorical | Salaried / Self-Employed | | Loan\_ID | String | Unique loan identifier | | Loan\_Type | Categorical | Home / Auto / Education / Personal | | Loan\_Amount | Numeric | Loan principal amount | | Interest\_Rate | Float | Annual % rate | | Tenure\_Months | Integer | Duration of loan in months | | Application\_Date | Date | Loan application date | | Approval\_Status | Categorical | Approved / Rejected | | EMI\_Amount | Numeric | Monthly repayment amount | | Disbursal\_Date | Date | Date loan amount credited | | EMIs\_Paid | Integer | Number of EMIs successfully paid | | Default\_Status | Categorical | Yes / No | | Outstanding\_Balance | Numeric | Remaining balance if loan active | |

## Data Dictionary

This data dictionary provides an initial description of each field in the dataset. The values will be synthetically generated to reflect realistic distributions. Additional derived fields (such as DPD - Days Past Due, Prepayment\_Flag, etc.) may be added in later iterations as per project scope.

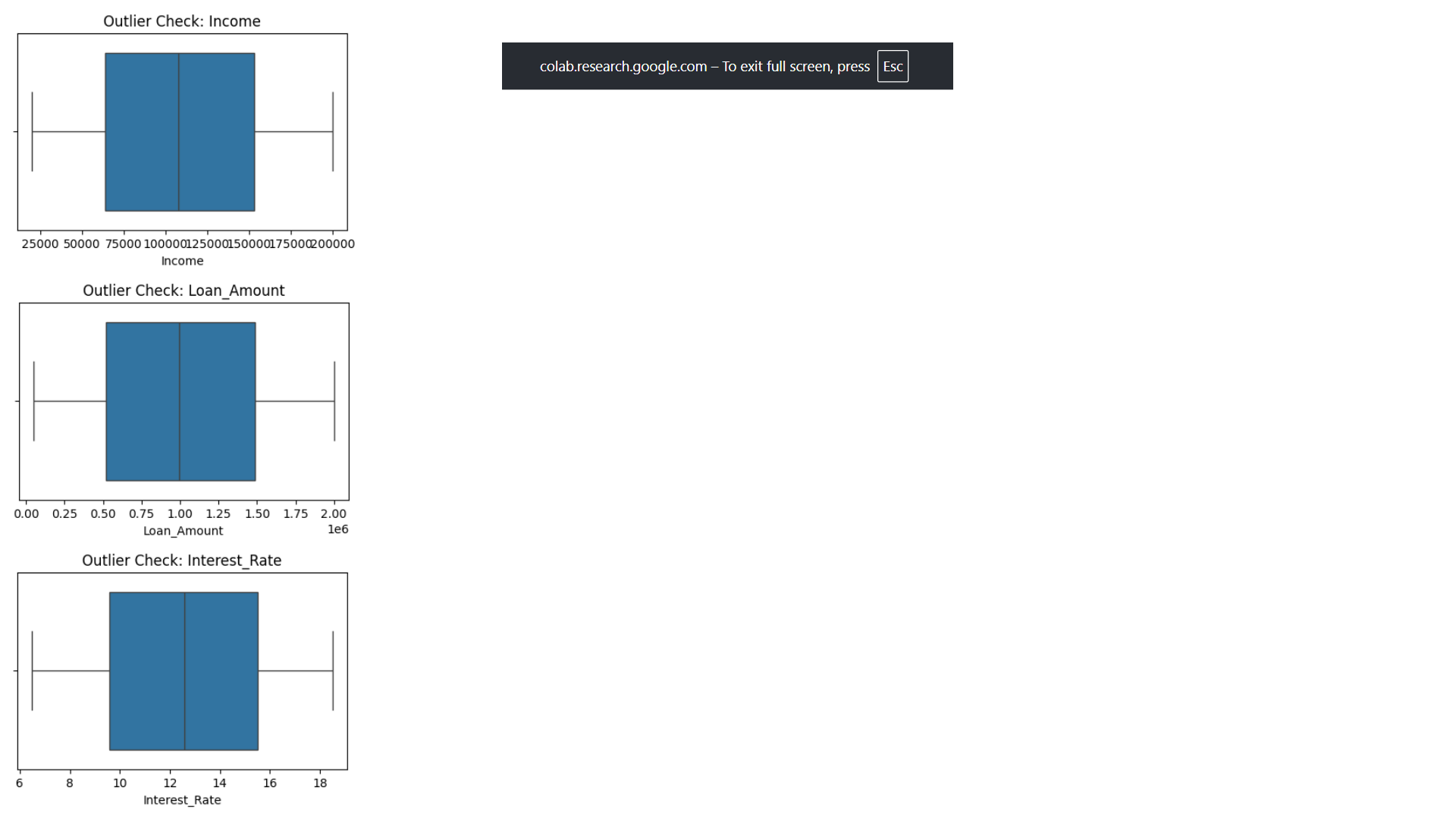
* **Customer ID:** Format: CUSTXXXX, Unique key
* **Loan ID:** Format: LOANXXXX, Unique per loan
* **Age:** Integer, range [18–70]
* **Income:** Annual income, range [₹30,000 – ₹30,00,000]
* **Interest Rate:** Float, range [5% – 25%]
* **Default Status:** Binary, Yes/No (default if EMIs\_Paid < Tenure\_Months and Outstanding Balance > 0)

Exploratory Data Analysis (EDA) & Summary Statistics

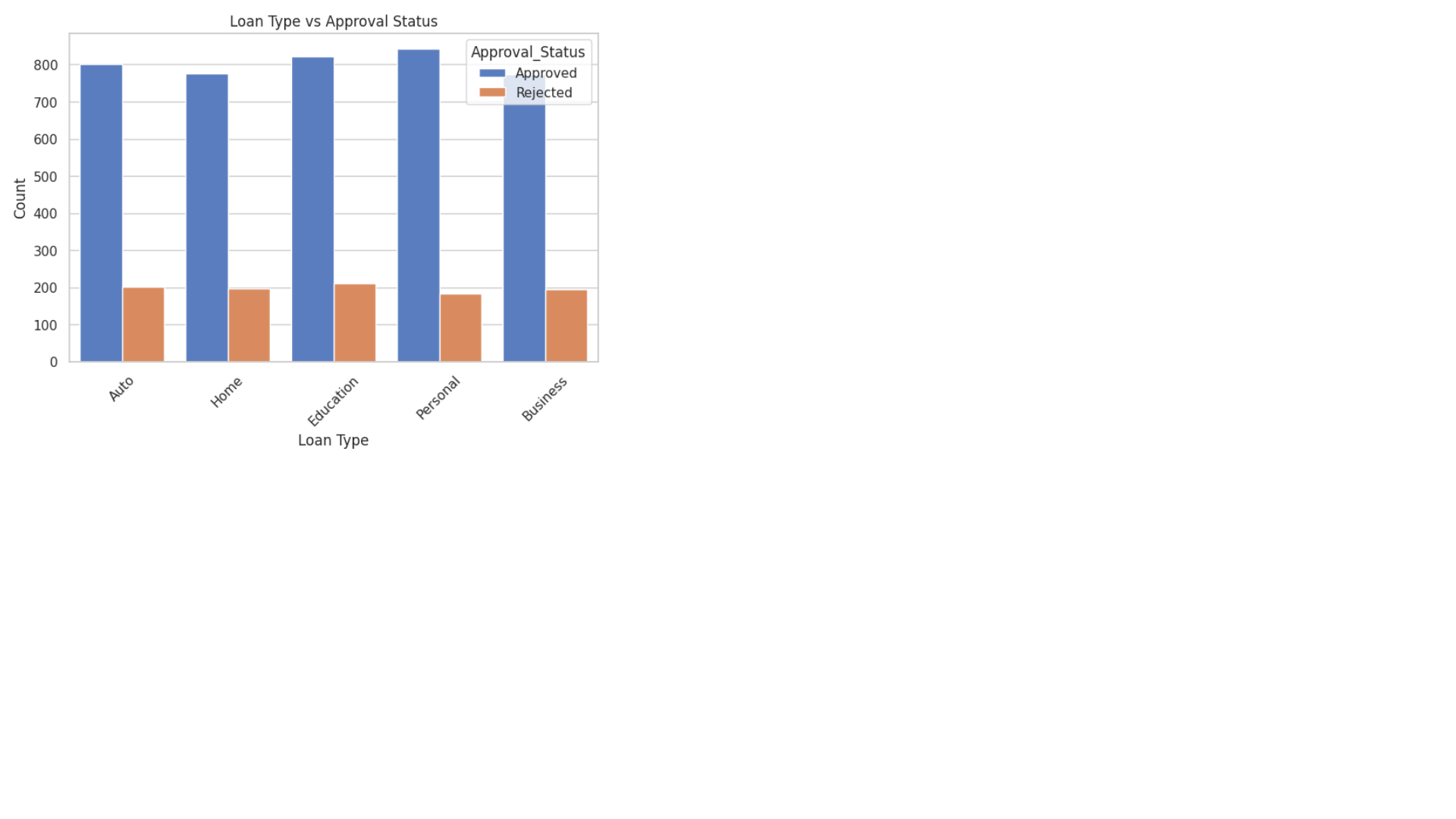
* Check for missing values, duplicates, and data types.
* Summary stats of numeric features (Loan Amount, EMI Amount, Income, etc.).
* Value counts of categorical features (Loan Type, Approval Status, Default Status).

## **Visualizations** (basic insights)

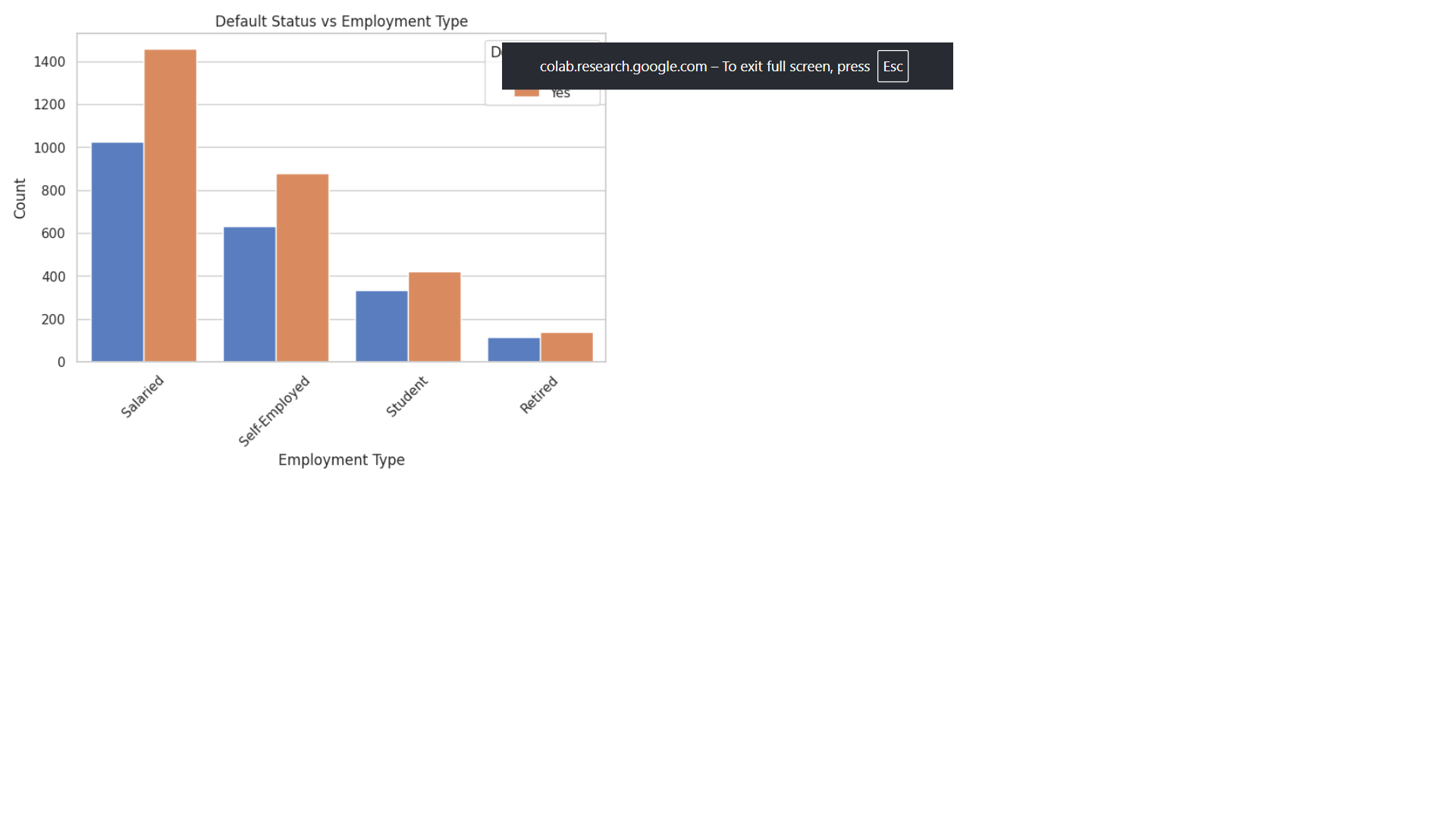
* Distribution of Loan Amount, Income, Interest Rate.



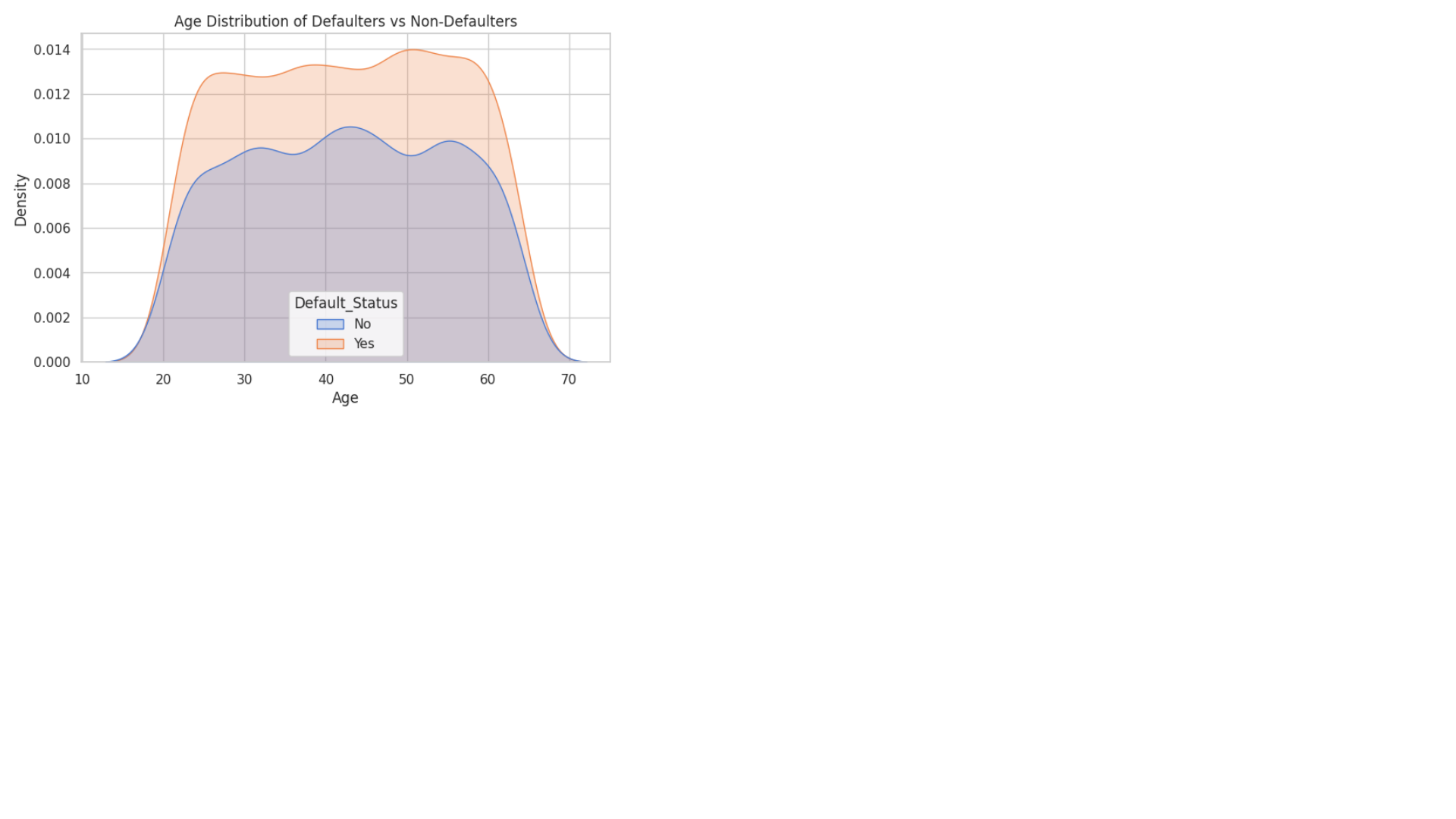
* Loan Type vs Approval Status (bar chart).



* Default Status vs Employment Type (bar chart).

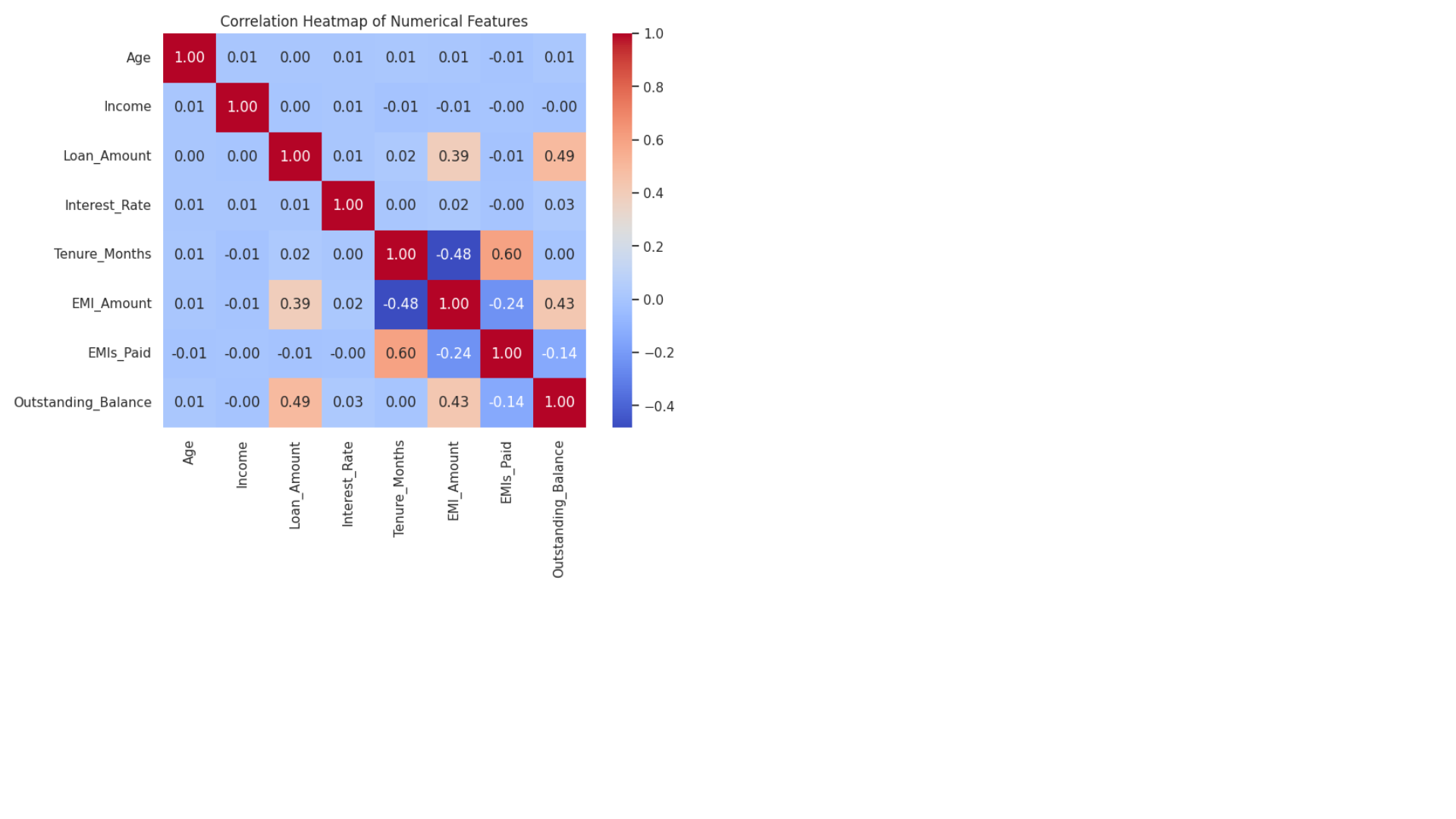


* Age distribution of defaulters vs non-defaulters



## Key Insights from Exploratory Data Analysis

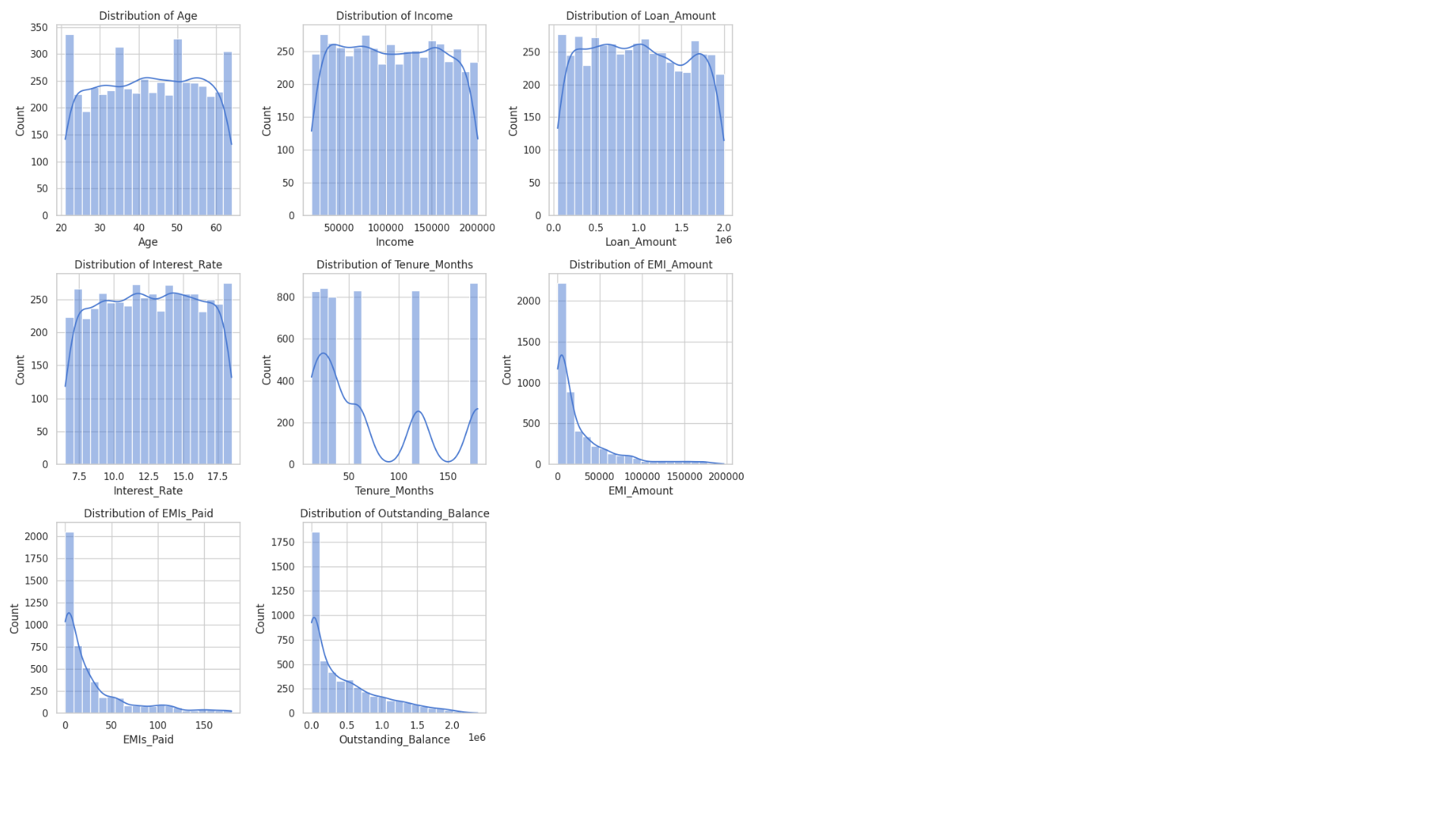
## **Correlation Analysis (Numerical Features):**



Numerical variables exhibit **low-to-moderate correlations**, with no strong multicollinearity detected.

Loan Amount and Income show a **moderate positive relationship**, indicating higher income applicants tend to apply for higher loan amounts.

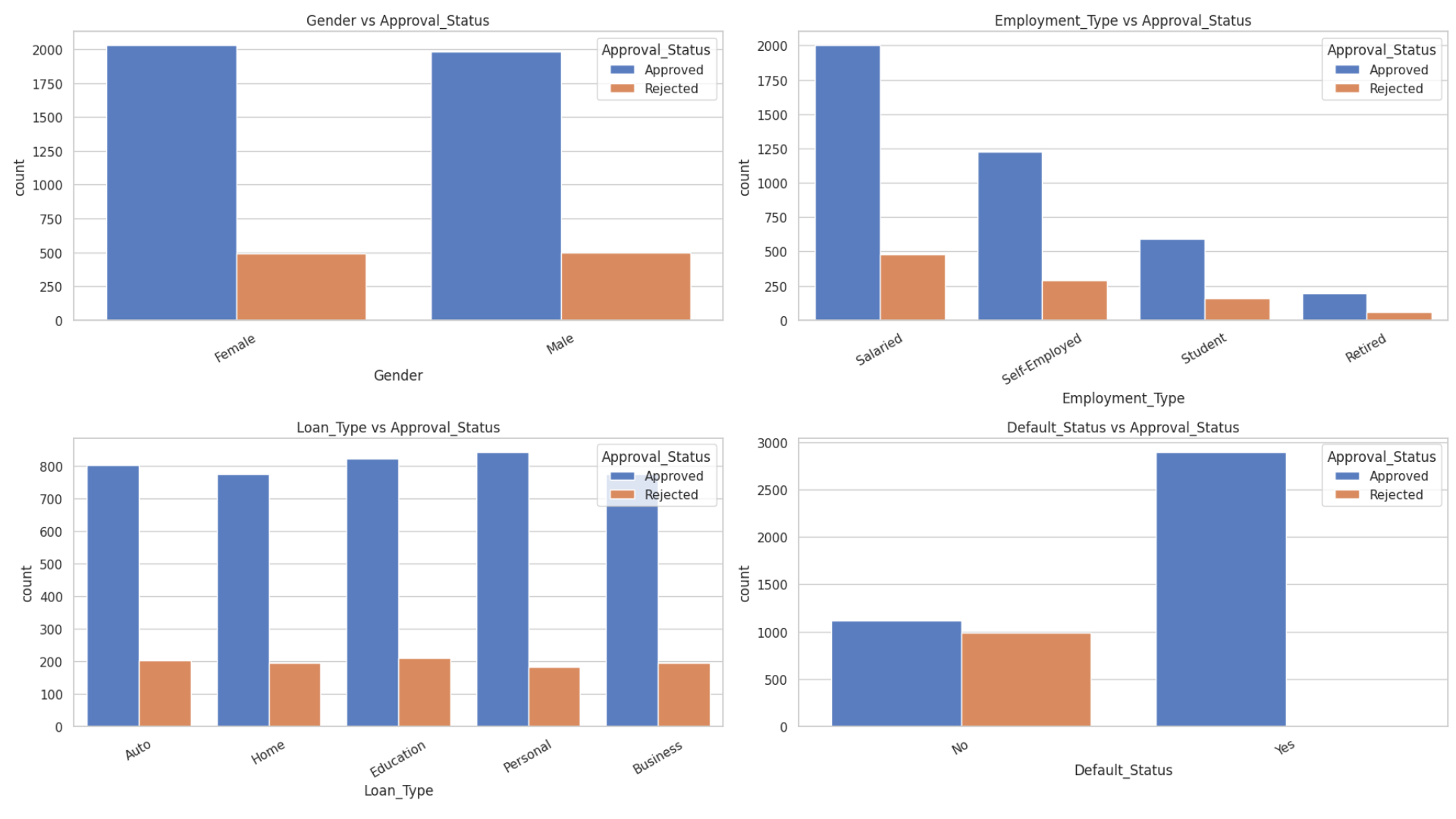
## **Numerical Distributions & Outliers:**



Variables like **Income** and **Loan Amount** are **right-skewed** and contain notable outliers.

Outliers may affect model performance and should be handled via **IQR/Winsorization** or scaling techniques.

## **Categorical Variables & Loan Approval:**

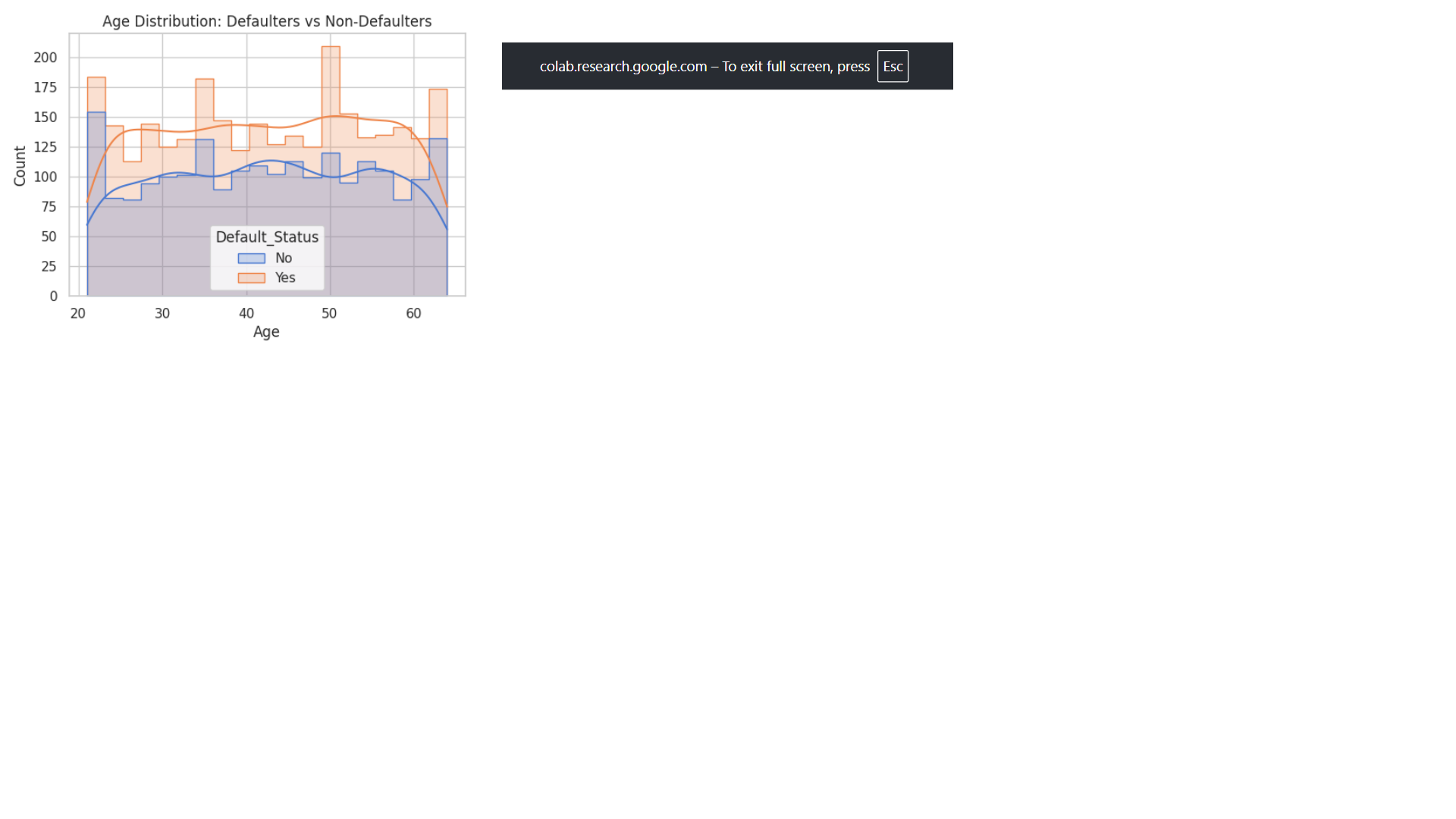


**Employment\_ Type:** Salaried applicants have a higher approval rate compared to self-employed.

**Education & Marital Status:** Variations observed, suggesting these features influence approval outcomes.

These categorical features are **informative predictors** and should be retained through encoding.

## **Default Risk Indicators:**



**Employment Stability:** Self-employed applicants show a **higher likelihood of default** than salaried ones.

**Age Groups:** Younger (<30) and older (>55) applicants exhibit slightly higher default risks, while mid-age groups (30–45) appear more stable.

These findings suggest that **employment stability and age segmentation** are critical factors in assessing loan default risk.

Overall Conclusion:  
The EDA highlights that both numerical and categorical features contribute valuable predictive power. While numerical features require outlier handling and scaling, categorical variables like employment type, education, and marital status strongly influence loan approval and default risk. These insights will guide feature engineering and model development in the later stages of the pipeline.

## Feature Engineering

To enhance predictive power, new variables were engineered from the raw dataset:

1. **Debt-to-Income Ratio** – Ratio of requested loan amount to applicant’s income. Helps assess repayment capacity.
2. **EMI Burden** – Proportion of EMI to income, indicating monthly affordability stress.
3. **Payment Progress** – Percentage of EMIs paid relative to total tenure, highlighting repayment discipline.
4. **Application-to-Disbursal Days** – Processing time between loan application and disbursal, capturing operational or fraud-related delays.
5. **Loan-to-Tenure Ratio** – Normalized monthly loan burden.
6. **Categorical Encoding** – Converted categorical features such as Employment Type, Loan Type, and Approval Status into machine-readable numeric variables.

These transformations created domain-driven features that improve both interpretability and predictive performance of subsequent machine learning models.

## Predictive Modeling

After preparing the feature-engineered dataset, machine learning models were applied to predict **loan default probability**.

1. **Data Splitting** – Dataset was divided into training (80%) and testing (20%) subsets with stratified sampling to maintain class balance.
2. **Feature Scaling** – StandardScaler was used to normalize feature values.
3. **Models Applied**:
   * **Logistic Regression** – Baseline interpretable classifier.
   * **Decision Tree Classifier** – Captures non-linear relationships and rule-based decisions.
   * **Random Forest Classifier** – Ensemble method providing higher accuracy and reduced overfitting.
4. **Evaluation Metrics**:
   * **Confusion Matrix** – To assess classification accuracy across Default vs Non-default.
   * **Classification Report** – Precision, Recall, F1-score for balanced evaluation.
   * **ROC-AUC Score** – Measures ability to discriminate between classes.
   * **ROC Curves** – Compared model performance visually.

**Findings**:

## ****Logistic Regression****

* **Accuracy:** 99.7%
* **Precision/Recall/F1:** Very high for both classes (No and Yes)
* **ROC-AUC:** 1.0 → perfect separation between defaulters and non-defaulters

Model is highly reliable and generalizes well on your test set.

## ****Decision Tree****

* **Accuracy:** 99.7% (same as Logistic Regression)
* **Precision/Recall/F1:** Also extremely high
* **ROC-AUC:** 0.997 → slight drop compared to Logistic Regression, possibly due to overfitting or the tree structure

Minor misclassifications: 1 false positive and 2 false negatives.

## ****Random Forest****

* **Accuracy:** 100%
* **Precision/Recall/F1:** Perfect across both classes
* **ROC-AUC:** 1.0 → perfect predictive power

Random Forest appears to be **the strongest model**, combining multiple trees to eliminate misclassifications.