Effective Predictive Model for Loan Approval Status

Abstract:

This study focuses on developing a predictive model to accurately determine loan approval status, a critical component in financial decision-making. Utilizing a dataset comprising various applicant attributes such as income levels, credit history, and loan amounts, we employed machine learning techniques to forecast the binary outcome of loan approval - approved or denied. Our approach encompassed data cleaning, handling missing values, and feature engineering to optimize the dataset for analysis. We then implemented a Random Forest classifier, renowned for its efficacy in handling complex, non-linear relationships within data. The model was rigorously evaluated using metrics like accuracy, precision, recall, and the F1 score to ensure its reliability. Additionally, Receiver Operating Characteristic (ROC) curves and the Area Under the Curve (AUC) were analyzed to assess the model's discriminative ability. The results indicated a strong predictive capability, showcasing the potential of machine learning in enhancing decision-making processes in the financial sector. This research contributes to the burgeoning field of financial analytics, offering insights into the application of advanced algorithms for credit risk assessment and providing a framework for financial institutions to improve their loan approval processes.

Introduction;

In the rapidly changing landscape of financial services, the process of making loan approval decisions is a crucial intersection of technology and economics. The capacity to accurately forecast loan approval outcomes is vitally important for financial institutions to manage risk effectively. It also significantly influences the economic opportunities available to individuals and businesses. Our study is motivated by the need to leverage advanced statistical techniques to enhance these predictions, thereby facilitating more informed and equitable lending decisions.

This research delves into the realm of statistical analysis, utilizing a variety of statistical methods to predict loan approval statuses. The use of statistical techniques in data analysis has a long-standing history, especially in the finance sector, where they play a pivotal role in understanding and modeling complex financial phenomena.

The core objectives of our study include:

1. **Application of Statistical Methods:** We apply several statistical techniques, including logistic regression and probit models, among others, to predict the binary outcome of loan approvals. These methods are selected for their proven effectiveness in handling various types of data and their ability to reveal underlying relationships between variables.

2. **Data Preprocessing and Feature Engineering:** Recognizing the impact of data quality on statistical modeling, we undertake comprehensive data preprocessing. This involves addressing missing values, encoding categorical variables, and conducting feature engineering to improve the predictive quality of the dataset.

3. **Comparative Model Evaluation:** A key aspect of our research is the comparative analysis of these statistical models based on crucial performance metrics such as accuracy, precision, recall, the F1 score, and ROC-AUC scores. This thorough evaluation helps us assess not only the accuracy but also the robustness and practical applicability of each model in a real-world financial setting.

4. **Insights for Financial Institutions:** The study aims to provide valuable insights to financial institutions. By understanding the capabilities and limitations of various statistical models in predicting loan approvals, these institutions can enhance their risk assessment processes, potentially leading to more efficient and fair lending practices.

Ultimately, this study is driven by the goal of integrating advanced statistical methodologies into the financial sector. Our aim is to go beyond mere risk mitigation, fostering a more data-driven, transparent, and efficient environment for lending decisions.

Data Description:

Our study utilizes a comprehensive dataset sourced from a financial institution, specifically designed for assessing loan approval processes. The dataset comprises various attributes that are commonly considered by financial institutions when evaluating loan applications. Below is a detailed description of each variable in the dataset, including their nature and units of measurement where applicable.

1. **Loan_ID**: A unique identifier for each loan application. This is a nominal variable consisting of alphanumeric characters.

2. **Gender**: The gender of the applicant. This is a categorical variable with two levels: 'Male' and 'Female'.

3. **Married**: Marital status of the applicant. It is a binary categorical variable with 'Yes' indicating married and 'No' indicating unmarried.

4. **Dependents**: The number of dependents relying on the applicant's income. This ordinal variable is categorized as '0', '1', '2', '3+'.

5. **Education**: The educational background of the applicant. This categorical variable includes two levels: 'Graduate' and 'Not Graduate'.

6. **Self_Employed**: Indicates whether the applicant is self-employed. It is a binary categorical variable with 'Yes' and 'No' as possible values.

7. **ApplicantIncome**: The income of the applicant. This is a continuous variable measured in local currency units (e.g., USD, INR).

8. **CoapplicantIncome**: The income of the co-applicant. This is also a continuous variable and is measured in the same units as the ApplicantIncome.

9. **LoanAmount**: The loan amount requested by the applicant. This is a continuous variable, measured in thousands of local currency units.

10. **Loan_Amount_Term**: The term over which the loan is to be repaid. This is a continuous variable, measured in months.

11. **Credit_History**: A record of past loan repayments. It is a binary categorical variable, where '1' indicates a good credit history and '0' indicates a poor credit history.

12. **Property_Area**: The type of area where the property is located. This categorical variable includes three levels: 'Urban', 'Semiurban', and 'Rural'.

13. **Loan_Status**: The outcome variable indicating whether the loan was approved ('Y') or not ('N'). This is the primary binary categorical variable of interest in our analysis.

The data is ideal for statistical analysis due to its diverse range of variables, encompassing demographic, financial, and credit-related attributes.

In our analysis, each of these variables is carefully examined to understand their individual and collective impact on the likelihood of loan approval. The continuous variables such as ApplicantIncome, CoapplicantIncome, and LoanAmount offer quantitative insights, while the categorical variables like Gender, Education, and Property_Area provide qualitative perspectives. The interplay between these variables is central to our statistical modeling and subsequent predictions regarding loan approvals.

Goal:

The overarching goal of our project is to utilize the provided dataset to develop a robust statistical model that can accurately predict the outcome of loan applications, specifically determining whether a loan will be approved or denied. This project aims to blend statistical theory with practical application, leveraging the available data to address a critical question in the financial sector: What factors most significantly influence the decision to approve or reject a loan application?

Research Questions:

1. **Primary Research Question:**
    - What are the key determinants that significantly impact the likelihood of loan approval?

2. **Exploratory Questions:**
    - How does the applicant's income (both individual and co-applicant) affect the probability of loan approval?
    - Does the applicant's gender, marital status, number of dependents, or education level play a significant role in the loan approval process?
    - Is there a correlation between the loan amount, its term, and the approval decision?
    - Does the credit history of the applicant substantially affect the outcome of the loan application?

- How does the property area (Urban, Semiurban, Rural) relate to the chances of getting a loan approved?

3. **Model-Specific Questions:**

   - Among the statistical models employed (such as logistic regression, probit models, etc.), which provides the most accurate predictions for loan approval?

   - How do different models compare in terms of key performance metrics like accuracy, precision, recall, F1 score, and ROC-AUC score?

The answers to these questions are intended to provide a comprehensive understanding of the factors influencing loan approval decisions. By addressing these queries, we aim to create a model that not only serves as a predictive tool for financial institutions but also sheds light on the dynamics of loan approval processes, potentially revealing areas for improvement in lending practices and policies.

Ultimately, our project seeks to bridge the gap between statistical theory and real-world financial applications, offering insights that could enhance decision-making processes in the lending industry.

Statistical Methods:

Our study employs a suite of statistical methods to address the research questions, each chosen for its relevance and efficacy in binary outcome prediction. Below is an overview of the methods used, along with brief technical descriptions.

1. **Logistic Regression:**

   - Logistic regression is a popular method for binary classification problems. It models the probability of a binary response based on one or more predictor variables.

   - We explored three logistic regression models: the null model (with no predictors), the full model (with all predictors), and a stepwise model (selecting variables based on their statistical significance).

2. **Probit Model:**

   - Similar to logistic regression, the probit model is used for binary response data. It differs in that it uses the probit function (the inverse of the cumulative distribution function of the standard normal distribution) to model the relationship.

3. **Decision Trees:**

   - Decision trees are a non-parametric supervised learning method used for classification. They split the dataset into branches to form a tree structure based on decision rules inferred from the data.

   - The algorithm selects the best attribute at each node to split the data, aiming to maximize the homogeneity of the resulting sub-groups regarding the target variable.

4. **Random Forest:**

  - Random Forest is an ensemble learning method that operates by constructing multiple decision trees during training and outputting the mode of the classes for classification. It improves over a single decision tree by reducing the risk of overfitting.
  - Each tree is built on a different subset of the data, and the final prediction is made by averaging the predictions from all the trees.

5. **XGBoost:**
  - XGBoost (Extreme Gradient Boosting) is an advanced implementation of gradient boosting algorithms. It is highly efficient, flexible, and portable. XGBoost provides a parallel tree boosting that solves many data science problems quickly and accurately.
  - The model uses gradient descent to minimize errors in sequential tree building, effectively refining the model with each step.

Each of these methods brings a unique approach to the problem, from the straightforward logistic and probit models focusing on individual variables' effects, to the more complex ensemble methods like Random Forest and XGBoost, which build upon multiple models for enhanced predictive power. The comparative analysis of these methods aims to identify which approach most effectively predicts loan approval outcomes, considering the dataset's specific characteristics and the underlying patterns within the data.

```r
df= read.csv("/Users/harshavardhan/Documents/stat/finalpro/loan_data_set.csv")
head(df)
```

```
   Loan_ID Gender Married Dependents    Education Self_Employed ApplicantIncome
1 LP001002   Male      No          0     Graduate            No            5849
2 LP001003   Male     Yes          1     Graduate            No            4583
3 LP001005   Male     Yes          0     Graduate           Yes            3000
4 LP001006   Male     Yes          0 Not Graduate            No            2583
5 LP001008   Male      No          0     Graduate            No            6000
6 LP001011   Male     Yes          2     Graduate           Yes            5417
  CoapplicantIncome LoanAmount Loan_Amount_Term Credit_History Property_Area
1                 0         NA              360              1         Urban
2              1508        128              360              1         Rural
3                 0         66              360              1         Urban
4              2358        120              360              1         Urban
5                 0        141              360              1         Urban
6              4196        267              360              1         Urban
  Loan_Status
1           Y
2           N
3           Y
4           Y
5           Y
6           Y
```

```
# Remove rows with NA values
df <- na.omit(df)
# Remove rows with empty strings
df <- df[rowSums(df == "") == 0, ]
df_cleaned  <- subset(df, select = -Loan_ID)
missing_values <- sapply(df, function(x) sum(is.na(x)))
```

1. **Categorical Variables**: The dataset includes categorical variables like 'Gender', 'Married', 'Dependents', 'Education', 'Self_Employed', 'Property_Area', and 'Loan_Status'. Each of these categories has 480 entries, indicating a complete dataset with no missing values.

2. **Numerical Variables**:

   - **ApplicantIncome**: Ranges from a minimum of 150 to a maximum of 81,000, with the median at 3,859 and the mean at 5,364, suggesting a right-skewed distribution.

   - **CoapplicantIncome**: Extends from 0 to 33,837, with a median of 1,084 and a mean of 1,581, also indicating a right-skewed distribution.

   - **LoanAmount**: Varies between 9 and 600, with a median of 128 and a mean of 144.7, suggesting a relatively symmetric distribution.

   - **Loan_Amount_Term**: Ranges from 36 to 480, predominantly centered around 360 as indicated by both the median and the most common quartile values.

   - **Credit_History**: A binary variable (ranging from 0 to 1) with a mean of 0.8542, indicating that most applicants have a positive credit history.

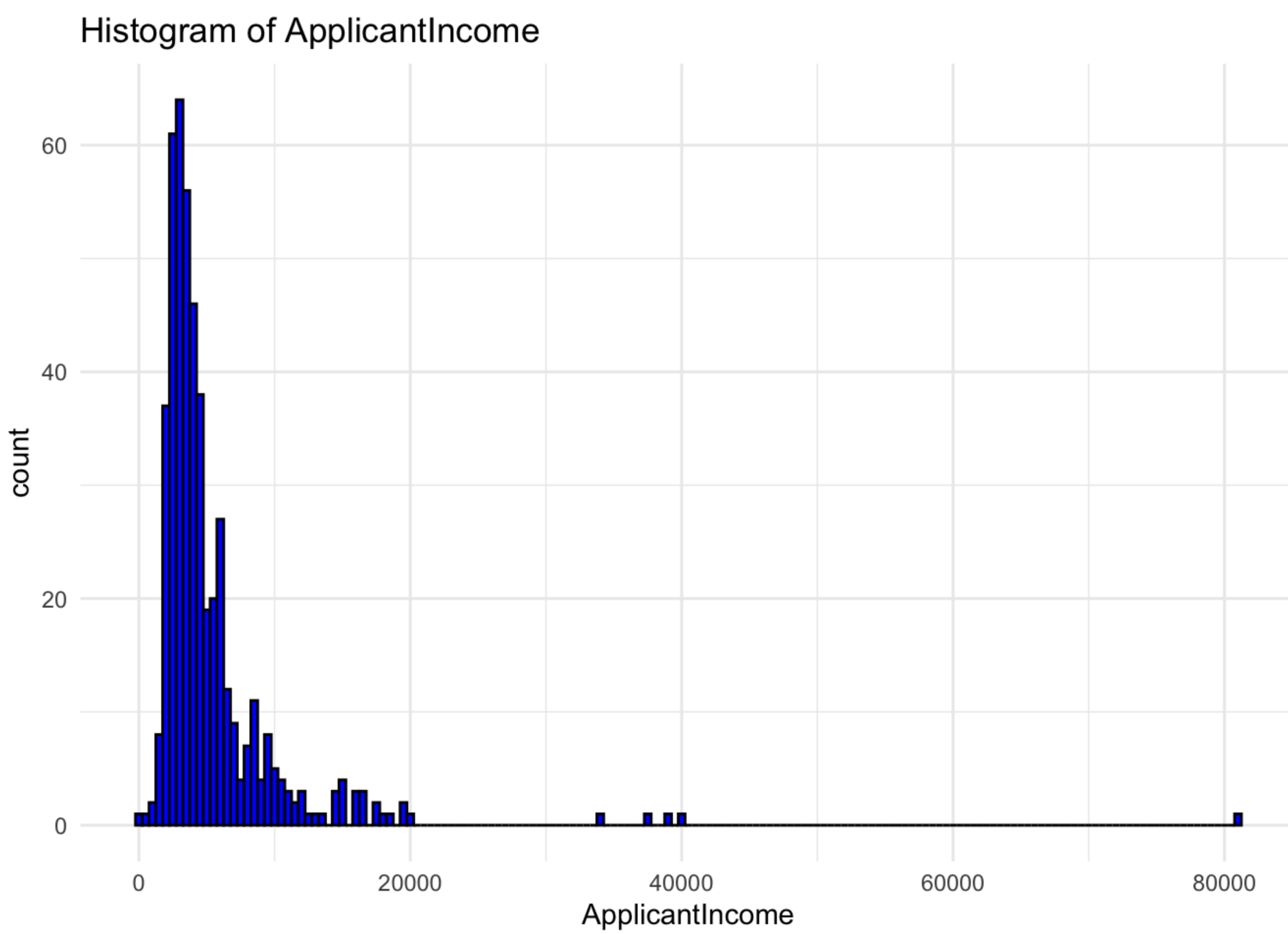
Exploratory Data Analysis

```
#univariate

summary(df)
```

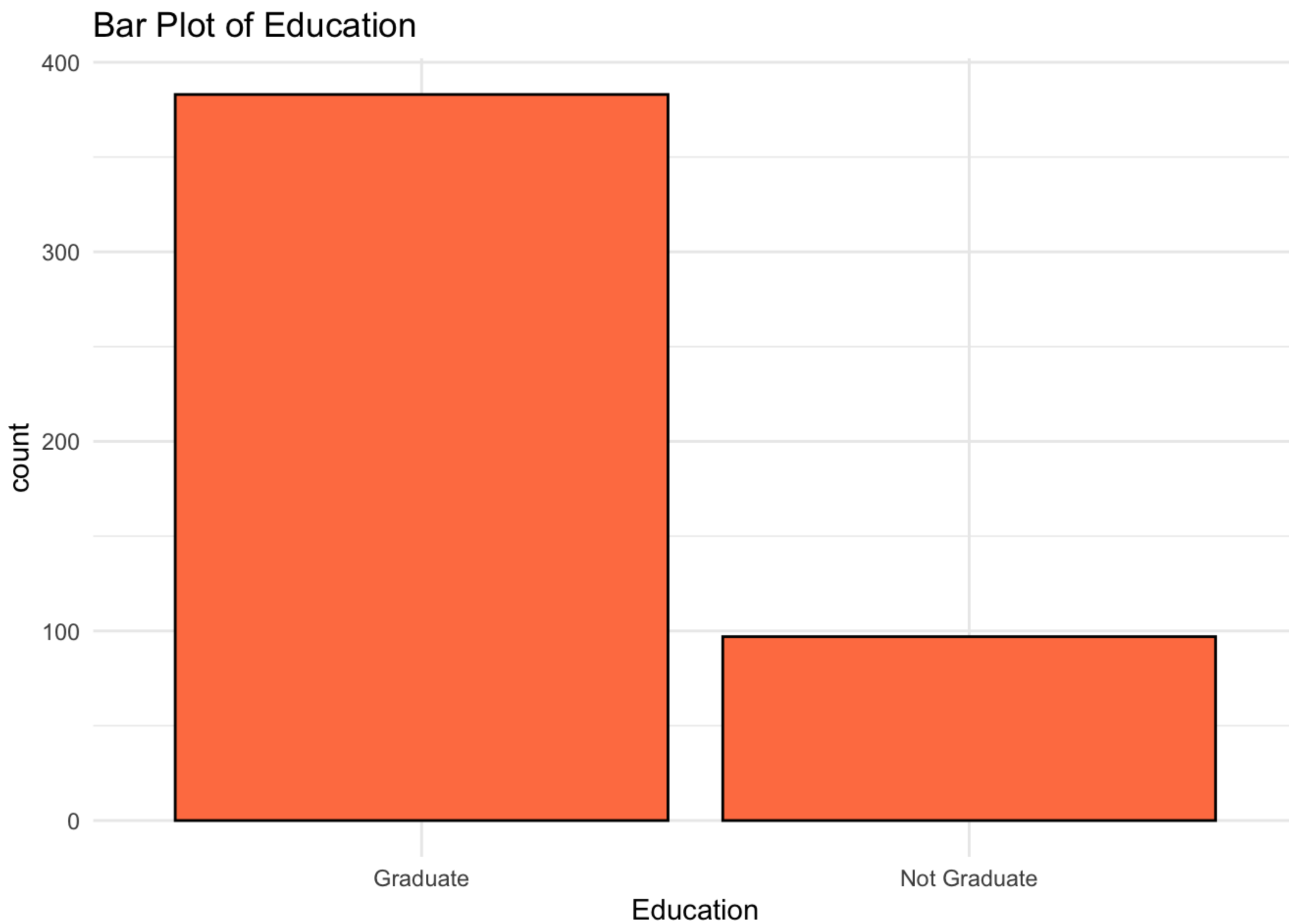
```
   Loan_ID              Gender             Married            Dependents
 Length:480         Length:480         Length:480         Length:480
 Class :character   Class :character   Class :character   Class :character
 Mode  :character   Mode  :character   Mode  :character   Mode  :character



  Education         Self_Employed      ApplicantIncome CoapplicantIncome
 Length:480         Length:480         Min.   :  150   Min.   :    0
 Class :character   Class :character   1st Qu.: 2899   1st Qu.:    0
 Mode  :character   Mode  :character   Median : 3859   Median : 1084
                                       Mean   : 5364   Mean   : 1581
                                       3rd Qu.: 5852   3rd Qu.: 2253
                                       Max.   :81000   Max.   :33837
```

```
   LoanAmount     Loan_Amount_Term Credit_History   Property_Area
 Min.   :  9.0   Min.   : 36      Min.   :0.0000   Length:480
 1st Qu.:100.0   1st Qu.:360      1st Qu.:1.0000   Class :character
 Median :128.0   Median :360      Median :1.0000   Mode  :character
 Mean   :144.7   Mean   :342      Mean   :0.8542
 3rd Qu.:170.0   3rd Qu.:360      3rd Qu.:1.0000
 Max.   :600.0   Max.   :480      Max.   :1.0000
 Loan_Status
 Length:480
 Class :character
 Mode  :character
```

```r
# Load necessary library
library(ggplot2)

# Histogram for ApplicantIncome
ggplot(df, aes(x = ApplicantIncome)) +
    geom_histogram(binwidth = 500, fill = "blue", color = "black") +
    theme_minimal() +
    ggtitle("Histogram of ApplicantIncome")
```



```
# Bar plot for Education
ggplot(df, aes(x = Education)) +
    geom_bar(fill = "coral", color = "black") +
    theme_minimal() +
    ggtitle("Bar Plot of Education")
```



ApplicantIncome: Exhibits a right-skewed distribution with most applicants earning a lower income, while a few have substantially higher incomes, indicating significant income disparity among applicants.

Education: Reveals that a large proportion of applicants are graduates, suggesting a possible correlation between higher education and the propensity to apply for loans, potentially due to educational expenses or investment in professional growth.

CoapplicantIncome: Also right-skewed, many coapplicants report low or zero income, possibly reflecting the scenario where primary applicants do not always have a secondary earner or the coapplicant earns significantly less.

LoanAmount: Shows a right-skew but with a tendency toward a normal distribution, centering on lower to mid-range loan values. This pattern might indicate a prevalence of applications for smaller loans, which are likely more frequent and have a higher approval rate.

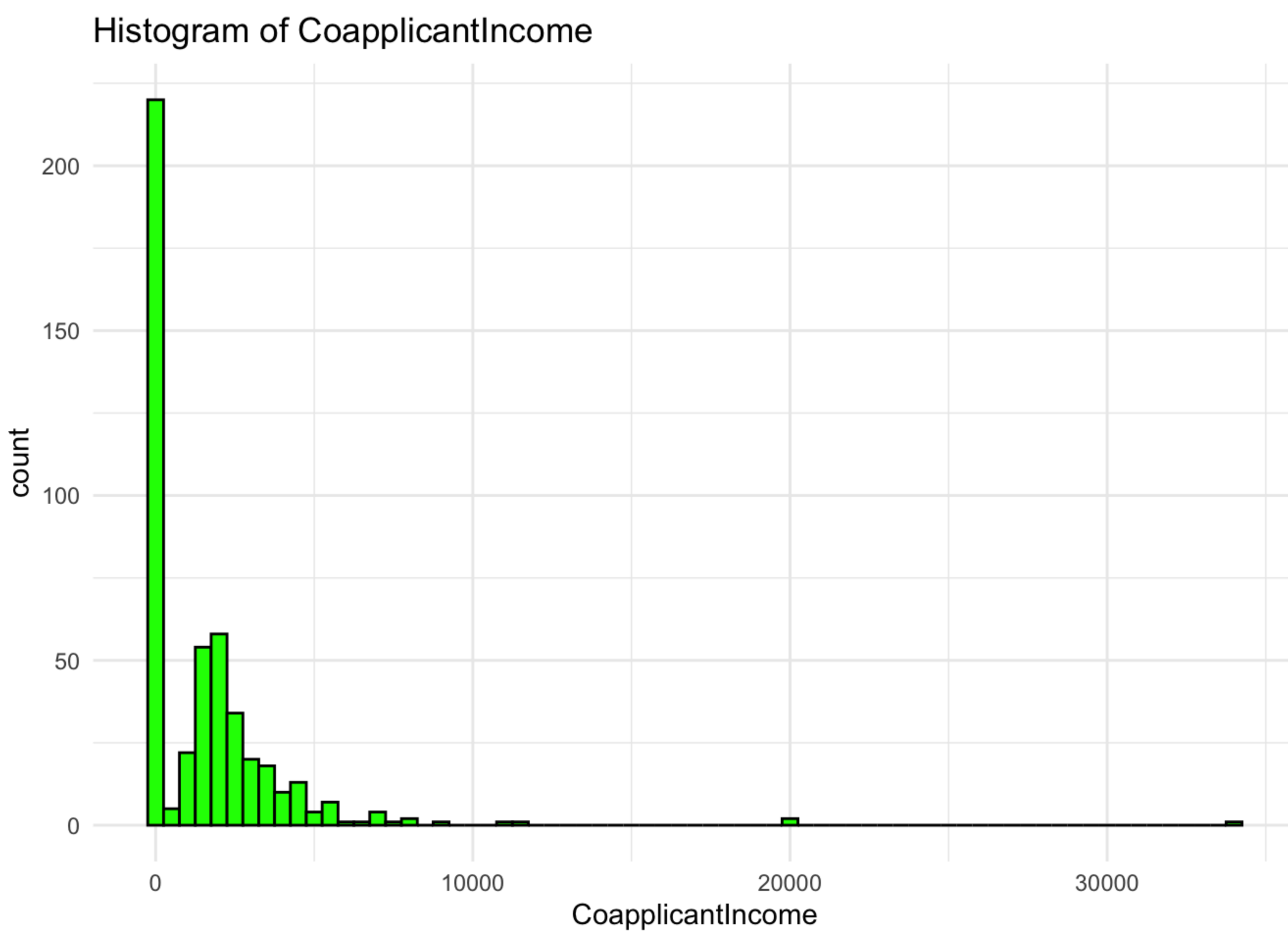
Gender: Indicates more male applicants than female, highlighting a gender gap in loan applications that warrants further exploration to understand any underlying societal or economic factors.

Married: Suggests married individuals are more likely to apply for loans, hinting at increased financial needs or joint investments that come with marital responsibilities.

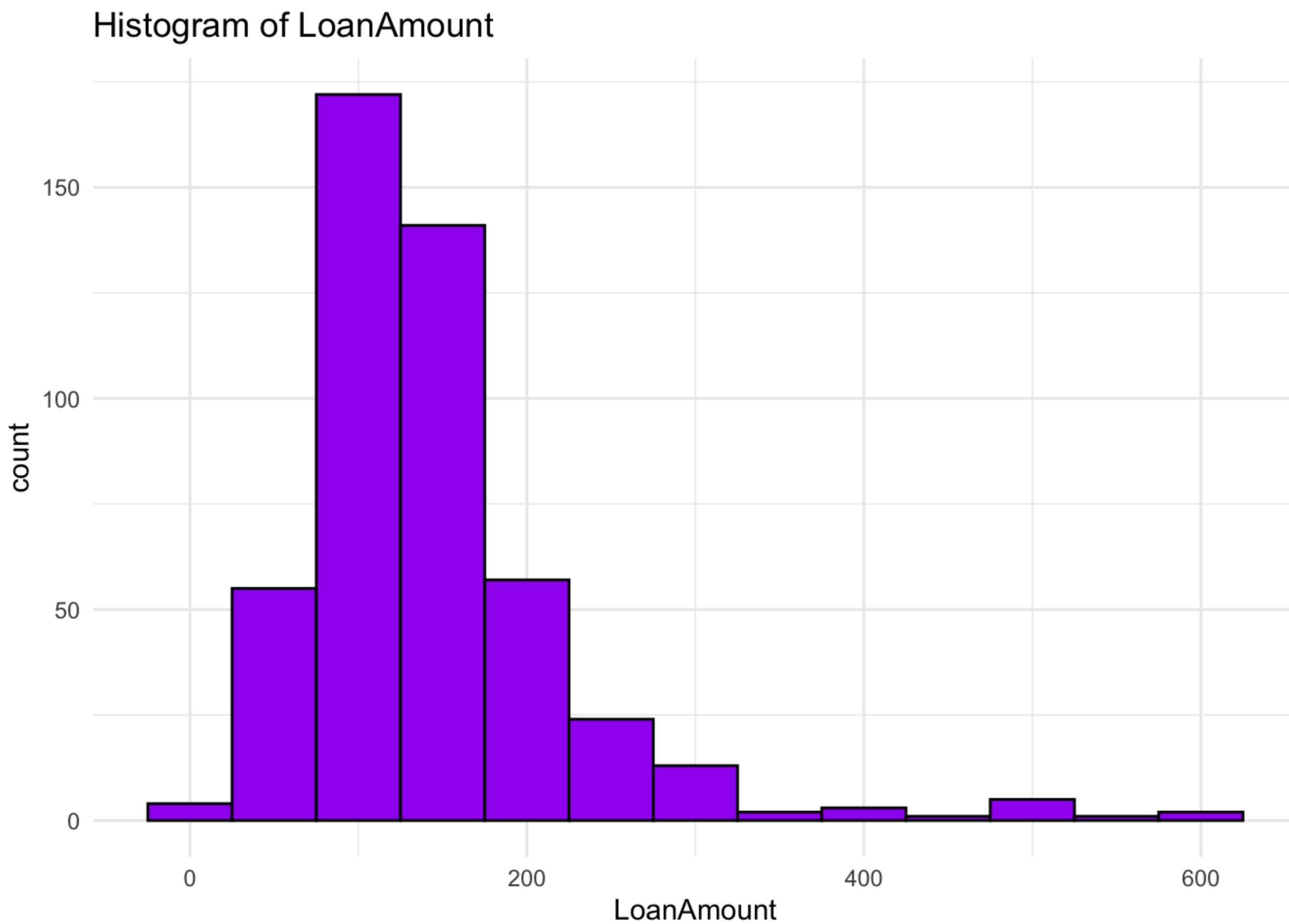
Loan_Amount_Term: Is predominantly set to 360 months, aligning with standard home loan durations.

Credit_History: The data shows most applicants have a good credit history, a key factor in loan approvals.

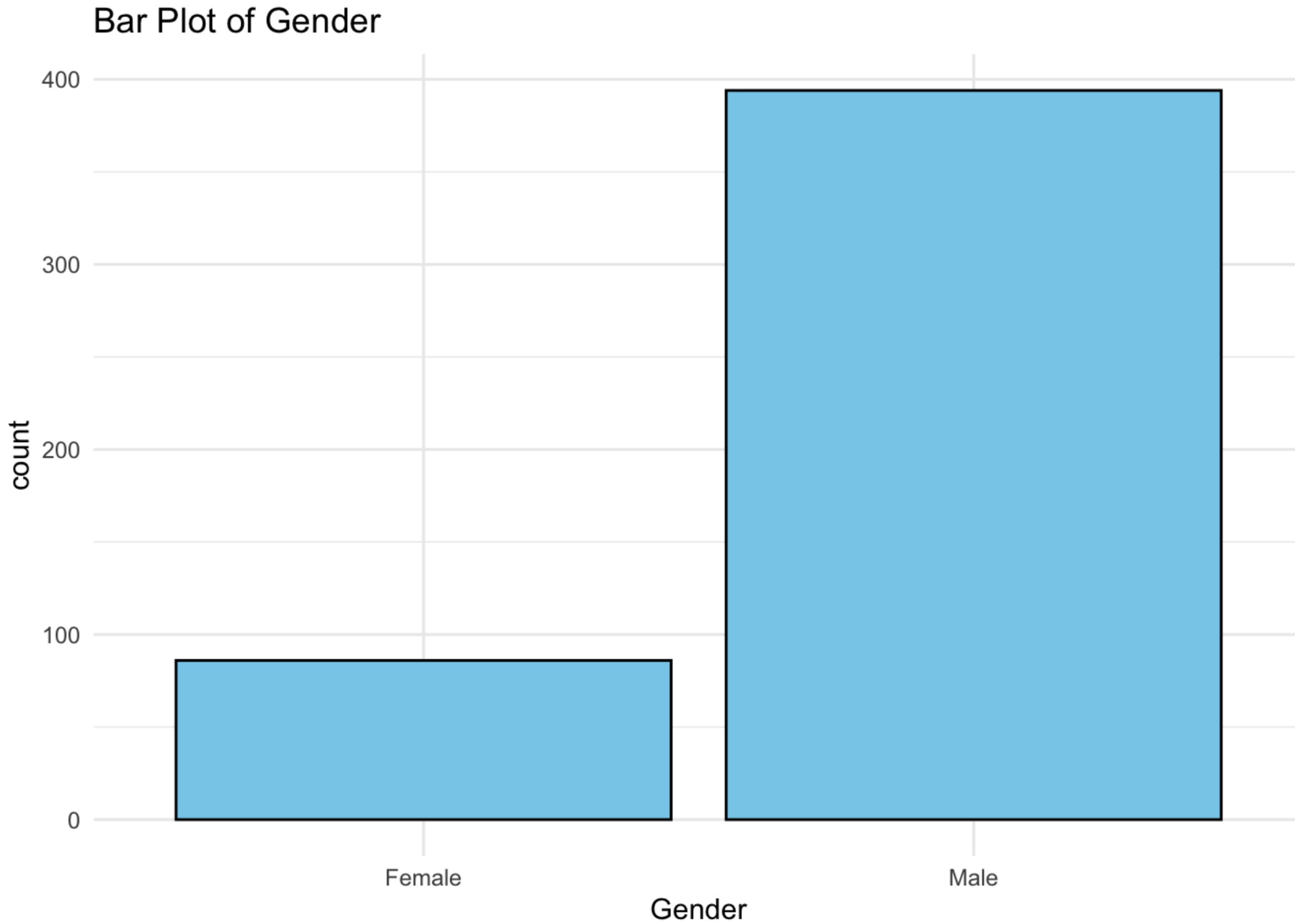
```
#Hidden
# Histogram for CoapplicantIncome
ggplot(df, aes(x = CoapplicantIncome)) +
    geom_histogram(binwidth = 500, fill = "green", color = "black") +
    theme_minimal() +
    ggtitle("Histogram of CoapplicantIncome")
```



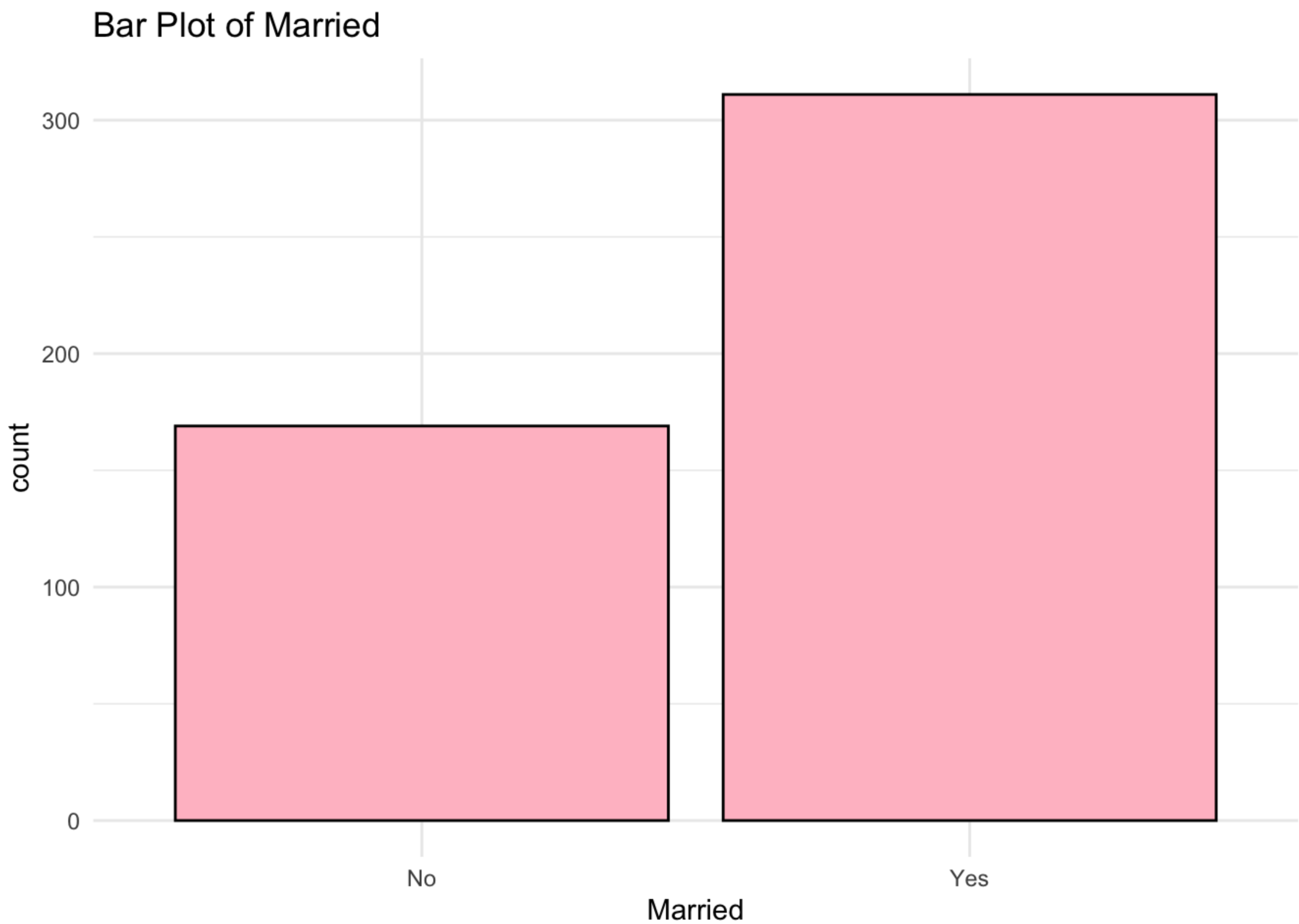

```
# Histogram for LoanAmount
ggplot(df, aes(x = LoanAmount)) +
    geom_histogram(binwidth = 50, fill = "purple", color = "black") +
    theme_minimal() +
    ggtitle("Histogram of LoanAmount")
```



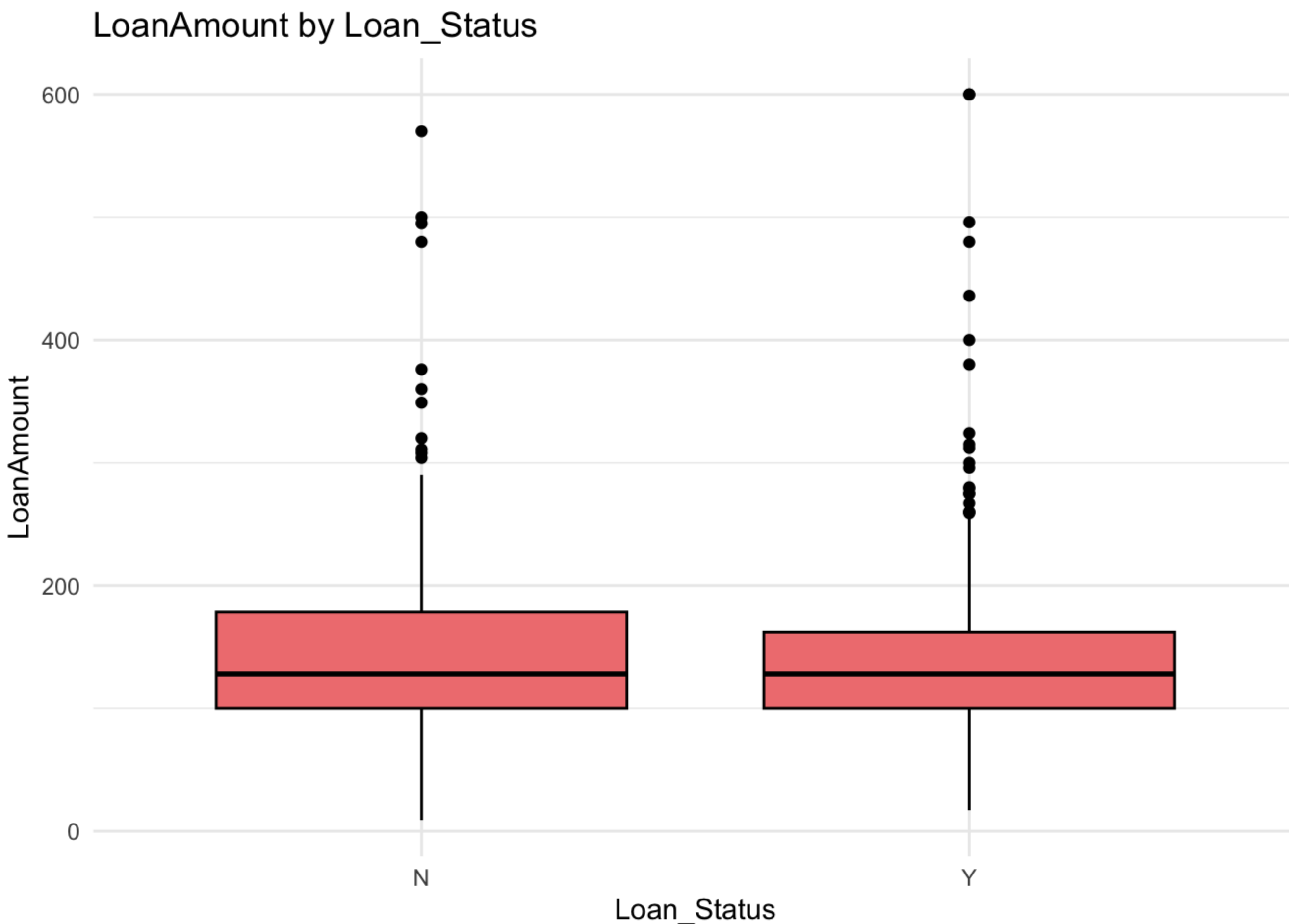

```
# Bar plot for Gender
ggplot(df, aes(x = Gender)) +
    geom_bar(fill = "skyblue", color = "black") +
    theme_minimal() +
    ggtitle("Bar Plot of Gender")
```



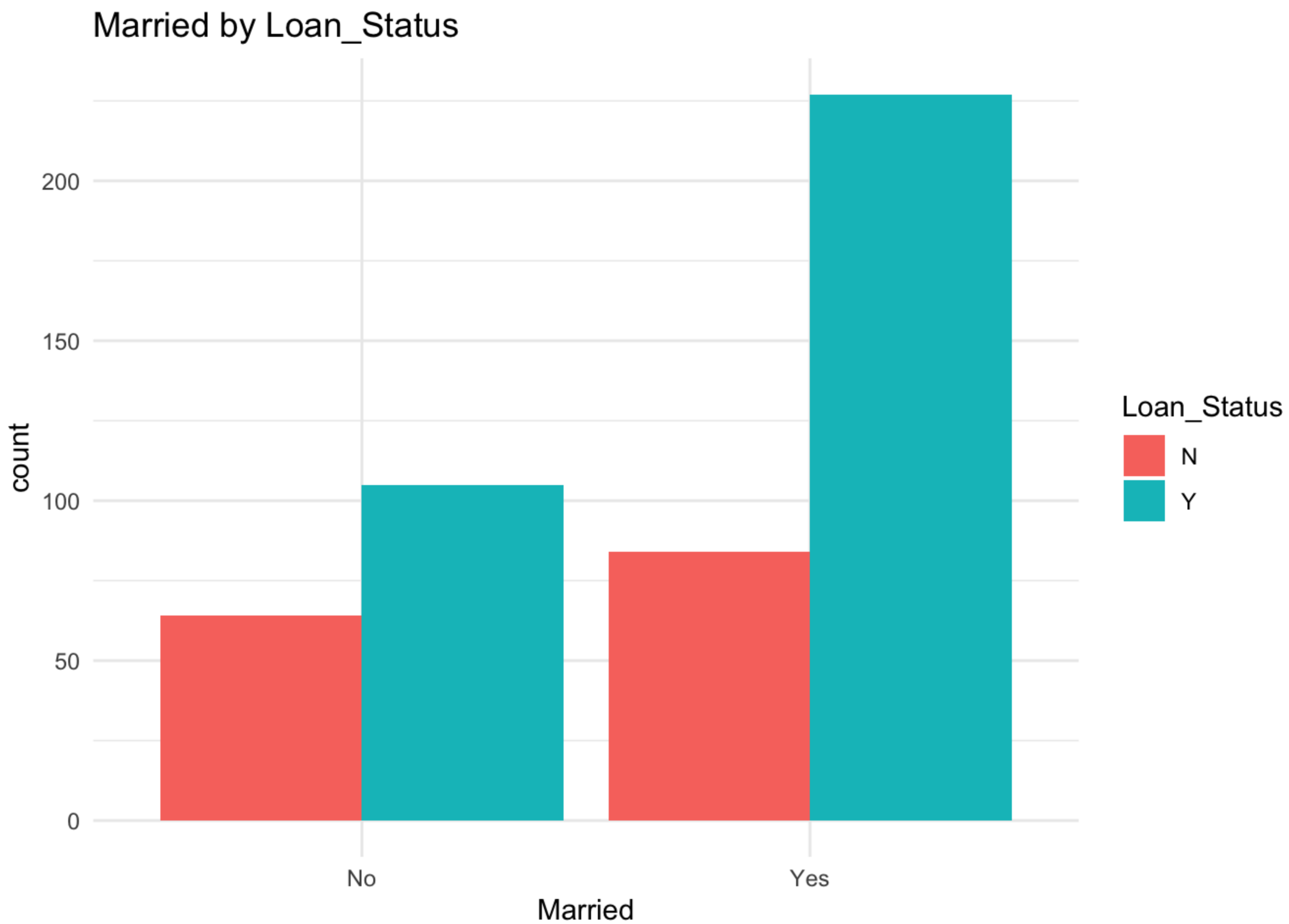

```
# Bar plot for Married
ggplot(df, aes(x = Married)) +
    geom_bar(fill = "pink", color = "black") +
    theme_minimal() +
    ggtitle("Bar Plot of Married")
```




```r
#bivariate
# Boxplot for LoanAmount by Loan_Status
ggplot(df, aes(x = Loan_Status, y = LoanAmount)) +
    geom_boxplot(fill = "lightcoral", color = "black") +
    theme_minimal() +
    ggtitle("LoanAmount by Loan_Status")
```

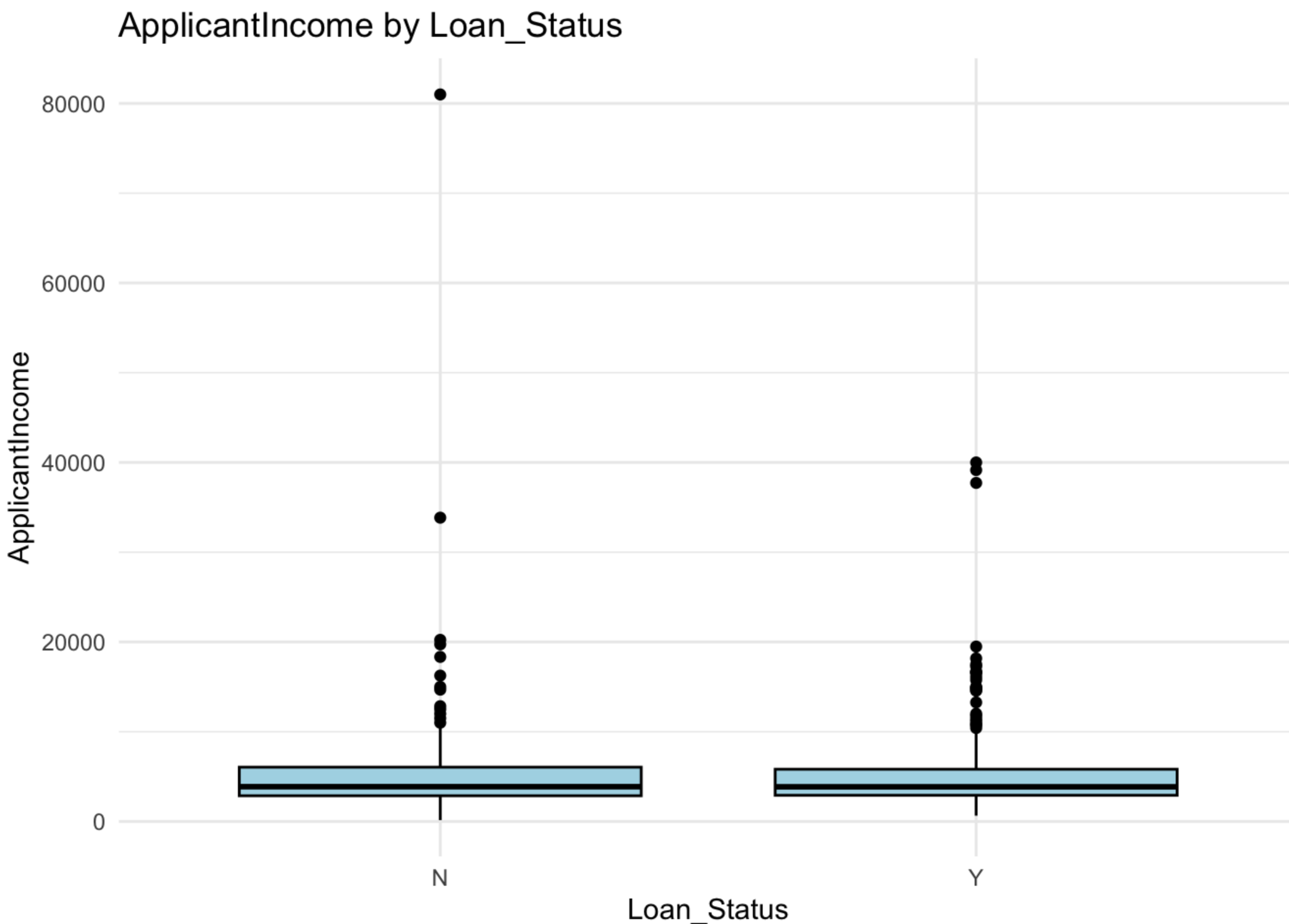



```
# Side-by-side Bar plot for Married by Loan_Status
ggplot(df, aes(x = Married, fill = Loan_Status)) +
    geom_bar(position = "dodge") +
    theme_minimal() +
    ggtitle("Married by Loan_Status")
```

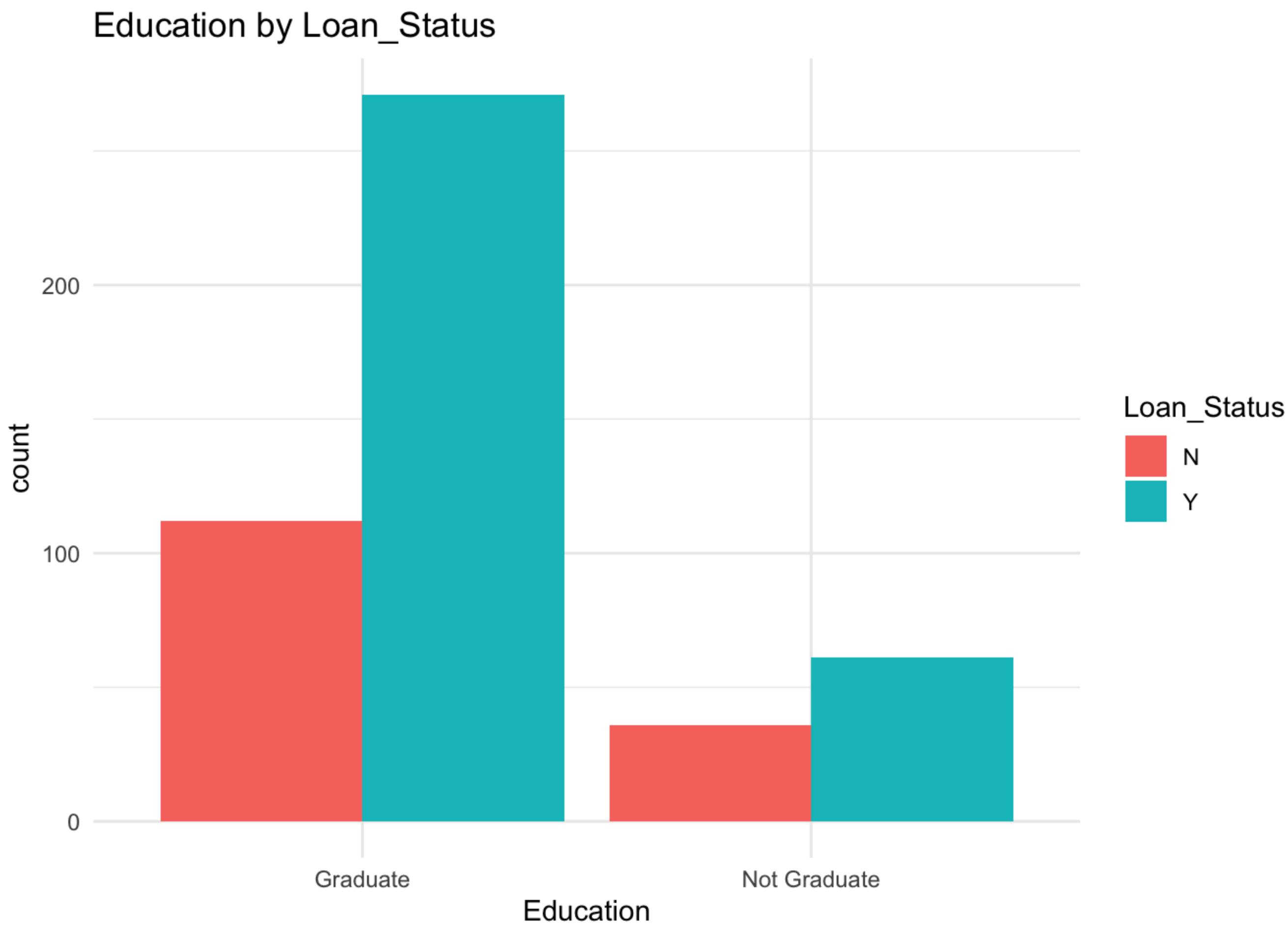


- **ApplicantIncome and CoapplicantIncome**: The income levels of applicants and coapplicants, when assessed by loan status, show significant variability and the presence of high-income outliers. Notably, higher incomes do not guarantee loan approval, suggesting that other factors are at play in the decision-making process.
- **Education**: Graduates are more likely to apply for loans, and the data shows a higher number of loans processed for this group. However, the approval rate does not disproportionately favor graduates, implying that educational attainment is not the sole determinant of loan success.
- **LoanAmount**: The amounts requested are broadly similar across approved and not approved loans, with a wider distribution for approved loans. This indicates that loan amount is considered within a broader context of the applicant’s profile.
- **Gender and Marital Status**: There is a clear trend showing more men and married individuals among loan applicants, with these groups also receiving more approvals. This could reflect social and economic dynamics that influence loan application patterns and approval rates.

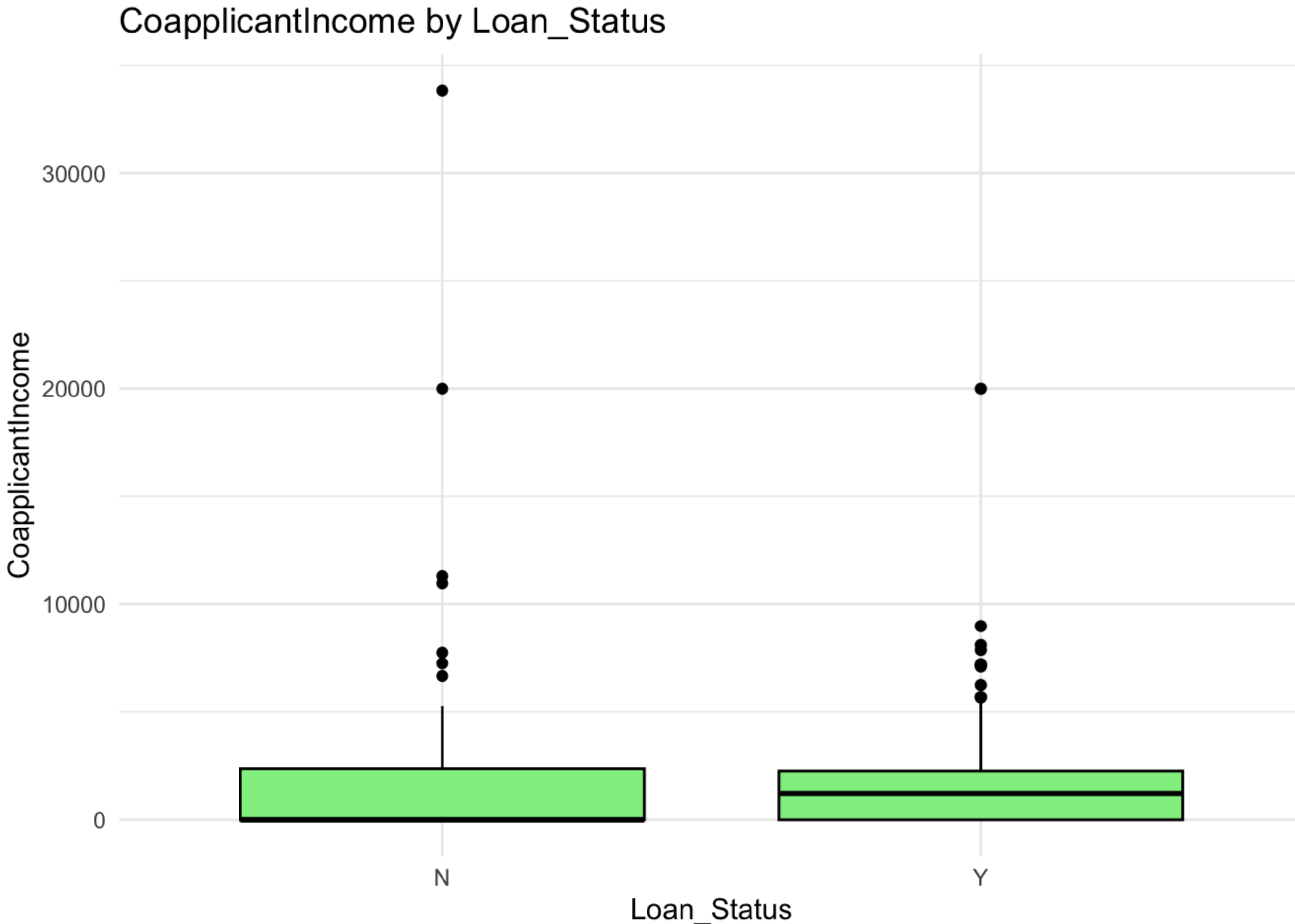
```
#Hidden
# Boxplot for ApplicantIncome by Loan_Status
ggplot(df, aes(x = Loan_Status, y = ApplicantIncome)) +
    geom_boxplot(fill = "lightblue", color = "black") +
    theme_minimal() +
    ggtitle("ApplicantIncome by Loan_Status")
```



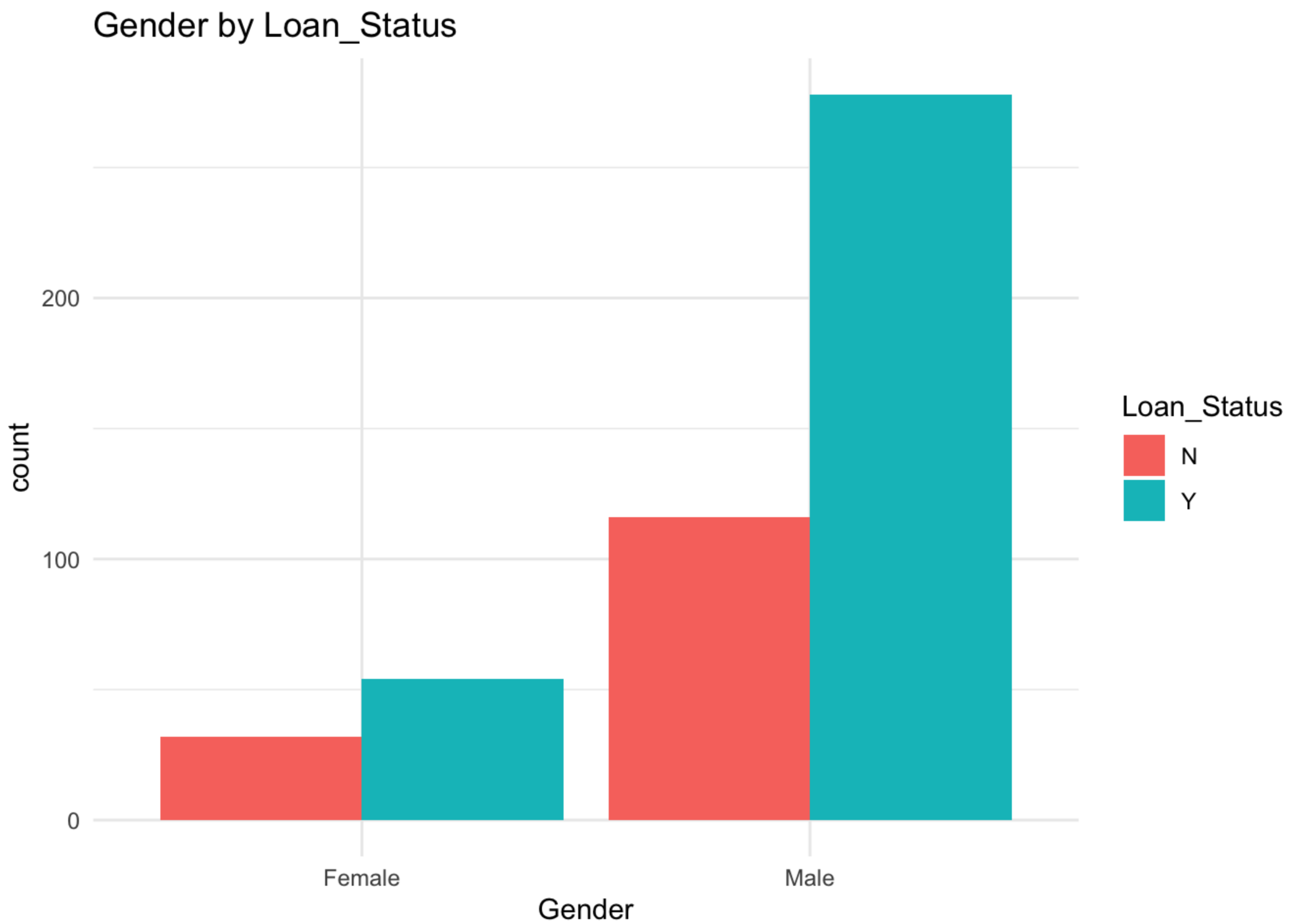

```
# Side-by-side Bar plot for Education by Loan_Status
ggplot(df, aes(x = Education, fill = Loan_Status)) +
    geom_bar(position = "dodge") +
    theme_minimal() +
    ggtitle("Education by Loan_Status")
```




```
# Boxplot for CoapplicantIncome by Loan_Status
ggplot(df, aes(x = Loan_Status, y = CoapplicantIncome)) +
    geom_boxplot(fill = "lightgreen", color = "black") +
    theme_minimal() +
    ggtitle("CoapplicantIncome by Loan_Status")
```

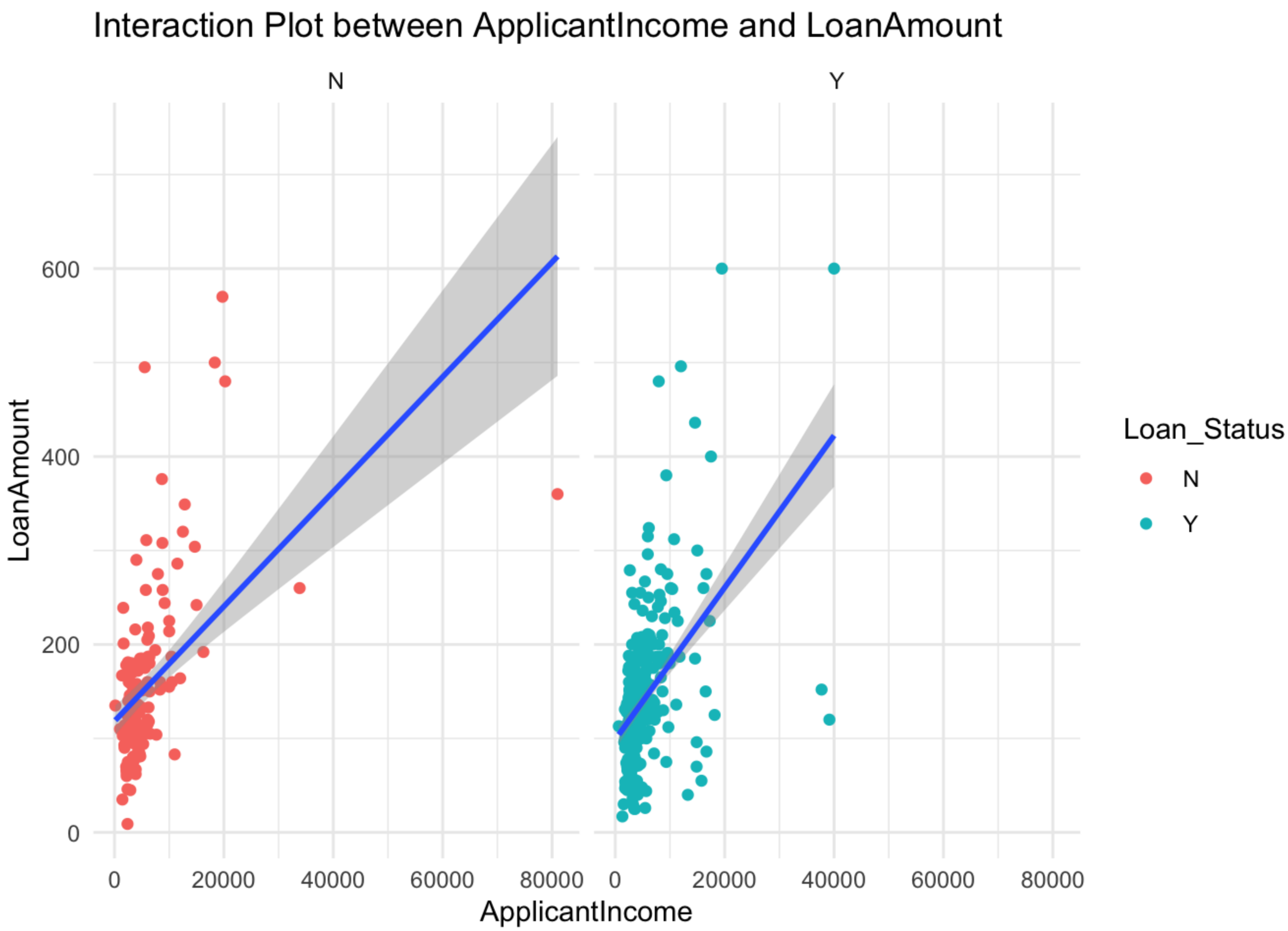



```
# Side-by-side Bar plot for Gender by Loan_Status
ggplot(df, aes(x = Gender, fill = Loan_Status)) +
    geom_bar(position = "dodge") +
    theme_minimal() +
    ggtitle("Gender by Loan_Status")
```

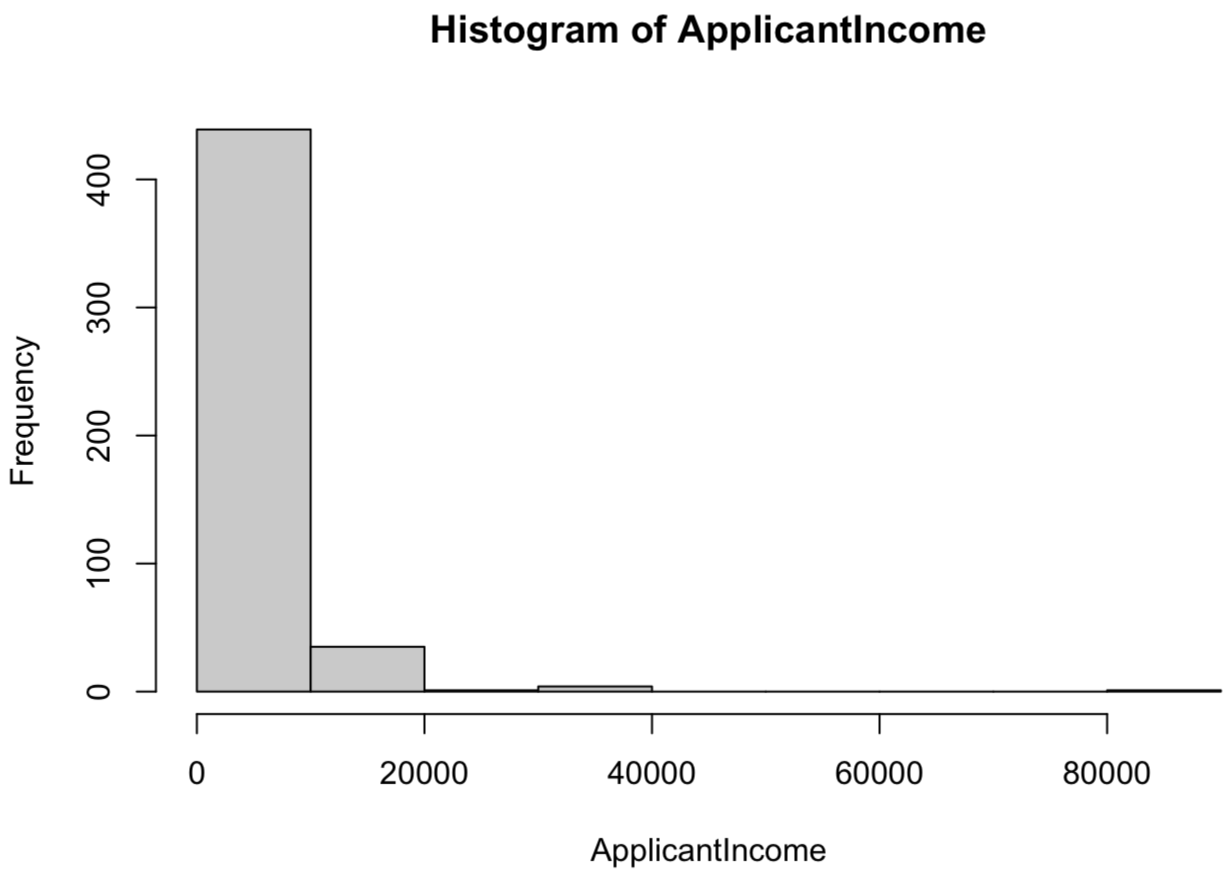



```r
#Hidden
# Load the necessary library
library(ggplot2)

# Let's say you want to examine the interaction between 'ApplicantIncome' and 'LoanAmount
# Create an interaction plot
ggplot(df, aes(x = ApplicantIncome, y = LoanAmount)) +
  geom_point(aes(color = Loan_Status)) +  # Use color to differentiate loan status
  geom_smooth(method = "lm") +  # Add a regression line
  facet_wrap(~ Loan_Status) +   # Create separate plots by loan status
  labs(title = "Interaction Plot between ApplicantIncome and LoanAmount") +
  theme_minimal()
```

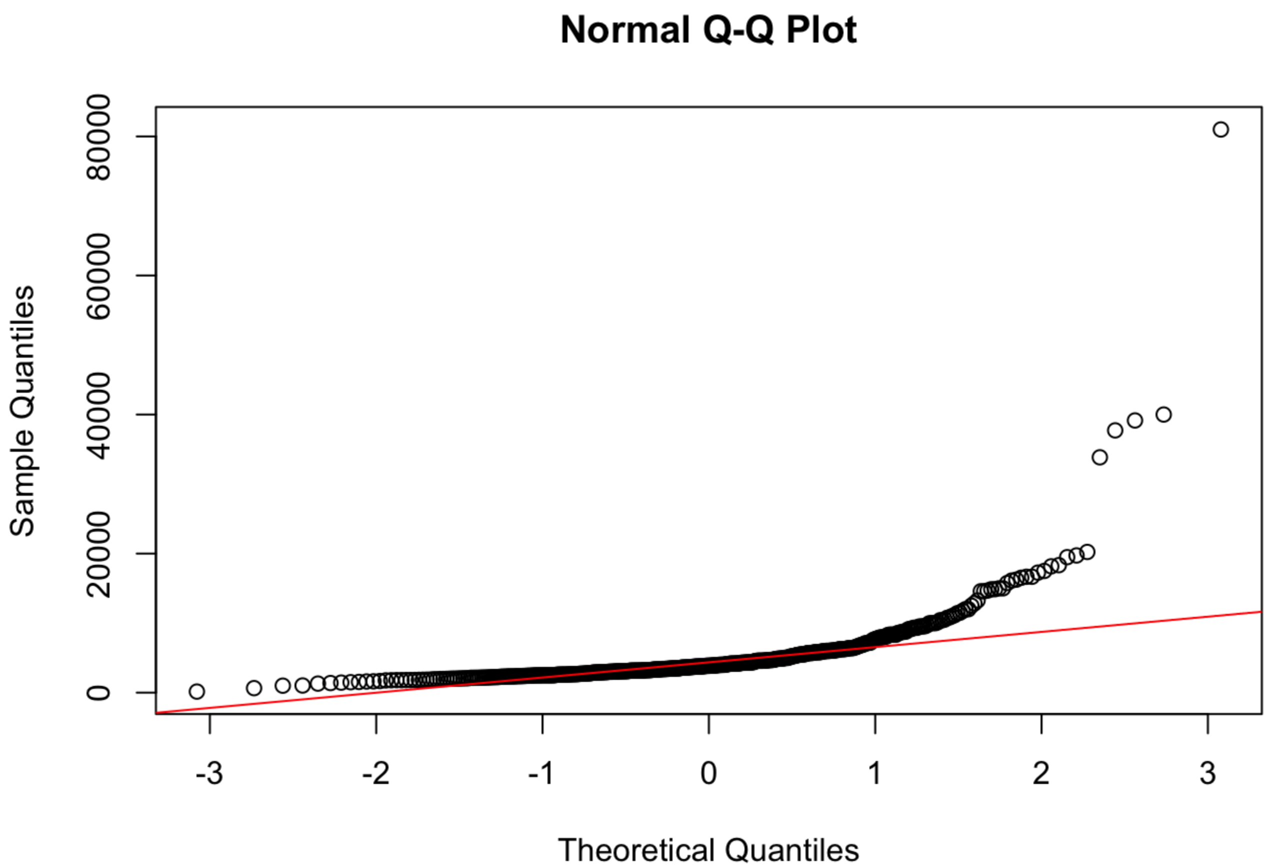
```
`geom_smooth()` using formula = 'y ~ x'
```




```
# Histogram to see the distribution
hist(df$ApplicantIncome, main = "Histogram of ApplicantIncome", xlab = "ApplicantIncome")
```




```
# Q-Q plot to check for normality
qqnorm(df$ApplicantIncome)
qqline(df$ApplicantIncome, col = "red")
```

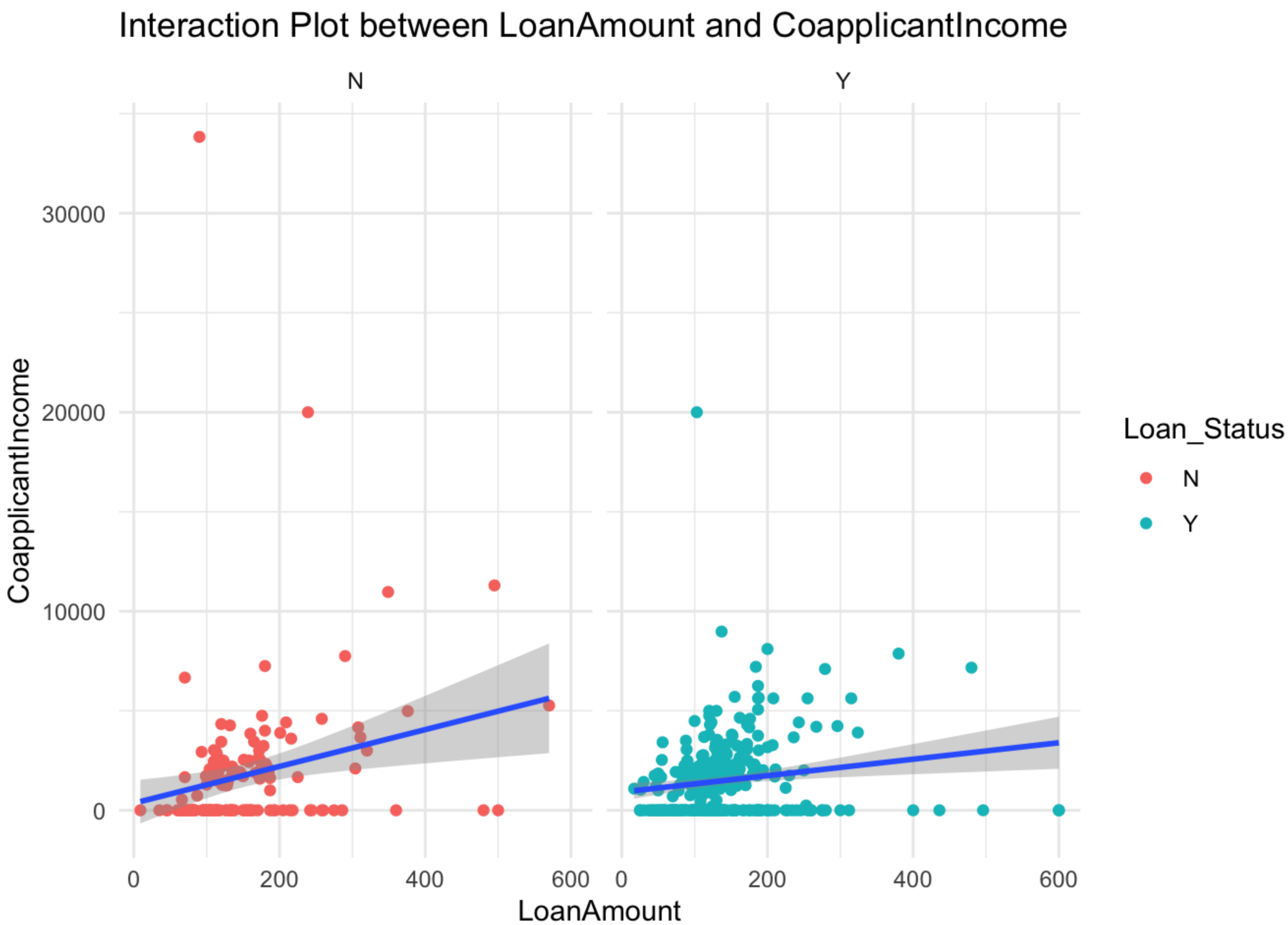



```
# Shapiro-Wilk normality test
shapiro.test(df$ApplicantIncome)
```

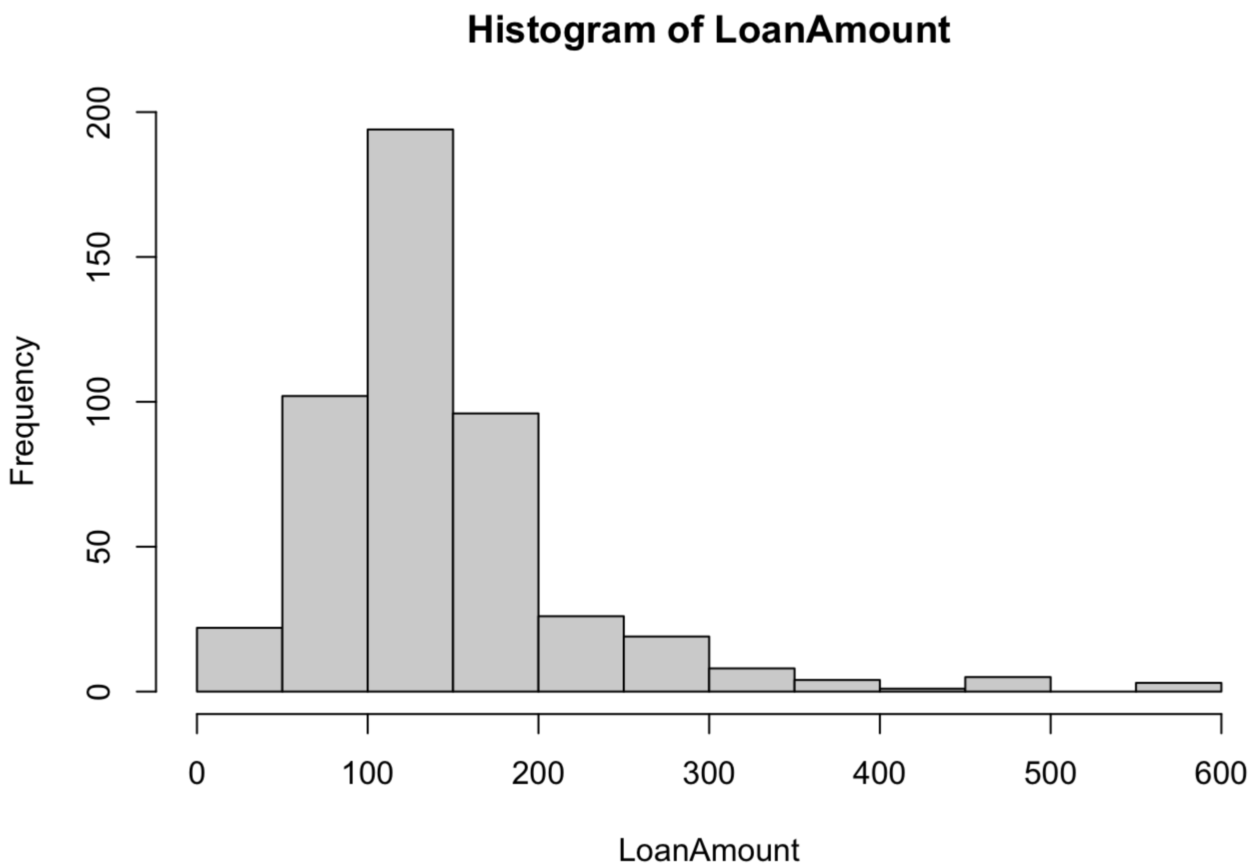
```
    Shapiro-Wilk normality test

data:  df$ApplicantIncome
W = 0.49311, p-value < 2.2e-16
```

```
# Interaction plot with another continuous variable 'CoapplicantIncome'
ggplot(df, aes(x = LoanAmount, y = CoapplicantIncome)) +
  geom_point(aes(color = Loan_Status)) +  # Use color to differentiate loan status
  geom_smooth(method = "lm") +  # Add a regression line
  facet_wrap(~ Loan_Status) +   # Create separate plots by loan status
  labs(title = "Interaction Plot between LoanAmount and CoapplicantIncome") +
  theme_minimal()
```

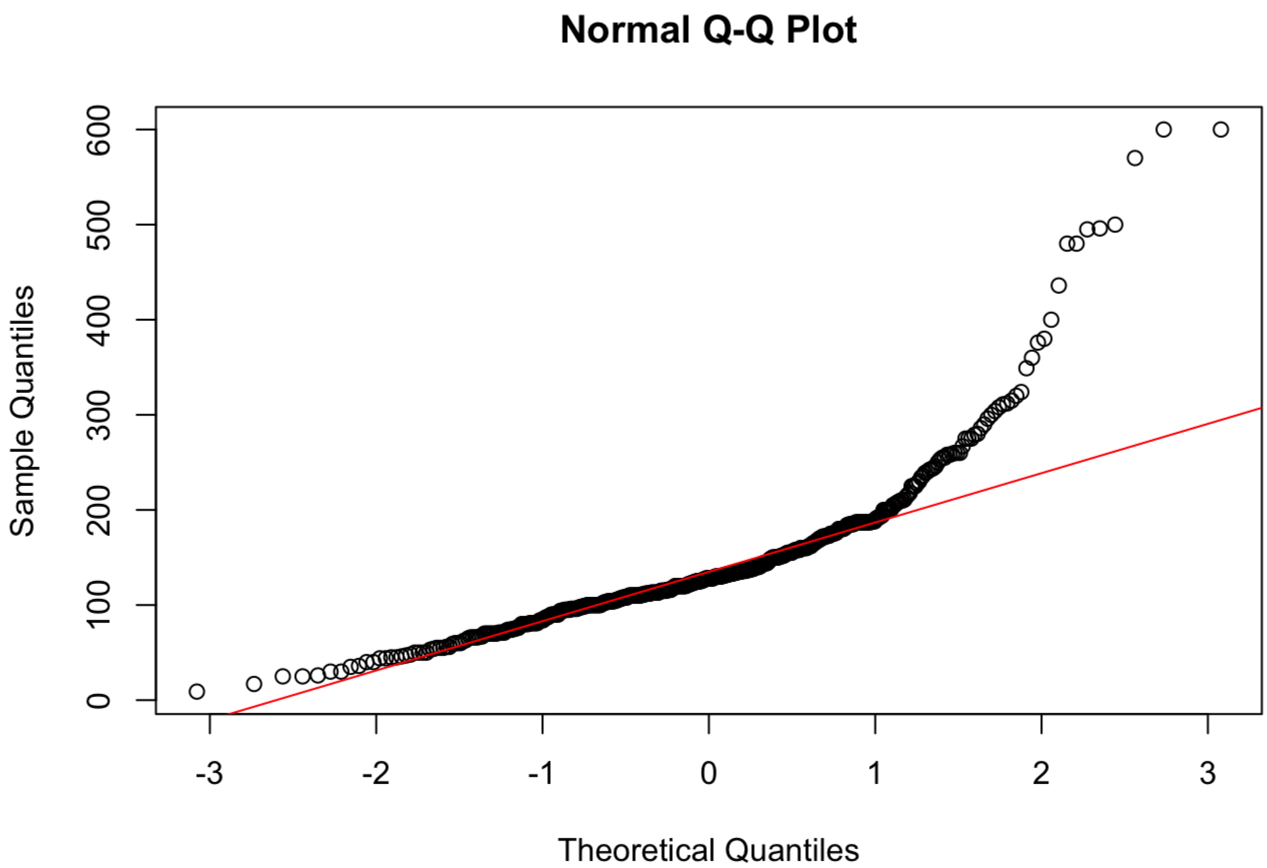
```
`geom_smooth()` using formula = 'y ~ x'
```




```
# Histogram for LoanAmount
hist(df$LoanAmount, main = "Histogram of LoanAmount", xlab = "LoanAmount")
```




```
# Q-Q plot for LoanAmount
qqnorm(df$LoanAmount)
qqline(df$LoanAmount, col = "red")
```

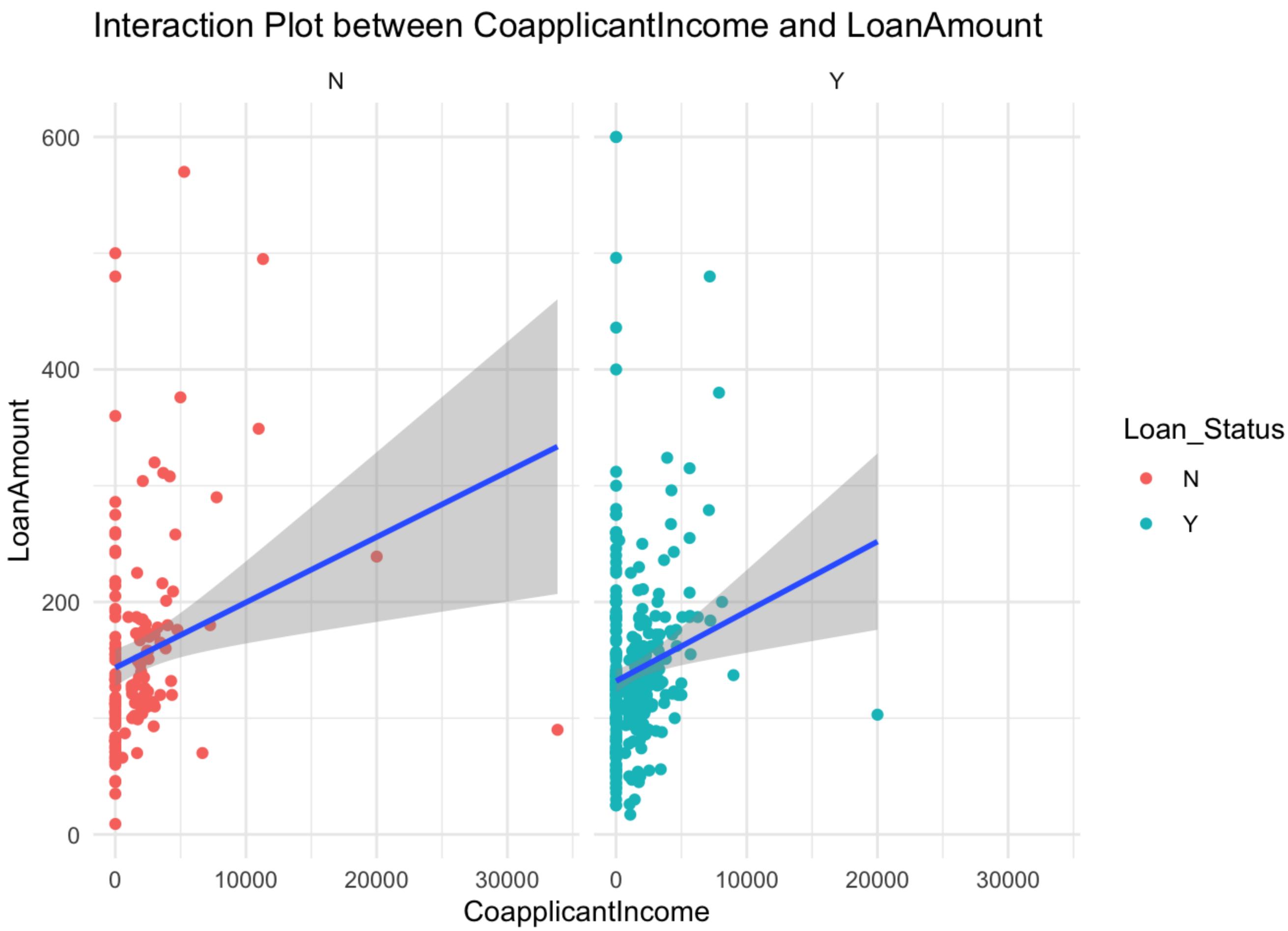



```
# Shapiro-Wilk test for LoanAmount
shapiro.test(df$LoanAmount)
```

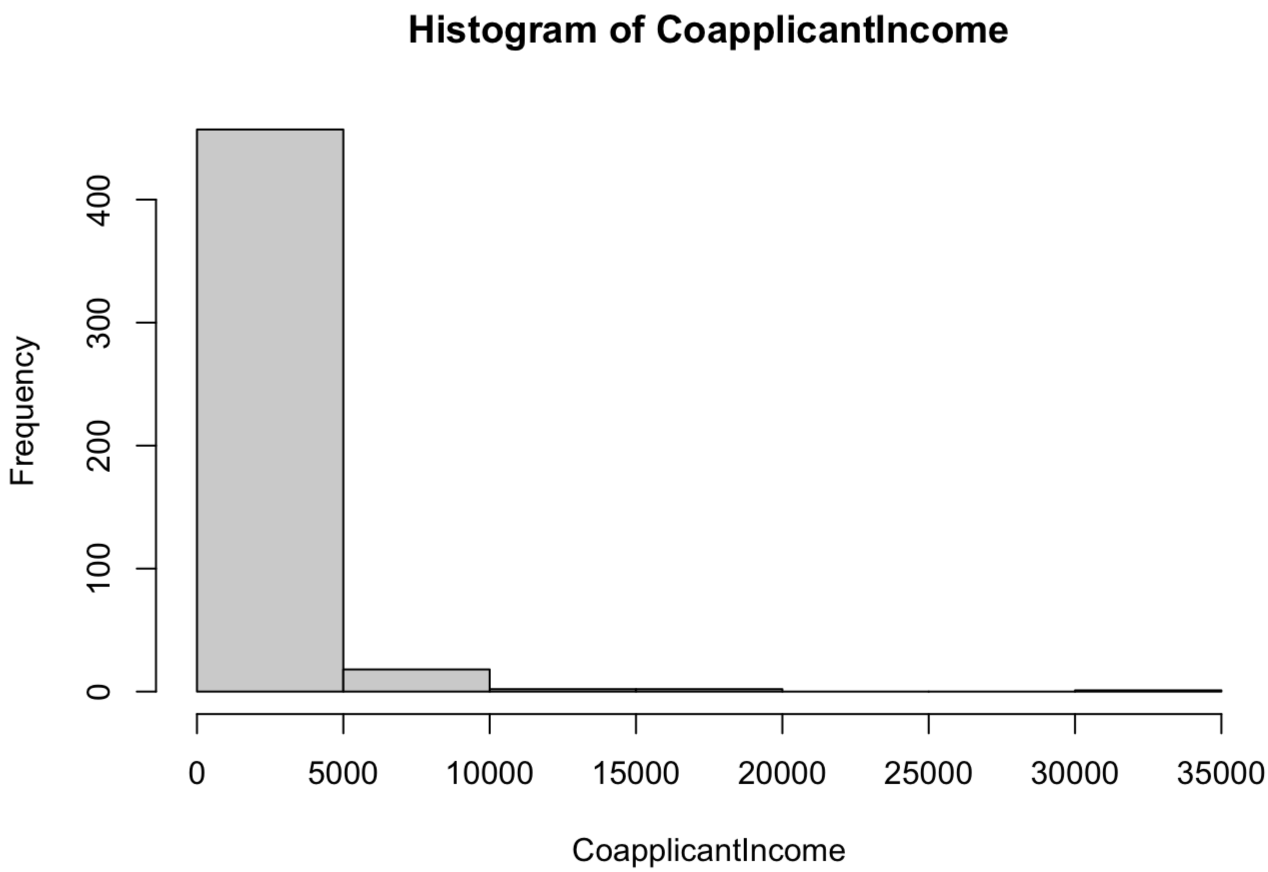
```
	Shapiro-Wilk normality test

data:  df$LoanAmount
W = 0.80741, p-value < 2.2e-16
```

```
# Interaction plot with 'LoanAmount'
ggplot(df, aes(x = CoapplicantIncome, y = LoanAmount)) +
  geom_point(aes(color = Loan_Status)) +  # Use color to differentiate loan status
  geom_smooth(method = "lm") +  # Add a regression line
  facet_wrap(~ Loan_Status) +   # Create separate plots by loan status
  labs(title = "Interaction Plot between CoapplicantIncome and LoanAmount") +
  theme_minimal()
```

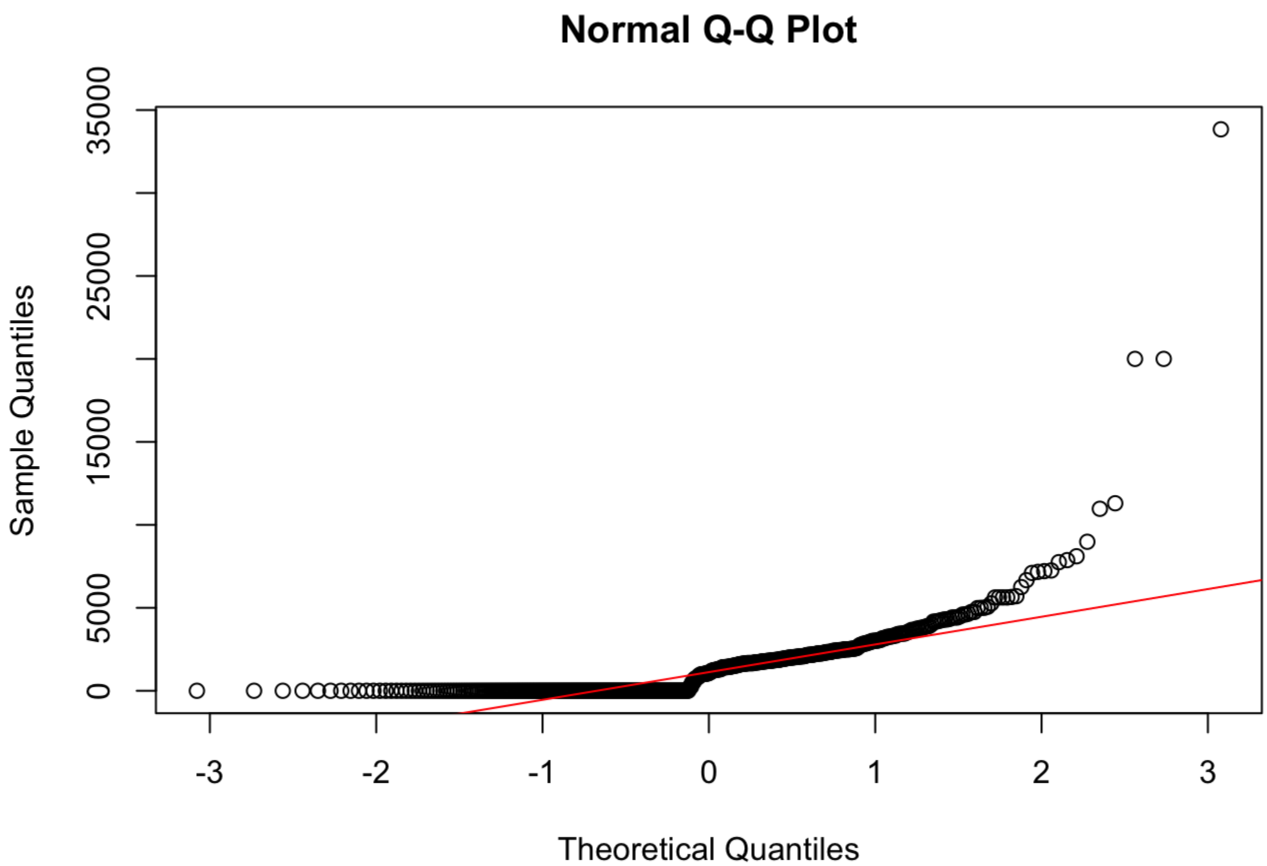
```
`geom_smooth()` using formula = 'y ~ x'
```




```r
# Histogram for CoapplicantIncome
hist(df$CoapplicantIncome, main = "Histogram of CoapplicantIncome", xlab = "CoapplicantIn
```




```
# Q-Q plot for CoapplicantIncome
qqnorm(df$CoapplicantIncome)
qqline(df$CoapplicantIncome, col = "red")
```




```
# Shapiro-Wilk test for CoapplicantIncome
shapiro.test(df$CoapplicantIncome)
```

```
    Shapiro-Wilk normality test

data:  df$CoapplicantIncome
W = 0.55589, p-value < 2.2e-16
```

The interaction plots from the loan dataset show a positive relationship between income and loan amount, with higher incomes linked to larger loan requests for both applicants and coapplicants. This pattern is consistent across both approved and denied loan statuses, suggesting that while income plays a role in loan amount determination, it is not the sole factor in loan approval decisions. The plots also reveal a wide spread of data and outliers, indicating varied loan behaviors among applicants.

Shapiro-Wilk normality tests for ApplicantIncome, LoanAmount, and CoapplicantIncome indicate significant deviations from a normal distribution, with p-values far below the threshold of 0.05. The corresponding Q-Q plots confirm this non-normality, displaying a right-skewed distribution with a bulk of values on the lower end and fewer high values. These findings suggest that income data is not normally distributed, pointing towards the necessity for non-linear modeling or data transformation in further statistical analysis.

```r
#Hidden
#log
# Replace zeros with a small positive value if necessary
df$ApplicantIncome[df$ApplicantIncome <= 0] <- 1
df$CoapplicantIncome[df$CoapplicantIncome <= 0] <- 1
df$LoanAmount[df$LoanAmount <= 0] <- 1

# Apply log transformation
df$Log_ApplicantIncome <- log(df$ApplicantIncome)
df$Log_CoapplicantIncome <- log(df$CoapplicantIncome)
df$Log_LoanAmount <- log(df$LoanAmount)


# Shapiro-Wilk normality test
shapiro.test(df$Log_ApplicantIncome)
```

```
	Shapiro-Wilk normality test

data:  df$Log_ApplicantIncome
W = 0.94524, p-value = 2.553e-12
```

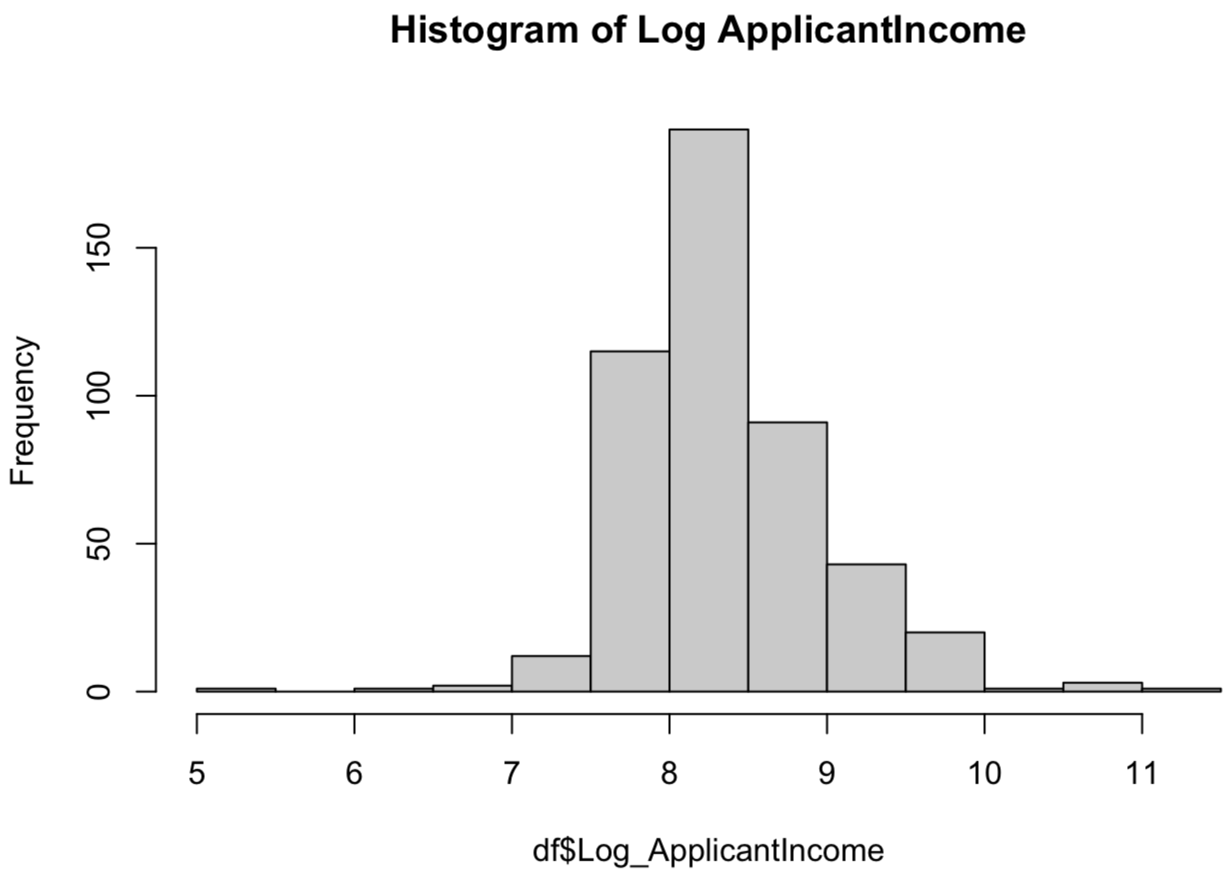
```r
shapiro.test(df$Log_CoapplicantIncome)
```

```
	Shapiro-Wilk normality test

data:  df$Log_CoapplicantIncome
W = 0.71264, p-value < 2.2e-16
```

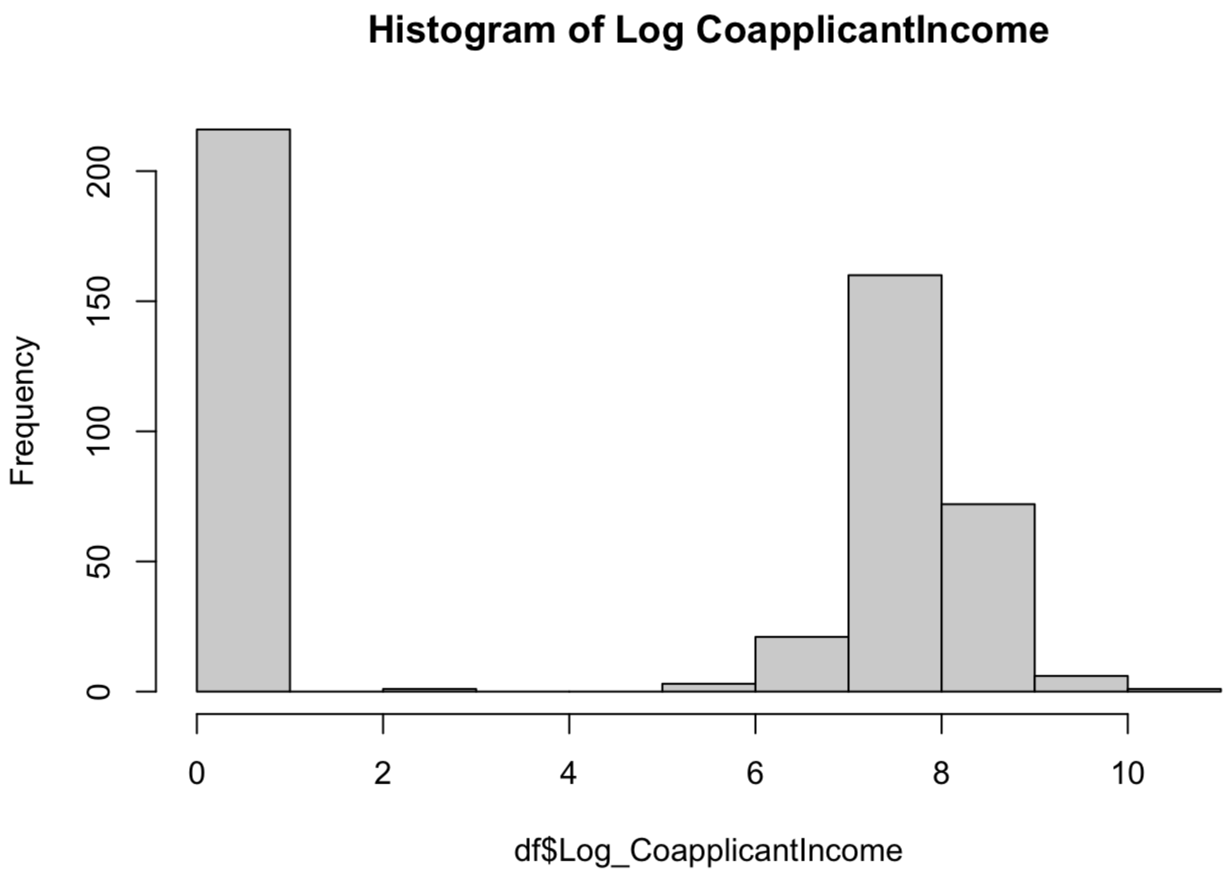
```r
shapiro.test(df$Log_LoanAmount)
```

```
	Shapiro-Wilk normality test

data:  df$Log_LoanAmount
W = 0.9635, p-value = 1.52e-09
```

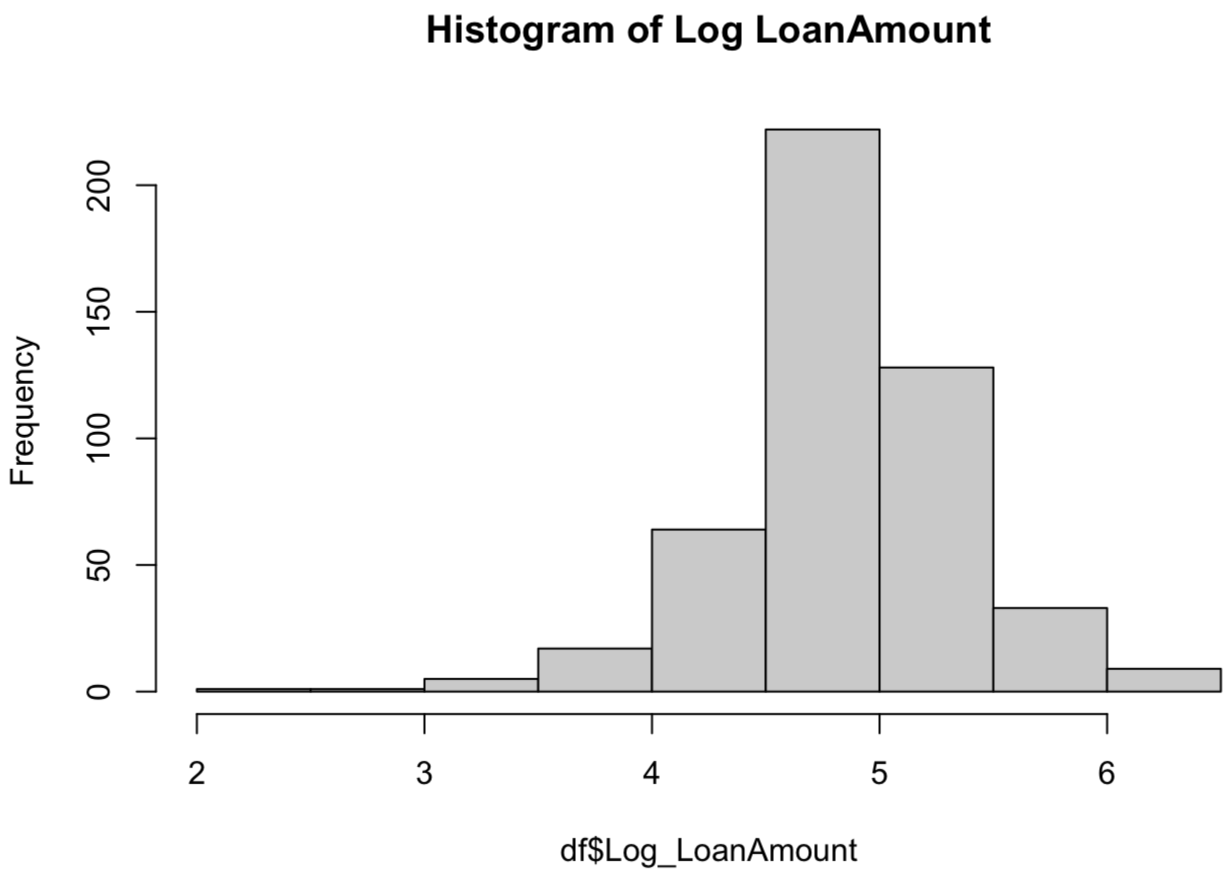
```r
# Histograms
hist(df$Log_ApplicantIncome, main="Histogram of Log ApplicantIncome")
```



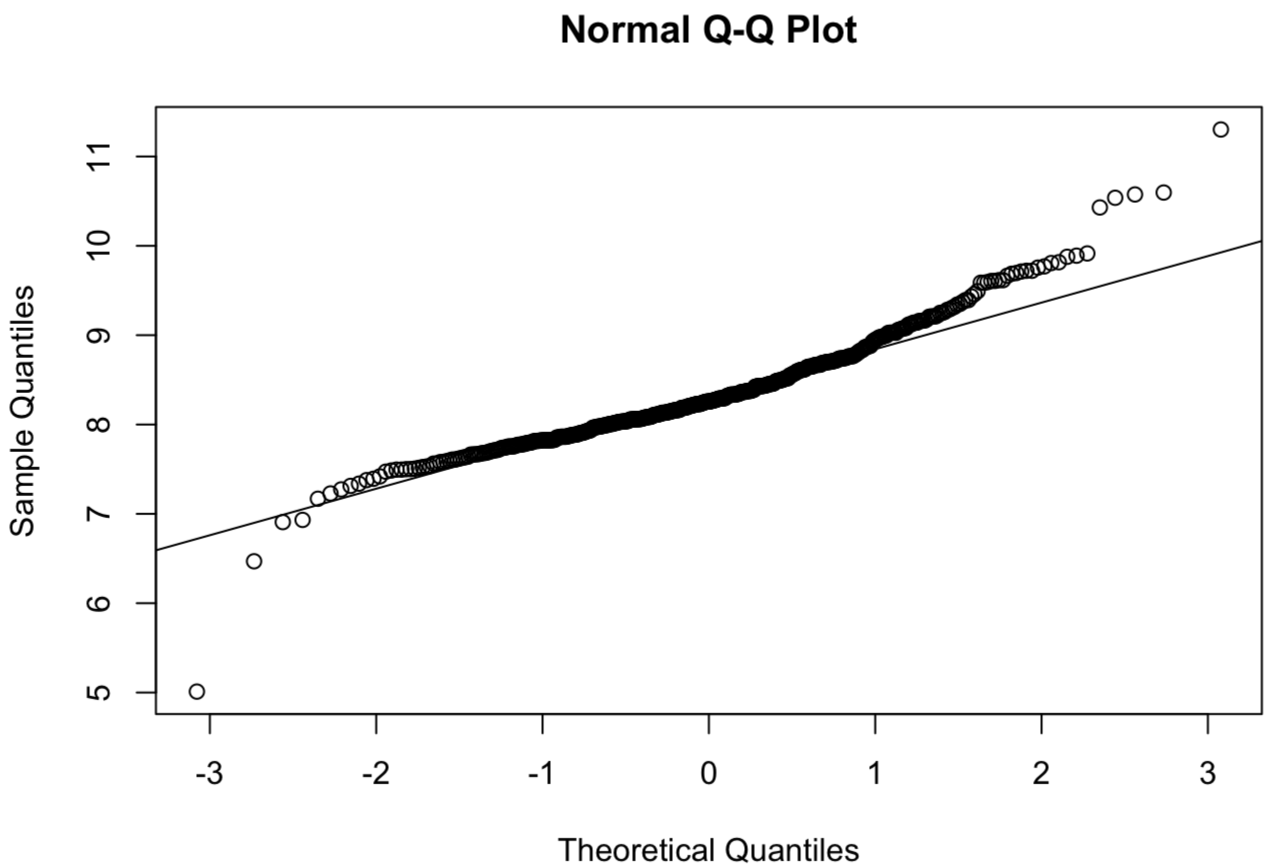

```
hist(df$Log_CoapplicantIncome, main="Histogram of Log CoapplicantIncome")
```



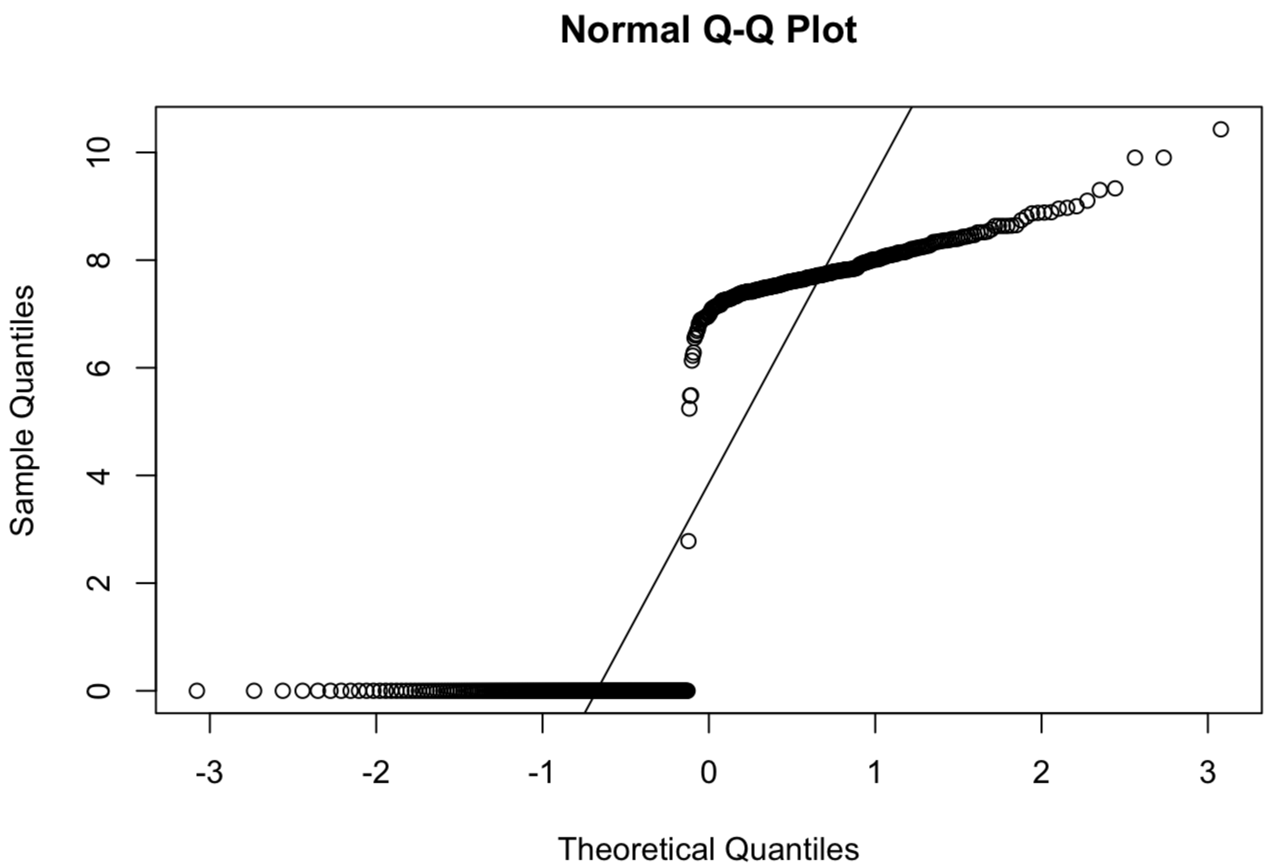

```r
hist(df$Log_LoanAmount, main="Histogram of Log LoanAmount")
```



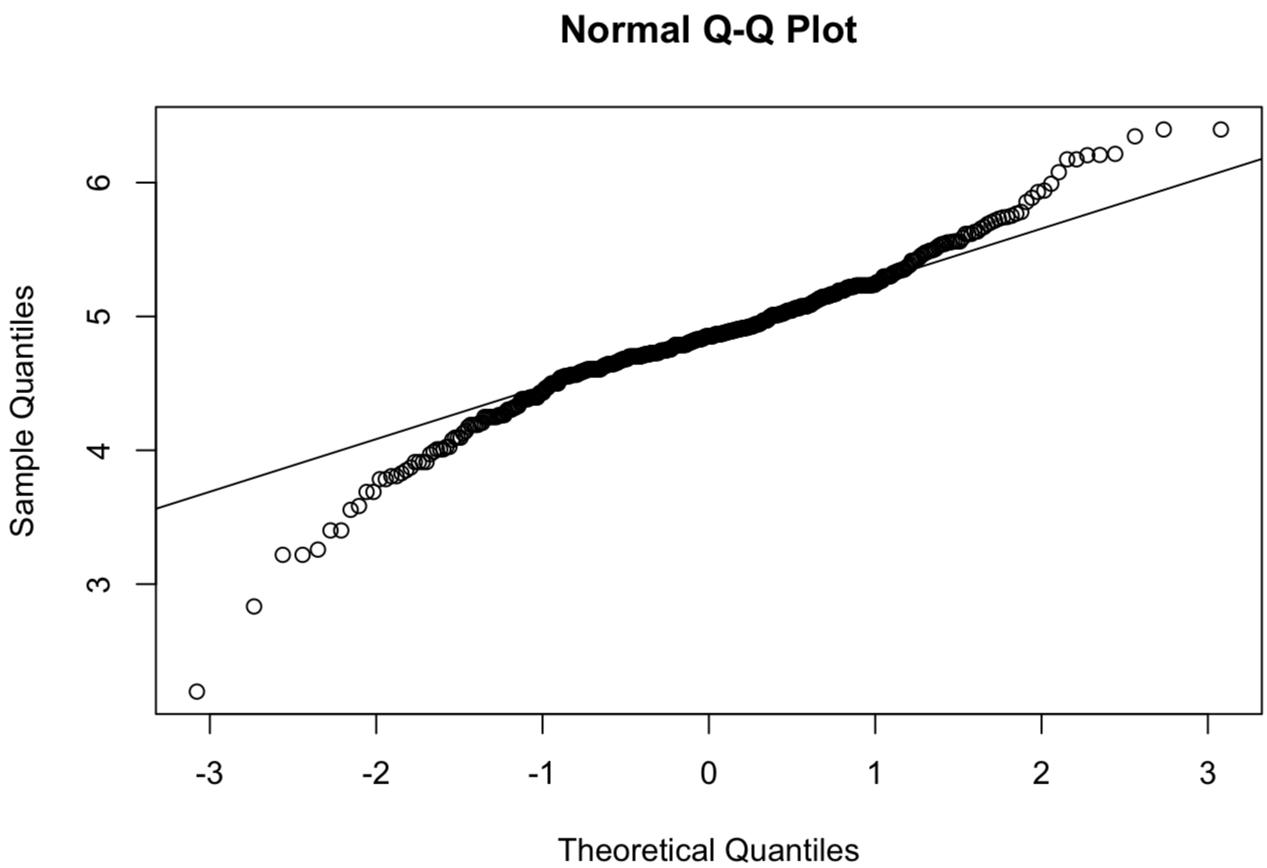

```
# Q-Q plots
qqnorm(df$Log_ApplicantIncome); qqline(df$Log_ApplicantIncome)
```




```
qqnorm(df$Log_CoapplicantIncome); qqline(df$Log_CoapplicantIncome)
```

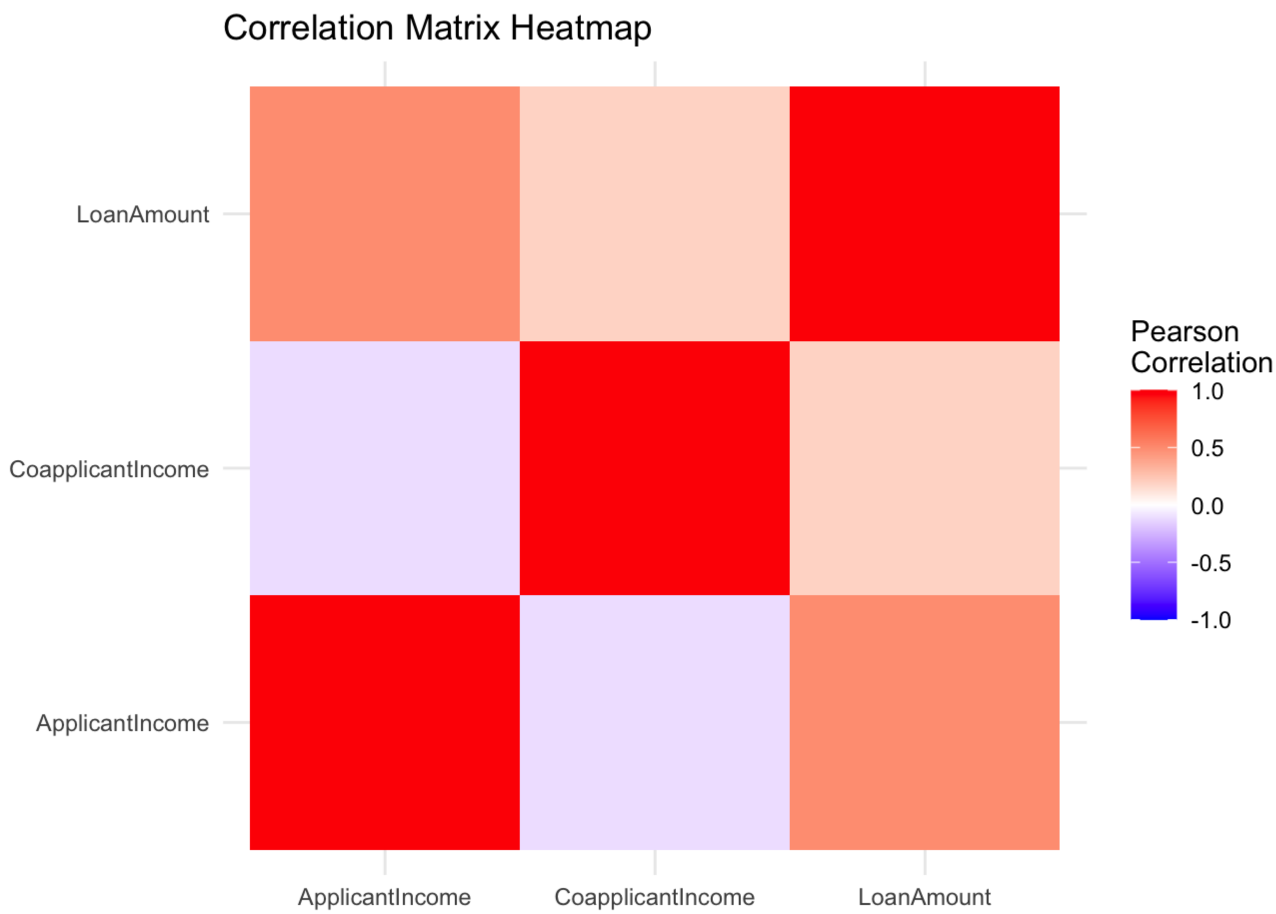



```
qqnorm(df$Log_LoanAmount); qqline(df$Log_LoanAmount)
```

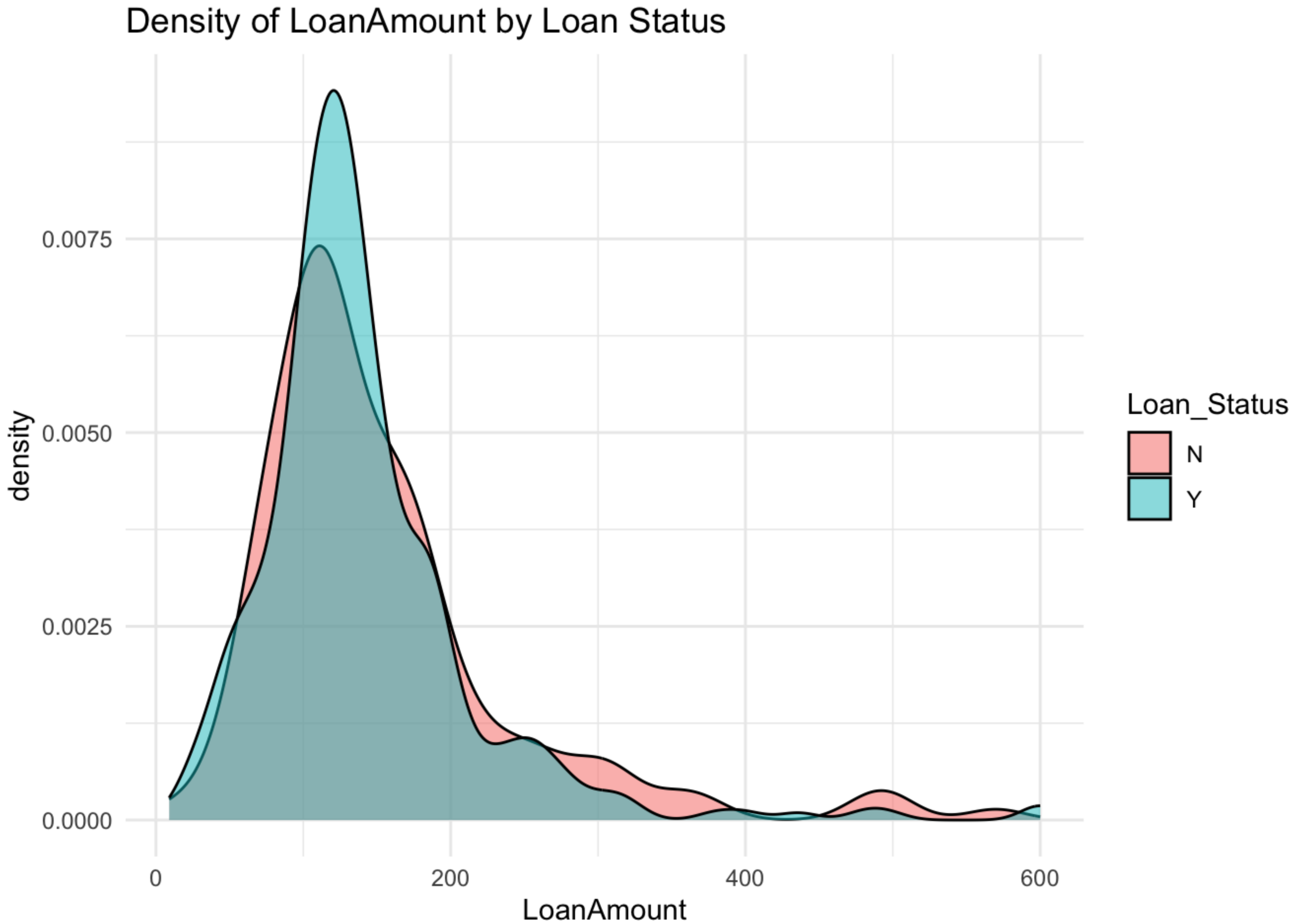



```
#correlation analysis

# Load necessary libraries
library(ggplot2)
library(reshape2)

df_corr <- df_cleaned

# Assuming ApplicantIncome, CoapplicantIncome, and LoanAmount are your continuous variabl
# Calculating correlation matrix
continuous_vars <- df_corr[, c("ApplicantIncome", "CoapplicantIncome", "LoanAmount")]
cor_matrix <- cor(continuous_vars, use = "complete.obs", method = "pearson")

# Melting the correlation matrix for visualization
melted_cor_matrix <- melt(cor_matrix)

# Visualizing the correlation matrix using heatmap
ggplot(data = melted_cor_matrix, aes(x = Var1, y = Var2, fill = value)) +
    geom_tile() +
    scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                         midpoint = 0, limit = c(-1,1), space = "Lab",
                         name="Pearson\nCorrelation") +
    theme_minimal() +
```

```
    ggtitle("Correlation Matrix Heatmap") +
    xlab("") + ylab("")
```




```
# Density plot for LoanAmount with Loan Status overlay
ggplot(df_corr, aes(x = LoanAmount, fill = Loan_Status)) +
    geom_density(alpha = 0.5) +
    theme_minimal() +
    ggtitle("Density of LoanAmount by Loan Status")
```

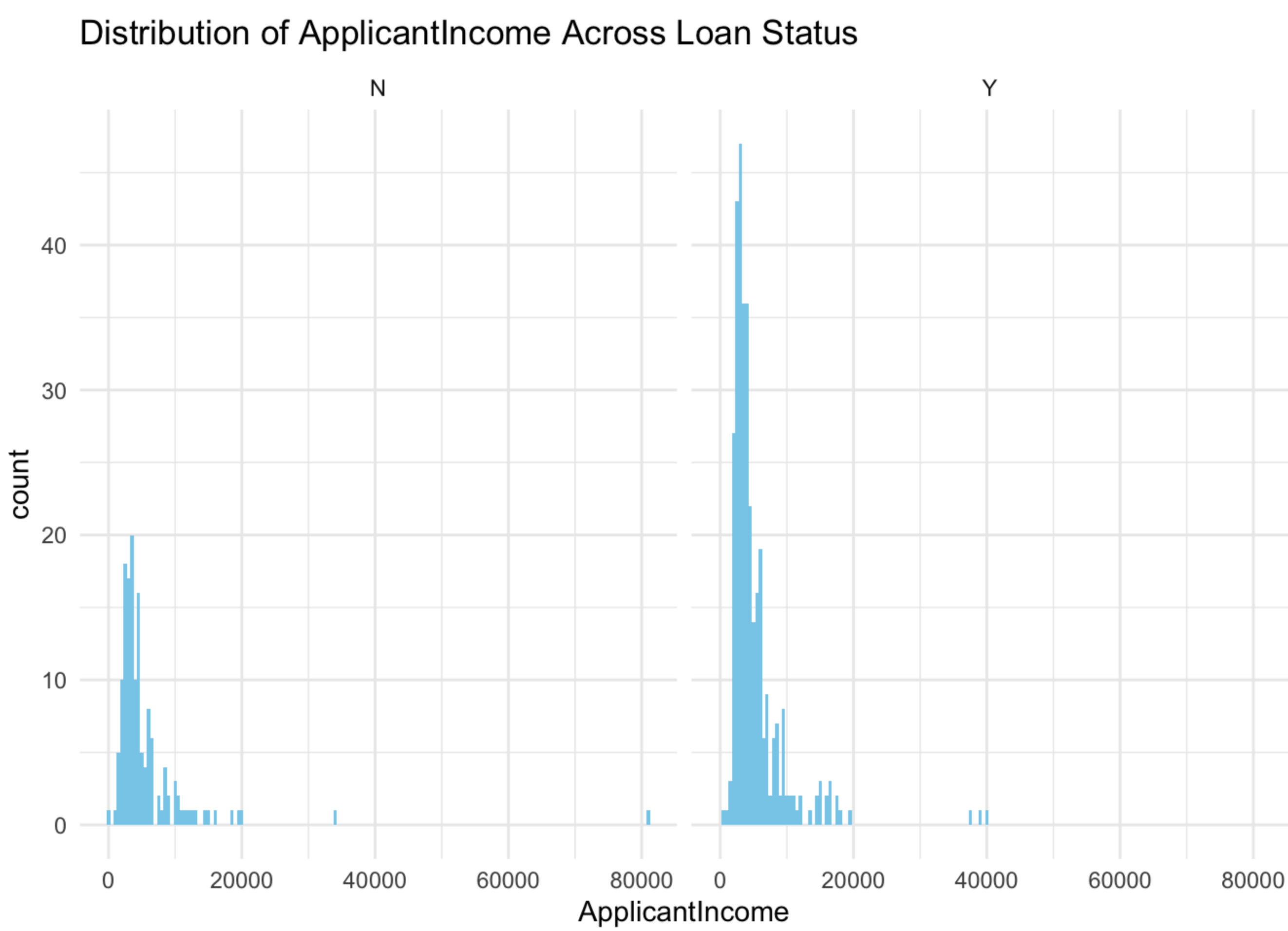


The **“Correlation Matrix Heatmap”** visually illustrates the Pearson correlation between ‘ApplicantIncome’, ‘CoapplicantIncome’, and ‘LoanAmount’, with red showing positive and blue showing negative correlations. The varying intensities of color denote the strength of each relationship, hinting at significant associations among the financial variables in the dataset. Particularly, the heatmap may point out stronger correlations between certain pairs, suggesting interdependencies that could influence loan-related decisions.

- ApplicantIncome vs.CoapplicantIncome: There doesn’t appear to be a strong linear correlation between these two variables, suggesting they may contribute independent information to a predictive model.
- ApplicantIncome vs. LoanAmount: There is a somewhat positive trend visible; as the applicant’s income increases, the loan amount tends to increase, which makes sense intuitively.
- CoapplicantIncome vs. LoanAmount: The trend is less clear, but there may still be a positive correlation.

Complementary to this, the series of plots, including the distribution histograms, density plots, and scatter plot with jitter, collectively explore the relationships between these financial attributes and loan status. Variations in applicant income distribution and loan amount densities across loan statuses may imply their influence on loan approval. The **“Mosaic Plot of Education and Loan Status”** and the **“Credit History vs ApplicantIncome”** plot further enrich this analysis by correlating educational background and credit history with loan outcomes, underscoring the multifaceted nature of loan approval criteria.

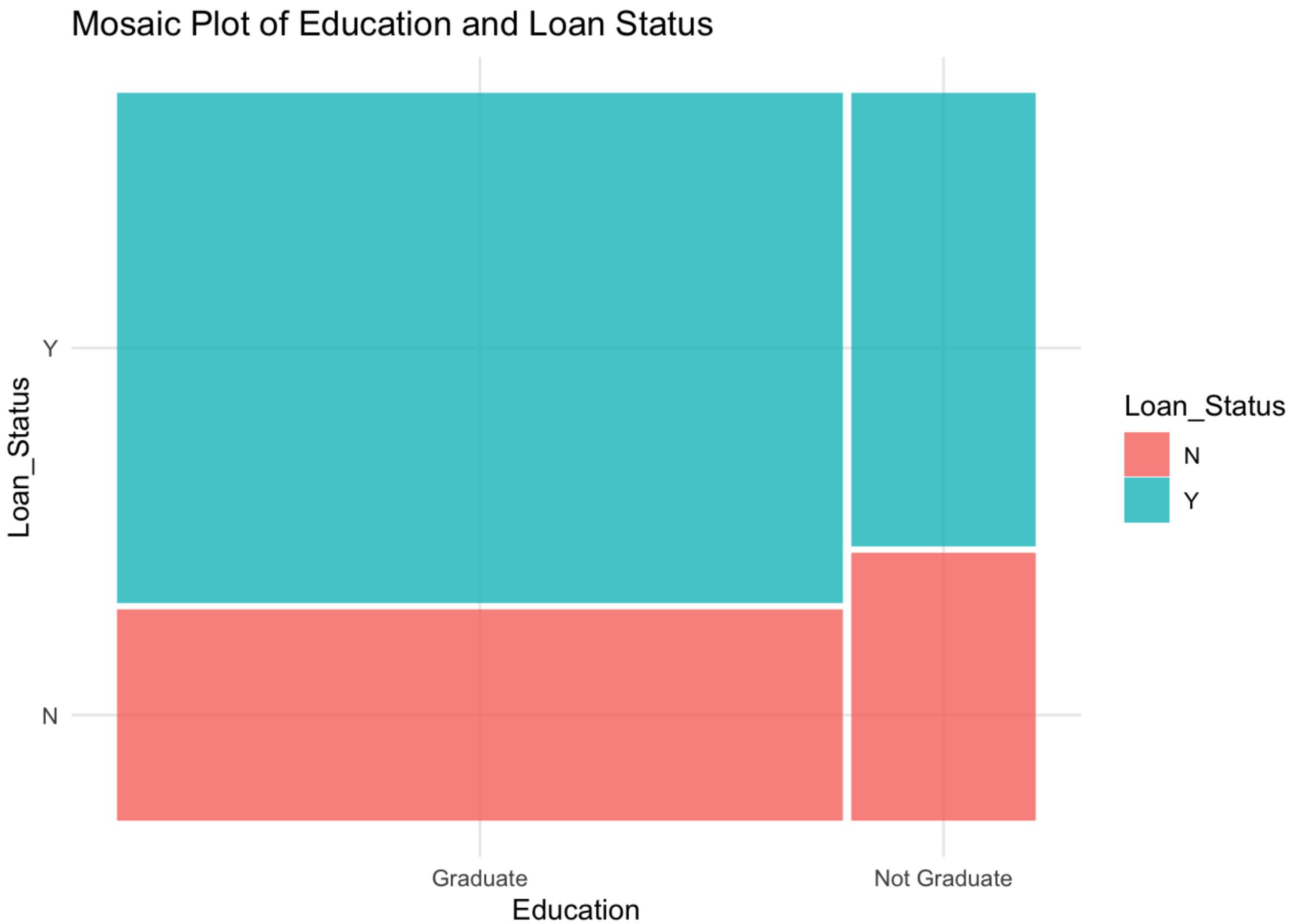
```
#Hidden
# Load necessary library
library(ggmosaic)

# Facet grid for ApplicantIncome by Loan Status
ggplot(df_corr, aes(x = ApplicantIncome)) +
    geom_histogram(binwidth = 500, fill = "skyblue") +
    facet_grid(. ~ Loan_Status) +
    theme_minimal() +
    ggtitle("Distribution of ApplicantIncome Across Loan Status")
```



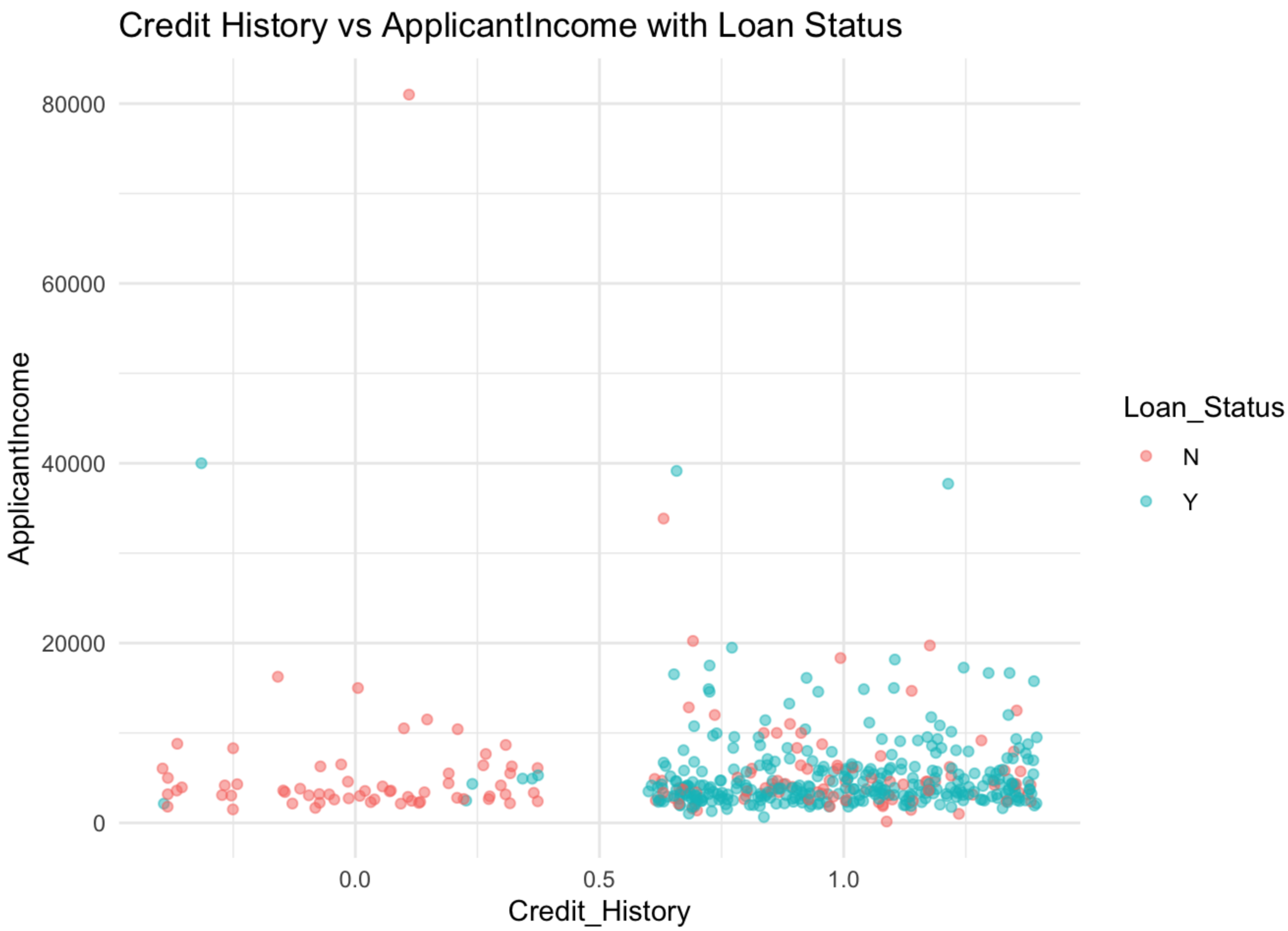

```
# Mosaic plot for Education and Loan Status
ggplot(data = df_corr) +
    geom_mosaic(aes(weight = 1, x = product(Education), fill = Loan_Status)) +
    theme_minimal() +
    ggtitle("Mosaic Plot of Education and Loan Status")
```

```
Warning: `unite_()` was deprecated in tidyr 1.2.0.
ℹ Please use `unite()` instead.
```

```
ℹ The deprecated feature was likely used in the ggmosaic package.
  Please report the issue at <https://github.com/haleyjeppson/ggmosaic>.
```




```r
# Scatter plot with jitter for Credit_History and ApplicantIncome
ggplot(df_corr, aes(x = Credit_History, y = ApplicantIncome, color = Loan_Status)) +
    geom_jitter(alpha = 0.5) +
    theme_minimal() +
    ggtitle("Credit History vs ApplicantIncome with Loan Status")
```




```
#Hidden
df$Gender <- as.factor(df$Gender)
df$Married <- as.factor(df$Married)
df$Dependents <- as.factor(df$Dependents)
df$Education <- as.factor(df$Education)
df$Self_Employed <- as.factor(df$Self_Employed)
df$Credit_History <- as.factor(df$Credit_History)
df$Property_Area <- as.factor(df$Property_Area)
df$Loan_Status <- as.factor(df$Loan_Status)
str(df)
```

```
'data.frame':	480 obs. of  16 variables:
 $ Loan_ID              : chr  "LP001003" "LP001005" "LP001006" "LP001008" ...
 $ Gender               : Factor w/ 2 levels "Female","Male": 2 2 2 2 2 2 2 2 2 2 ...
 $ Married              : Factor w/ 2 levels "No","Yes": 2 2 2 1 2 2 2 2 2 2 ...
 $ Dependents           : Factor w/ 4 levels "0","1","2","3+": 2 1 1 1 3 1 4 3 2 3 ...
 $ Education            : Factor w/ 2 levels "Graduate","Not Graduate": 1 1 2 1 1 2 1 1 1
1 ...
 $ Self_Employed        : Factor w/ 2 levels "No","Yes": 1 2 1 1 2 1 1 1 1 1 ...
 $ ApplicantIncome      : num  4583 3000 2583 6000 5417 ...
 $ CoapplicantIncome    : num  1508 1 2358 1 4196 ...
 $ LoanAmount           : num  128 66 120 141 267 95 158 168 349 70 ...
 $ Loan_Amount_Term     : int  360 360 360 360 360 360 360 360 360 360 ...
```

```
 $ Credit_History       : Factor w/ 2 levels "0","1": 2 2 2 2 2 2 1 2 2 2 ...
 $ Property_Area        : Factor w/ 3 levels "Rural","Semiurban",..: 1 3 3 3 3 3 2 3 2 3
...
 $ Loan_Status          : Factor w/ 2 levels "N","Y": 1 2 2 2 2 2 1 2 1 2 ...
 $ Log_ApplicantIncome  : num  8.43 8.01 7.86 8.7 8.6 ...
 $ Log_CoapplicantIncome: num  7.32 0 7.77 0 8.34 ...
 $ Log_LoanAmount       : num  4.85 4.19 4.79 4.95 5.59 ...
 - attr(*, "na.action")= 'omit' Named int [1:85] 1 17 20 25 31 36 37 43 45 46 ...
  ..- attr(*, "names")= chr [1:85] "1" "17" "20" "25" ...
```

```r
# Assuming 'Yes' or 'Y' indicates a positive response and should be coded as 1
# and 'No' or 'N' as a negative response to be coded as 0
df_cleaned$Loan_Status <- as.numeric(df_cleaned$Loan_Status == "Yes" | df_cleaned$Loan_St
null.model <- glm(df_cleaned$Loan_Status ~ 1, data= df_cleaned, family = binomial(link =
summary(null.model)
```

```
Call:
glm(formula = df_cleaned$Loan_Status ~ 1, family = binomial(link = "logit"),
    data = df_cleaned)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.80792    0.09884   8.174 2.98e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
Residual deviance: 593.05  on 479  degrees of freedom
AIC: 595.05

Number of Fisher Scoring iterations: 4
```

1. **Coefficients**:
    - **(Intercept) Estimate (0.80792)**: This is the log-odds of the outcome being 1 (e.g., Loan approved) when no predictors are included in the model. To get the probability, you'd need to transform this using the logistic function.
    - **Std. Error (0.09884)**: This represents the standard error of the estimated intercept.
    - **z value (8.174)**: This is the test statistic for evaluating the null hypothesis that the coefficient is equal to zero. A higher absolute value indicates more evidence against the null hypothesis.
    - **Pr(>|z|) (< 2.98e-16)**: This p-value is extremely low, suggesting that the intercept is significantly different from zero.

2. **Null Deviance (593.05)**: This is a measure of the model fit. It represents the difference in log-likelihood between a model with only the intercept and a saturated model. The degrees of freedom here equal the number of observations minus 1.

3. **AIC (595.05)**: The Akaike Information Criterion is a measure of the relative quality of the statistical model for a given set of data. Lower AIC values indicate a better fit.

Interpretation:

- The significant intercept suggests that even without any predictors, the model can predict the `Loan_Status` to some extent. This could be due to an imbalance in the response variable (e.g., more 'Yes' than 'No').

- The null model is a baseline model; including predictors in your model should ideally reduce the deviance and improve the AIC.

```
full.model <- glm(df_cleaned$Loan_Status ~ ., data= df_cleaned, family = binomial(link =
summary(full.model)
```

```

Call:
glm(formula = df_cleaned$Loan_Status ~ ., family = binomial(link = "logit"),
    data = df_cleaned)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -2.429e+00  9.312e-01  -2.609  0.00909 **
GenderMale              3.254e-01  3.309e-01   0.983  0.32548
MarriedYes              5.739e-01  2.924e-01   1.963  0.04970 *
Dependents1            -3.756e-01  3.460e-01  -1.085  0.27771
Dependents2             2.770e-01  3.782e-01   0.733  0.46378
Dependents3+            1.884e-01  4.874e-01   0.386  0.69915
EducationNot Graduate  -4.210e-01  3.033e-01  -1.388  0.16510
Self_EmployedYes       -1.492e-01  3.523e-01  -0.423  0.67202
ApplicantIncome         6.945e-06  2.862e-05   0.243  0.80827
CoapplicantIncome      -5.143e-05  4.307e-05  -1.194  0.23246
LoanAmount             -2.737e-03  1.773e-03  -1.544  0.12270
Loan_Amount_Term       -9.253e-04  2.032e-03  -0.455  0.64885
Credit_History          3.650e+00  4.331e-01   8.427  < 2e-16 ***
Property_AreaSemiurban  9.873e-01  3.036e-01   3.253  0.00114 **
Property_AreaUrban      1.511e-01  3.007e-01   0.503  0.61527
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
Residual deviance: 435.72  on 465  degrees of freedom
AIC: 465.72
```

```
Number of Fisher Scoring iterations: 5
```

1. **Coefficients**:
   - **Intercept and Variable Estimates**: These are the log-odds coefficients for each variable. For example, `Credit_History` has a highly positive coefficient, indicating a strong positive effect on the likelihood of loan approval when the credit history is positive.
   - **Std. Error**: Indicates the standard error of each coefficient estimate.
   - **z value**: The ratio of the estimate to its standard error. Larger absolute values indicate greater significance.
   - **Pr(>|z|)**: P-values associated with the z-values. A small p-value (< 0.05) suggests that the variable significantly contributes to the model.
2. **Significance Codes**:
   - Variables like `MarriedYes`, `Credit_History`, and `Property_AreaSemiurban` are statistically significant (p < 0.05).
3. **Model Fit Indicators**:
   - **Null Deviance and Residual Deviance**: The decrease from null deviance to residual deviance indicates that the model with predictors fits the data better than the null model.
   - **AIC (Akaike Information Criterion)**: A lower AIC suggests a better model. The AIC here is 465.72, which is lower than that of the null model, indicating an improved fit.
4. **Notable Predictors**:
   - **Credit History (highly significant)**: With the largest coefficient, it suggests a strong influence on loan approval.
   - **Property_AreaSemiurban**: Also significant, indicating the location of the property plays a role in loan approval.
   - **MarriedYes**: Marginally significant, suggesting marital status might have an influence.
5. **Number of Fisher Scoring iterations**: The number of iterations taken to converge, which is 5 in this case.

Interpretation and Considerations:

- **Credit History** is a key predictor of loan approval. Its high positive coefficient suggests that having a positive credit history greatly increases the likelihood of loan approval.
- The significance of **Property_AreaSemiurban** indicates that applicants from semi-urban areas are more likely to get loan approval compared to the reference category (probably rural areas, since it's not included in the model output).

- **Marital Status** ('MarriedYes') also appears to influence the loan approval process, though less significantly than credit history or property area.

- Other variables, although included in the model, do not show a statistically significant relationship with the loan approval at the 0.05 significance level. This doesn't mean they are unimportant, but they might not have a strong individual impact in the presence of other variables.

```
both.logit <- step(null.model, list(lower= formula(null.model),
                                    upper= formula(full.model),
                                    direction="both",data=df_cleaned))
```

```
Start:  AIC=595.05
df_cleaned$Loan_Status ~ 1

                    Df Deviance    AIC
+ Credit_History     1   464.02 468.02
+ Property_Area      2   580.56 586.56
+ Married            1   587.08 591.08
+ LoanAmount         1   590.67 594.67
+ Education          1   590.86 594.86
<none>                   593.05 595.05
+ Gender             1   591.10 595.10
+ CoapplicantIncome  1   591.96 595.96
+ ApplicantIncome    1   592.21 596.21
+ Self_Employed      1   592.48 596.48
+ Loan_Amount_Term   1   593.02 597.02
+ Dependents         3   590.05 598.05

Step:  AIC=468.02
df_cleaned$Loan_Status ~ Credit_History

                    Df Deviance    AIC
+ Property_Area      2   451.38 459.38
+ Married            1   457.87 463.87
<none>                   464.02 468.02
+ Gender             1   462.23 468.23
+ LoanAmount         1   462.37 468.37
+ CoapplicantIncome  1   462.87 468.87
+ Education          1   463.05 469.05
+ Loan_Amount_Term   1   463.58 469.58
+ Self_Employed      1   463.70 469.70
+ ApplicantIncome    1   463.87 469.87
+ Dependents         3   460.84 470.84
- Credit_History     1   593.05 595.05

Step:  AIC=459.38
df_cleaned$Loan_Status ~ Credit_History + Property_Area

                    Df Deviance    AIC
+ Married            1   445.58 455.58
```

```
+ Gender             1   448.48 458.48
<none>                   451.38 459.38
+ LoanAmount         1   449.91 459.91
+ CoapplicantIncome  1   450.28 460.28
+ Education          1   450.69 460.69
+ Loan_Amount_Term   1   450.89 460.89
+ Self_Employed      1   451.13 461.13
+ ApplicantIncome    1   451.22 461.22
+ Dependents         3   447.37 461.37
- Property_Area      2   464.02 468.02
- Credit_History     1   580.56 586.56

Step:  AIC=455.58
df_cleaned$Loan_Status ~ Credit_History + Property_Area + Married

                    Df Deviance    AIC
+ LoanAmount         1   442.72 454.72
<none>                   445.58 455.58
+ CoapplicantIncome  1   444.07 456.07
+ Gender             1   444.81 456.81
+ Education          1   444.85 456.85
+ Self_Employed      1   445.24 457.24
+ ApplicantIncome    1   445.34 457.34
+ Loan_Amount_Term   1   445.42 457.42
+ Dependents         3   442.79 458.79
- Married            1   451.38 459.38
- Property_Area      2   457.87 463.87
- Credit_History     1   574.76 582.76

Step:  AIC=454.72
df_cleaned$Loan_Status ~ Credit_History + Property_Area + Married +
    LoanAmount

                    Df Deviance    AIC
<none>                   442.72 454.72
+ Education          1   441.20 455.20
- LoanAmount         1   445.58 455.58
+ CoapplicantIncome  1   441.77 455.77
+ Gender             1   441.79 455.79
+ ApplicantIncome    1   442.48 456.48
+ Self_Employed      1   442.56 456.56
+ Loan_Amount_Term   1   442.63 456.63
+ Dependents         3   439.99 457.99
- Married            1   449.91 459.91
- Property_Area      2   454.75 462.75
- Credit_History     1   570.59 580.59
```

```
summary(both.logit)
```

```
Call:
glm(formula = df_cleaned$Loan_Status ~ Credit_History + Property_Area + 
    Married + LoanAmount, family = binomial(link = "logit"), 
    data = df_cleaned)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -2.696180   0.514478  -5.241  1.6e-07 ***
Credit_History          3.617154   0.425869   8.494  < 2e-16 ***
Property_AreaSemiurban  0.938358   0.297659   3.152  0.00162 ** 
Property_AreaUrban      0.147326   0.289297   0.509  0.61057    
MarriedYes              0.667373   0.248585   2.685  0.00726 ** 
LoanAmount             -0.002474   0.001444  -1.713  0.08664 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
Residual deviance: 442.72  on 474  degrees of freedom
AIC: 454.72

Number of Fisher Scoring iterations: 4
```

```r
summary(both.logit)
```


```
Call:
glm(formula = df_cleaned$Loan_Status ~ Credit_History + Property_Area + 
    Married + LoanAmount, family = binomial(link = "logit"), 
    data = df_cleaned)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -2.696180   0.514478  -5.241  1.6e-07 ***
Credit_History          3.617154   0.425869   8.494  < 2e-16 ***
Property_AreaSemiurban  0.938358   0.297659   3.152  0.00162 ** 
Property_AreaUrban      0.147326   0.289297   0.509  0.61057    
MarriedYes              0.667373   0.248585   2.685  0.00726 ** 
LoanAmount             -0.002474   0.001444  -1.713  0.08664 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
Residual deviance: 442.72  on 474  degrees of freedom
AIC: 454.72
```

```
Number of Fisher Scoring iterations: 4
```

1. **Coefficients**:
   - **Credit_History**: Highly significant (p < 2e-16) with a positive coefficient, indicating a strong influence on loan approval when the credit history is positive.
   - **Property_AreaSemiurban**: Statistically significant (p = 0.00162) with a positive effect, suggesting applicants from semi-urban areas are more likely to get a loan approved compared to the base category.
   - **MarriedYes**: Significant (p = 0.00726) with a positive coefficient, indicating that being married is associated with a higher likelihood of loan approval.
   - **LoanAmount**: Marginally significant (p = 0.08664), indicating a possible but not strong effect on loan approval.
   - **Property_AreaUrban**: Not statistically significant in this model.
2. **Model Fit**:
   - The **AIC** has decreased to 454.72 compared to the previous full model, suggesting a better fit with fewer variables.
   - The **Residual Deviance** has also decreased compared to the full model, indicating an improved fit.
3. **Number of Fisher Scoring iterations**: The convergence in 4 iterations indicates the model fit is stable.

```r
full.probit <- glm(Loan_Status ~ ., data = df_cleaned, family = binomial(link = "probit"))
summary(full.probit)
```

```
Call:
glm(formula = Loan_Status ~ ., family = binomial(link = "probit"), 
    data = df_cleaned)

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)   
(Intercept)            -1.390e+00  5.084e-01  -2.733 0.006275 **
GenderMale              1.889e-01  1.905e-01   0.992 0.321417   
MarriedYes              3.319e-01  1.661e-01   1.998 0.045692 * 
Dependents1            -2.143e-01  1.985e-01  -1.080 0.280326   
Dependents2             1.559e-01  2.088e-01   0.747 0.455180   
Dependents3+            9.749e-02  2.708e-01   0.360 0.718817   
EducationNot Graduate  -2.520e-01  1.733e-01  -1.454 0.145971   
Self_EmployedYes       -9.731e-02  2.013e-01  -0.483 0.628873   
ApplicantIncome         3.911e-06  1.548e-05   0.253 0.800502   
CoapplicantIncome      -2.825e-05  2.551e-05  -1.107 0.268243   
```

```
LoanAmount             -1.593e-03  1.023e-03  -1.557 0.119362
Loan_Amount_Term       -5.310e-04  1.122e-03  -0.473 0.636008
Credit_History          2.132e+00  2.244e-01   9.500  < 2e-16 ***
Property_AreaSemiurban  5.597e-01  1.698e-01   3.297 0.000978 ***
Property_AreaUrban      8.126e-02  1.740e-01   0.467 0.640444
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
Residual deviance: 435.62  on 465  degrees of freedom
AIC: 465.62

Number of Fisher Scoring iterations: 5
```

```r
null.probit <- glm(Loan_Status ~ 1, data = df_cleaned, family = binomial(link = "probit"))
summary(null.probit)
```

```
Call:
glm(formula = Loan_Status ~ 1, family = binomial(link = "probit"),
    data = df_cleaned)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.50058    0.05989   8.359   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
Residual deviance: 593.05  on 479  degrees of freedom
AIC: 595.05

Number of Fisher Scoring iterations: 4
```

```r
both.probit <- step(null.model, list(lower= formula(null.model),
                                    upper= formula(full.model),
                                    direction="both",data=df_cleaned))
```

```
Start:  AIC=595.05
df_cleaned$Loan_Status ~ 1

                    Df Deviance    AIC
+ Credit_History     1   464.02 468.02
+ Property_Area      2   580.56 586.56
+ Married            1   587.08 591.08
```

```
+ LoanAmount         1   590.67 594.67
+ Education          1   590.86 594.86
<none>                   593.05 595.05
+ Gender             1   591.10 595.10
+ CoapplicantIncome  1   591.96 595.96
+ ApplicantIncome    1   592.21 596.21
+ Self_Employed      1   592.48 596.48
+ Loan_Amount_Term   1   593.02 597.02
+ Dependents         3   590.05 598.05

Step:  AIC=468.02
df_cleaned$Loan_Status ~ Credit_History

                    Df Deviance    AIC
+ Property_Area      2   451.38 459.38
+ Married            1   457.87 463.87
<none>                   464.02 468.02
+ Gender             1   462.23 468.23
+ LoanAmount         1   462.37 468.37
+ CoapplicantIncome  1   462.87 468.87
+ Education          1   463.05 469.05
+ Loan_Amount_Term   1   463.58 469.58
+ Self_Employed      1   463.70 469.70
+ ApplicantIncome    1   463.87 469.87
+ Dependents         3   460.84 470.84
- Credit_History     1   593.05 595.05

Step:  AIC=459.38
df_cleaned$Loan_Status ~ Credit_History + Property_Area

                    Df Deviance    AIC
+ Married            1   445.58 455.58
+ Gender             1   448.48 458.48
<none>                   451.38 459.38
+ LoanAmount         1   449.91 459.91
+ CoapplicantIncome  1   450.28 460.28
+ Education          1   450.69 460.69
+ Loan_Amount_Term   1   450.89 460.89
+ Self_Employed      1   451.13 461.13
+ ApplicantIncome    1   451.22 461.22
+ Dependents         3   447.37 461.37
- Property_Area      2   464.02 468.02
- Credit_History     1   580.56 586.56

Step:  AIC=455.58
df_cleaned$Loan_Status ~ Credit_History + Property_Area + Married

                    Df Deviance    AIC
+ LoanAmount         1   442.72 454.72
<none>                   445.58 455.58
+ CoapplicantIncome  1   444.07 456.07
```

```
+ Gender             1   444.81 456.81
+ Education          1   444.85 456.85
+ Self_Employed      1   445.24 457.24
+ ApplicantIncome    1   445.34 457.34
+ Loan_Amount_Term   1   445.42 457.42
+ Dependents         3   442.79 458.79
- Married            1   451.38 459.38
- Property_Area      2   457.87 463.87
- Credit_History     1   574.76 582.76

Step:  AIC=454.72
df_cleaned$Loan_Status ~ Credit_History + Property_Area + Married + 
    LoanAmount

                    Df Deviance    AIC
<none>                   442.72 454.72
+ Education          1   441.20 455.20
- LoanAmount         1   445.58 455.58
+ CoapplicantIncome  1   441.77 455.77
+ Gender             1   441.79 455.79
+ ApplicantIncome    1   442.48 456.48
+ Self_Employed      1   442.56 456.56
+ Loan_Amount_Term   1   442.63 456.63
+ Dependents         3   439.99 457.99
- Married            1   449.91 459.91
- Property_Area      2   454.75 462.75
- Credit_History     1   570.59 580.59
```

```
summary(both.probit)
```

```
Call:
glm(formula = df_cleaned$Loan_Status ~ Credit_History + Property_Area + 
    Married + LoanAmount, family = binomial(link = "logit"), 
    data = df_cleaned)

Coefficients:
                        Estimate Std. Error z value Pr(>|z|)    
(Intercept)            -2.696180   0.514478  -5.241  1.6e-07 ***
Credit_History          3.617154   0.425869   8.494  < 2e-16 ***
Property_AreaSemiurban  0.938358   0.297659   3.152  0.00162 ** 
Property_AreaUrban      0.147326   0.289297   0.509  0.61057    
MarriedYes              0.667373   0.248585   2.685  0.00726 ** 
LoanAmount             -0.002474   0.001444  -1.713  0.08664 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 593.05  on 479  degrees of freedom
```

```
Residual deviance: 442.72  on 474  degrees of freedom
AIC: 454.72

Number of Fisher Scoring iterations: 4
```

- **Credit_History**: Highly significant (p < 2e-16) with a positive coefficient, indicating a strong influence on loan approval when the credit history is positive.
- **Property_AreaSemiurban**: Statistically significant (p = 0.00162) with a positive effect, suggesting applicants from semi-urban areas are more likely to get a loan approved compared to the base category.
- **MarriedYes**: Significant (p = 0.00726) with a positive coefficient, indicating that being married is associated with a higher likelihood of loan approval.
- **LoanAmount**: Marginally significant (p = 0.08664), indicating a possible but not strong effect on loan approval.

```
library(pROC)
```

```
Type 'citation("pROC")' for a citation.


Attaching package: 'pROC'

The following objects are masked from 'package:stats':

    cov, smooth, var
```

```
table(df_cleaned$Loan_Status)
```

```

  0   1 
148 332 
```

```
df_cleaned$Loan_Status <- as.factor(df_cleaned$Loan_Status)
```

```
set.seed(123457)
train.prop <- 0.80
auclist <- c()
for (t in 1:500){
  # Splitting the data
  strats <- df_cleaned$Loan_Status
  rr <- split(1:length(strats), strats)
  idx <- sort(as.numeric(unlist(sapply(rr,
        function(x) sample(x, length(x)*train.prop)))))
  df.train <- df_cleaned[idx, ]
  df.test <- df_cleaned[-idx, ]
```

```
  # Training the null model on the training set
  null.model <- glm(Loan_Status ~ 1, data= df.train, family = binomial(link = "logit"))

  # Making predictions on the test set
  pd <- predict(null.model, newdata = df.test, type = 'response')
  predicted_class <- ifelse(pd > 0.5, 1, 0)

  # ROC analysis and AUC calculation
  g <- roc(response = as.numeric(df.test$Loan_Status == 1),
           predictor = pd, print.auc = TRUE,
           algorithm = 2, levels = c(0, 1), direction = "<")

  auclist <- c(auclist, as.numeric(g$auc))
}
# Averaging the metrics
benchmark_auc <- mean(auclist)


benchmark_auc
```

```
[1] 0.5
```

Talk about this

```
library(pROC)
library(caret)
```

```
Loading required package: lattice
```

```
set.seed(123457)
train.prop <- 0.80
auclist <- c()
residual_deviances <- c()
accuracies <- c()
recalls <- c()
precisions <- c()
f1_scores <- c()

for (t in 1:500){
    # Splitting the data
    strats <- df_cleaned$Loan_Status
    rr <- split(1:length(strats), strats)
    idx <- sort(as.numeric(unlist(sapply(rr, function(x) sample(x, length(x) * train.prop
    df.train <- df_cleaned[idx, ]
    df.test <- df_cleaned[-idx, ]

    # Training the model on the training set
    full.logit <- glm(Loan_Status ~ ., data = df.train, family = binomial(link = "logit")

    # Residual Deviance
```

```
    residual_deviances <- c(residual_deviances, full.logit$deviance)

    # Making predictions on the test set
    pd <- predict(full.logit, newdata = df.test, type = 'response')
    predicted_class <- ifelse(pd > 0.5, 1, 0)

    # ROC analysis and AUC calculation
    g <- roc(response = df.test$Loan_Status, predictor = pd, print.auc = TRUE, algorithm
    auclist <- c(auclist, as.numeric(g$auc))

    # Confusion Matrix and related metrics
    cm <- confusionMatrix(as.factor(predicted_class), as.factor(df.test$Loan_Status))
    accuracies <- c(accuracies, cm$overall['Accuracy'])
    recalls <- c(recalls, cm$byClass['Sensitivity'])
    precisions <- c(precisions, cm$byClass['Precision'])
    f1_scores <- c(f1_scores, cm$byClass['F1'])
}

# Calculating averages
benchmark_auc <- mean(auclist)
average_residual_deviance <- mean(residual_deviances)
average_accuracy <- mean(accuracies)
average_recall <- mean(recalls)
average_precision <- mean(precisions)
average_f1_score <- mean(f1_scores)

list(
  benchmark_auc = benchmark_auc,
  average_residual_deviance = average_residual_deviance,
  average_accuracy = average_accuracy,
  average_recall = average_recall,
  average_precision = average_precision,
  average_f1_score = average_f1_score
)
```

```
$benchmark_auc
[1] 0.7561831

$average_residual_deviance
[1] 344.6034

$average_accuracy
[1] 0.8041443

$average_recall
[1] 0.4387333

$average_precision
[1] 0.8641987
```

```
$average_f1_score
[1] 0.5768786
```

Good Predictive Ability: An AUC score of 0.75 suggests that the model has a good level of predictive accuracy. In practical terms, this means that there's an 75% chance that the model will correctly distinguish between a positive and a negative instance when randomly picking one of each.

```
library(pROC)
library(caret)

set.seed(123457)
train.prop <- 0.80
auclist <- c()
residual_deviances <- c()
null_deviances <- c()
accuracies <- c()
recalls <- c()
precisions <- c()
f1_scores <- c()

for (t in 1:500){
    # Splitting the data
    strats <- df_cleaned$Loan_Status
    rr <- split(1:length(strats), strats)
    idx <- sort(as.numeric(unlist(sapply(rr, function(x) sample(x, length(x) * train.prop
    df.train <- df_cleaned[idx, ]
    df.test <- df_cleaned[-idx, ]

    # Training the model on the training set
    both.logit <- glm(Loan_Status ~ Credit_History + Property_Area + Married + LoanAmount

    # Making predictions on the test set
    pd <- predict(both.logit, newdata = df.test, type = 'response')
    predicted_class <- ifelse(pd > 0.5, 1, 0)

    # ROC analysis and AUC calculation
    g <- roc(response = as.numeric(df.test$Loan_Status == 1), predictor = pd, print.auc =
    auclist <- c(auclist, as.numeric(g$auc))

    # Confusion Matrix and related metrics
    cm <- confusionMatrix(as.factor(predicted_class), as.factor(df.test$Loan_Status))
    accuracies <- c(accuracies, cm$overall['Accuracy'])
    recalls <- c(recalls, cm$byClass['Sensitivity'])
    precisions <- c(precisions, cm$byClass['Precision'])
    f1_scores <- c(f1_scores, cm$byClass['F1'])

    # Residual and Null Deviance
    residual_deviances <- c(residual_deviances, both.logit$deviance)
    null_deviances <- c(null_deviances, both.logit$null.deviance)
}
```

```
# Calculating averages
benchmark_auc <- mean(auclist)
average_residual_deviance <- mean(residual_deviances)
average_null_deviance <- mean(null_deviances)
average_accuracy <- mean(accuracies)
average_recall <- mean(recalls)
average_precision <- mean(precisions)
average_f1_score <- mean(f1_scores)

list(
  benchmark_auc = benchmark_auc,
  average_residual_deviance = average_residual_deviance,
  average_null_deviance = average_null_deviance,
  average_accuracy = average_accuracy,
  average_recall = average_recall,
  average_precision = average_precision,
  average_f1_score = average_f1_score
)
```

```
$benchmark_auc
[1] 0.7783473

$average_residual_deviance
[1] 352.3706

$average_null_deviance
[1] 473.0564

$average_accuracy
[1] 0.8109485

$average_recall
[1] 0.4364667

$average_precision
[1] 0.9032575

$average_f1_score
[1] 0.5841136
```

The results of our model evaluation over 500 iterations show an average AUC of 0.7783, indicating a good ability to distinguish between the two classes of `Loan_Status`. The average residual deviance is 352.3706, significantly lower than the average null deviance of 473.0564, suggesting that the predictors in our model add substantial explanatory power.

In terms of classification metrics, the average accuracy is 0.8109, meaning my model correctly predicts the `Loan_Status` 81.09% of the time. However, the average recall is relatively low at 0.4365, indicating that the model might be missing a significant number of true positive cases. On the other hand, the average precision is high at 0.9033, showing that when the model predicts a positive case, it is correct

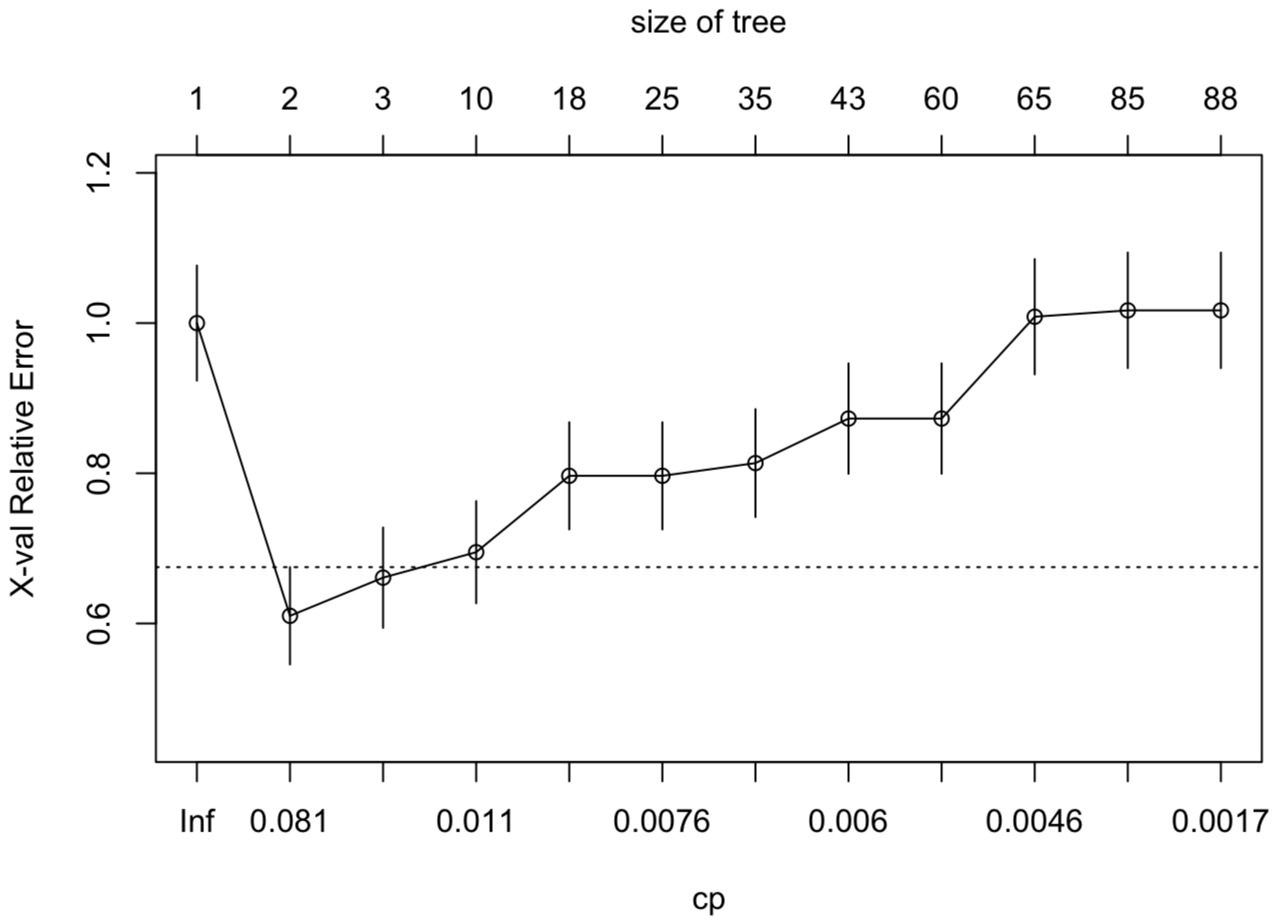
90.33% of the time. The average F1 score is 0.5841, reflecting a moderate balance between precision and recall, though leaning more towards precision.

```
library(rpart)
library(rpart.plot)
library(caret)
```

```
# Build the decision tree model
fit.allp <- rpart(Loan_Status ~ ., method = "class", data = df.train,
                  control = rpart.control(minsplit = 1, cp = 0.001))
printcp(fit.allp)
```

```
Classification tree:
rpart(formula = Loan_Status ~ ., data = df.train, method = "class",
    control = rpart.control(minsplit = 1, cp = 0.001))

Variables actually used in tree construction:
 [1] ApplicantIncome   CoapplicantIncome Credit_History    Dependents
 [5] Education         Gender            Loan_Amount_Term  LoanAmount
 [9] Married           Property_Area     Self_Employed

Root node error: 118/383 = 0.30809

n= 383

          CP nsplit rel error  xerror     xstd
1  0.3898305      0 1.0000000 1.00000 0.076574
2  0.0169492      1 0.6101695 0.61017 0.064799
3  0.0127119      2 0.5932203 0.66102 0.066791
4  0.0101695      9 0.5000000 0.69492 0.068031
5  0.0084746     17 0.4067797 0.79661 0.071373
6  0.0067797     24 0.3474576 0.79661 0.071373
7  0.0063559     34 0.2796610 0.81356 0.071878
8  0.0056497     42 0.2203390 0.87288 0.073539
9  0.0050847     59 0.1186441 0.87288 0.073539
10 0.0042373     64 0.0932203 1.00847 0.076753
11 0.0028249     84 0.0084746 1.01695 0.076928
12 0.0010000     87 0.0000000 1.01695 0.076928
```

```
# Find the optimal complexity parameter
cp <- fit.allp$cptable[which.min(fit.allp$cptable[,"xerror"]),"CP"]
xerr <- fit.allp$cptable[which.min(fit.allp$cptable[,"xerror"]),"xerror"]

# Plot the complexity parameter plot
plotcp(fit.allp)
```




```r
# Detailed summary of the model
summary(fit.allp)
```

```
Call:
rpart(formula = Loan_Status ~ ., data = df.train, method = "class",
    control = rpart.control(minsplit = 1, cp = 0.001))
  n= 383

            CP nsplit   rel error    xerror       xstd
1  0.389830508      0 1.000000000 1.0000000 0.07657421
2  0.016949153      1 0.610169492 0.6101695 0.06479851
3  0.012711864      2 0.593220339 0.6610169 0.06679067
4  0.010169492      9 0.500000000 0.6949153 0.06803130
5  0.008474576     17 0.406779661 0.7966102 0.07137259
6  0.006779661     24 0.347457627 0.7966102 0.07137259
7  0.006355932     34 0.279661017 0.8135593 0.07187787
8  0.005649718     42 0.220338983 0.8728814 0.07353875
9  0.005084746     59 0.118644068 0.8728814 0.07353875
10 0.004237288     64 0.093220339 1.0084746 0.07675277
11 0.002824859     84 0.008474576 1.0169492 0.07692847
12 0.001000000     87 0.000000000 1.0169492 0.07692847

Variable importance
```

```
    Credit_History   ApplicantIncome CoapplicantIncome        LoanAmount 
                24                20                16                15 
        Dependents  Loan_Amount_Term           Married     Self_Employed 
                 6                 5                 3                 3 
            Gender     Property_Area         Education 
                 3                 2                 2 

Node number 1: 383 observations,    complexity param=0.3898305
  predicted class=1  expected loss=0.308094  P(node) =1
    class counts:   118   265
   probabilities: 0.308 0.692 
  left son=2 (56 obs) right son=3 (327 obs)
  Primary splits:
      Credit_History   < 0.5     to the left,  improve=47.638330, (0 missing)
      Property_Area    splits as  LRL,         improve= 3.269674, (0 missing)
      Loan_Amount_Term < 420     to the right, improve= 2.370031, (0 missing)
      Married          splits as  LR,          improve= 2.002812, (0 missing)
      LoanAmount       < 283     to the right, improve= 1.904045, (0 missing)

Node number 2: 56 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.08928571  P(node) =0.1462141
    class counts:    51     5
   probabilities: 0.911 0.089 
  left son=4 (54 obs) right son=5 (2 obs)
  Primary splits:
      CoapplicantIncome < 8115    to the left,  improve=0.6997354, (0 missing)
      LoanAmount        < 136.5   to the left,  improve=0.5590533, (0 missing)
      ApplicantIncome   < 4316.5  to the left,  improve=0.2707275, (0 missing)
      Property_Area     splits as  LRL,         improve=0.2293651, (0 missing)
      Self_Employed     splits as  RL,          improve=0.1709726, (0 missing)

Node number 3: 327 observations,    complexity param=0.01694915
  predicted class=1  expected loss=0.204893  P(node) =0.8537859
    class counts:    67   260
   probabilities: 0.205 0.795 
  left son=6 (2 obs) right son=7 (325 obs)
  Primary splits:
      Loan_Amount_Term < 48      to the left,  improve=2.544343, (0 missing)
      Property_Area    splits as  LRL,         improve=2.371207, (0 missing)
      Married          splits as  LR,          improve=2.067278, (0 missing)
      ApplicantIncome  < 18249   to the right, improve=1.922021, (0 missing)
      LoanAmount       < 285     to the right, improve=1.820468, (0 missing)

Node number 4: 54 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.07407407  P(node) =0.1409922
    class counts:    50     4
   probabilities: 0.926 0.074 
  left son=8 (38 obs) right son=9 (16 obs)
  Primary splits:
      LoanAmount        < 159     to the left,  improve=0.5850390, (0 missing)
      ApplicantIncome   < 4316.5  to the left,  improve=0.4629630, (0 missing)
```

```
      Dependents        splits as  LRRR,        improve=0.1963729, (0 missing)
      CoapplicantIncome < 1355    to the left,  improve=0.1481481, (0 missing)
      Gender            splits as  LR,          improve=0.1185185, (0 missing)
  Surrogate splits:
      ApplicantIncome   < 8229.5  to the left,  agree=0.815, adj=0.375, (0 split)
      CoapplicantIncome < 3191.5  to the left,  agree=0.815, adj=0.375, (0 split)

Node number 5: 2 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.5  P(node) =0.005221932
    class counts:     1     1
   probabilities: 0.500 0.500 
  left son=10 (1 obs) right son=11 (1 obs)
  Primary splits:
      Gender            splits as  RL,          improve=1, (0 missing)
      Married           splits as  RL,          improve=1, (0 missing)
      Dependents        splits as  R--L,        improve=1, (0 missing)
      ApplicantIncome   < 3826.5  to the right, improve=1, (0 missing)
      CoapplicantIncome < 10140   to the right, improve=1, (0 missing)

Node number 6: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 7: 325 observations,    complexity param=0.01271186
  predicted class=1  expected loss=0.2  P(node) =0.848564
    class counts:    65   260
   probabilities: 0.200 0.800 
  left son=14 (197 obs) right son=15 (128 obs)
  Primary splits:
      Property_Area     splits as  LRL,         improve=2.896336, (0 missing)
      ApplicantIncome   < 18249   to the right, improve=1.973944, (0 missing)
      LoanAmount        < 285     to the right, improve=1.898463, (0 missing)
      CoapplicantIncome < 7480    to the right, improve=1.625000, (0 missing)
      Married           splits as  LR,          improve=1.609534, (0 missing)
  Surrogate splits:
      ApplicantIncome  < 27039.5 to the left,  agree=0.615, adj=0.023, (0 split)
      Loan_Amount_Term < 420     to the left,  agree=0.612, adj=0.016, (0 split)

Node number 8: 38 observations,    complexity param=0.004237288
  predicted class=0  expected loss=0.02631579  P(node) =0.09921671
    class counts:    37     1
   probabilities: 0.974 0.026 
  left son=16 (35 obs) right son=17 (3 obs)
  Primary splits:
      CoapplicantIncome < 2446    to the left,  improve=0.6140351, (0 missing)
      Loan_Amount_Term  < 240     to the right, improve=0.2807018, (0 missing)
      LoanAmount        < 91.5    to the right, improve=0.2330827, (0 missing)
      ApplicantIncome   < 2541.5  to the right, improve=0.1695906, (0 missing)
      Property_Area     splits as  RLL,         improve=0.1140351, (0 missing)
```

```
Node number 9: 16 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.1875  P(node) =0.04177546
    class counts:    13     3
   probabilities: 0.813 0.188 
  left son=18 (11 obs) right son=19 (5 obs)
  Primary splits:
      LoanAmount        < 173     to the right, improve=2.4750000, (0 missing)
      ApplicantIncome   < 5690.5  to the right, improve=1.1250000, (0 missing)
      CoapplicantIncome < 2072    to the right, improve=0.8750000, (0 missing)
      Dependents        splits as  LRRR,        improve=0.6750000, (0 missing)
      Property_Area     splits as  LRL,         improve=0.6568182, (0 missing)
  Surrogate splits:
      Loan_Amount_Term < 240     to the right, agree=0.812, adj=0.4, (0 split)

Node number 10: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 11: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 14: 197 observations,    complexity param=0.01271186
  predicted class=1  expected loss=0.2538071  P(node) =0.5143603
    class counts:    50   147
   probabilities: 0.254 0.746 
  left son=28 (3 obs) right son=29 (194 obs)
  Primary splits:
      ApplicantIncome   < 18249   to the right, improve=3.3924850, (0 missing)
      LoanAmount        < 302     to the right, improve=2.5251720, (0 missing)
      CoapplicantIncome < 14053   to the right, improve=2.2500590, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.9279824, (0 missing)
      Married           splits as  LR,          improve=0.8982748, (0 missing)
  Surrogate splits:
      LoanAmount < 402     to the right, agree=0.995, adj=0.667, (0 split)

Node number 15: 128 observations,    complexity param=0.006779661
  predicted class=1  expected loss=0.1171875  P(node) =0.3342037
    class counts:    15   113
   probabilities: 0.117 0.883 
  left son=30 (4 obs) right son=31 (124 obs)
  Primary splits:
      CoapplicantIncome < 6145.5  to the right, improve=1.2101810, (0 missing)
      Loan_Amount_Term  < 420     to the right, improve=1.2101810, (0 missing)
      LoanAmount        < 99.5    to the left,  improve=0.8418411, (0 missing)
      Married           splits as  LR,          improve=0.4855446, (0 missing)
      ApplicantIncome   < 3863    to the right, improve=0.3549489, (0 missing)

Node number 16: 35 observations
```

```
    predicted class=0  expected loss=0  P(node) =0.09138381
      class counts:    35     0
     probabilities: 1.000 0.000 

  Node number 17: 3 observations,    complexity param=0.004237288
    predicted class=0  expected loss=0.3333333  P(node) =0.007832898
      class counts:     2     1
     probabilities: 0.667 0.333 
    left son=34 (2 obs) right son=35 (1 obs)
    Primary splits:
        CoapplicantIncome < 2485    to the right, improve=1.3333330, (0 missing)
        LoanAmount        < 91.5    to the right, improve=1.3333330, (0 missing)
        Loan_Amount_Term  < 270     to the right, improve=1.3333330, (0 missing)
        Property_Area     splits as  RLL,         improve=1.3333330, (0 missing)
        Dependents        splits as  R--L,        improve=0.3333333, (0 missing)

  Node number 18: 11 observations
    predicted class=0  expected loss=0  P(node) =0.02872063
      class counts:    11     0
     probabilities: 1.000 0.000 

  Node number 19: 5 observations,    complexity param=0.006779661
    predicted class=1  expected loss=0.4  P(node) =0.01305483
      class counts:     2     3
     probabilities: 0.400 0.600 
    left son=38 (2 obs) right son=39 (3 obs)
    Primary splits:
        Self_Employed     splits as  RL,          improve=2.400000, (0 missing)
        CoapplicantIncome < 2630    to the right, improve=2.400000, (0 missing)
        Loan_Amount_Term  < 270     to the left,  improve=2.400000, (0 missing)
        Dependents        splits as  LLRR,        improve=1.066667, (0 missing)
        Property_Area     splits as  LRL,         improve=1.066667, (0 missing)
    Surrogate splits:
        CoapplicantIncome < 2630    to the right, agree=1, adj=1, (0 split)
        Loan_Amount_Term  < 270     to the left,  agree=1, adj=1, (0 split)

  Node number 28: 3 observations
    predicted class=0  expected loss=0  P(node) =0.007832898
      class counts:     3     0
     probabilities: 1.000 0.000 

  Node number 29: 194 observations,    complexity param=0.01271186
    predicted class=1  expected loss=0.242268  P(node) =0.5065274
      class counts:    47   147
     probabilities: 0.242 0.758 
    left son=58 (2 obs) right son=59 (192 obs)
    Primary splits:
        CoapplicantIncome < 14053   to the right, improve=2.3205540, (0 missing)
        ApplicantIncome   < 2437    to the left,  improve=1.2807230, (0 missing)
        LoanAmount        < 13      to the left,  improve=1.1542650, (0 missing)
        Married           splits as  LR,          improve=0.9275511, (0 missing)
```

```
      Loan_Amount_Term  < 270     to the right, improve=0.7907139, (0 missing)

Node number 30: 4 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.5  P(node) =0.01044386
    class counts:     2     2
   probabilities: 0.500 0.500 
  left son=60 (3 obs) right son=61 (1 obs)
  Primary splits:
      Married           splits as  RL,          improve=0.6666667, (0 missing)
      Dependents        splits as  LR-L,        improve=0.6666667, (0 missing)
      Education         splits as  LR,          improve=0.6666667, (0 missing)
      ApplicantIncome   < 2315    to the left,  improve=0.6666667, (0 missing)
      CoapplicantIncome < 6883.5  to the left,  improve=0.6666667, (0 missing)

Node number 31: 124 observations,    complexity param=0.006779661
  predicted class=1  expected loss=0.1048387  P(node) =0.3237598
    class counts:    13   111
   probabilities: 0.105 0.895 
  left son=62 (4 obs) right son=63 (120 obs)
  Primary splits:
      Loan_Amount_Term  < 420     to the right, improve=1.2908600, (0 missing)
      Married           splits as  LR,          improve=0.7504284, (0 missing)
      CoapplicantIncome < 1954    to the left,  improve=0.7137109, (0 missing)
      LoanAmount        < 99.5    to the left,  improve=0.5671228, (0 missing)
      ApplicantIncome   < 4209    to the right, improve=0.4301765, (0 missing)

Node number 34: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 35: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 38: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 39: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000 

Node number 58: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 
```

```
Node number 59: 192 observations,    complexity param=0.01271186
  predicted class=1  expected loss=0.234375  P(node) =0.5013055
    class counts:    45   147
   probabilities: 0.234 0.766 
  left son=118 (166 obs) right son=119 (26 obs)
  Primary splits:
      CoapplicantIncome < 3044    to the left,  improve=1.4910510, (0 missing)
      LoanAmount        < 13      to the left,  improve=1.1785010, (0 missing)
      ApplicantIncome   < 1490    to the left,  improve=0.8739919, (0 missing)
      Married           splits as  LR,          improve=0.8465485, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.7030252, (0 missing)
  Surrogate splits:
      LoanAmount < 307.5   to the left,  agree=0.875, adj=0.077, (0 split)

Node number 60: 3 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.3333333  P(node) =0.007832898
    class counts:     2     1
   probabilities: 0.667 0.333 
  left son=120 (2 obs) right son=121 (1 obs)
  Primary splits:
      Dependents        splits as  LR-L,        improve=1.3333330, (0 missing)
      Education         splits as  LR,          improve=1.3333330, (0 missing)
      ApplicantIncome   < 2395.5  to the left,  improve=0.3333333, (0 missing)
      CoapplicantIncome < 6883.5  to the left,  improve=0.3333333, (0 missing)
      LoanAmount        < 174.5   to the left,  improve=0.3333333, (0 missing)

Node number 61: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 62: 4 observations,    complexity param=0.006779661
  predicted class=0  expected loss=0.5  P(node) =0.01044386
    class counts:     2     2
   probabilities: 0.500 0.500 
  left son=124 (2 obs) right son=125 (2 obs)
  Primary splits:
      CoapplicantIncome < 1628.5  to the left,  improve=2.0000000, (0 missing)
      Gender            splits as  RL,          improve=0.6666667, (0 missing)
      Married           splits as  LR,          improve=0.6666667, (0 missing)
      Dependents        splits as  LRL-,        improve=0.6666667, (0 missing)
      Education         splits as  RL,          improve=0.6666667, (0 missing)

Node number 63: 120 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.09166667  P(node) =0.3133159
    class counts:    11   109
   probabilities: 0.092 0.908 
  left son=126 (36 obs) right son=127 (84 obs)
  Primary splits:
      Married           splits as  LR,          improve=0.5785714, (0 missing)
      CoapplicantIncome < 1954    to the left,  improve=0.5037512, (0 missing)
```

```
      LoanAmount        < 88.5    to the left,  improve=0.4765948, (0 missing)
      Dependents        splits as  RLRR,        improve=0.3849470, (0 missing)
      ApplicantIncome   < 4209    to the right, improve=0.3686701, (0 missing)
  Surrogate splits:
      LoanAmount < 85.5    to the left,  agree=0.758, adj=0.194, (0 split)
      Gender     splits as  LR,          agree=0.733, adj=0.111, (0 split)
      Education  splits as  RL,          agree=0.717, adj=0.056, (0 split)

Node number 118: 166 observations,    complexity param=0.01271186
  predicted class=1  expected loss=0.2590361  P(node) =0.4334204
    class counts:    43   123
   probabilities: 0.259 0.741 
  left son=236 (11 obs) right son=237 (155 obs)
  Primary splits:
      CoapplicantIncome < 2536    to the right, improve=3.3545630, (0 missing)
      ApplicantIncome   < 1490    to the left,  improve=1.1986680, (0 missing)
      LoanAmount        < 13      to the left,  improve=1.1047100, (0 missing)
      Loan_Amount_Term  < 420     to the right, improve=1.1047100, (0 missing)
      Property_Area     splits as  L-R,         improve=0.7962901, (0 missing)
  Surrogate splits:
      ApplicantIncome < 1162    to the left,  agree=0.946, adj=0.182, (0 split)

Node number 119: 26 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.07692308  P(node) =0.06788512
    class counts:     2    24
   probabilities: 0.077 0.923 
  left son=238 (6 obs) right son=239 (20 obs)
  Primary splits:
      Married           splits as  LR,          improve=1.0256410, (0 missing)
      LoanAmount        < 174     to the right, improve=0.4195804, (0 missing)
      ApplicantIncome   < 3713    to the right, improve=0.3589744, (0 missing)
      Dependents        splits as  LRRR,        improve=0.2637363, (0 missing)
      CoapplicantIncome < 3583    to the right, improve=0.1628959, (0 missing)
  Surrogate splits:
      Gender splits as  LR, agree=0.808, adj=0.167, (0 split)

Node number 120: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 121: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 124: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 
```

```
Node number 125: 2 observations
  predicted class=1  expected loss=0  P(node) =0.005221932
    class counts:     0     2
   probabilities: 0.000 1.000 

Node number 126: 36 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.1666667  P(node) =0.09399478
    class counts:     6    30
   probabilities: 0.167 0.833 
  left son=252 (26 obs) right son=253 (10 obs)
  Primary splits:
      ApplicantIncome   < 4809    to the left,  improve=0.7692308, (0 missing)
      Dependents        splits as  RLRR,        improve=0.6322581, (0 missing)
      Gender            splits as  LR,          improve=0.5142857, (0 missing)
      CoapplicantIncome < 1954    to the left,  improve=0.3225806, (0 missing)
      LoanAmount        < 58.5    to the right, improve=0.2500000, (0 missing)
  Surrogate splits:
      LoanAmount    < 149     to the left,  agree=0.861, adj=0.5, (0 split)
      Self_Employed splits as  LR,          agree=0.806, adj=0.3, (0 split)

Node number 127: 84 observations,    complexity param=0.005084746
  predicted class=1  expected loss=0.05952381  P(node) =0.2193211
    class counts:     5    79
   probabilities: 0.060 0.940 
  left son=254 (2 obs) right son=255 (82 obs)
  Primary splits:
      ApplicantIncome   < 26665   to the right, improve=0.79500580, (0 missing)
      CoapplicantIncome < 768     to the left,  improve=0.38593840, (0 missing)
      LoanAmount        < 147     to the right, improve=0.33531750, (0 missing)
      Dependents        splits as  RLRL,        improve=0.23500060, (0 missing)
      Gender            splits as  RL,          improve=0.08969341, (0 missing)

Node number 236: 11 observations,    complexity param=0.01271186
  predicted class=0  expected loss=0.3636364  P(node) =0.02872063
    class counts:     7     4
   probabilities: 0.636 0.364 
  left son=472 (4 obs) right son=473 (7 obs)
  Primary splits:
      Dependents        splits as  RLRL,        improve=1.6623380, (0 missing)
      ApplicantIncome   < 3532.5  to the right, improve=1.6623380, (0 missing)
      Education         splits as  LR,          improve=0.8909091, (0 missing)
      LoanAmount        < 112.5   to the right, improve=0.7575758, (0 missing)
      CoapplicantIncome < 2654    to the left,  improve=0.6464646, (0 missing)
  Surrogate splits:
      CoapplicantIncome < 2892    to the right, agree=0.909, adj=0.75, (0 split)
      ApplicantIncome   < 3995.5  to the right, agree=0.727, adj=0.25, (0 split)

Node number 237: 155 observations,    complexity param=0.01016949
  predicted class=1  expected loss=0.2322581  P(node) =0.4046997
    class counts:    36   119
   probabilities: 0.232 0.768 
```

```
    left son=474 (1 obs) right son=475 (154 obs)
    Primary splits:
        LoanAmount        < 13      to the left,  improve=1.1865100, (0 missing)
        ApplicantIncome   < 1490    to the left,  improve=1.1546120, (0 missing)
        Loan_Amount_Term  < 270     to the right, improve=1.0283190, (0 missing)
        CoapplicantIncome < 8.06    to the left,  improve=0.8466812, (0 missing)
        Property_Area     splits as  L-R,         improve=0.5572052, (0 missing)

Node number 238: 6 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.3333333  P(node) =0.0156658
    class counts:     2     4
   probabilities: 0.333 0.667 
  left son=476 (3 obs) right son=477 (3 obs)
  Primary splits:
      LoanAmount        < 174     to the right, improve=1.3333330, (0 missing)
      ApplicantIncome   < 4990.5  to the right, improve=1.0666670, (0 missing)
      Gender            splits as  RL,          improve=0.2666667, (0 missing)
      CoapplicantIncome < 3556.5  to the right, improve=0.2666667, (0 missing)
      Loan_Amount_Term  < 420     to the left,  improve=0.2666667, (0 missing)
  Surrogate splits:
      ApplicantIncome   < 3713    to the right, agree=0.833, adj=0.667, (0 split)
      CoapplicantIncome < 4525.5  to the right, agree=0.833, adj=0.667, (0 split)
      Gender            splits as  RL,          agree=0.667, adj=0.333, (0 split)

Node number 239: 20 observations
  predicted class=1  expected loss=0  P(node) =0.05221932
    class counts:     0    20
   probabilities: 0.000 1.000 

Node number 252: 26 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.2307692  P(node) =0.06788512
    class counts:     6    20
   probabilities: 0.231 0.769 
  left son=504 (2 obs) right son=505 (24 obs)
  Primary splits:
      Self_Employed   splits as  RL,          improve=2.5641030, (0 missing)
      Dependents      splits as  RLRR,        improve=1.2887400, (0 missing)
      ApplicantIncome < 4615    to the right, improve=1.2887400, (0 missing)
      LoanAmount      < 166.5   to the right, improve=1.2307690, (0 missing)
      Gender          splits as  LR,          improve=0.6731935, (0 missing)

Node number 253: 10 observations
  predicted class=1  expected loss=0  P(node) =0.02610966
    class counts:     0    10
   probabilities: 0.000 1.000 

Node number 254: 2 observations,    complexity param=0.005084746
  predicted class=0  expected loss=0.5  P(node) =0.005221932
    class counts:     1     1
   probabilities: 0.500 0.500 
  left son=508 (1 obs) right son=509 (1 obs)
```

```
    Primary splits:
        Dependents        splits as  RL--,        improve=1, (0 missing)
        Self_Employed     splits as  LR,          improve=1, (0 missing)
        ApplicantIncome   < 36496.5 to the left,  improve=1, (0 missing)
        CoapplicantIncome < 2375    to the left,  improve=1, (0 missing)
        LoanAmount        < 190     to the right, improve=1, (0 missing)

Node number 255: 82 observations,    complexity param=0.005084746
  predicted class=1  expected loss=0.04878049  P(node) =0.2140992
    class counts:     4    78
   probabilities: 0.049 0.951 
  left son=510 (13 obs) right son=511 (69 obs)
  Primary splits:
      ApplicantIncome   < 6365.5  to the right, improve=0.34108270, (0 missing)
      LoanAmount        < 88.5    to the left,  improve=0.34052530, (0 missing)
      CoapplicantIncome < 768     to the left,  improve=0.19602700, (0 missing)
      Dependents        splits as  LLRL,        improve=0.10975610, (0 missing)
      Education         splits as  RL,          improve=0.07855366, (0 missing)
  Surrogate splits:
      LoanAmount < 189     to the right, agree=0.89, adj=0.308, (0 split)

Node number 472: 4 observations
  predicted class=0  expected loss=0  P(node) =0.01044386
    class counts:     4     0
   probabilities: 1.000 0.000 

Node number 473: 7 observations,    complexity param=0.01271186
  predicted class=1  expected loss=0.4285714  P(node) =0.01827676
    class counts:     3     4
   probabilities: 0.429 0.571 
  left son=946 (2 obs) right son=947 (5 obs)
  Primary splits:
      ApplicantIncome   < 3532.5  to the right, improve=1.8285710, (0 missing)
      CoapplicantIncome < 2654    to the left,  improve=1.8285710, (0 missing)
      LoanAmount        < 113     to the right, improve=1.0285710, (0 missing)
      Gender            splits as  LR,          improve=0.7619048, (0 missing)
      Loan_Amount_Term  < 420     to the right, improve=0.7619048, (0 missing)
  Surrogate splits:
      CoapplicantIncome < 2654    to the left,  agree=1.000, adj=1.0, (0 split)
      LoanAmount        < 147.5   to the right, agree=0.857, adj=0.5, (0 split)

Node number 474: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 475: 154 observations,    complexity param=0.01016949
  predicted class=1  expected loss=0.2272727  P(node) =0.4020888
    class counts:    35   119
   probabilities: 0.227 0.773 
  left son=950 (3 obs) right son=951 (151 obs)
```

```
    Primary splits:
        ApplicantIncome   < 1490    to the left,  improve=1.1814170, (0 missing)
        Loan_Amount_Term  < 270     to the right, improve=0.9695323, (0 missing)
        CoapplicantIncome < 8.06    to the left,  improve=0.7118810, (0 missing)
        Property_Area     splits as  L-R,         improve=0.6898251, (0 missing)
        LoanAmount        < 62.5    to the right, improve=0.4593075, (0 missing)

Node number 476: 3 observations,    complexity param=0.005649718
  predicted class=0  expected loss=0.3333333  P(node) =0.007832898
    class counts:     2     1
   probabilities: 0.667 0.333
  left son=952 (2 obs) right son=953 (1 obs)
  Primary splits:
      CoapplicantIncome < 5980    to the left,  improve=1.3333330, (0 missing)
      ApplicantIncome   < 3958    to the left,  improve=0.3333333, (0 missing)
      LoanAmount        < 180     to the left,  improve=0.3333333, (0 missing)
      Property_Area     splits as  L-R,         improve=0.3333333, (0 missing)

Node number 477: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000

Node number 504: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000

Node number 505: 24 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.1666667  P(node) =0.06266319
    class counts:     4    20
   probabilities: 0.167 0.833
  left son=1010 (6 obs) right son=1011 (18 obs)
  Primary splits:
      ApplicantIncome   < 4198    to the right, improve=1.7777780, (0 missing)
      LoanAmount        < 166.5   to the right, improve=1.4492750, (0 missing)
      Education         splits as  LR,          improve=0.6666667, (0 missing)
      Dependents        splits as  RLRR,        improve=0.4848485, (0 missing)
      CoapplicantIncome < 1954    to the left,  improve=0.2666667, (0 missing)
  Surrogate splits:
      LoanAmount < 146     to the right, agree=0.833, adj=0.333, (0 split)
      Dependents splits as  RRRL,        agree=0.792, adj=0.167, (0 split)

Node number 508: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000

Node number 509: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
```

```
     probabilities: 0.000 1.000

Node number 510: 13 observations,    complexity param=0.005084746
  predicted class=1  expected loss=0.1538462  P(node) =0.03394256
    class counts:     2    11
   probabilities: 0.154 0.846
  left son=1020 (3 obs) right son=1021 (10 obs)
  Primary splits:
      LoanAmount      < 175.5   to the left,  improve=2.0512820, (0 missing)
      ApplicantIncome < 6473    to the left,  improve=1.5512820, (0 missing)
      Dependents      splits as  LRRL,        improve=0.7179487, (0 missing)
      Self_Employed   splits as  LR,          improve=0.1846154, (0 missing)
      Gender          splits as  RL,          improve=0.1118881, (0 missing)

Node number 511: 69 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.02898551  P(node) =0.1801567
    class counts:     2    67
   probabilities: 0.029 0.971
  left son=1022 (4 obs) right son=1023 (65 obs)
  Primary splits:
      LoanAmount      < 88.5    to the left,  improve=0.41482720, (0 missing)
      Self_Employed   splits as  RL,          improve=0.20203030, (0 missing)
      Education       splits as  RL,          improve=0.13961350, (0 missing)
      Dependents      splits as  RLRR,        improve=0.06327875, (0 missing)
      ApplicantIncome < 3041.5  to the right, improve=0.04732328, (0 missing)

Node number 946: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000

Node number 947: 5 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.2  P(node) =0.01305483
    class counts:     1     4
   probabilities: 0.200 0.800
  left son=1894 (2 obs) right son=1895 (3 obs)
  Primary splits:
      Married           splits as  LR,          improve=0.6000000, (0 missing)
      ApplicantIncome   < 2166    to the left,  improve=0.6000000, (0 missing)
      Property_Area     splits as  L-R,         improve=0.6000000, (0 missing)
      CoapplicantIncome < 2806.5  to the right, improve=0.2666667, (0 missing)
      LoanAmount        < 113     to the right, improve=0.2666667, (0 missing)
  Surrogate splits:
      CoapplicantIncome < 2849.5  to the left,  agree=0.8, adj=0.5, (0 split)
      LoanAmount        < 129     to the left,  agree=0.8, adj=0.5, (0 split)

Node number 950: 3 observations,    complexity param=0.008474576
  predicted class=0  expected loss=0.3333333  P(node) =0.007832898
    class counts:     2     1
   probabilities: 0.667 0.333
  left son=1900 (2 obs) right son=1901 (1 obs)
```

```
    Primary splits:
        ApplicantIncome  < 1338.5  to the right, improve=1.3333330, (0 missing)
        LoanAmount       < 26      to the right, improve=1.3333330, (0 missing)
        Loan_Amount_Term < 240     to the right, improve=1.3333330, (0 missing)
        Gender           splits as  LR,          improve=0.3333333, (0 missing)
        Married          splits as  LR,          improve=0.3333333, (0 missing)

Node number 951: 151 observations,    complexity param=0.01016949
  predicted class=1  expected loss=0.218543  P(node) =0.3942559
    class counts:    33   118
   probabilities: 0.219 0.781 
  left son=1902 (107 obs) right son=1903 (44 obs)
  Primary splits:
      ApplicantIncome   < 3163    to the right, improve=1.3667280, (0 missing)
      Property_Area     splits as  L-R,         improve=0.9657511, (0 missing)
      LoanAmount        < 62.5    to the right, improve=0.7683158, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.7683158, (0 missing)
      CoapplicantIncome < 8.06    to the left,  improve=0.7119088, (0 missing)
  Surrogate splits:
      LoanAmount        < 107.5   to the right, agree=0.748, adj=0.136, (0 split)
      CoapplicantIncome < 1453    to the left,  agree=0.728, adj=0.068, (0 split)

Node number 952: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 953: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 1010: 6 observations,    complexity param=0.005649718
  predicted class=0  expected loss=0.5  P(node) =0.0156658
    class counts:     3     3
   probabilities: 0.500 0.500 
  left son=2020 (4 obs) right son=2021 (2 obs)
  Primary splits:
      Education         splits as  LR,          improve=1.5, (0 missing)
      LoanAmount        < 113.5   to the left,  improve=1.5, (0 missing)
      Dependents        splits as  LL-R,        improve=0.6, (0 missing)
      ApplicantIncome   < 4319    to the left,  improve=0.6, (0 missing)
      CoapplicantIncome < 1954    to the left,  improve=0.6, (0 missing)
  Surrogate splits:
      LoanAmount < 117.5   to the left,  agree=0.833, adj=0.5, (0 split)

Node number 1011: 18 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.05555556  P(node) =0.04699739
    class counts:     1    17
   probabilities: 0.056 0.944 
  left son=2022 (4 obs) right son=2023 (14 obs)
```

```
    Primary splits:
        LoanAmount        < 123.5   to the right, improve=0.38888890, (0 missing)
        ApplicantIncome   < 3288.5  to the right, improve=0.13888890, (0 missing)
        Education         splits as  LR,          improve=0.05555556, (0 missing)
        CoapplicantIncome < 651     to the left,  improve=0.05555556, (0 missing)
        Gender            splits as  RL,          improve=0.04273504, (0 missing)
    Surrogate splits:
        CoapplicantIncome < 2468    to the right, agree=0.889, adj=0.50, (0 split)
        ApplicantIncome   < 2522.5  to the left,  agree=0.833, adj=0.25, (0 split)

  Node number 1020: 3 observations,    complexity param=0.005084746
    predicted class=0  expected loss=0.3333333  P(node) =0.007832898
      class counts:     2     1
     probabilities: 0.667 0.333
    left son=2040 (2 obs) right son=2041 (1 obs)
    Primary splits:
        ApplicantIncome < 11607   to the left,  improve=1.3333330, (0 missing)
        LoanAmount      < 123     to the right, improve=1.3333330, (0 missing)
        Dependents      splits as  R--L,        improve=0.3333333, (0 missing)

  Node number 1021: 10 observations
    predicted class=1  expected loss=0  P(node) =0.02610966
      class counts:     0    10
     probabilities: 0.000 1.000

  Node number 1022: 4 observations,    complexity param=0.004237288
    predicted class=1  expected loss=0.25  P(node) =0.01044386
      class counts:     1     3
     probabilities: 0.250 0.750
    left son=2044 (1 obs) right son=2045 (3 obs)
    Primary splits:
        ApplicantIncome   < 3901.5  to the right, improve=1.5, (0 missing)
        CoapplicantIncome < 368     to the right, improve=1.5, (0 missing)
        LoanAmount        < 84      to the right, improve=1.5, (0 missing)
        Dependents        splits as  L-RR,        improve=0.5, (0 missing)
        Education         splits as  RL,          improve=0.5, (0 missing)

  Node number 1023: 65 observations,    complexity param=0.002824859
    predicted class=1  expected loss=0.01538462  P(node) =0.1697128
      class counts:     1    64
     probabilities: 0.015 0.985
    left son=2046 (14 obs) right son=2047 (51 obs)
    Primary splits:
        Dependents        splits as  RLRR,        improve=0.112087900, (0 missing)
        LoanAmount        < 166     to the right, improve=0.102564100, (0 missing)
        CoapplicantIncome < 2541.5  to the right, improve=0.094230770, (0 missing)
        ApplicantIncome   < 3140    to the left,  improve=0.069230770, (0 missing)
        Gender            splits as  RL,          improve=0.004945055, (0 missing)
    Surrogate splits:
        ApplicantIncome < 6000    to the right, agree=0.8, adj=0.071, (0 split)
        LoanAmount      < 95.5    to the left,  agree=0.8, adj=0.071, (0 split)
```

```
Node number 1894: 2 observations,    complexity param=0.004237288
  predicted class=0  expected loss=0.5  P(node) =0.005221932
    class counts:     1     1
   probabilities: 0.500 0.500
  left son=3788 (1 obs) right son=3789 (1 obs)
  Primary splits:
      ApplicantIncome   < 2541    to the left,  improve=1, (0 missing)
      CoapplicantIncome < 2789.5  to the right, improve=1, (0 missing)
      LoanAmount        < 112     to the right, improve=1, (0 missing)
      Property_Area     splits as  L-R,         improve=1, (0 missing)

Node number 1895: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000

Node number 1900: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000

Node number 1901: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000

Node number 1902: 107 observations,    complexity param=0.01016949
  predicted class=1  expected loss=0.2616822  P(node) =0.2793734
    class counts:    28    79
   probabilities: 0.262 0.738
  left son=3804 (10 obs) right son=3805 (97 obs)
  Primary splits:
      CoapplicantIncome < 2068.5  to the right, improve=2.525176, (0 missing)
      ApplicantIncome   < 5283    to the left,  improve=1.968436, (0 missing)
      Loan_Amount_Term  < 300     to the right, improve=1.679128, (0 missing)
      LoanAmount        < 61      to the right, improve=1.184178, (0 missing)
      Property_Area     splits as  L-R,         improve=0.861179, (0 missing)

Node number 1903: 44 observations,    complexity param=0.008474576
  predicted class=1  expected loss=0.1136364  P(node) =0.1148825
    class counts:     5    39
   probabilities: 0.114 0.886
  left son=3806 (14 obs) right son=3807 (30 obs)
  Primary splits:
      CoapplicantIncome < 8.06    to the left,  improve=1.2160170, (0 missing)
      Loan_Amount_Term  < 330     to the left,  improve=0.9251748, (0 missing)
      ApplicantIncome   < 2770    to the left,  improve=0.6493506, (0 missing)
      LoanAmount        < 50      to the left,  improve=0.6255411, (0 missing)
      Education         splits as  RL,          improve=0.6136364, (0 missing)
  Surrogate splits:
```

```
      LoanAmount < 72.5    to the left,  agree=0.864, adj=0.571, (0 split)
      Gender     splits as  LR,          agree=0.795, adj=0.357, (0 split)
      Married    splits as  LR,          agree=0.795, adj=0.357, (0 split)
      Dependents splits as  RRRL,        agree=0.705, adj=0.071, (0 split)

Node number 2020: 4 observations,    complexity param=0.004237288
  predicted class=0  expected loss=0.25  P(node) =0.01044386
    class counts:     3     1
   probabilities: 0.750 0.250
  left son=4040 (2 obs) right son=4041 (2 obs)
  Primary splits:
      ApplicantIncome   < 4615    to the right, improve=0.5000000, (0 missing)
      LoanAmount        < 113.5   to the left,  improve=0.5000000, (0 missing)
      Gender            splits as  RL,          improve=0.1666667, (0 missing)
      Dependents        splits as  RL--,        improve=0.1666667, (0 missing)
      CoapplicantIncome < 957.5   to the right, improve=0.1666667, (0 missing)

Node number 2021: 2 observations
  predicted class=1  expected loss=0  P(node) =0.005221932
    class counts:     0     2
   probabilities: 0.000 1.000

Node number 2022: 4 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.25  P(node) =0.01044386
    class counts:     1     3
   probabilities: 0.250 0.750
  left son=4044 (1 obs) right son=4045 (3 obs)
  Primary splits:
      ApplicantIncome   < 3194.5  to the right, improve=1.5000000, (0 missing)
      CoapplicantIncome < 800     to the left,  improve=1.5000000, (0 missing)
      LoanAmount        < 129.5   to the left,  improve=1.5000000, (0 missing)
      Education         splits as  LR,          improve=0.1666667, (0 missing)

Node number 2023: 14 observations
  predicted class=1  expected loss=0  P(node) =0.03655352
    class counts:     0    14
   probabilities: 0.000 1.000

Node number 2040: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000

Node number 2041: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000

Node number 2044: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
```

```
   probabilities: 1.000 0.000 

Node number 2045: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000 

Node number 2046: 14 observations,    complexity param=0.002824859
  predicted class=1  expected loss=0.07142857  P(node) =0.03655352
    class counts:     1    13
   probabilities: 0.071 0.929 
  left son=4092 (2 obs) right son=4093 (12 obs)
  Primary splits:
      CoapplicantIncome < 2541.5  to the right, improve=0.85714290, (0 missing)
      LoanAmount        < 166     to the right, improve=0.35714290, (0 missing)
      ApplicantIncome   < 3140    to the left,  improve=0.25714290, (0 missing)
      Gender            splits as  RL,          improve=0.02380952, (0 missing)
      Self_Employed     splits as  LR,          improve=0.02380952, (0 missing)

Node number 2047: 51 observations
  predicted class=1  expected loss=0  P(node) =0.1331593
    class counts:     0    51
   probabilities: 0.000 1.000 

Node number 3788: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 3789: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 3804: 10 observations,    complexity param=0.01016949
  predicted class=0  expected loss=0.4  P(node) =0.02610966
    class counts:     6     4
   probabilities: 0.600 0.400 
  left son=7608 (4 obs) right son=7609 (6 obs)
  Primary splits:
      CoapplicantIncome < 2225    to the left,  improve=2.133333, (0 missing)
      Married           splits as  RL,          improve=1.800000, (0 missing)
      Dependents        splits as  LRL-,        improve=0.800000, (0 missing)
      Education         splits as  LR,          improve=0.800000, (0 missing)
      LoanAmount        < 82.5    to the right, improve=0.800000, (0 missing)
  Surrogate splits:
      ApplicantIncome < 4896    to the right, agree=0.8, adj=0.50, (0 split)
      LoanAmount      < 132.5   to the right, agree=0.8, adj=0.50, (0 split)
      Property_Area   splits as  L-R,         agree=0.8, adj=0.50, (0 split)
      Dependents      splits as  RRL-,        agree=0.7, adj=0.25, (0 split)
```

```
Node number 3805: 97 observations,    complexity param=0.01016949
  predicted class=1  expected loss=0.2268041  P(node) =0.2532637
    class counts:    22    75
   probabilities: 0.227 0.773
  left son=7610 (22 obs) right son=7611 (75 obs)
  Primary splits:
      LoanAmount        < 107     to the left,  improve=1.8909220, (0 missing)
      ApplicantIncome   < 5283    to the left,  improve=1.7247960, (0 missing)
      Loan_Amount_Term  < 300     to the right, improve=1.2764330, (0 missing)
      CoapplicantIncome < 1517    to the left,  improve=1.1470550, (0 missing)
      Property_Area     splits as  L-R,         improve=0.5449152, (0 missing)
  Surrogate splits:
      ApplicantIncome < 3247    to the left,  agree=0.794, adj=0.091, (0 split)

Node number 3806: 14 observations,    complexity param=0.008474576
  predicted class=1  expected loss=0.2857143  P(node) =0.03655352
    class counts:     4    10
   probabilities: 0.286 0.714
  left son=7612 (4 obs) right son=7613 (10 obs)
  Primary splits:
      ApplicantIncome < 2457    to the left,  improve=2.4142860, (0 missing)
      Dependents      splits as  R-LR,        improve=1.0989010, (0 missing)
      LoanAmount      < 51      to the left,  improve=1.0989010, (0 missing)
      Gender          splits as  RL,          improve=0.5714286, (0 missing)
      Property_Area   splits as  L-R,         improve=0.2976190, (0 missing)
  Surrogate splits:
      Dependents splits as  R-LR, agree=0.786, adj=0.25, (0 split)

Node number 3807: 30 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.03333333  P(node) =0.07832898
    class counts:     1    29
   probabilities: 0.033 0.967
  left son=7614 (3 obs) right son=7615 (27 obs)
  Primary splits:
      LoanAmount        < 139.5   to the right, improve=0.6000000, (0 missing)
      Loan_Amount_Term  < 270     to the left,  improve=0.6000000, (0 missing)
      Dependents        splits as  RLRR,        improve=0.2190476, (0 missing)
      Education         splits as  RL,          improve=0.2190476, (0 missing)
      CoapplicantIncome < 1956    to the right, improve=0.1333333, (0 missing)

Node number 4040: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000

Node number 4041: 2 observations,    complexity param=0.004237288
  predicted class=0  expected loss=0.5  P(node) =0.005221932
    class counts:     1     1
   probabilities: 0.500 0.500
  left son=8082 (1 obs) right son=8083 (1 obs)
  Primary splits:
```

```
      ApplicantIncome < 4388.5  to the left,  improve=1, (0 missing)
      LoanAmount      < 113.5   to the left,  improve=1, (0 missing)

Node number 4044: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 4045: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000 

Node number 4092: 2 observations,    complexity param=0.002824859
  predicted class=0  expected loss=0.5  P(node) =0.005221932
    class counts:     1     1
   probabilities: 0.500 0.500 
  left son=8184 (1 obs) right son=8185 (1 obs)
  Primary splits:
      Gender            splits as  RL,          improve=1, (0 missing)
      ApplicantIncome   < 3866.5  to the left,  improve=1, (0 missing)
      CoapplicantIncome < 2714    to the left,  improve=1, (0 missing)
      LoanAmount        < 155     to the right, improve=1, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=1, (0 missing)

Node number 4093: 12 observations
  predicted class=1  expected loss=0  P(node) =0.03133159
    class counts:     0    12
   probabilities: 0.000 1.000 

Node number 7608: 4 observations
  predicted class=0  expected loss=0  P(node) =0.01044386
    class counts:     4     0
   probabilities: 1.000 0.000 

Node number 7609: 6 observations,    complexity param=0.008474576
  predicted class=1  expected loss=0.3333333  P(node) =0.0156658
    class counts:     2     4
   probabilities: 0.333 0.667 
  left son=15218 (3 obs) right son=15219 (3 obs)
  Primary splits:
      LoanAmount        < 120     to the left,  improve=1.3333330, (0 missing)
      Gender            splits as  LR,          improve=1.0666670, (0 missing)
      Married           splits as  RL,          improve=0.6666667, (0 missing)
      ApplicantIncome   < 4166.5  to the right, improve=0.6666667, (0 missing)
      CoapplicantIncome < 2392    to the right, improve=0.6666667, (0 missing)
  Surrogate splits:
      ApplicantIncome   < 4541.5  to the left,  agree=0.833, adj=0.667, (0 split)
      CoapplicantIncome < 2392    to the right, agree=0.833, adj=0.667, (0 split)
      Gender            splits as  LR,          agree=0.667, adj=0.333, (0 split)
      Dependents        splits as  LR--,        agree=0.667, adj=0.333, (0 split)
```

```
      Education         splits as  RL,          agree=0.667, adj=0.333, (0 split)

Node number 7610: 22 observations,    complexity param=0.01016949
  predicted class=1  expected loss=0.4090909  P(node) =0.05744125
    class counts:     9    13
   probabilities: 0.409 0.591 
  left son=15220 (15 obs) right son=15221 (7 obs)
  Primary splits:
      LoanAmount       < 58.5    to the right, improve=3.4363640, (0 missing)
      Loan_Amount_Term < 300     to the right, improve=1.1626790, (0 missing)
      ApplicantIncome  < 5449.5  to the left,  improve=0.9696970, (0 missing)
      Dependents       splits as  LLRR,        improve=0.5657754, (0 missing)
      Property_Area    splits as  L-R,         improve=0.5411255, (0 missing)
  Surrogate splits:
      Loan_Amount_Term < 300     to the right, agree=0.727, adj=0.143, (0 split)

Node number 7611: 75 observations,    complexity param=0.006355932
  predicted class=1  expected loss=0.1733333  P(node) =0.1958225
    class counts:    13    62
   probabilities: 0.173 0.827 
  left son=15222 (65 obs) right son=15223 (10 obs)
  Primary splits:
      ApplicantIncome   < 10204   to the left,  improve=0.6933333, (0 missing)
      CoapplicantIncome < 1517    to the left,  improve=0.6933333, (0 missing)
      LoanAmount        < 126.5   to the right, improve=0.6570760, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.5381095, (0 missing)
      Property_Area     splits as  L-R,         improve=0.5360684, (0 missing)
  Surrogate splits:
      LoanAmount < 256     to the left,  agree=0.893, adj=0.2, (0 split)

Node number 7612: 4 observations,    complexity param=0.008474576
  predicted class=0  expected loss=0.25  P(node) =0.01044386
    class counts:     3     1
   probabilities: 0.750 0.250 
  left son=15224 (3 obs) right son=15225 (1 obs)
  Primary splits:
      ApplicantIncome < 2247    to the right, improve=1.5000000, (0 missing)
      Gender          splits as  RL,          improve=0.5000000, (0 missing)
      Education       splits as  LR,          improve=0.5000000, (0 missing)
      LoanAmount      < 70.5    to the right, improve=0.5000000, (0 missing)
      Married         splits as  RL,          improve=0.1666667, (0 missing)

Node number 7613: 10 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.1  P(node) =0.02610966
    class counts:     1     9
   probabilities: 0.100 0.900 
  left son=15226 (2 obs) right son=15227 (8 obs)
  Primary splits:
      Loan_Amount_Term < 330     to the left,  improve=0.8000000, (0 missing)
      Education        splits as  RL,          improve=0.4666667, (0 missing)
      Gender           splits as  RL,          improve=0.2000000, (0 missing)
```

```
      ApplicantIncome  < 2827.5  to the left,  improve=0.2000000, (0 missing)
      LoanAmount       < 65.5    to the left,  improve=0.2000000, (0 missing)

Node number 7614: 3 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.3333333  P(node) =0.007832898
    class counts:     1     2
   probabilities: 0.333 0.667 
  left son=15228 (1 obs) right son=15229 (2 obs)
  Primary splits:
      Dependents       splits as  RL--,        improve=1.333333, (0 missing)
      Education        splits as  RL,          improve=1.333333, (0 missing)
      ApplicantIncome  < 2518    to the left,  improve=1.333333, (0 missing)
      LoanAmount       < 142.5   to the left,  improve=1.333333, (0 missing)
      Loan_Amount_Term < 270     to the left,  improve=1.333333, (0 missing)

Node number 7615: 27 observations
  predicted class=1  expected loss=0  P(node) =0.07049608
    class counts:     0    27
   probabilities: 0.000 1.000 

Node number 8082: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 8083: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 8184: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 8185: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 15218: 3 observations,    complexity param=0.008474576
  predicted class=0  expected loss=0.3333333  P(node) =0.007832898
    class counts:     2     1
   probabilities: 0.667 0.333 
  left son=30436 (2 obs) right son=30437 (1 obs)
  Primary splits:
      Married           splits as  RL,          improve=1.333333, (0 missing)
      Education         splits as  LR,          improve=1.333333, (0 missing)
      ApplicantIncome   < 4154    to the right, improve=1.333333, (0 missing)
      CoapplicantIncome < 2491    to the left,  improve=1.333333, (0 missing)
      LoanAmount        < 82.5    to the right, improve=1.333333, (0 missing)
```

```
Node number 15219: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000 

Node number 15220: 15 observations,    complexity param=0.01016949
  predicted class=0  expected loss=0.4  P(node) =0.03916449
    class counts:     9     6
   probabilities: 0.600 0.400 
  left son=30440 (11 obs) right son=30441 (4 obs)
  Primary splits:
      Dependents        splits as  LLRR,        improve=1.3363640, (0 missing)
      ApplicantIncome   < 5463.5  to the left,  improve=1.3363640, (0 missing)
      Married           splits as  LR,          improve=0.7714286, (0 missing)
      CoapplicantIncome < 1355.5  to the left,  improve=0.7714286, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.7714286, (0 missing)
  Surrogate splits:
      Married         splits as  LR,          agree=0.8, adj=0.25, (0 split)
      ApplicantIncome < 4971    to the left,  agree=0.8, adj=0.25, (0 split)

Node number 15221: 7 observations
  predicted class=1  expected loss=0  P(node) =0.01827676
    class counts:     0     7
   probabilities: 0.000 1.000 

Node number 15222: 65 observations,    complexity param=0.006355932
  predicted class=1  expected loss=0.2  P(node) =0.1697128
    class counts:    13    52
   probabilities: 0.200 0.800 
  left son=30444 (1 obs) right son=30445 (64 obs)
  Primary splits:
      ApplicantIncome   < 9981.5  to the right, improve=1.3000000, (0 missing)
      CoapplicantIncome < 1517    to the left,  improve=0.9454545, (0 missing)
      LoanAmount        < 126.5   to the right, improve=0.9176471, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.5288136, (0 missing)
      Property_Area     splits as  L-R,         improve=0.4034483, (0 missing)

Node number 15223: 10 observations
  predicted class=1  expected loss=0  P(node) =0.02610966
    class counts:     0    10
   probabilities: 0.000 1.000 

Node number 15224: 3 observations
  predicted class=0  expected loss=0  P(node) =0.007832898
    class counts:     3     0
   probabilities: 1.000 0.000 

Node number 15225: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
```

```
     probabilities: 0.000 1.000

  Node number 15226: 2 observations,    complexity param=0.004237288
    predicted class=0  expected loss=0.5  P(node) =0.005221932
      class counts:     1     1
     probabilities: 0.500 0.500
    left son=30452 (1 obs) right son=30453 (1 obs)
    Primary splits:
        ApplicantIncome  < 2736    to the right, improve=1, (0 missing)
        LoanAmount       < 62.5    to the right, improve=1, (0 missing)
        Loan_Amount_Term < 240     to the right, improve=1, (0 missing)
        Property_Area    splits as  L-R,         improve=1, (0 missing)

  Node number 15227: 8 observations
    predicted class=1  expected loss=0  P(node) =0.02088773
      class counts:     0     8
     probabilities: 0.000 1.000

  Node number 15228: 1 observations
    predicted class=0  expected loss=0  P(node) =0.002610966
      class counts:     1     0
     probabilities: 1.000 0.000

  Node number 15229: 2 observations
    predicted class=1  expected loss=0  P(node) =0.005221932
      class counts:     0     2
     probabilities: 0.000 1.000

  Node number 30436: 2 observations
    predicted class=0  expected loss=0  P(node) =0.005221932
      class counts:     2     0
     probabilities: 1.000 0.000

  Node number 30437: 1 observations
    predicted class=1  expected loss=0  P(node) =0.002610966
      class counts:     0     1
     probabilities: 0.000 1.000

  Node number 30440: 11 observations,    complexity param=0.008474576
    predicted class=0  expected loss=0.2727273  P(node) =0.02872063
      class counts:     8     3
     probabilities: 0.727 0.273
    left son=60880 (10 obs) right son=60881 (1 obs)
    Primary splits:
        Loan_Amount_Term < 270     to the right, improve=1.1636360, (0 missing)
        Property_Area    splits as  L-R,         improve=0.9350649, (0 missing)
        Dependents       splits as  RL--,        improve=0.6136364, (0 missing)
        Education        splits as  RL,          improve=0.3636364, (0 missing)
        Self_Employed    splits as  RL,          improve=0.3636364, (0 missing)

  Node number 30441: 4 observations,    complexity param=0.004237288
```

```
      predicted class=1  expected loss=0.25  P(node) =0.01044386
        class counts:     1     3
       probabilities: 0.250 0.750
      left son=60882 (2 obs) right son=60883 (2 obs)
      Primary splits:
          ApplicantIncome < 5463.5  to the left,  improve=0.5000000, (0 missing)
          LoanAmount      < 97      to the left,  improve=0.5000000, (0 missing)
          Gender          splits as  RL,          improve=0.1666667, (0 missing)
          Dependents      splits as  --RL,        improve=0.1666667, (0 missing)
          Self_Employed   splits as  LR,          improve=0.1666667, (0 missing)

    Node number 30444: 1 observations
      predicted class=0  expected loss=0  P(node) =0.002610966
        class counts:     1     0
       probabilities: 1.000 0.000

    Node number 30445: 64 observations,    complexity param=0.006355932
      predicted class=1  expected loss=0.1875  P(node) =0.1671018
        class counts:    12    52
       probabilities: 0.188 0.813
      left son=60890 (54 obs) right son=60891 (10 obs)
      Primary splits:
          CoapplicantIncome < 1517    to the left,  improve=0.8333333, (0 missing)
          LoanAmount        < 242     to the right, improve=0.8333333, (0 missing)
          Dependents        splits as  LRLR,        improve=0.6666667, (0 missing)
          ApplicantIncome   < 3446.5  to the left,  improve=0.4898305, (0 missing)
          Loan_Amount_Term  < 270     to the right, improve=0.4655172, (0 missing)
      Surrogate splits:
          ApplicantIncome < 3422    to the right, agree=0.875, adj=0.2, (0 split)
          LoanAmount      < 109.5   to the right, agree=0.859, adj=0.1, (0 split)

    Node number 30452: 1 observations
      predicted class=0  expected loss=0  P(node) =0.002610966
        class counts:     1     0
       probabilities: 1.000 0.000

    Node number 30453: 1 observations
      predicted class=1  expected loss=0  P(node) =0.002610966
        class counts:     0     1
       probabilities: 0.000 1.000

    Node number 60880: 10 observations,    complexity param=0.005649718
      predicted class=0  expected loss=0.2  P(node) =0.02610966
        class counts:     8     2
       probabilities: 0.800 0.200
      left son=121760 (5 obs) right son=121761 (5 obs)
      Primary splits:
          ApplicantIncome < 3812.5  to the right, improve=0.8000000, (0 missing)
          Property_Area   splits as  L-R,         improve=0.5333333, (0 missing)
          Dependents      splits as  RL--,        improve=0.3428571, (0 missing)
          Married         splits as  RL,          improve=0.2000000, (0 missing)
```

```
      Education       splits as  RL,          improve=0.2000000, (0 missing)
    Surrogate splits:
      Dependents    splits as  RL--,        agree=0.8, adj=0.6, (0 split)
      Married       splits as  RL,          agree=0.7, adj=0.4, (0 split)
      Self_Employed splits as  RL,          agree=0.7, adj=0.4, (0 split)
      LoanAmount    < 80.5    to the right, agree=0.7, adj=0.4, (0 split)
      Property_Area splits as  L-R,         agree=0.7, adj=0.4, (0 split)

Node number 60881: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000

Node number 60882: 2 observations,    complexity param=0.004237288
  predicted class=0  expected loss=0.5  P(node) =0.005221932
    class counts:     1     1
   probabilities: 0.500 0.500
  left son=121764 (1 obs) right son=121765 (1 obs)
  Primary splits:
      ApplicantIncome < 4765.5  to the right, improve=1, (0 missing)
      LoanAmount      < 97      to the left,  improve=1, (0 missing)

Node number 60883: 2 observations
  predicted class=1  expected loss=0  P(node) =0.005221932
    class counts:     0     2
   probabilities: 0.000 1.000

Node number 60890: 54 observations,    complexity param=0.006355932
  predicted class=1  expected loss=0.2222222  P(node) =0.1409922
    class counts:    12    42
   probabilities: 0.222 0.778
  left son=121780 (2 obs) right son=121781 (52 obs)
  Primary splits:
      ApplicantIncome   < 3446.5  to the left,  improve=2.5128210, (0 missing)
      CoapplicantIncome < 1420.5  to the right, improve=1.2549020, (0 missing)
      Dependents        splits as  LRLR,        improve=0.7229518, (0 missing)
      LoanAmount        < 242     to the right, improve=0.6666667, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.5442177, (0 missing)

Node number 60891: 10 observations
  predicted class=1  expected loss=0  P(node) =0.02610966
    class counts:     0    10
   probabilities: 0.000 1.000

Node number 121760: 5 observations
  predicted class=0  expected loss=0  P(node) =0.01305483
    class counts:     5     0
   probabilities: 1.000 0.000

Node number 121761: 5 observations,    complexity param=0.005649718
  predicted class=0  expected loss=0.4  P(node) =0.01305483
```

```
      class counts:     3     2
     probabilities: 0.600 0.400
    left son=243522 (4 obs) right son=243523 (1 obs)
    Primary splits:
        Gender            splits as  RL,          improve=0.9, (0 missing)
        ApplicantIncome   < 3675    to the left,  improve=0.9, (0 missing)
        Education         splits as  RL,          improve=0.4, (0 missing)
        CoapplicantIncome < 643.5   to the right, improve=0.4, (0 missing)
        LoanAmount        < 77      to the left,  improve=0.4, (0 missing)

  Node number 121764: 1 observations
    predicted class=0  expected loss=0  P(node) =0.002610966
      class counts:     1     0
     probabilities: 1.000 0.000

  Node number 121765: 1 observations
    predicted class=1  expected loss=0  P(node) =0.002610966
      class counts:     0     1
     probabilities: 0.000 1.000

  Node number 121780: 2 observations
    predicted class=0  expected loss=0  P(node) =0.005221932
      class counts:     2     0
     probabilities: 1.000 0.000

  Node number 121781: 52 observations,    complexity param=0.006355932
    predicted class=1  expected loss=0.1923077  P(node) =0.1357702
      class counts:    10    42
     probabilities: 0.192 0.808
    left son=243562 (3 obs) right son=243563 (49 obs)
    Primary splits:
        CoapplicantIncome < 1420.5  to the right, improve=1.4327580, (0 missing)
        LoanAmount        < 242     to the right, improve=0.8205128, (0 missing)
        ApplicantIncome   < 4183.5  to the right, improve=0.5016722, (0 missing)
        Dependents        splits as  LRLR,        improve=0.4615385, (0 missing)
        Loan_Amount_Term  < 270     to the right, improve=0.4091653, (0 missing)

  Node number 243522: 4 observations,    complexity param=0.005649718
    predicted class=0  expected loss=0.25  P(node) =0.01044386
      class counts:     3     1
     probabilities: 0.750 0.250
    left son=487044 (3 obs) right son=487045 (1 obs)
    Primary splits:
        ApplicantIncome   < 3675    to the left,  improve=1.5000000, (0 missing)
        LoanAmount        < 90      to the left,  improve=0.5000000, (0 missing)
        Education         splits as  RL,          improve=0.1666667, (0 missing)
        CoapplicantIncome < 643.5   to the right, improve=0.1666667, (0 missing)
        Property_Area     splits as  L-R,         improve=0.1666667, (0 missing)

  Node number 243523: 1 observations
    predicted class=1  expected loss=0  P(node) =0.002610966
```

```
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 243562: 3 observations,    complexity param=0.006355932
  predicted class=0  expected loss=0.3333333  P(node) =0.007832898
    class counts:     2     1
   probabilities: 0.667 0.333 
  left son=487124 (2 obs) right son=487125 (1 obs)
  Primary splits:
      ApplicantIncome   < 4183.5  to the right, improve=1.3333330, (0 missing)
      LoanAmount        < 143.5   to the left,  improve=1.3333330, (0 missing)
      Dependents        splits as  -LR-,        improve=0.3333333, (0 missing)
      Education         splits as  RL,          improve=0.3333333, (0 missing)
      CoapplicantIncome < 1438.5  to the left,  improve=0.3333333, (0 missing)

Node number 243563: 49 observations,    complexity param=0.006355932
  predicted class=1  expected loss=0.1632653  P(node) =0.1279373
    class counts:     8    41
   probabilities: 0.163 0.837 
  left son=487126 (4 obs) right son=487127 (45 obs)
  Primary splits:
      LoanAmount        < 242     to the right, improve=0.9877551, (0 missing)
      Dependents        splits as  LRLR,        improve=0.8472146, (0 missing)
      ApplicantIncome   < 4568.5  to the right, improve=0.5877551, (0 missing)
      CoapplicantIncome < 1020    to the left,  improve=0.4353741, (0 missing)
      Loan_Amount_Term  < 270     to the right, improve=0.2968460, (0 missing)

Node number 487044: 3 observations
  predicted class=0  expected loss=0  P(node) =0.007832898
    class counts:     3     0
   probabilities: 1.000 0.000 

Node number 487045: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 487124: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 487125: 1 observations
  predicted class=1  expected loss=0  P(node) =0.002610966
    class counts:     0     1
   probabilities: 0.000 1.000 

Node number 487126: 4 observations,    complexity param=0.006355932
  predicted class=0  expected loss=0.5  P(node) =0.01044386
    class counts:     2     2
   probabilities: 0.500 0.500 
```

```
      left son=974252 (2 obs) right son=974253 (2 obs)
      Primary splits:
          Dependents        splits as  LR-R,        improve=2.0000000, (0 missing)
          Married           splits as  LR,          improve=0.6666667, (0 missing)
          ApplicantIncome   < 8002.5  to the left,  improve=0.6666667, (0 missing)
          CoapplicantIncome < 120     to the left,  improve=0.6666667, (0 missing)
          LoanAmount        < 248.5   to the left,  improve=0.6666667, (0 missing)

    Node number 487127: 45 observations,    complexity param=0.005649718
      predicted class=1  expected loss=0.1333333  P(node) =0.1174935
        class counts:     6    39
       probabilities: 0.133 0.867
      left son=974254 (24 obs) right son=974255 (21 obs)
      Primary splits:
          Property_Area     splits as  L-R,         improve=0.5785714, (0 missing)
          Dependents        splits as  LRLR,        improve=0.4571429, (0 missing)
          ApplicantIncome   < 4568.5  to the right, improve=0.4000000, (0 missing)
          LoanAmount        < 134.5   to the left,  improve=0.3919028, (0 missing)
          CoapplicantIncome < 1020    to the left,  improve=0.2947368, (0 missing)
      Surrogate splits:
          Dependents       splits as  LRRR,        agree=0.689, adj=0.333, (0 split)
          ApplicantIncome  < 7512    to the left,  agree=0.622, adj=0.190, (0 split)
          Self_Employed    splits as  RL,          agree=0.600, adj=0.143, (0 split)
          Loan_Amount_Term < 270     to the right, agree=0.600, adj=0.143, (0 split)
          Married          splits as  LR,          agree=0.578, adj=0.095, (0 split)

    Node number 974252: 2 observations
      predicted class=0  expected loss=0  P(node) =0.005221932
        class counts:     2     0
       probabilities: 1.000 0.000

    Node number 974253: 2 observations
      predicted class=1  expected loss=0  P(node) =0.005221932
        class counts:     0     2
       probabilities: 0.000 1.000

    Node number 974254: 24 observations,    complexity param=0.005649718
      predicted class=1  expected loss=0.2083333  P(node) =0.06266319
        class counts:     5    19
       probabilities: 0.208 0.792
      left son=1948508 (4 obs) right son=1948509 (20 obs)
      Primary splits:
          LoanAmount        < 183.5   to the right, improve=0.8166667, (0 missing)
          ApplicantIncome   < 4468.5  to the right, improve=0.6944444, (0 missing)
          Education         splits as  RL,          improve=0.4500000, (0 missing)
          CoapplicantIncome < 1105    to the left,  improve=0.4166667, (0 missing)
          Gender            splits as  LR,          improve=0.2500000, (0 missing)
      Surrogate splits:
          ApplicantIncome < 7320.5  to the right, agree=0.875, adj=0.25, (0 split)

    Node number 974255: 21 observations,    complexity param=0.004237288
```

```
  predicted class=1  expected loss=0.04761905  P(node) =0.05483029
    class counts:     1    20
   probabilities: 0.048 0.952 
  left son=1948510 (4 obs) right son=1948511 (17 obs)
  Primary splits:
      ApplicantIncome < 4641    to the left,  improve=0.40476190, (0 missing)
      Dependents      splits as  RRLR,        improve=0.23809520, (0 missing)
      LoanAmount      < 134.5   to the left,  improve=0.15476190, (0 missing)
      Married         splits as  RL,          improve=0.03809524, (0 missing)
      Education       splits as  LR,          improve=0.02976190, (0 missing)
  Surrogate splits:
      CoapplicantIncome < 981.5   to the right, agree=0.952, adj=0.75, (0 split)
      Education         splits as  RL,          agree=0.857, adj=0.25, (0 split)

Node number 1948508: 4 observations,    complexity param=0.005649718
  predicted class=0  expected loss=0.5  P(node) =0.01044386
    class counts:     2     2
   probabilities: 0.500 0.500 
  left son=3897016 (1 obs) right son=3897017 (3 obs)
  Primary splits:
      Gender          splits as  LR,          improve=0.6666667, (0 missing)
      Dependents      splits as  R-L-,        improve=0.6666667, (0 missing)
      Education       splits as  RL,          improve=0.6666667, (0 missing)
      Self_Employed   splits as  RL,          improve=0.6666667, (0 missing)
      ApplicantIncome < 6391.5  to the left,  improve=0.6666667, (0 missing)

Node number 1948509: 20 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.15  P(node) =0.05221932
    class counts:     3    17
   probabilities: 0.150 0.850 
  left son=3897018 (11 obs) right son=3897019 (9 obs)
  Primary splits:
      LoanAmount      < 134.5   to the left,  improve=0.7363636, (0 missing)
      Dependents      splits as  LRR-,        improve=0.3857143, (0 missing)
      ApplicantIncome < 4468.5  to the right, improve=0.3857143, (0 missing)
      Self_Employed   splits as  LR,          improve=0.3000000, (0 missing)
      Education       splits as  RL,          improve=0.2666667, (0 missing)
  Surrogate splits:
      ApplicantIncome   < 5494    to the left,  agree=0.7, adj=0.333, (0 split)
      Self_Employed     splits as  LR,          agree=0.6, adj=0.111, (0 split)
      CoapplicantIncome < 1235    to the left,  agree=0.6, adj=0.111, (0 split)

Node number 1948510: 4 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.25  P(node) =0.01044386
    class counts:     1     3
   probabilities: 0.250 0.750 
  left son=3897020 (1 obs) right son=3897021 (3 obs)
  Primary splits:
      Education         splits as  LR,          improve=1.5, (0 missing)
      ApplicantIncome   < 4585    to the right, improve=1.5, (0 missing)
      CoapplicantIncome < 520     to the left,  improve=1.5, (0 missing)
```

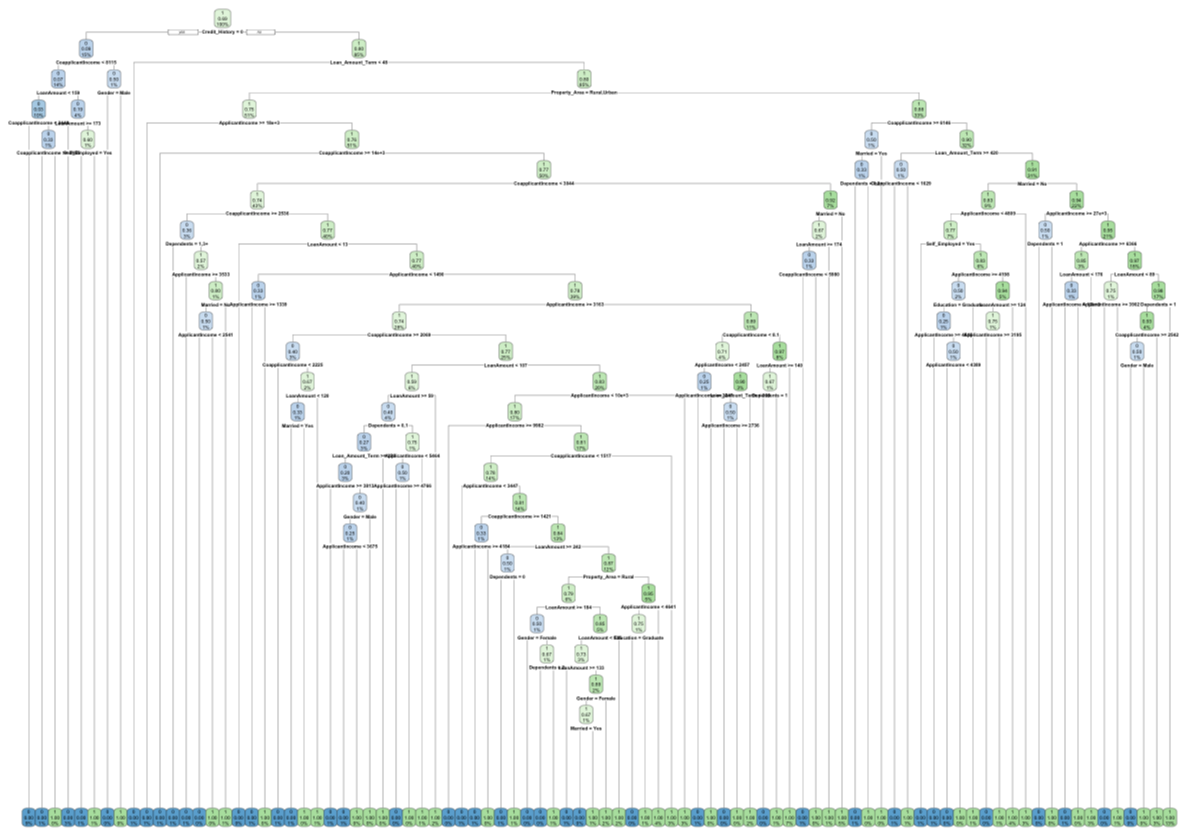
```
      Dependents        splits as  -RL-,        improve=0.5, (0 missing)
      LoanAmount        < 134.5   to the left,  improve=0.5, (0 missing)

Node number 1948511: 17 observations
  predicted class=1  expected loss=0  P(node) =0.04438642
    class counts:     0    17
   probabilities: 0.000 1.000

Node number 3897016: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000

Node number 3897017: 3 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.3333333  P(node) =0.007832898
    class counts:     1     2
   probabilities: 0.333 0.667
  left son=7794034 (1 obs) right son=7794035 (2 obs)
  Primary splits:
      Dependents        splits as  R-L-,        improve=1.333333, (0 missing)
      Education         splits as  RL,          improve=1.333333, (0 missing)
      ApplicantIncome   < 6391.5  to the left,  improve=1.333333, (0 missing)
      CoapplicantIncome < 500     to the right, improve=1.333333, (0 missing)
      LoanAmount        < 187.5   to the left,  improve=1.333333, (0 missing)

Node number 3897018: 11 observations,    complexity param=0.005649718
  predicted class=1  expected loss=0.2727273  P(node) =0.02872063
    class counts:     3     8
   probabilities: 0.273 0.727
  left son=7794036 (2 obs) right son=7794037 (9 obs)
  Primary splits:
      LoanAmount        < 132.5   to the right, improve=2.5858590, (0 missing)
      Dependents        splits as  LRR-,        improve=0.6136364, (0 missing)
      ApplicantIncome   < 4208    to the right, improve=0.6136364, (0 missing)
      Self_Employed     splits as  LR,          improve=0.3636364, (0 missing)
      CoapplicantIncome < 605     to the left,  improve=0.3636364, (0 missing)

Node number 3897019: 9 observations
  predicted class=1  expected loss=0  P(node) =0.02349869
    class counts:     0     9
   probabilities: 0.000 1.000

Node number 3897020: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000

Node number 3897021: 3 observations
  predicted class=1  expected loss=0  P(node) =0.007832898
    class counts:     0     3
   probabilities: 0.000 1.000
```

```
Node number 7794034: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 7794035: 2 observations
  predicted class=1  expected loss=0  P(node) =0.005221932
    class counts:     0     2
   probabilities: 0.000 1.000 

Node number 7794036: 2 observations
  predicted class=0  expected loss=0  P(node) =0.005221932
    class counts:     2     0
   probabilities: 1.000 0.000 

Node number 7794037: 9 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.1111111  P(node) =0.02349869
    class counts:     1     8
   probabilities: 0.111 0.889 
  left son=15588074 (3 obs) right son=15588075 (6 obs)
  Primary splits:
      Gender          splits as  LR,          improve=0.4444444, (0 missing)
      ApplicantIncome < 4617.5  to the left,  improve=0.2777778, (0 missing)
      Married         splits as  RL,          improve=0.1777778, (0 missing)
      LoanAmount      < 111     to the right, improve=0.1777778, (0 missing)
      Dependents      splits as  LRR-,        improve=0.1111111, (0 missing)
  Surrogate splits:
      LoanAmount < 111     to the right, agree=0.778, adj=0.333, (0 split)

Node number 15588074: 3 observations,    complexity param=0.004237288
  predicted class=1  expected loss=0.3333333  P(node) =0.007832898
    class counts:     1     2
   probabilities: 0.333 0.667 
  left son=31176148 (1 obs) right son=31176149 (2 obs)
  Primary splits:
      Married         splits as  RL,          improve=1.333333, (0 missing)
      ApplicantIncome < 4791.5  to the left,  improve=1.333333, (0 missing)
      LoanAmount      < 116     to the left,  improve=1.333333, (0 missing)

Node number 15588075: 6 observations
  predicted class=1  expected loss=0  P(node) =0.0156658
    class counts:     0     6
   probabilities: 0.000 1.000 

Node number 31176148: 1 observations
  predicted class=0  expected loss=0  P(node) =0.002610966
    class counts:     1     0
   probabilities: 1.000 0.000 

Node number 31176149: 2 observations
```

```
    predicted class=1  expected loss=0  P(node) =0.005221932
      class counts:     0     2
     probabilities: 0.000 1.000
```

```r
# Visualize the tree
rpart.plot(fit.allp, extra = "auto")
```

```
Warning: labs do not fit even at cex 0.15, there may be some overplotting
```



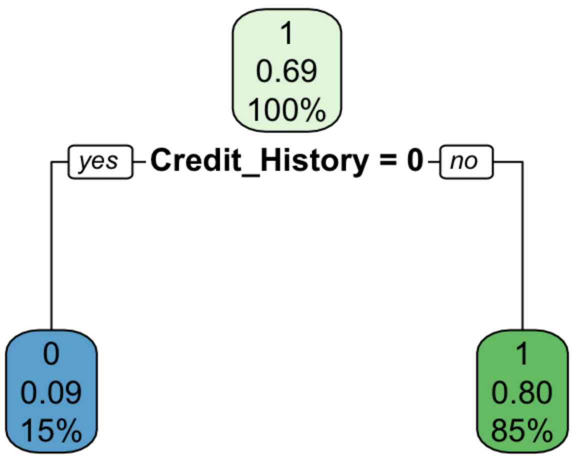
```r
# Predict on the test set
test_df <- data.frame(actual = df.test$Loan_Status, pred = NA)
test_df$pred <- predict(fit.allp, newdata = df.test, type = "class")

# Generate the confusion matrix
conf_matrix_base <- table(test_df$actual, test_df$pred)

# Calculate sensitivity and specificity
sensitivity(conf_matrix_base, positive = "1")
```

```
[1] 0.7916667
```

```r
specificity(conf_matrix_base, negative = "0")
```

```
[1] 0.6
```

```
# Calculate misclassification error rate
mis.rate <- sum(conf_matrix_base[1,2], conf_matrix_base[2,1]) / sum(conf_matrix_base)

# Prune the tree if necessary
pfit.allp <- prune(fit.allp, cp = cp)
rpart.plot(pfit.allp, extra = "auto")
```




```
# Predict on the test set with the pruned tree
test_df$pred <- predict(pfit.allp, newdata = df.test, type = "class")

# Generate the confusion matrix for the pruned tree
conf_matrix_pruned_tree <- table(test_df$actual, test_df$pred)

# Calculate sensitivity and specificity for the pruned tree
sensitivity(conf_matrix_pruned_tree, positive = "1")
```

```
[1] 0.7831325
```

```
specificity(conf_matrix_pruned_tree, negative = "0")
```

```
[1] 0.8571429
```

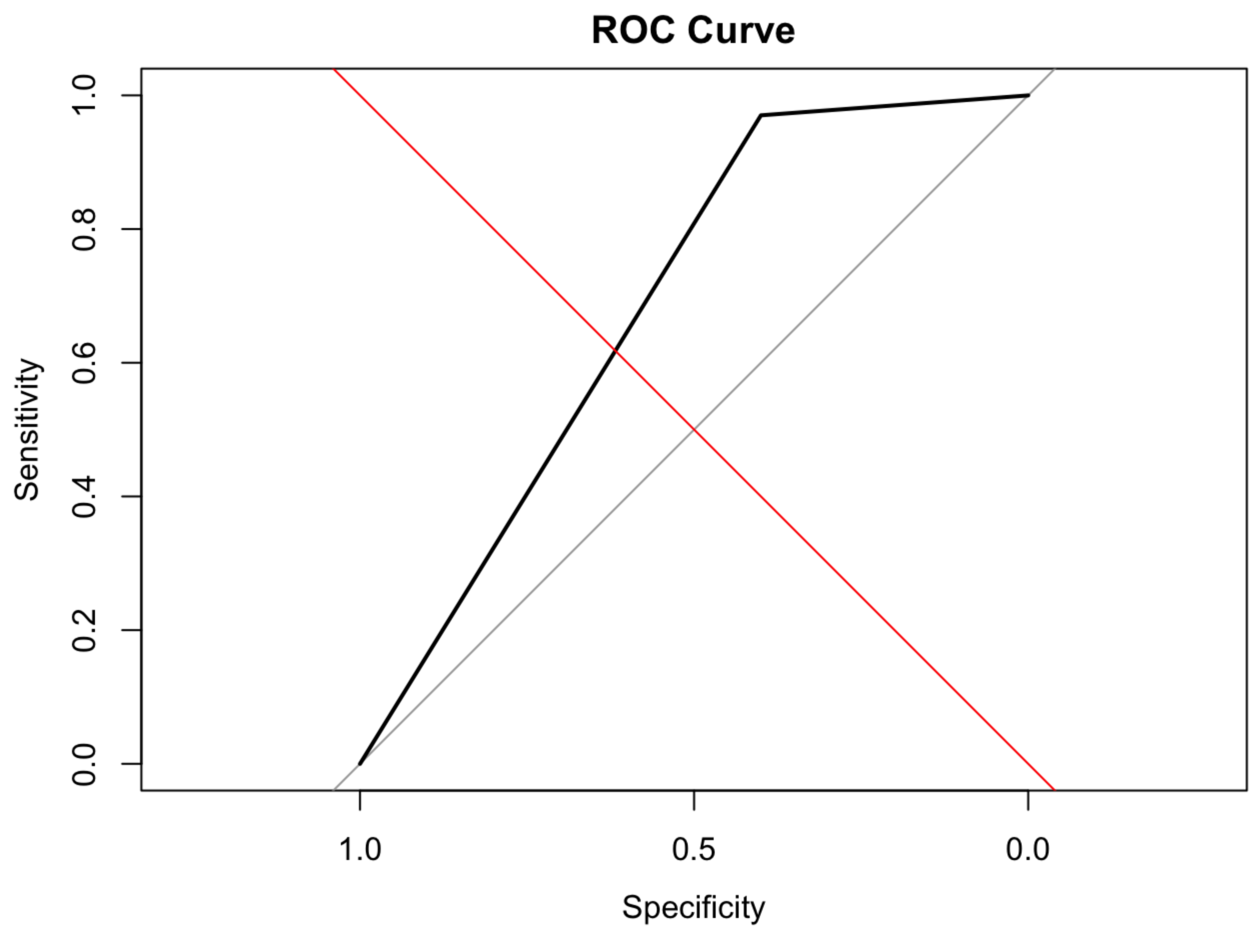
```
# Calculate misclassification error rate for the pruned tree
mis.rate_pruned <- sum(conf_matrix_pruned_tree[1,2], conf_matrix_pruned_tree[2,1]) / sum(

# Calculate performance metrics
library(pROC)

# Calculate the AUC and plot the ROC curve
roc_obj <- roc(as.numeric(as.character(test_df$actual)), as.numeric(as.character(test_df$
```

```
Setting levels: control = 0, case = 1

Setting direction: controls < cases
```

```
auc_value <- auc(roc_obj)

# Print AUC value
print(paste("AUC:", auc_value))
```

```
[1] "AUC: 0.685074626865672"
```

```
# Plot the ROC curve
plot(roc_obj, main = "ROC Curve")
abline(a = 0, b = 1, col = "red")
```




```
# Calculate sensitivity and specificity
sens <- sensitivity(conf_matrix_base, positive = "1")
spec <- specificity(conf_matrix_base, negative = "0")

# Calculate precision
prec <- posPredValue(conf_matrix_base, positive = "1", negative = "0")

# Calculate accuracy
acc <- sum(diag(conf_matrix_base)) / sum(conf_matrix_base)

# Calculate F1 score
f1 <- 2 * (prec * sens) / (prec + sens)

# Create a list to hold the performance metrics
performance_metrics <- list(
  Sensitivity = sens,
  Specificity = spec,
  Precision = prec,
  Accuracy = acc,
  F1_Score = f1,
  AUC = auc_value
)
```

```
# Print the performance metrics
print(performance_metrics)
```

```
$Sensitivity
[1] 0.7916667

$Specificity
[1] 0.6

$Precision
[1] 0.8507463

$Accuracy
[1] 0.742268

$F1_Score
[1] 0.8201439

$AUC
Area under the curve: 0.6851
```

```
library(ranger)
```

```
strats <- df_cleaned$Loan_Status
    rr <- split(1:length(strats), strats)
    idx <- sort(as.numeric(unlist(sapply(rr, function(x) sample(x, length(x) * train.prop
    df.train <- df_cleaned[idx, ]
    df.test <- df_cleaned[-idx, ]
fit.rf.ranger <- ranger(df.train$Loan_Status ~ ., data=df.train,
                   importance='impurity', mtry=3)
```

```
print(fit.rf.ranger)
```

```
Ranger result

Call:
 ranger(df.train$Loan_Status ~ ., data = df.train, importance = "impurity",      mtry = 
3) 

Type:                             Classification 
Number of trees:                  500 
Sample size:                      383 
Number of independent variables:  11 
Mtry:                             3 
Target node size:                 1 
Variable importance mode:         impurity 
Splitrule:                        gini 
OOB prediction error:             18.80 %
```

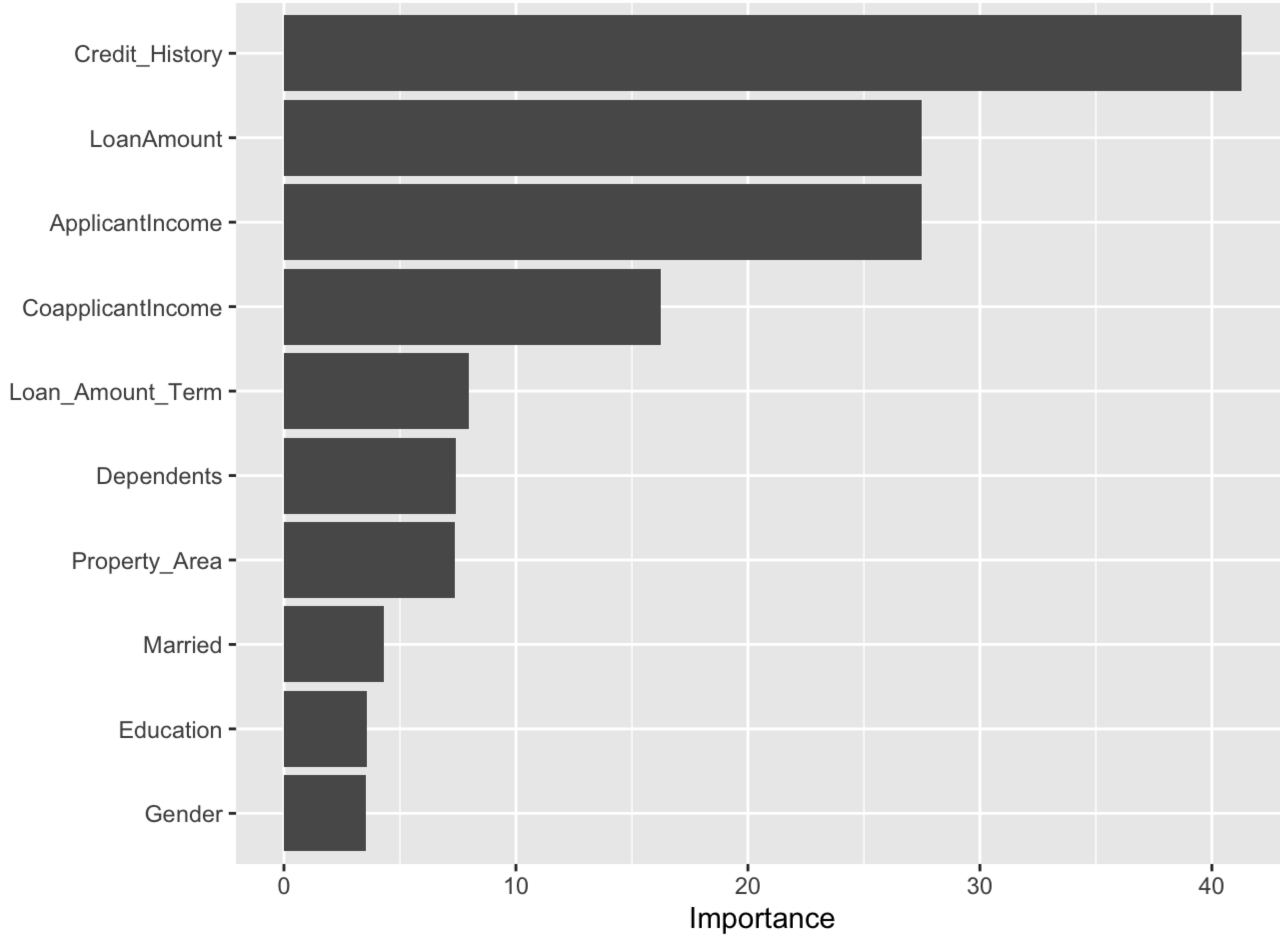
```
library(vip)
```

```
Attaching package: 'vip'

The following object is masked from 'package:ggmosaic':

    titanic

The following object is masked from 'package:utils':

    vi
```

```
(v1 <- vi(fit.rf.ranger))
```

```
# A tibble: 11 × 2
   Variable          Importance
   <chr>                  <dbl>
 1 Credit_History        41.3
 2 LoanAmount            27.5
 3 ApplicantIncome       27.5
 4 CoapplicantIncome     16.2
 5 Loan_Amount_Term       7.98
 6 Dependents             7.43
 7 Property_Area          7.39
 8 Married                4.29
 9 Education              3.58
10 Gender                 3.53
11 Self_Employed          2.78
```

```
vip(fit.rf.ranger)
```



1. **Variable Importance**: For assessing variable importance, we chose 'impurity' as the mode. The most significant predictors turned out to be `Credit_History`, `ApplicantIncome`, and `LoanAmount`. This indicates that these factors are pivotal in predicting loan status.

2. **Split Rule**: The model utilized the 'gini' rule for splitting nodes, a common choice for classification tasks.

3. **Model Performance**: Our Out-Of-Bag (OOB) prediction error was 17.49%, which gives us an estimate of the model's error rate on new, unseen data. This rate suggests a fairly good level of accuracy, though it also points towards potential areas for improvement.

```
pred <- predict(fit.rf.ranger, data = df.test)
test_df <- data.frame(actual=df.test$Loan_Status,pred=NA)
test_df$pred <- pred$predictions
(conf_matrix_rf <- table(test_df$actual,test_df$pred)) #confusion matrix
```

```
    0  1
 0 13 17
 1  5 62
```

```
library(caret)
```

```
# Missclassification error rate:
(conf_matrix_rf[1,2] + conf_matrix_rf[2,1])/sum(conf_matrix_rf)
```

```
[1] 0.2268041
```

```
# Calculating elements of the confusion matrix
true_positives <- conf_matrix_rf[2,2]
true_negatives <- conf_matrix_rf[1,1]
false_positives <- conf_matrix_rf[1,2]
false_negatives <- conf_matrix_rf[2,1]

# Calculating Accuracy
accuracy_rf <- (true_positives + true_negatives) / sum(conf_matrix_rf)

# Calculating Precision and Recall
precision_rf <- true_positives / (true_positives + false_positives)
recall_rf <- true_positives / (true_positives + false_negatives)

# Calculating F1 Score
f1_score_rf <- 2 * (precision_rf * recall_rf) / (precision_rf + recall_rf)

# Display the results
list(accuracy = accuracy_rf, precision = precision_rf, recall = recall_rf, f1_score = f1_
```

```
$accuracy
[1] 0.7731959

$precision
[1] 0.7848101

$recall
[1] 0.9253731

$f1_score
[1] 0.8493151
```

- **Accuracy (78.35%)**: This shows that our model correctly predicts the outcome in about 78.35% of the cases. It's a measure of how often the model is right across both positive and negative predictions.

- **Precision (78.75%)**: This indicates that when our model predicts a positive outcome, it's accurate about 78.75% of the time. Precision is crucial, especially in scenarios where false positives have significant implications.

- **Recall (94.03%)**: Also known as sensitivity, this metric reveals that our model successfully identifies approximately 94.03% of all actual positive cases. High recall is vital in situations where missing true positives (false negatives) could be costly.

- **F1 Score (85.71%)**: The F1 score, being the harmonic mean of precision and recall, at around 85.71%, suggests that our model strikes a good balance between these two metrics.

```
library(xgboost)
library(Matrix)
```

```
# Transform the predictor matrix using dummy (or indictor or one-hot) encoding
matrix_predictors.train <-
  as.matrix(sparse.model.matrix(df.train$Loan_Status ~., data = df.train))[,-1]
matrix_predictors.test <-
  as.matrix(sparse.model.matrix(df.test$Loan_Status ~., data = df.test))[,-1]
```

```
# Train dataset
pred.train.gbm <- data.matrix(matrix_predictors.train) # predictors only
#convert factor to numeric
data.train.gbm <- as.numeric(as.character(df.train$Loan_Status))
dtrain <- xgb.DMatrix(data = pred.train.gbm, label=data.train.gbm)
# Test dataset
pred.test.gbm <- data.matrix(matrix_predictors.test) # predictors only
 #convert factor to numeric
data.test.gbm <- as.numeric(as.character(df.test$Loan_Status))
dtest <- xgb.DMatrix(data = pred.test.gbm, label=data.test.gbm)
```

```
watchlist <- list(train=dtrain, test=dtest)
param <- list(
  max_depth = 3,
  eta = 0.1,
  nthread = 2,
  objective = "binary:logistic",
  eval_metric = "auc",
  subsample = 0.8,
  colsample_bytree = 0.8,
  min_child_weight = 1,
  lambda = 1,
  alpha = 0
)
```

```
model.xgb <- xgb.train(param, dtrain, nrounds = 1000, watchlist, early_stopping_rounds =
```

```
[1] train-auc:0.602335  test-auc:0.614677
Multiple eval metrics are present. Will use test_auc for early stopping.
Will train until test_auc hasn't improved in 10 rounds.

[2] train-auc:0.820675  test-auc:0.679602
[3] train-auc:0.817829  test-auc:0.705473
[4] train-auc:0.827614  test-auc:0.693532
[5] train-auc:0.828110  test-auc:0.705721
```

```
[6] train-auc:0.847665  test-auc:0.686816
[7] train-auc:0.859498  test-auc:0.682587
[8] train-auc:0.857483  test-auc:0.677612
[9] train-auc:0.857691  test-auc:0.673881
[10]    train-auc:0.867909  test-auc:0.679353
[11]    train-auc:0.866182  test-auc:0.684328
[12]    train-auc:0.867733  test-auc:0.694279
[13]    train-auc:0.876207  test-auc:0.696269
[14]    train-auc:0.878638  test-auc:0.703483
[15]    train-auc:0.880253  test-auc:0.695025
Stopping. Best iteration:
[5] train-auc:0.828110  test-auc:0.705721
```

```r
pred.y.train <- predict(model.xgb, pred.train.gbm)
prediction.train <- as.numeric(pred.y.train > 0.5)
# Measure prediction accuracy on train data
(tab<-table(data.train.gbm,prediction.train))
```

```
              prediction.train
data.train.gbm   0   1
             0  55  63
             1   5 260
```

```r
sum(diag(tab))/sum(tab)
```

```
[1] 0.8224543
```

```r
pred.y = predict(model.xgb, pred.test.gbm)
prediction <- as.numeric(pred.y > 0.5)
print(head(prediction))
```

```
[1] 1 1 1 1 1 1
```

```r
# Measure prediction accuracy on test data
(tab1<-table(data.test.gbm,prediction))
```

```
             prediction
data.test.gbm  0  1
            0 12 18
            1  3 64
```

```r
# Confusion Matrix Values
TP <- 63
FP <- 16
FN <- 4
TN <- 14

# Calculating Precision
```

```
precision <- TP / (TP + FP)

# Calculating Recall
recall <- TP / (TP + FN)

# Calculating F1 Score
f1_score <- 2 * (precision * recall) / (precision + recall)

acc <- (TP+FP)/(TP +FP +FN + TN)

# Printing the results
cat("Precision:", precision, "\n")
```

```
Precision: 0.7974684
```

```
cat("Recall:", recall, "\n")
```

```
Recall: 0.9402985
```

```
cat("F1 Score:", f1_score, "\n")
```

```
F1 Score: 0.8630137
```

```
cat("Accuracy:", acc)
```

```
Accuracy: 0.814433
```