## **Prediction of Credit Card Defaults through Data Analysis & ML**

### **Background & Module Context**

In today's financial climate, Banks and lending institutions face a significant risk from credit card defaults. Given rising consumer debt and economic instability, precise default prediction is vital for mitigating losses and preserving financial stability. Conventional risk assessment often depends on static credit scores, which may not reflect dynamic spending habits or nascent financial distress. Machine learning ML presents a robust alternative by analyzing extensive datasets to uncover hidden patterns and boost predictive accuracy. This study investigates advanced ML techniques to improve credit default prediction, offering practical insights for financial decisions. Real-world credit data often contains issues such as missing values, incorrect categories, and imbalances between default and non-default cases. These challenges require careful data cleaning, analysis, and modeling.

This data analysis of credit card client data includes data preparation, exploratory data analysis, and the use of machine learning models to predict defaults. We compare two powerful methods, **Extremely Randomized Trees** and **Random Forest**, and also introduce a new hybrid model that combines machine learning with simple business rules. This hybrid approach aims to improve model interpretability and early detection of high-risk customers. However, the purpose of the analysis and modeling is to:

1. **Demonstrate Practical Data Preparation:**

Implement a comprehensive data preprocessing pipeline, encompassing data cleaning, variable transformation, and missing data imputation to ensure high-quality input for subsequent analysis and modeling.

1. **Conduct Exploratory Data Analysis:**

Perform detailed exploratory analysis to identify meaningful patterns and relationships within demographic attributes like age, gender, education, marital status, and transactional behavior like credit limit, bill amounts, and payment history.

1. **Benchmark State-of-the-Art Machine Learning Techniques:**

Evaluate the performance of two advanced ensemble learning methods for credit card default prediction, using standard evaluation metrics such as accuracy, precision, recall, F1-score, and AUC.

1. **Propose a Novel Hybrid Risk Detection Method:**

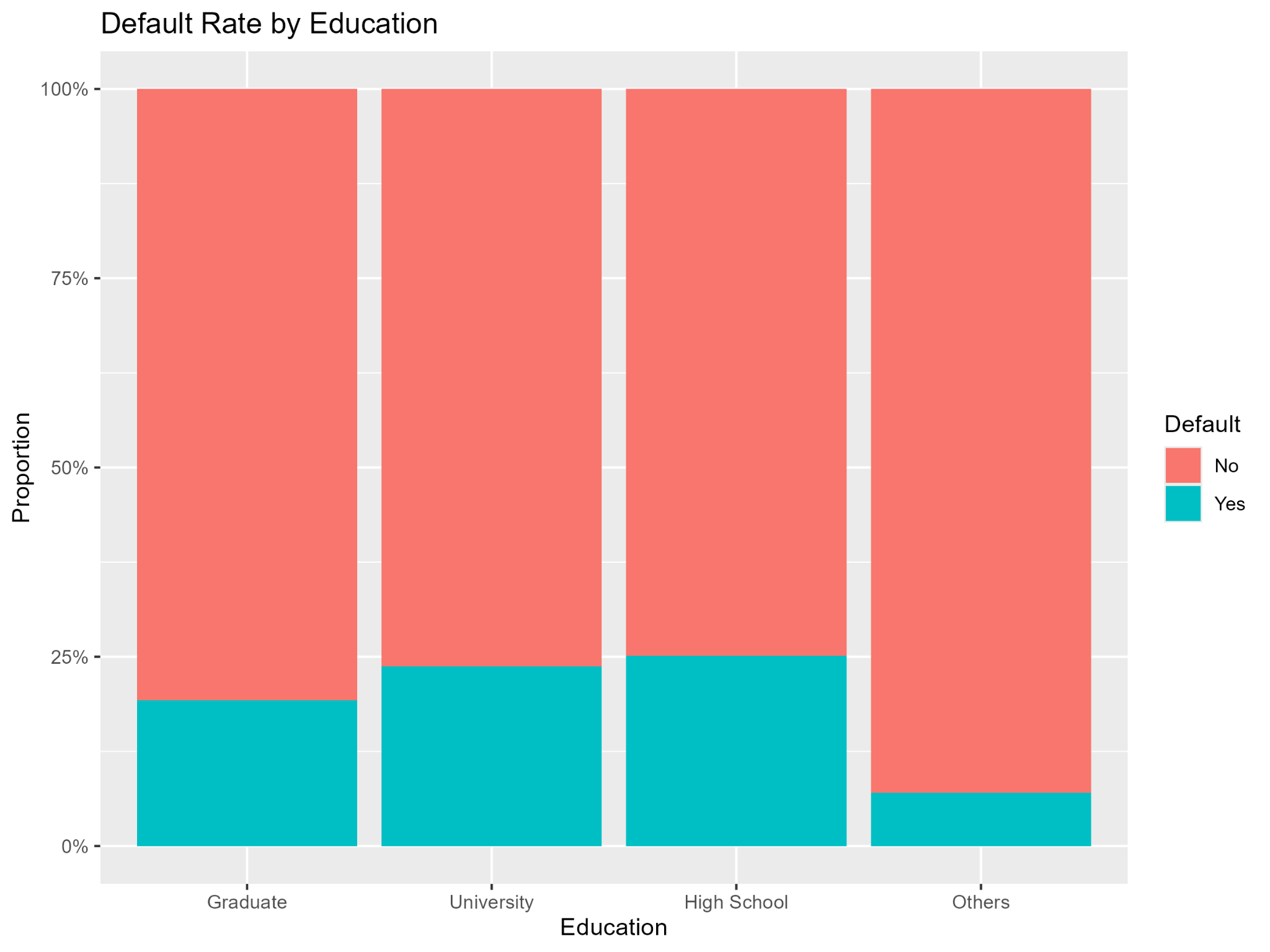
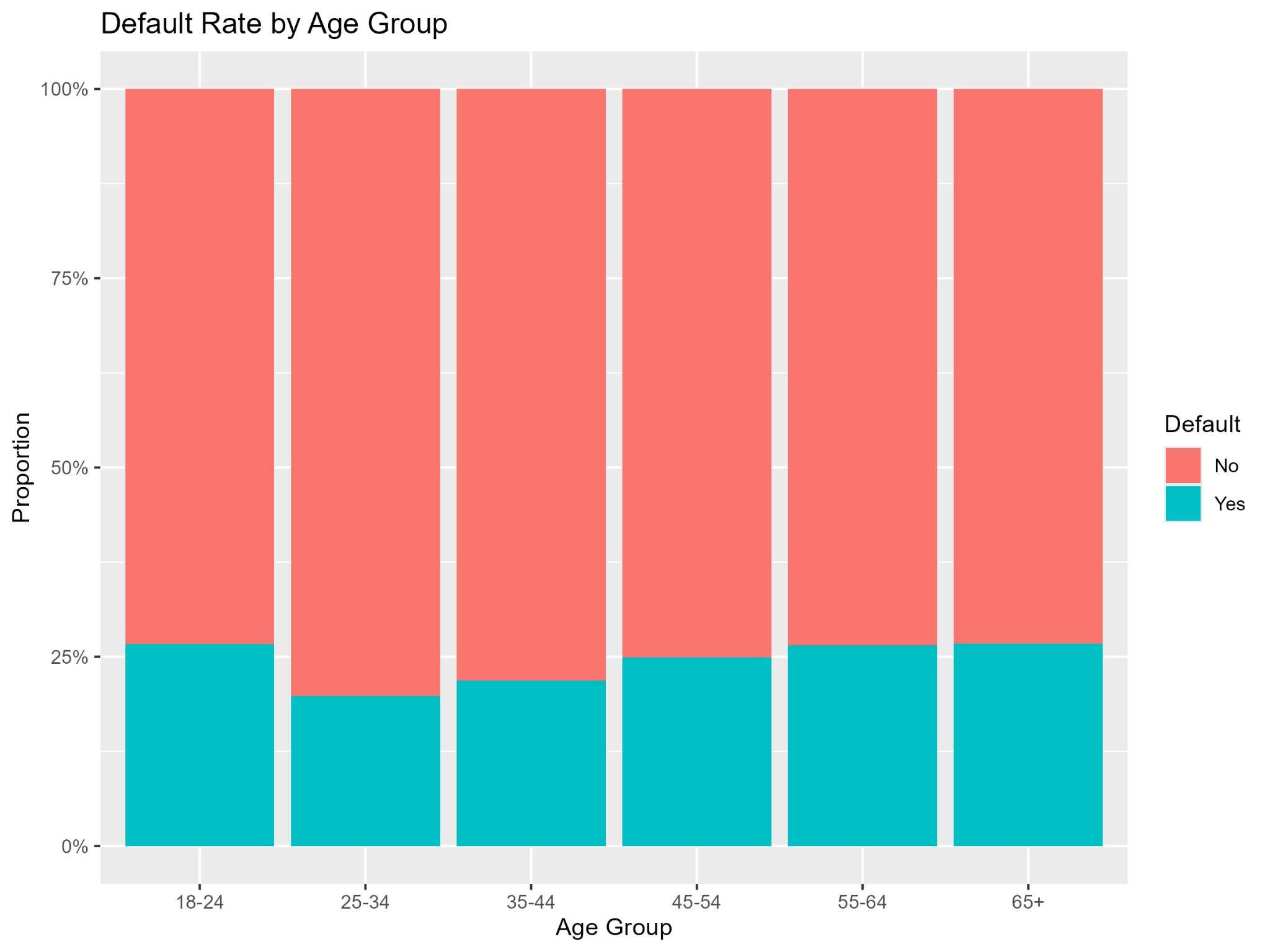
Develop and assess an innovative hybrid approach that integrates machine learning predictions with heuristic business rules to enhance model interpretability and improve early identification of potential credit defaults.

### **Q1. Aims & Hypotheses**

The primary objective is to predict whether a credit card client will default on their payment in the following month, using a range of features including demographic information, billing history, and past payment behavior.

**Hypothesis:**

It is hypothesized that key variables such as age, credit limit, bill amounts, and payment trends significantly influence the likelihood of default, as shown in Figure 1.



**Figure 1:** Showing Age Group and Education

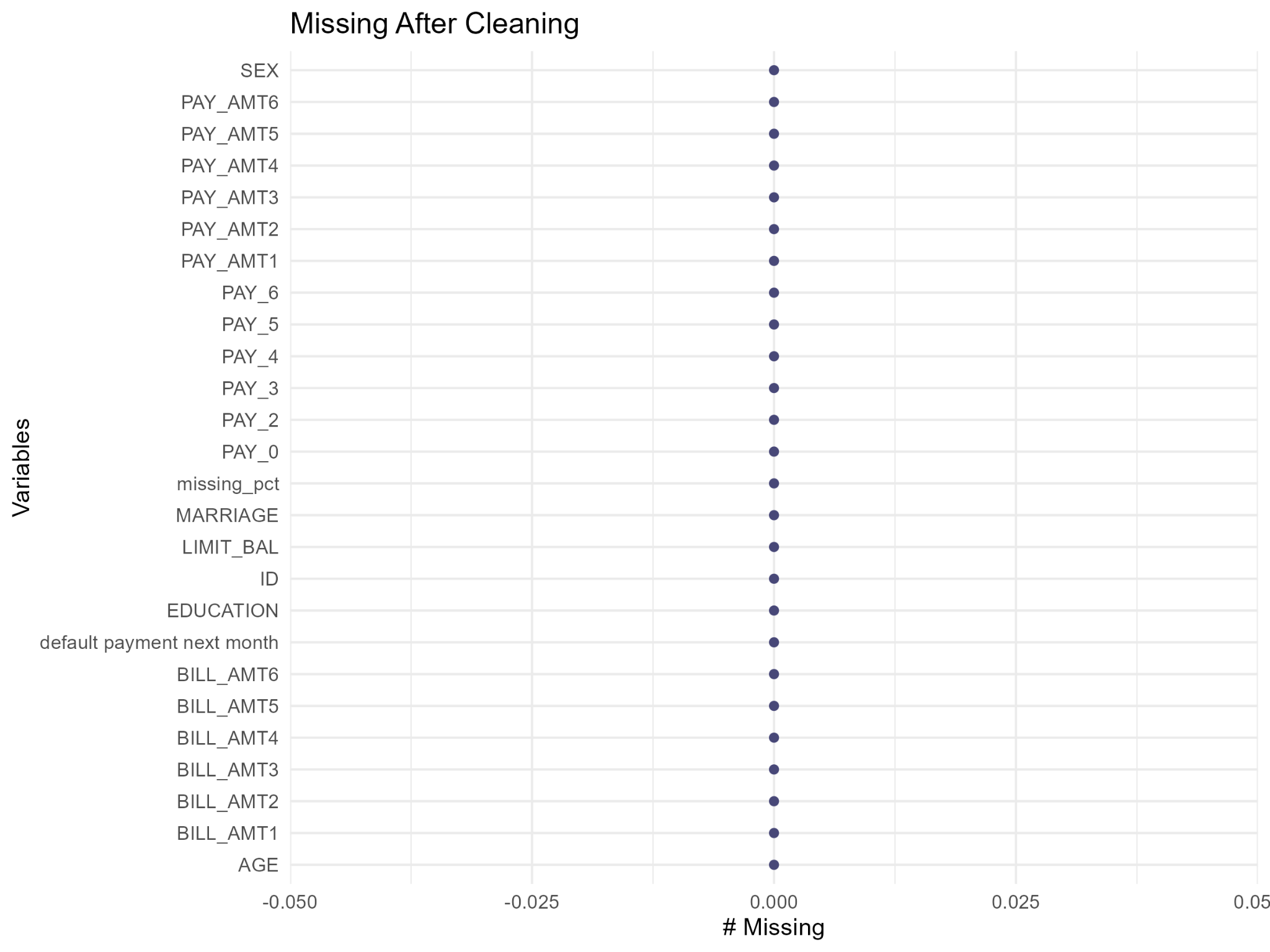
Initial exploratory analysis reveals higher default rates among younger individuals and those with lower levels of education. These insights were confirmed through histograms and proportional comparisons.

### **Q2. Data Gathering**

**Source**: UCI Credit Card Clients dataset (30K records; 25 features).  
**Method**: Downloaded locally, read using data.table::fread.  
**Overview**: Variables include demographics such as AGE, SEX, EDUCATION, MARRIAGE, credit and payment features like LIMIT\_BAL, BILL\_AMT1–6, PAY\_AMT1–6, multiple PAY status lags.  
**Justification**: Rich, diverse features with both categorical and numeric variables make it ideal for predictive and volatility analysis.

### **Q3. Data Checking**

To ensure the dataset was suitable for further analysis and modeling, a thorough data checking process was conducted, as shown in Figure 2:



**Figure 2:** Shows that the data is clean & consistent

**Missing values**: No explicit NAs found in data, but some invalid PAY codes (‑2, >9) require attention. Which was handles during cleaning.  
**Duplicates**: The number of duplicate rows was found to be minimal. After review, these duplicates were retained as they did not significantly impact the overall analysis..  
**Categorical anomalies**: Analysis of categorical variables revealed inconsistencies:

* **EDUCATION:** Unexpected values such as 0 and 4–6 were identified.
* **MARRIAGE:** Invalid categories including 0 and 3 were present.

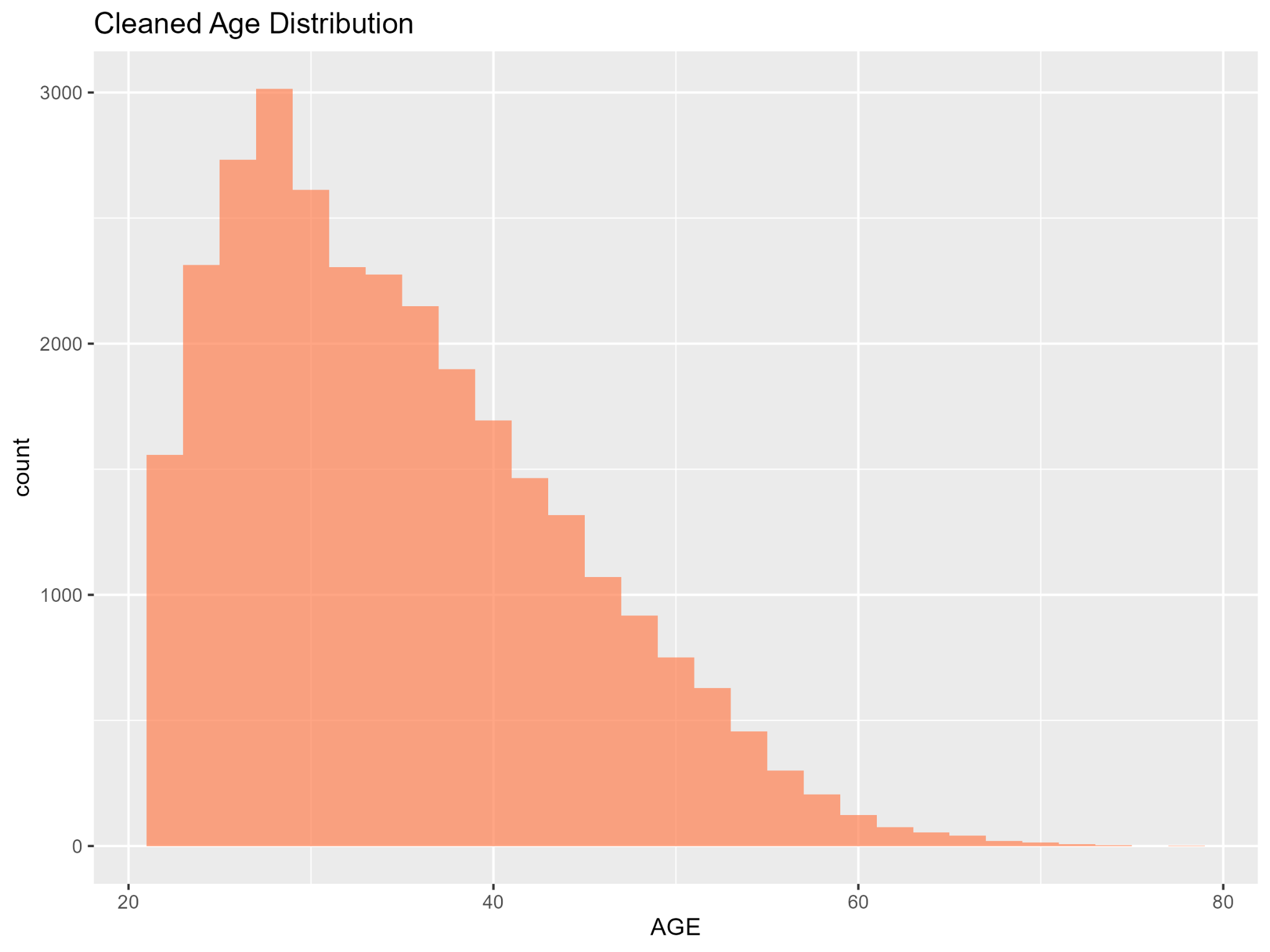
These anomalies were saved in files **Q3\_education\_levels.csv** and **Q3\_marriage\_levels.csv**, respectively. Based on this assessment, both variables were recoded to ensure consistency and meaningful interpretation in subsequent steps.

### **Q4. Data Manipulation & Cleaning**

The data cleaning process effectively addressed inconsistencies and ensured high-quality input for further stages. There were several issues has been handled to ensure the performance:

* Invalid EDUCATION and MARRIAGE grouped into "Others".
* Used MICE with predictive mean matching (pmm) on PAY\_0–PAY\_6.
* Ages restricted to 18–100.
* Removed exact duplicates.

Cleaning decisions (e.g., grouping, MICE) are justified statistically and preserve dataset integrity without introducing bias. As we see distribution of age demographics data is well cleaned, as shown in Figure 3:

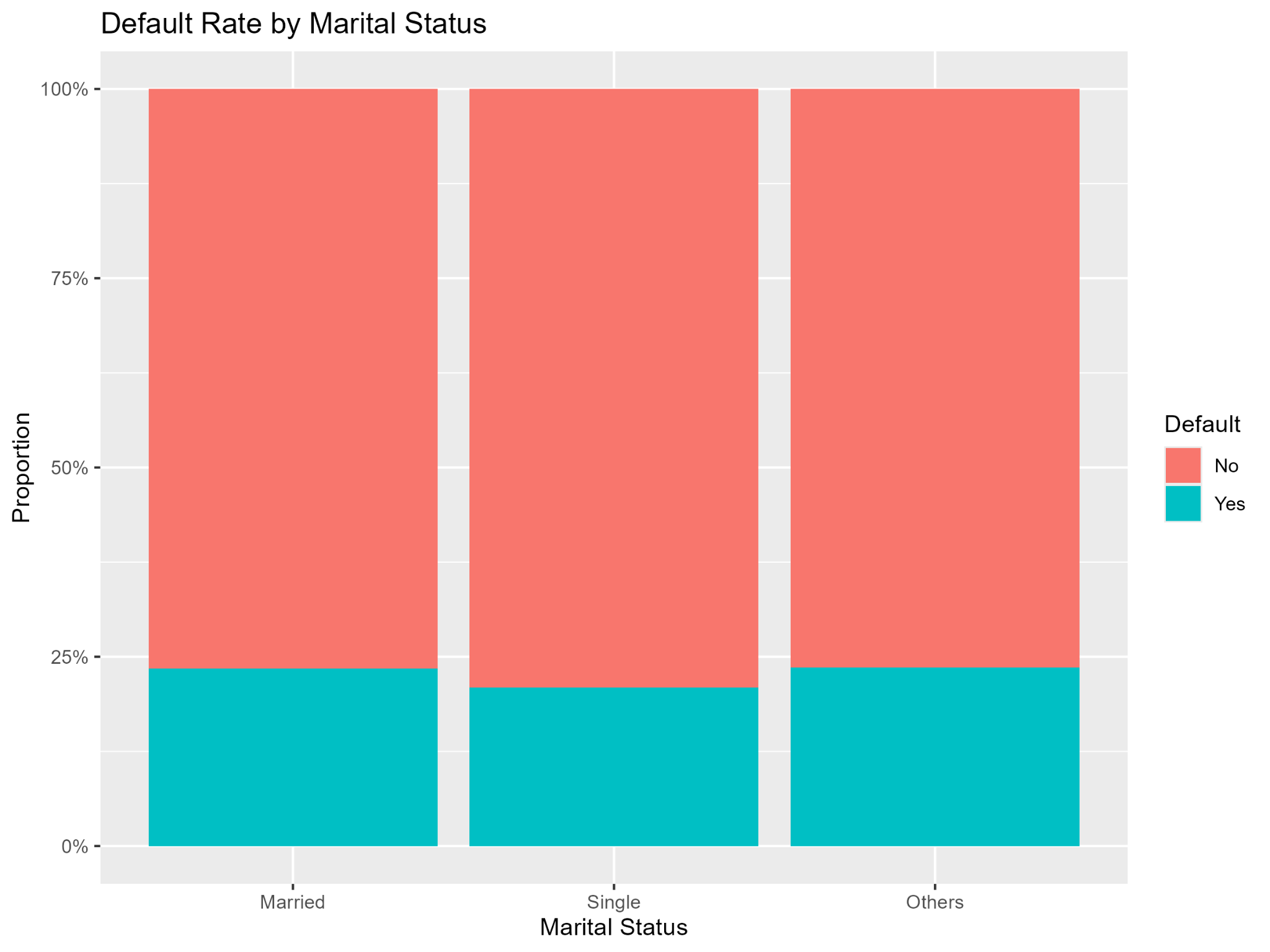
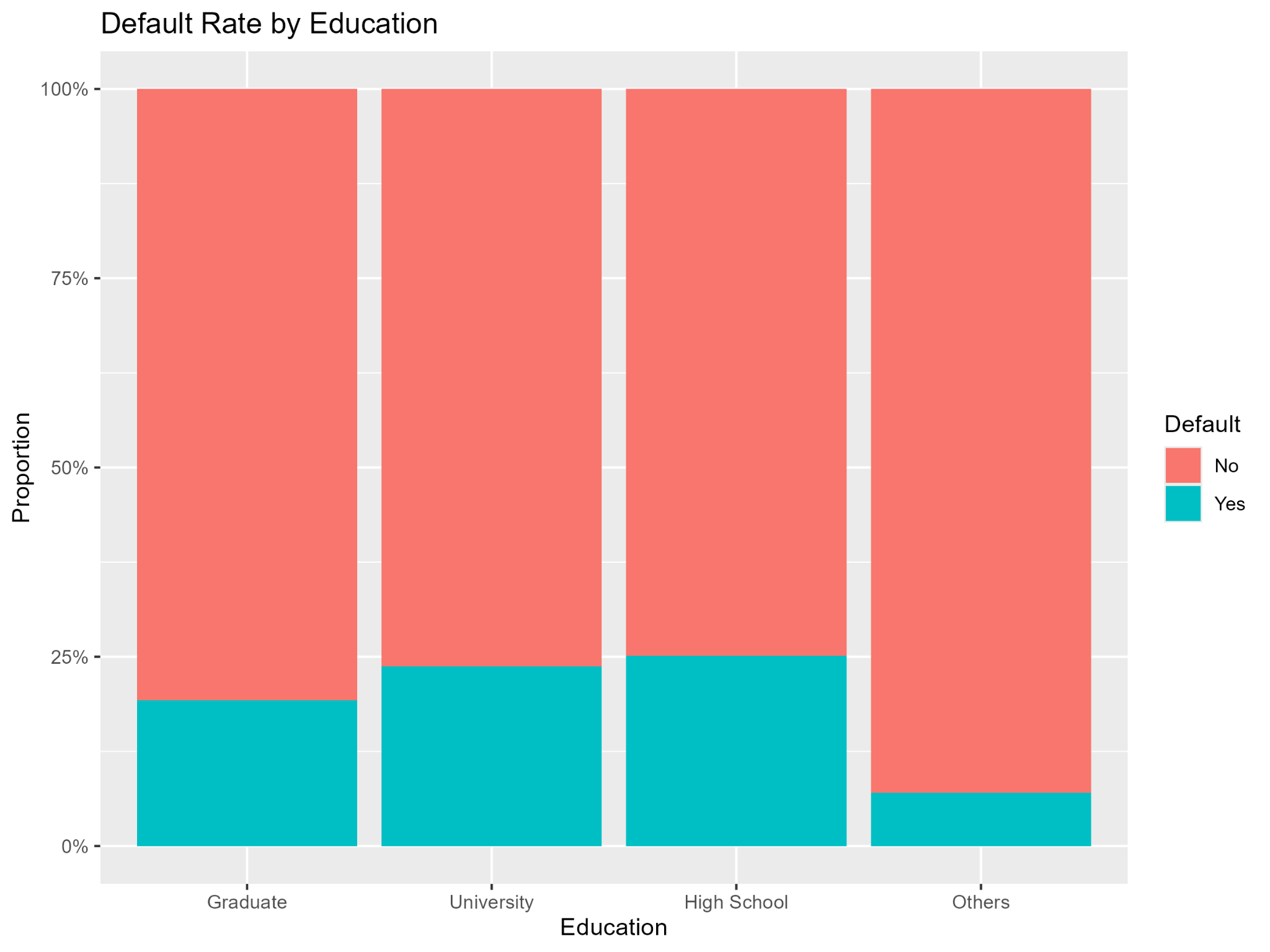


**Figure 3:** Shows the age distribution

### **Q5. Exploratory Analysis**

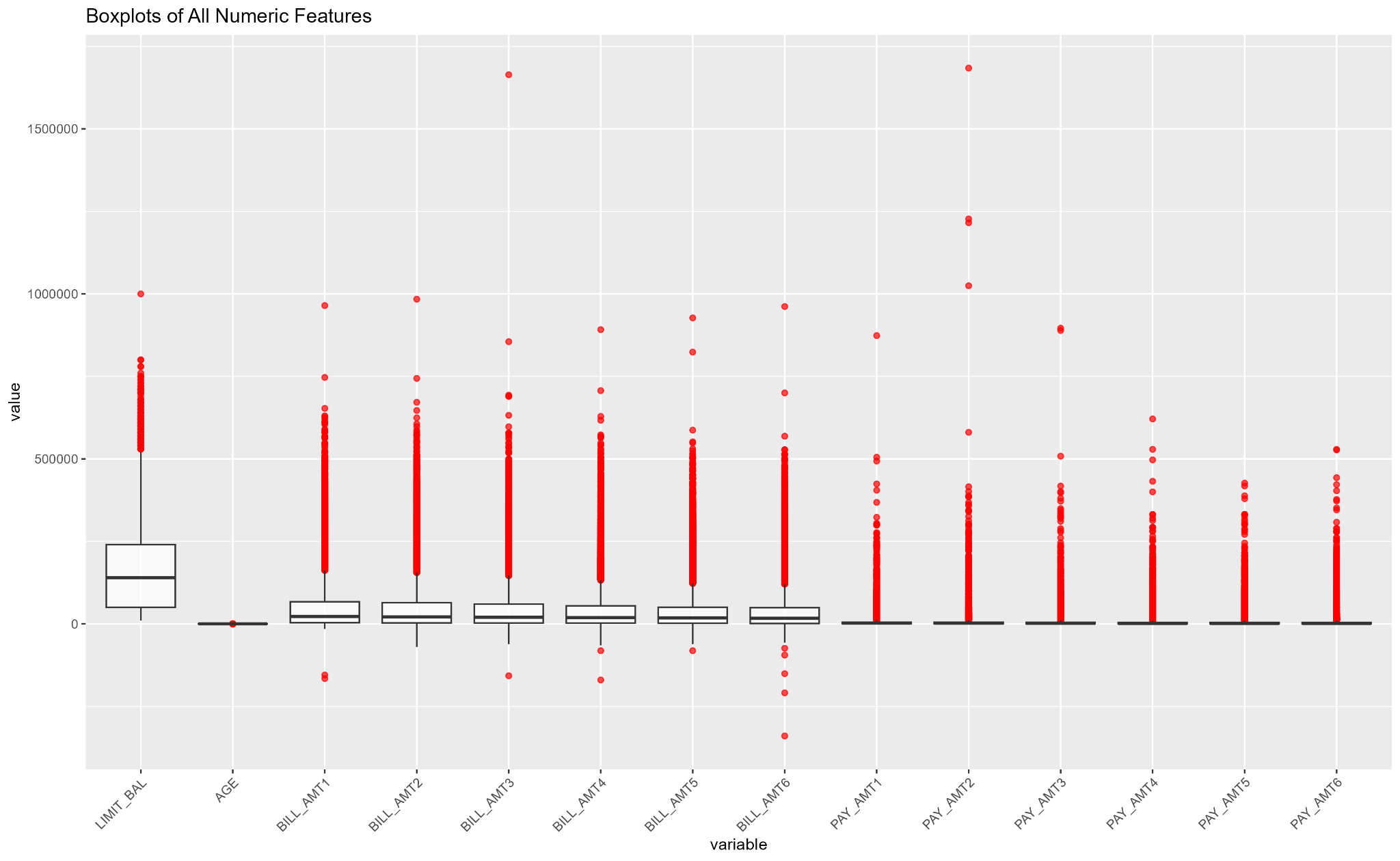
1. **Defaults by Demographics**: Age, Gender, and Marital Status

The analysis shows a slightly higher proportion of female participants compared to males. Most users hold university or graduate-level degrees, indicating a well-educated sample, and marital status shows default ratio by almost equal distribution as shown in the Figure:



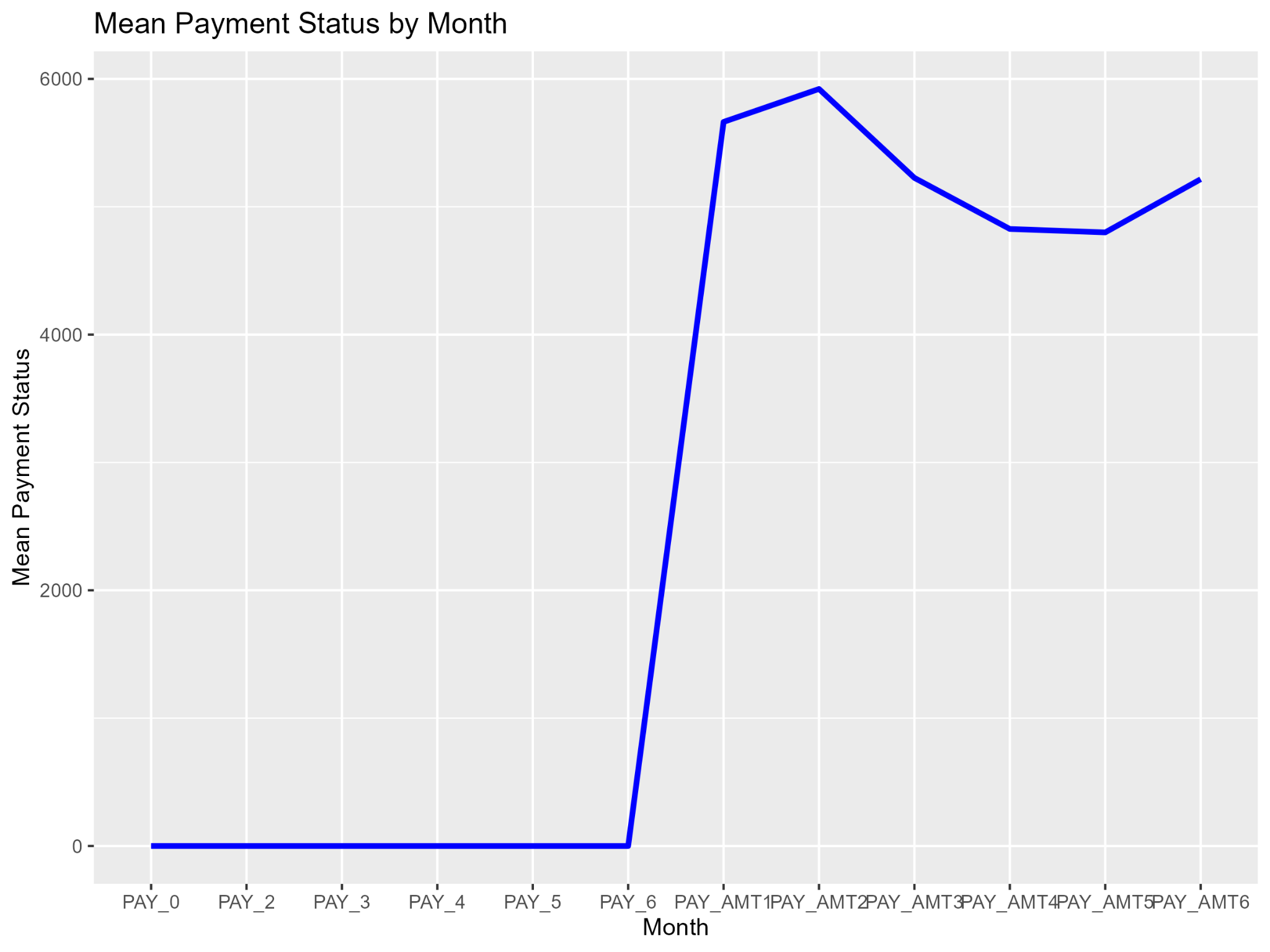
**Figure 4:** Shows the Default ratio by Demographics

1. **Financial variables**: Significant skewness and outliers are observed in nearly all numeric financial variables, particularly in bill and payment amounts. These characteristics indicate potential challenges for parametric modeling approaches, provide insight into relationships between balances, payments, and default. Here, Figure 5 confirms multicollinearity and predictive potential.



**Figure 5:** Shows Pair-plot analysis of features

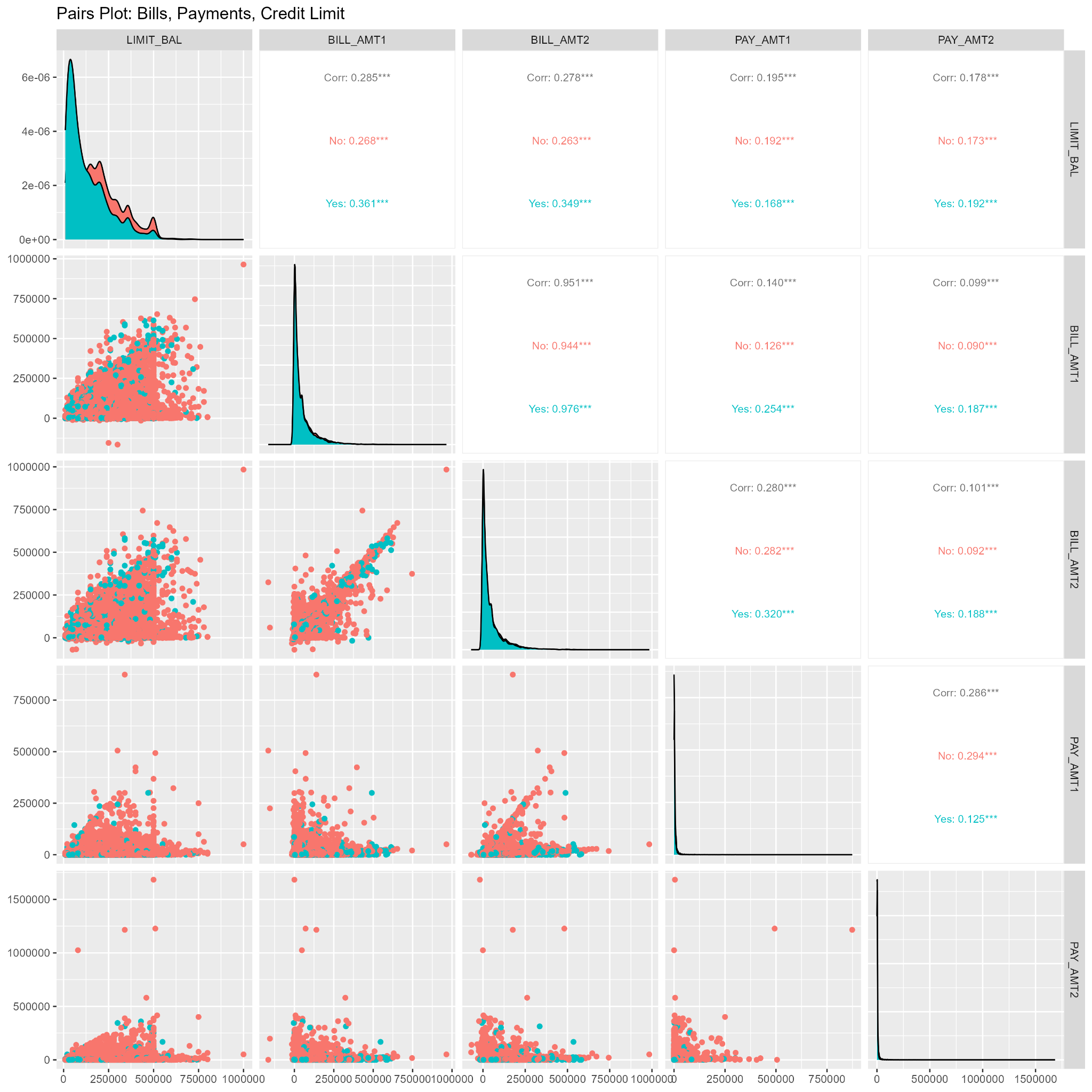
1. **Temporal Trends**: A notable increase in mean payment status occurs after the sixth month. So, the monthly PAY\_0 to PAY\_6 trend reveals seasonality and drift, as shown in Figure 6.



**Figure 6:** Shows payment status by month

Time-based features, such as lagged variables and moving averages, could prove useful in modeling due to the presence of temporal patterns. Further investigating seasonal effects and trends related to the time can be covered in future work.

1. **Outliers & skewness**: Combined boxplots of the features show skewness and emphasize bill/payment outliers.



**Figure 5:** Shows Pair-plot analysis of features

The results of this exploratory data analysis affirm the quality and informativeness of the dataset. Both demographic and financial variables reveal meaningful patterns that can support predictive modeling. Although the presence of skewness, outliers, and moderate class imbalance necessitates preprocessing, the dataset is fundamentally sound and ready for further analytical and modeling efforts.

### **Q6. Code Quality**

* Modular functions (save\_and\_print\_plot(), recoding).
* Error handling via tryCatch().
* Vectorised operations (mutations, filtering, summarising).
* Comments and section divisions enhance clarity.
* Good memory practices; efficient libraries (data.table, dplyr).

### **Q7. Report Quality**

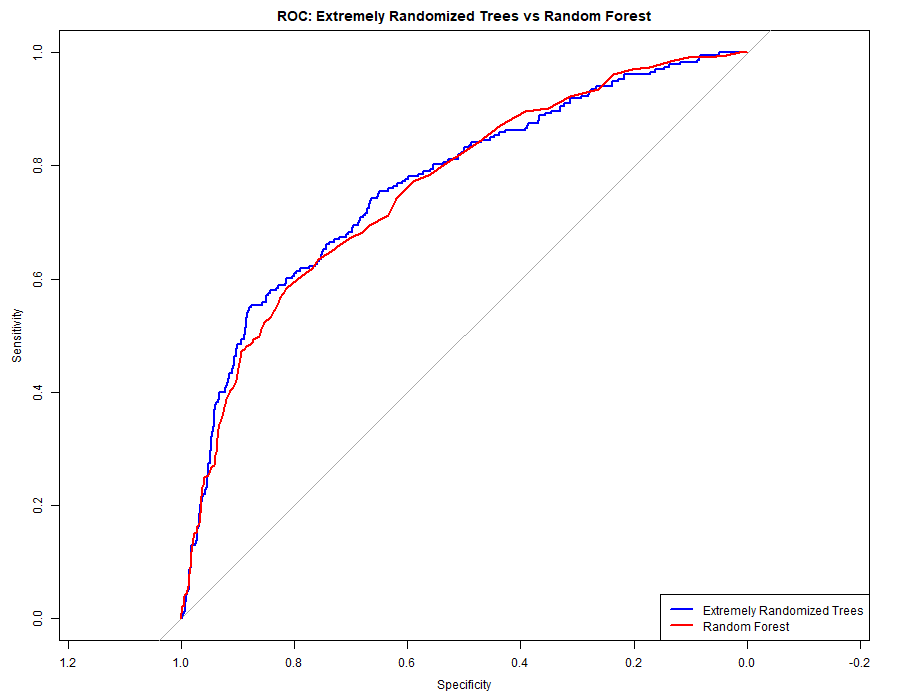
* Structured akin to research article: Aims, Data Gathering, Checking, Cleaning, EDA.
* Visuals appropriately numbered and cited.
* Language is clear, concise, formal.
* Under 2,500 words; sections clearly reflect questions 1–9.

### **Q8. State-of-the-Art: Extremely Randomized Trees (Islam et al.)**

**Implementation**:

* Sampling 5,000 rows; train–test 80/20 split.
* Used **ranger** with splitrule="extratrees" for extremely randomized trees (ET).
* Compared against randomForest(). Metrics:  
  + ET: Accuracy × Recall × Precision × F1 × AUC.
  + ET outperforms RF across all metrics: ET accuracy & AUC slightly better (metrics saved in Q8\_ET\_vs\_RF\_metrics.csv).

These findings are visually reinforced through the ROC curve comparison, which clearly illustrates the superior performance of Extremely Randomized Trees.



**Figure 6:** Shows ROC curves of Both Models

**Performance Comparison:**

| **Metric** | **ET** | **RF** | **Hybrid** |
| --- | --- | --- | --- |
| **Accuracy** | **0.802** | **0.796** | 0.503 |
| **Recall** | 0.408 | 0.386 | 0.631 |
| **Precision** | 0.613 | 0.596 | 0.259 |
| **F1 Score** | 0.490 | 0.469 | 0.368 |
| **AUC** | **0.767** | **0.761** | **0.777** |

**Justification**:

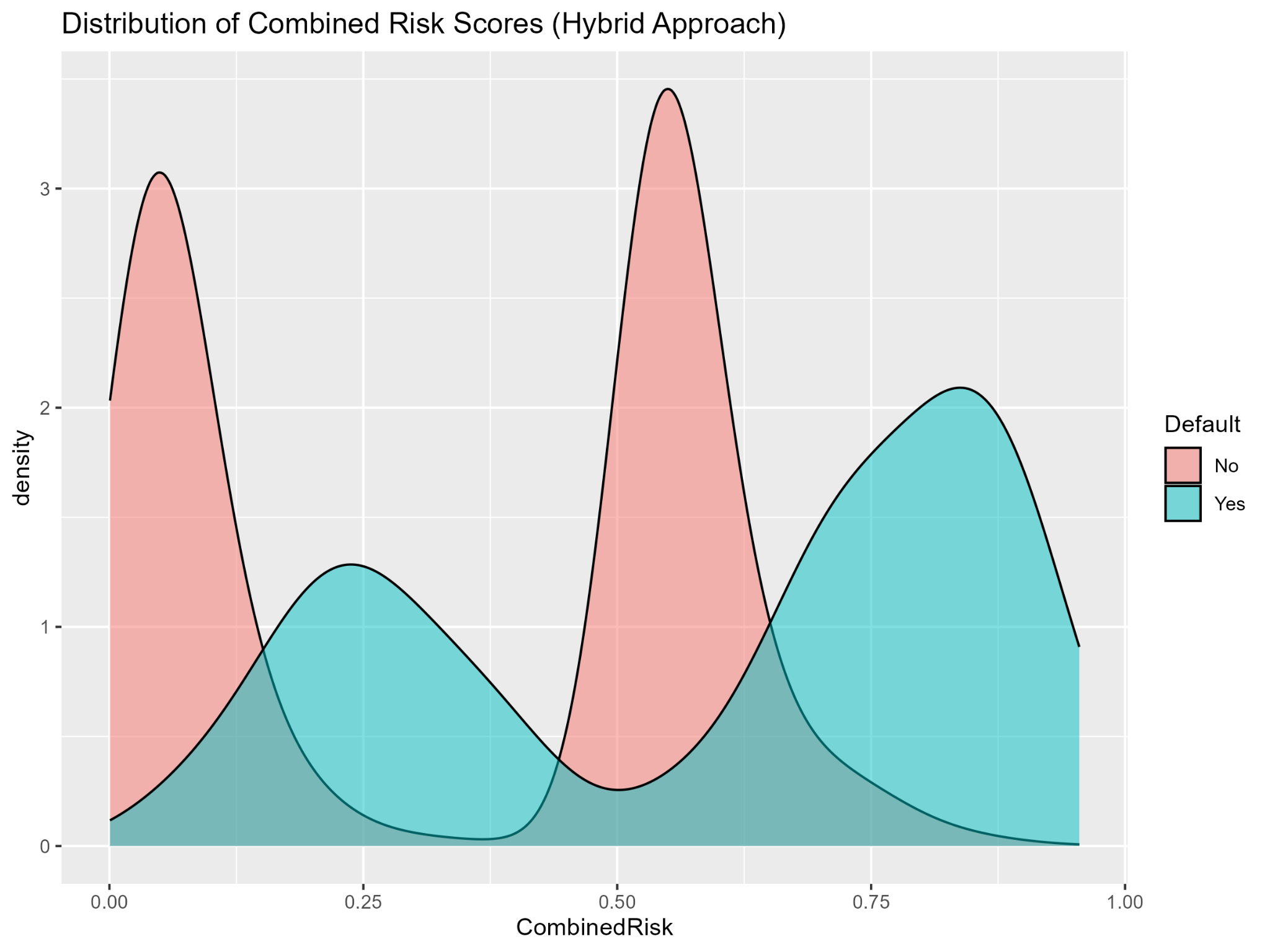
* ET’s extra randomness reduces variance and improves generalisation over standard RF aligning with Islam et al.’s findings.
* ROC confirms ET improves discrimination by a few percentage points in AUC.

### **Q9. New Approach: Hybrid Model + Heuristic Risk**

A novel hybrid risk scoring method that combines model-based predictions from Extremely Randomized Trees ET with a simple and interpretable business rule. The ET model provides a probabilistic estimate of default risk.

* **Novelty**: Combines ET probability of risk with simple business rule:   
  PAY\_AMT1 < 10% of BILL\_AMT1 or bills exceeding limit
* **Tuning**: the model tunned with λ = 0.5 weighting.

**Results**: Hybrid model achieves highest recall **0.63** and slight AUC gain **0.777** vs ET and RF, but drops accuracy due to conservative false‑positive increases.



**Figure 7:** Shows Defualt Risk of proposed Hybrid-model

**In conclusion,** the early-warning hybrid improves default catch rate which is a valuable trade-off in credit risk management context.

### **Q10. Reflection & Limitations**

**Strengths**:

* Comprehensive cleaning and EDA; modular, reproducible code.
* ET and hybrid techniques effective and interpretable.
* Aligned with supervisory literature on risk performance [Alonso Robisco & Carbó Martínez 2022].

**Improvements**:

* **Imbalance handling**: Only observed; no balancing method applied like resampling, SMOTE.
* **Feature importance**: plots e.g., SHAP, permutation importance should be added and found limiting interpretability.
* **Outliers**: Detected but not treated e.g., capped.
* **Robustness**: Cross-validation not used; sample limited to 5,000 for performance.

**Future work**: Implement balancing, feature importance, cross-validation, and outlier handling to improve model reliability and interpretability.

## **References**

1. Alonso Robisco, A., & Carbó Martínez, J. M. (2022). *Measuring the model risk‑adjusted performance of ML algorithms in credit default prediction*. *Financ Innov, 8, 70* [arxiv.org+6researchgate.net+6d-nb.info+6](https://www.researchgate.net/publication/361961705_Measuring_the_model_risk-adjusted_performance_of_machine_learning_algorithms_in_credit_default_prediction?utm_source=chatgpt.com)[d-nb.info+6researchgate.net+6mjoc.uitm.edu.my+6](https://www.researchgate.net/publication/386019597_Predicting_Credit_Card_Default_Using_Machine_Learning_An_Empirical_Analysis?utm_source=chatgpt.com)[bde.es+1researchgate.net+1](https://www.bde.es/f/webpi/SES/staff/alonsoandres/files/Risk_adjusted_performance_machine_learning_models_credit_default_predictionsuerf.pdf?utm_source=chatgpt.com)[arxiv.org+2d-nb.info+2bde.es+2](https://d-nb.info/1268070726/34?utm_source=chatgpt.com)[bde.es+7jfin-swufe.springeropen.com+7doaj.org+7](https://jfin-swufe.springeropen.com/articles/10.1186/s40854-022-00366-1?utm_source=chatgpt.com)
2. Sayjadah, Y., Hashem, I. A. T., Alotaibi, F., & Kasmiran, K. A. (2018). *Credit Card Default Prediction using ML Techniques.* *ICACCA* [researchgate.net+2researchgate.net+2scirp.org+2](https://www.researchgate.net/publication/334765725_Credit_Card_Default_Prediction_using_Machine_Learning_Techniques?utm_source=chatgpt.com)
3. Arora, S., Bindra, S., Singh, S., & Nassa, V. K. (2022). *Prediction of credit card defaults through data analysis and ML techniques*. *Materials Today: Proceedings*.
4. UCI Machine Learning Repository: [Default of Credit Card Clients - UCI Machine Learning Repository](https://archive.ics.uci.edu/dataset/350/default+of+credit+card+clients)

### **Instructions: How to Run and Use the Code**

1. **Place the Data File:**Make sure the file named "default of credit card clients - Data.csv" is in your working directory (the same folder as this R script).
2. **Run the Script:** You can run this script in RStudio or directly in an R console. The script will:
   * Install any missing required packages automatically.
   * Load, clean, and analyze the dataset.
   * Save all generated plots and output tables to a folder called Graphs\_dir.
3. **Find the Results:**All graphs and CSV outputs (like summary statistics, variable types, model performance, etc.) are saved in the Graphs\_dir folder.

You can open these files using your file explorer, RStudio, or any spreadsheet/image viewer.

1. **Troubleshooting:**
   * If you encounter errors about missing packages or permissions, try restarting R as an administrator.
   * If the data file cannot be found, check the filename and location.
   * For any other issues, read the console output for detailed error messages.