Final Project

Michael Ritacco

2024-04-14

Import Data and Required Libraries

```
suppressPackageStartupMessages({
  library(ggplot2)
  library(dplyr)
  library(caret)
  library(randomForest)
  library(MLmetrics)
  library(PRROC)
  library(xgboost)
  library(reshape2)
  library(tidyr)
})
```

```
## Warning: package 'ggplot2' was built under R version 4.3.1
```

```
## Warning: package 'dplyr' was built under R version 4.3.1
```

```
## Warning: package 'MLmetrics' was built under R version 4.3.1
```

```
## Warning: package 'xgboost' was built under R version 4.3.1
```

```
# Read in the data
df = read.csv('loan_default.csv')

# Output the first 5 rows of the dataset
head(df)
```

```
##       LoanID Age Income LoanAmount CreditScore MonthsEmployed NumCreditLines
## 1 I38PQUQS96  56  85994      50587         520             80              4
## 2 HPSK72WA7R  69  50432     124440         458             15              1
## 3 C1OZ6DPJ8Y  46  84208     129188         451             26              3
## 4 V2KKSFM3UN  32  31713      44799         743              0              3
## 5 EY08JDHTZP  60  20437       9139         633              8              4
## 6 A9S62RQ7US  25  90298      90448         720             18              2
##   InterestRate LoanTerm DTIRatio   Education EmploymentType MaritalStatus
## 1        15.23       36     0.44  Bachelor's      Full-time      Divorced
## 2         4.81       60     0.68    Master's      Full-time       Married
```

```
## 3        21.17       24     0.31    Master's     Unemployed      Divorced
## 4         7.07       24     0.23 High School      Full-time       Married
## 5         6.51       48     0.73  Bachelor's     Unemployed      Divorced
## 6        22.72       24     0.10 High School     Unemployed        Single
##   HasMortgage HasDependents LoanPurpose HasCoSigner Default
## 1         Yes           Yes       Other         Yes       0
## 2          No            No       Other         Yes       0
## 3         Yes           Yes        Auto          No       1
## 4          No            No    Business          No       0
## 5          No           Yes        Auto          No       0
## 6         Yes            No    Business         Yes       1
```

Data Preprocessing

Data Cleaning

```
colnames(df)
```

```
##  [1] "LoanID"         "Age"            "Income"         "LoanAmount"
##  [5] "CreditScore"    "MonthsEmployed" "NumCreditLines" "InterestRate"
##  [9] "LoanTerm"       "DTIRatio"       "Education"      "EmploymentType"
## [13] "MaritalStatus"  "HasMortgage"    "HasDependents"  "LoanPurpose"
## [17] "HasCoSigner"    "Default"
```

```
# Removing the index column
df = df[-1]
```

Examining Dataset Structure

```
# Examine the structure of the dataset
str(df)
```

```
## 'data.frame':    255347 obs. of  17 variables:
##  $ Age           : int  56 69 46 32 60 25 38 56 36 40 ...
##  $ Income        : int  85994 50432 84208 31713 20437 90298 111188 126802 42053 132784 ...
##  $ LoanAmount    : int  50587 124440 129188 44799 9139 90448 177025 155511 92357 228510 ...
##  $ CreditScore   : int  520 458 451 743 633 720 429 531 827 480 ...
##  $ MonthsEmployed: int  80 15 26 0 8 18 80 67 83 114 ...
##  $ NumCreditLines: int  4 1 3 3 4 2 1 4 1 4 ...
##  $ InterestRate  : num  15.23 4.81 21.17 7.07 6.51 ...
##  $ LoanTerm      : int  36 60 24 24 48 24 12 60 48 48 ...
##  $ DTIRatio      : num  0.44 0.68 0.31 0.23 0.73 0.1 0.16 0.43 0.2 0.33 ...
##  $ Education     : chr  "Bachelor's" "Master's" "Master's" "High School" ...
##  $ EmploymentType: chr  "Full-time" "Full-time" "Unemployed" "Full-time" ...
##  $ MaritalStatus : chr  "Divorced" "Married" "Divorced" "Married" ...
##  $ HasMortgage   : chr  "Yes" "No" "Yes" "No" ...
##  $ HasDependents : chr  "Yes" "No" "Yes" "No" ...
##  $ LoanPurpose   : chr  "Other" "Other" "Auto" "Business" ...
##  $ HasCoSigner   : chr  "Yes" "Yes" "No" "No" ...
##  $ Default       : int  0 0 1 0 0 1 0 0 1 0 ...
```

```r
# Get the total number of observations
n = nrow(df)

# Get the number of columns
p = ncol(df)

# Output the number of observations and columns
cat("There are", n, 'observations and ', p, 'columns in our dataset.')
```

```
## There are 255347 observations and  17 columns in our dataset.
```

Check for Missing Values

```r
# Check for missing values across columns
colSums(is.na(df))
```

```
##            Age         Income     LoanAmount    CreditScore MonthsEmployed
##              0              0              0              0              0
## NumCreditLines   InterestRate       LoanTerm       DTIRatio      Education
##              0              0              0              0              0
## EmploymentType  MaritalStatus    HasMortgage  HasDependents    LoanPurpose
##              0              0              0              0              0
##    HasCoSigner        Default
##              0              0
```

Encode Categorical Variables

```r
# Define vector of continuous variables
continuous_vars = c(
  'Age', 'Income', 'LoanAmount', 'CreditScore',
  'MonthsEmployed', 'InterestRate', 'DTIRatio'
)

# Define vector of categorical variables
categorical_vars = c(
  'Education', 'EmploymentType', 'MaritalStatus',
  'HasMortgage', 'HasDependents', 'LoanPurpose', 'HasCoSigner'
)

# Define vector of ordinal variables
ordinal_vars = c(
  'NumCreditLines', 'LoanTerm'
)

target = 'Default'

for (var in categorical_vars) {
  df[[var]] = as.factor(df[[var]])
}
```

```r
df$Default = factor(df$Default, levels = c('1', '0'), labels = c('Default', 'NonDefault'))
```

Exploratory Data Analysis

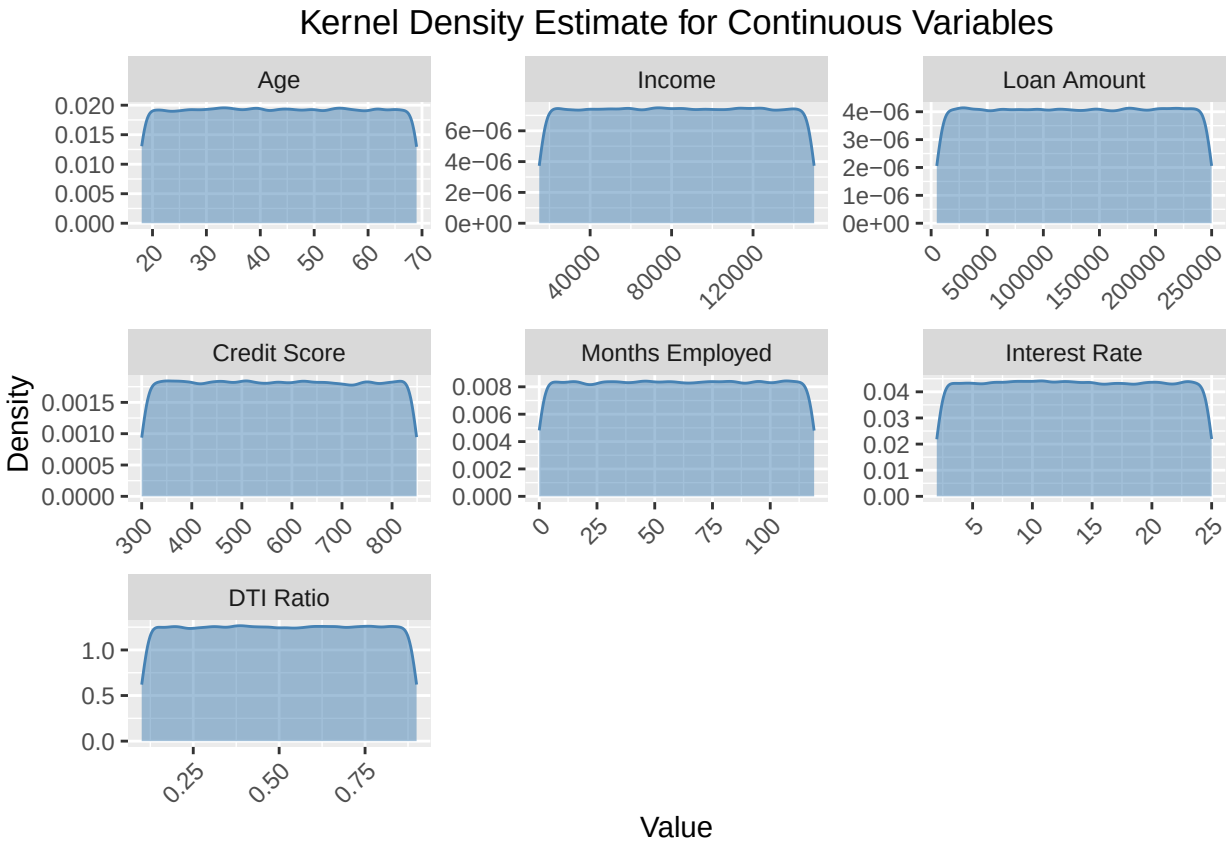
Distribution of Continuous Variables

```r
# Calculate the summary statistics
summary(df)
```

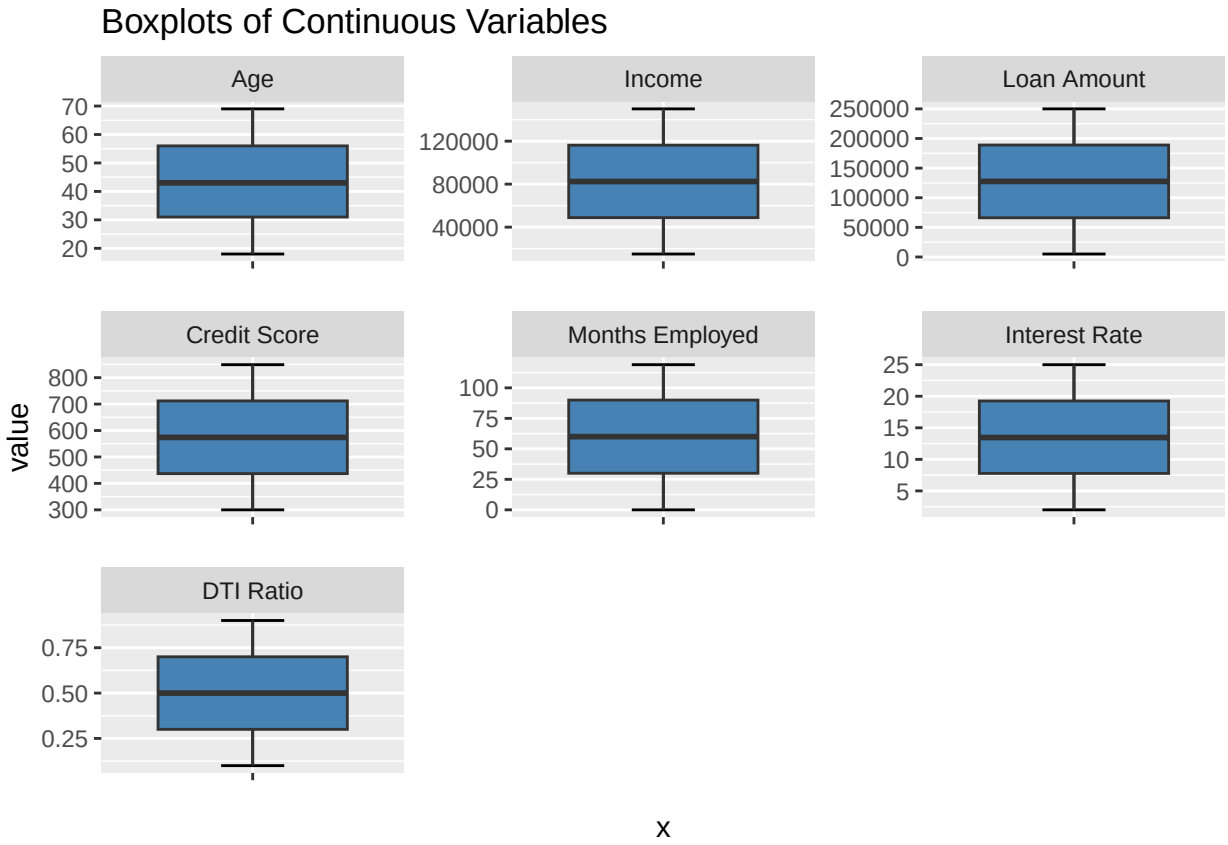
```
##       Age            Income         LoanAmount      CreditScore
##  Min.   :18.0   Min.   : 15000   Min.   :  5000   Min.   :300.0
##  1st Qu.:31.0   1st Qu.: 48826   1st Qu.: 66156   1st Qu.:437.0
##  Median :43.0   Median : 82466   Median :127556   Median :574.0
##  Mean   :43.5   Mean   : 82499   Mean   :127579   Mean   :574.3
##  3rd Qu.:56.0   3rd Qu.:116219   3rd Qu.:188985   3rd Qu.:712.0
##  Max.   :69.0   Max.   :149999   Max.   :249999   Max.   :849.0
##  MonthsEmployed   NumCreditLines   InterestRate      LoanTerm
##  Min.   :  0.00   Min.   :1.000   Min.   : 2.00   Min.   :12.00
##  1st Qu.: 30.00   1st Qu.:2.000   1st Qu.: 7.77   1st Qu.:24.00
##  Median : 60.00   Median :2.000   Median :13.46   Median :36.00
##  Mean   : 59.54   Mean   :2.501   Mean   :13.49   Mean   :36.03
##  3rd Qu.: 90.00   3rd Qu.:3.000   3rd Qu.:19.25   3rd Qu.:48.00
##  Max.   :119.00   Max.   :4.000   Max.   :25.00   Max.   :60.00
##     DTIRatio            Education           EmploymentType   MaritalStatus
##  Min.   :0.1000   Bachelor's :64366   Full-time    :63656   Divorced:85033
##  1st Qu.:0.3000   High School:63903   Part-time    :64161   Married :85302
##  Median :0.5000   Master's   :63541   Self-employed:63706   Single  :85012
##  Mean   :0.5002   PhD        :63537   Unemployed   :63824
##  3rd Qu.:0.7000
##  Max.   :0.9000
##  HasMortgage  HasDependents    LoanPurpose    HasCoSigner        Default
##  No :127670   No :127605    Auto     :50844   No :127646   Default   : 29653
##  Yes:127677   Yes:127742    Business :51298   Yes:127701   NonDefault:225694
##                             Education:51005
##                             Home     :51286
##                             Other    :50914
##
```

```r
# Reshape the dataframe to long format
df_long = melt(df, measure.vars = continuous_vars)

var_labels = c(
  Age = 'Age', Income = 'Income', LoanAmount = 'Loan Amount',
  CreditScore = 'Credit Score', MonthsEmployed = 'Months Employed',
  NumCreditLines = 'Number of Credit Lines', InterestRate = 'Interest Rate',
  LoanTerm = 'Loan Term', DTIRatio = 'DTI Ratio', Education = 'Education',
  EmploymentType = 'Employment Type', MaritalStatus = 'Marital Status',
  HasMortgage = 'Has Mortgage', HasDependents = 'Has Dependents',
  LoanPurpose = 'Loan Purpose', HasCoSigner = 'Has Co Signer',
```

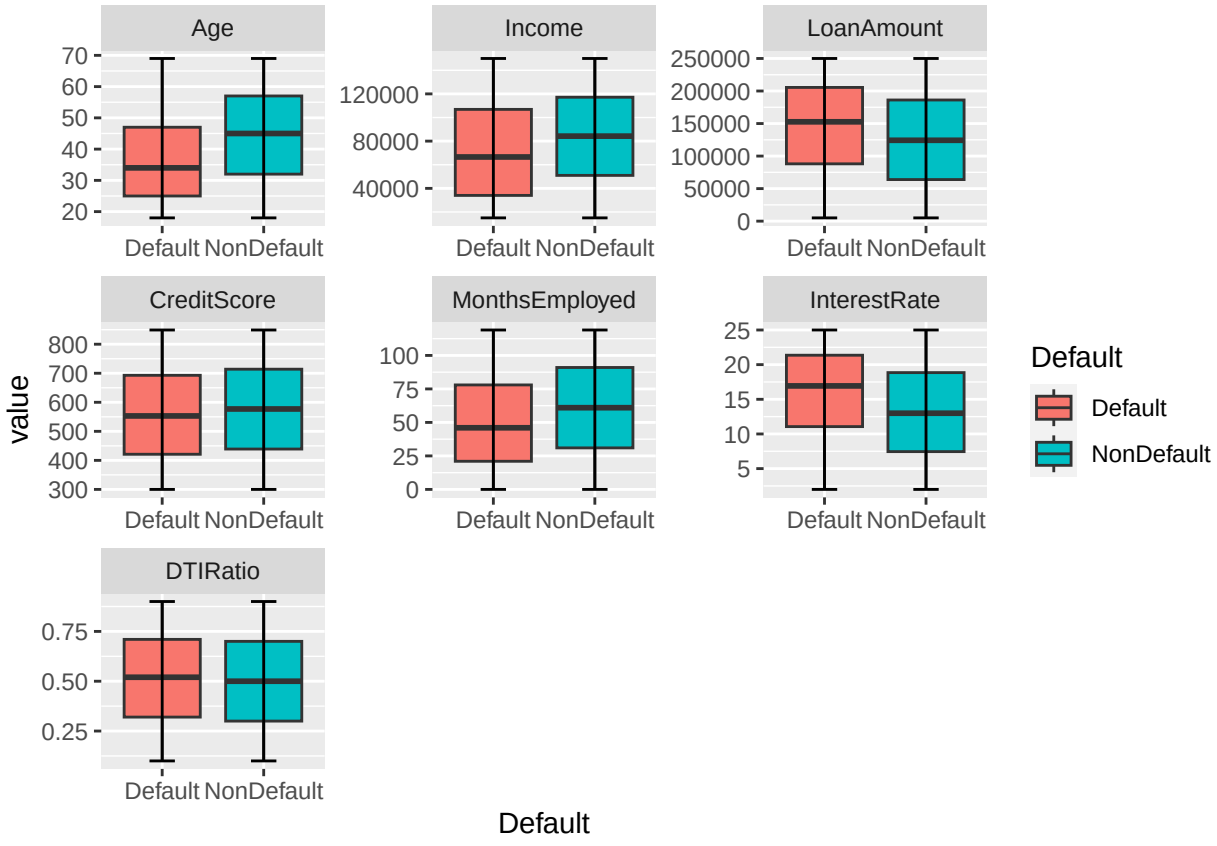
```
  Default = 'Loan Status'
)

# KDE plots for continuous variables
ggplot(df_long, aes(x = value)) +
  geom_density(color = 'steelblue', fill = 'steelblue', alpha = 0.5) +
  facet_wrap(~ variable, scales = 'free', ncol = 3,
             labeller = labeller(variable = var_labels)) +
  labs(title = 'Kernel Density Estimate for Continuous Variables',
       x = 'Value', y = 'Density') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1),
        plot.title = element_text(hjust = 0.5))
```



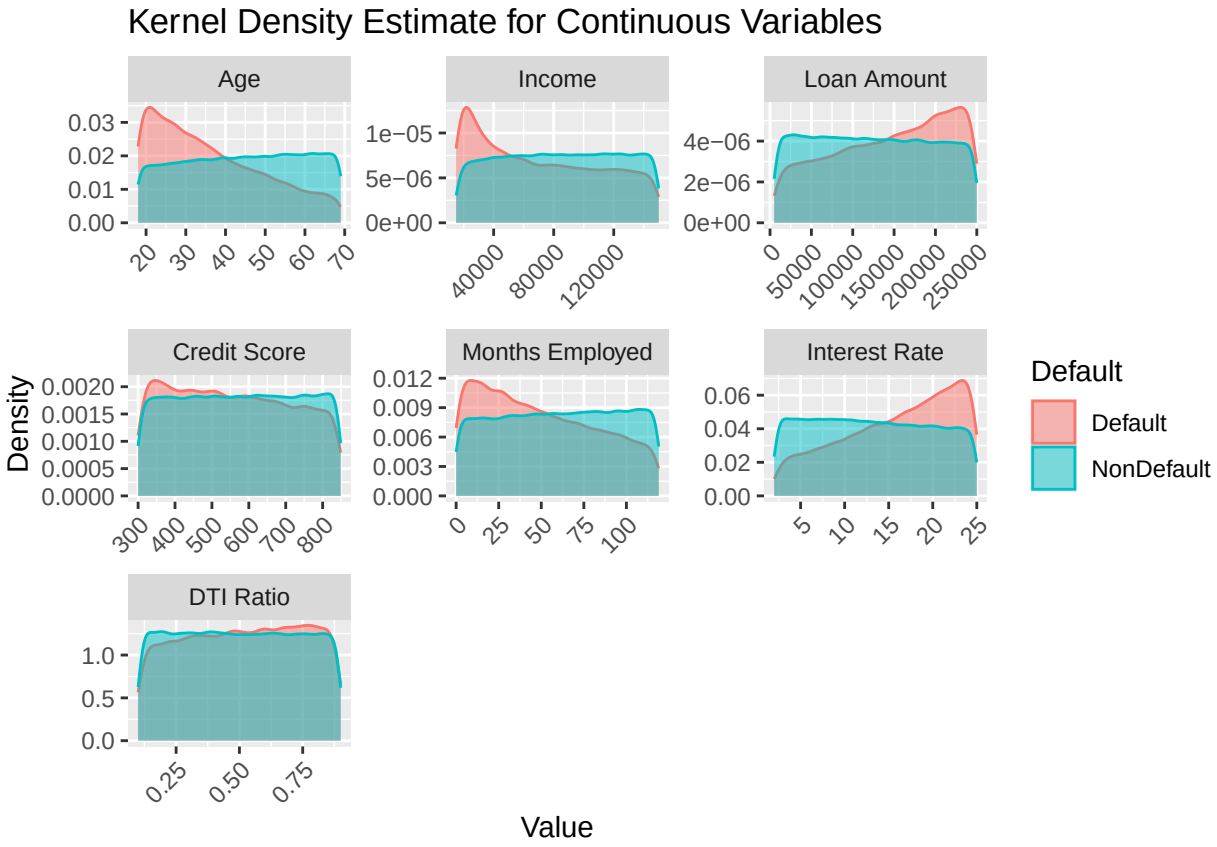

```
ggplot(df_long, aes(x = '', y = value)) +
  stat_boxplot(geom = 'errorbar',
               width = 0.25) +
  geom_boxplot(fill = 'steelblue') +
  facet_wrap(~ variable, scales = 'free',
             labeller = labeller(variable = var_labels), ncol = 3) +
  labs(title = 'Boxplots of Continuous Variables')
```




```
ggplot(df_long, aes(x = Default, y = value, fill = Default)) +
  geom_boxplot() +
  stat_boxplot(geom = 'errorbar',
               width = 0.25) +
  facet_wrap(~ variable, scales = 'free')
```




```
ggplot(df_long, aes(x = value, color = Default, fill = Default)) +
  geom_density(alpha = 0.5) +  # Adjust transparency with alpha
  facet_wrap(~ variable, scales = 'free',
             labeller = labeller(variable = var_labels), ncol = 3) +  # Use facet_wrap to create indivi
  labs(title = 'Kernel Density Estimate for Continuous Variables',
       x = 'Value',
       y = 'Density') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Distribution of Categorical Variables

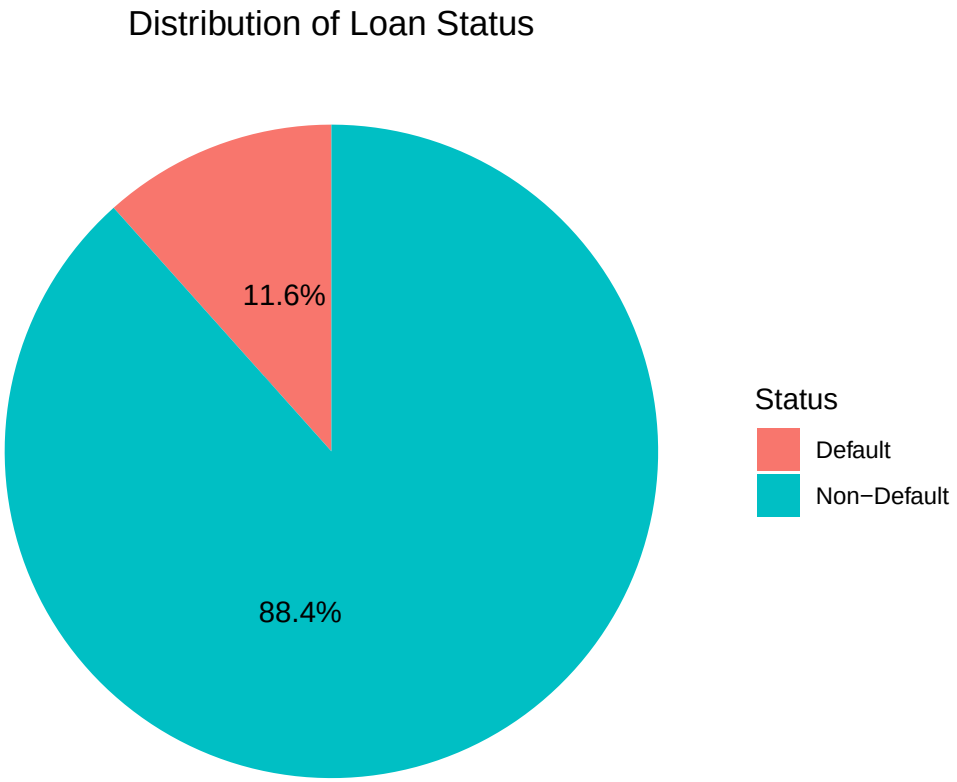
Distribution of Target Variable

```
default_counts = table(df$Default)
default_proportions = (default_counts / n) * 100

default_df = data.frame(
  Status = c('Default', 'Non-Default'),
  Count = as.vector(default_counts),
  Percent = as.vector(default_proportions)
  )
default_df
```

```
##         Status  Count  Percent
## 1      Default  29653 11.61282
## 2 Non-Default 225694 88.38718
```

```
ggplot(data = default_df, aes(x = '', y = Percent, fill = Status)) +
  geom_bar(stat = 'identity', width = 1) +
  coord_polar(theta = 'y') +
  geom_text(aes(label = sprintf('%.1f%%', Percent)), position = position_stack(vjust = 0.6)) +
  theme_void() +
```

```
  labs(title = 'Distribution of Loan Status') +
  theme(plot.title = element_text(hjust = 0.5))
```

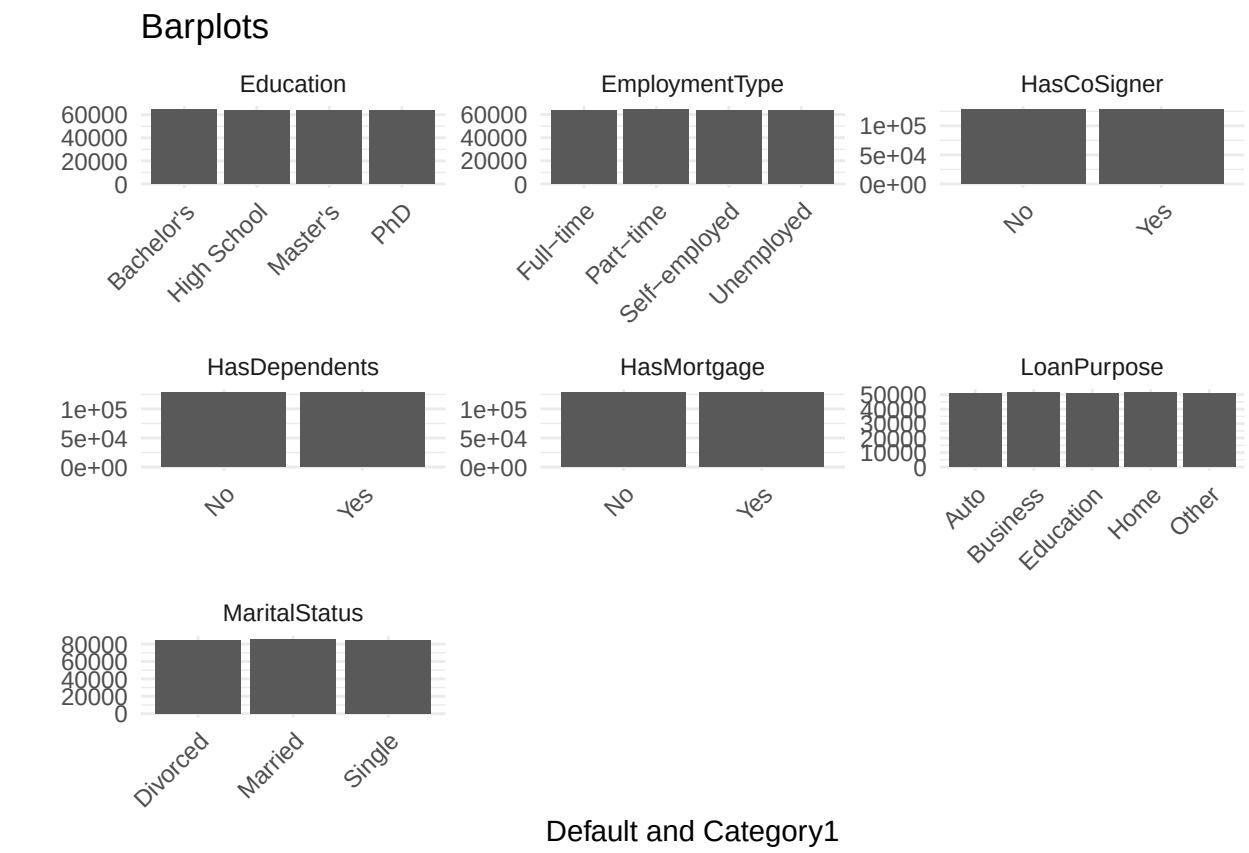



```
summary(df[categorical_vars])
```

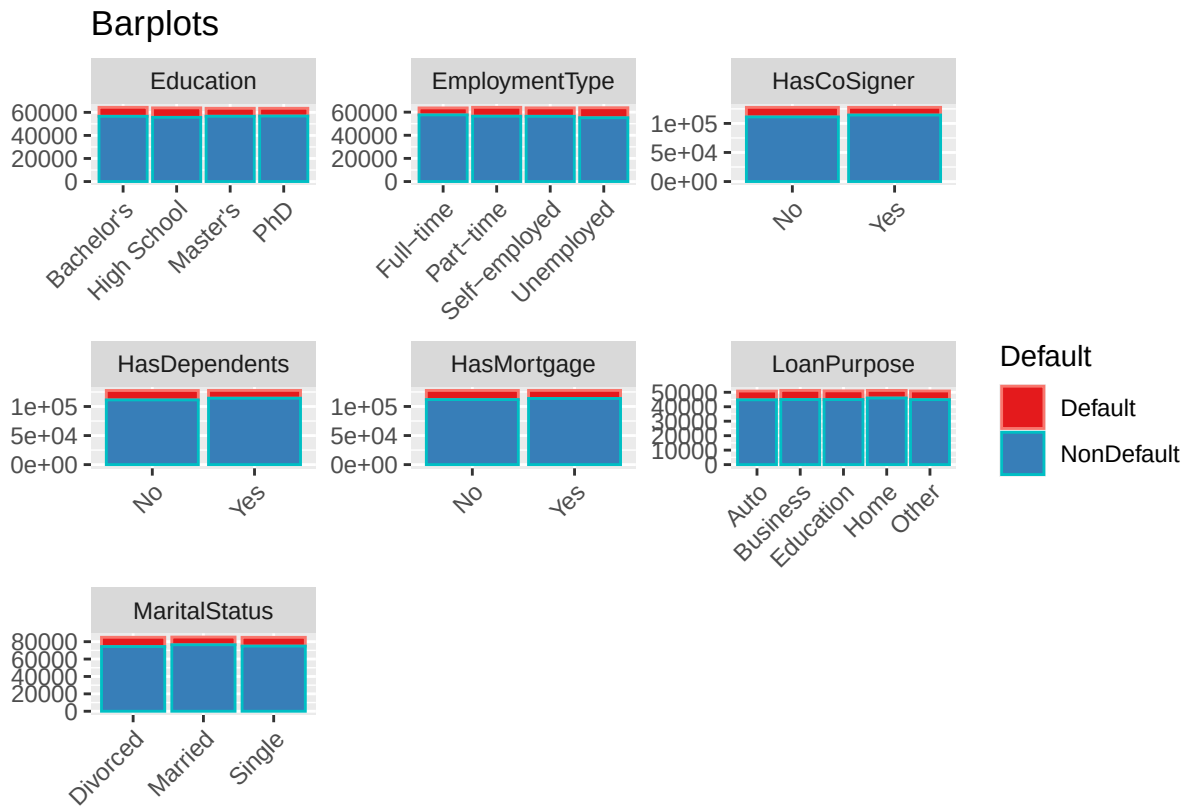
```
##        Education           EmploymentType   MaritalStatus   HasMortgage
##  Bachelor's :64366   Full-time     :63656   Divorced:85033   No :127670
##  High School:63903   Part-time     :64161   Married :85302   Yes:127677
##  Master's   :63541   Self-employed:63706   Single  :85012
##  PhD        :63537   Unemployed    :63824
##
##  HasDependents    LoanPurpose    HasCoSigner
##  No :127605    Auto     :50844   No :127646
##  Yes:127742    Business :51298   Yes:127701
##                Education:51005
##                Home     :51286
##                Other    :50914
```

```
df_long_cat = df[c(categorical_vars, target)] %>%
  pivot_longer(cols = -Default, names_to = 'variable', values_to = 'value')

ggplot(data = df_long_cat) +
  geom_bar(aes(x = value)) +
  labs(title = 'Barplots', x = 'Default and Category1', y = '') +
```

```
  facet_wrap( ~ variable, scales = 'free') +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = 'Set1')
```




```
ggplot(data = df_long_cat) +
  geom_bar(aes(x = value, fill = Default, color = Default)) +
  labs(title = 'Barplots', x = 'Default and Category1', y = '') +
  facet_wrap( ~ variable, scales = 'free') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_brewer(palette = 'Set1')
```



Main Data Analysis

Spliting the data into train/validation/test sets

Model Fitting with Down Sampling

Model Tuning and Validation

Model Evaluation

Performce Metrics

Variable Importance