midterm.rmd

2025-03-24

```
# Install and load the necessary package (for Excel files, but not required for CSV)
# install.packages("readxl")  # Uncomment this line if you're dealing with Excel files
# library(readxl)  # Uncomment this line if you're dealing with Excel files

# Set the working directory to where the CSV file is located (optional)
# This command sets the folder where R will look for my files:
setwd("C:/Users/prash/Downloads/")  # Update this path to your directory if needed

# Read the CSV file
# This can loads the dataset from the CSV file into R.
my_data <- read.csv("data science missdata.csv")  # Ensure file name is correct

# View all rows of the dataset (This will print the entire dataset)
#This prints all the data, which helps us quickly scan if everything loaded correctly.
print(my_data)  # Display the entire dataset with all 106 rows
```

```
##      Loan_ID Gender Married Dependents    Education Self_Employed
## 1   LP001002   Male      No          0     Graduate            No
## 2   LP001003   Male     Yes          1     Graduate            No
## 3   LP001005   Male     Yes          0     Graduate           Yes
## 4   LP001006   Male     Yes          0 Not Graduate            No
## 5   LP001008   Male      No          0     Graduate            No
## 6   LP001011   Male     Yes          2     Graduate           Yes
## 7   LP001013   Male     Yes          0 Not Graduate            No
## 8   LP001014   Male     Yes         3+     Graduate            No
## 9   LP001018   Male     Yes          2     Graduate            No
## 10  LP001020   Male     Yes          1     Graduate            No
## 11  LP001024   Male     Yes          2     Graduate            No
## 12  LP001027   Male     Yes          2     Graduate
## 13  LP001028   Male     Yes          2     Graduate            No
## 14  LP001029   Male      No          0     Graduate            No
## 15  LP001030   Male     Yes          2     Graduate            No
## 16  LP001032   Male      No          0     Graduate            No
## 17  LP001034   Male      No          1 Not Graduate            No
## 18  LP001036 Female      No          0     Graduate            No
## 19  LP001038   Male     Yes          0 Not Graduate            No
## 20  LP001041   Male     Yes          0     Graduate
## 21  LP001043   Male     Yes          0 Not Graduate            No
## 22  LP001046   Male     Yes          1     Graduate            No
## 23  LP001047   Male     Yes          0 Not Graduate            No
## 24  LP001050            Yes          2 Not Graduate            No
## 25  LP001052   Male     Yes          1     Graduate
## 26  LP001066   Male     Yes          0     Graduate           Yes
## 27  LP001068   Male     Yes          0     Graduate            No
## 28  LP001073   Male     Yes          2 Not Graduate            No
## 29  LP001086   Male      No          0 Not Graduate            No
## 30  LP001087 Female      No          2     Graduate
## 31  LP001091   Male     Yes          1     Graduate
## 32  LP001095   Male      No          0     Graduate            No
## 33  LP001097   Male      No          1     Graduate           Yes
## 34  LP001098   Male     Yes          0     Graduate            No
## 35  LP001100   Male      No         3+     Graduate            No
## 36  LP001106   Male     Yes          0     Graduate            No
## 37  LP001109   Male     Yes          0     Graduate            No
## 38  LP001112 Female     Yes          0     Graduate            No
## 39  LP001114   Male      No          0     Graduate            No
## 40  LP001116   Male      No          0 Not Graduate            No
## 41  LP001119   Male      No          0     Graduate            No
## 42  LP001120   Male      No          0     Graduate            No
## 43  LP001123   Male     Yes          0     Graduate            No
## 44  LP001131   Male     Yes          0     Graduate            No
## 45  LP001136   Male     Yes          0 Not Graduate           Yes
## 46  LP001137 Female      No          0     Graduate            No
## 47  LP001138   Male     Yes          1     Graduate            No
## 48  LP001144   Male     Yes          0     Graduate            No
## 49  LP001146 Female     Yes          0     Graduate            No
## 50  LP001151 Female      No          0     Graduate            No
## 51  LP001155 Female     Yes          0 Not Graduate            No
## 52  LP001157 Female      No          0     Graduate            No
## 53  LP001164 Female      No          0     Graduate            No
## 54  LP001179   Male     Yes          2     Graduate            No
```

```
## 55  LP001186 Female     Yes          1     Graduate           Yes
## 56  LP001194   Male     Yes          2     Graduate            No
## 57  LP001195   Male     Yes          0     Graduate            No
## 58  LP001197   Male     Yes          0     Graduate            No
## 59  LP001198   Male     Yes          1     Graduate            No
## 60  LP001199   Male     Yes          2 Not Graduate            No
## 61  LP001205   Male     Yes          0     Graduate            No
## 62  LP001206   Male     Yes         3+     Graduate            No
## 63  LP001207   Male     Yes          0 Not Graduate           Yes
## 64  LP001213   Male     Yes          1     Graduate            No
## 65  LP001222 Female      No          0     Graduate            No
## 66  LP001225   Male     Yes          0     Graduate            No
## 67  LP001228   Male      No          0 Not Graduate            No
## 68  LP001233   Male     Yes          1     Graduate            No
## 69  LP001238   Male     Yes         3+ Not Graduate           Yes
## 70  LP001241 Female      No          0     Graduate            No
## 71  LP001243   Male     Yes          0     Graduate            No
## 72  LP001245   Male     Yes          2 Not Graduate           Yes
## 73  LP001248   Male      No          0     Graduate            No
## 74  LP001250   Male     Yes         3+ Not Graduate            No
## 75  LP001253   Male     Yes         3+     Graduate           Yes
## 76  LP001255   Male      No          0     Graduate            No
## 77  LP001256   Male      No          0     Graduate            No
## 78  LP001259   Male     Yes          1     Graduate           Yes
## 79  LP001263   Male     Yes         3+     Graduate            No
## 80  LP001264   Male     Yes         3+ Not Graduate           Yes
## 81  LP001265 Female      No          0     Graduate            No
## 82  LP001266   Male     Yes          1     Graduate           Yes
## 83  LP001267 Female     Yes          2     Graduate            No
## 84  LP001273   Male     Yes          0     Graduate            No
## 85  LP001275   Male     Yes          1     Graduate            No
## 86  LP001279   Male      No          0     Graduate            No
## 87  LP001280   Male     Yes          2 Not Graduate            No
## 88  LP001282   Male     Yes          0     Graduate            No
## 89  LP001289   Male      No          0     Graduate            No
## 90  LP001310   Male     Yes          0     Graduate            No
## 91  LP001316   Male     Yes          0     Graduate            No
## 92  LP001318   Male     Yes          2     Graduate            No
## 93  LP001319   Male     Yes          2 Not Graduate            No
## 94  LP001322   Male      No          0     Graduate            No
## 95  LP001325   Male      No          0 Not Graduate            No
## 96  LP001326   Male      No          0     Graduate
## 97  LP001327 Female     Yes          0     Graduate            No
## 98  LP001333   Male     Yes          0     Graduate            No
## 99  LP001334   Male     Yes          0 Not Graduate            No
## 100 LP001343   Male     Yes          0     Graduate            No
## 101 LP001345   Male     Yes          2 Not Graduate            No
## 102 LP001349   Male      No          0     Graduate            No
## 103 LP001350   Male     Yes                Graduate            No
## 104 LP001356   Male     Yes          0     Graduate            No
## 105 LP001357   Male                        Graduate            No
## 106            Male     Yes          1     Graduate            No
## 107 LP001369   Male     Yes          2     Graduate            No
##     ApplicantIncome CoapplicantIncome LoanAmount Loan_Amount_Term
## 1              5849                  0         NA              360
## 2              4583               1508        128              360
```

```
## 3          3000              0        66          360
## 4          2583           2358       120          360
## 5          6000              0       141          360
## 6          5417           4196       267          360
## 7          2333           1516        95          360
## 8          3036           2504       158          360
## 9          4006           1526       168          360
## 10        12841          10968       349          360
## 11         3200            700        70          360
## 12         2500           1840       109          360
## 13         3073           8106       200          360
## 14         1853           2840       114          360
## 15         1299           1086        17          120
## 16         4950              0       125          360
## 17         3596              0       100          240
## 18         3510              0        76          360
## 19         4887              0       133          360
## 20         2600           3500       115           NA
## 21         7660              0       104          360
## 22         5955           5625       315          360
## 23         2600           1911       116          360
## 24         3365           1917       112          360
## 25         3717           2925       151          360
## 26         9560              0       191          360
## 27         2799           2253       122          360
## 28         4226           1040       110          360
## 29         1442              0        35          360
## 30         3750           2083       120          360
## 31         4166           3369       201          360
## 32         3167              0        74          360
## 33         4692              0       106          360
## 34         3500           1667       114          360
## 35        12500           3000       320          360
## 36         2275           2067        NA          360
## 37         1828           1330       100           NA
## 38         3667           1459       144          360
## 39         4166           7210       184          360
## 40         3748           1668       110          360
## 41         3600              0        80          360
## 42         1800           1213        47          360
## 43         2400              0        75          360
## 44         3941           2336       134          360
## 45         4695              0        96           NA
## 46         3410              0        88           NA
## 47         5649              0        44          360
## 48         5821              0       144          360
## 49         2645           3440       120          360
## 50         4000           2275       144          360
## 51         1928           1644       100          360
## 52         3086              0       120          360
## 53         4230              0       112          360
## 54         4616              0       134          360
## 55        11500              0       286          360
## 56         2708           1167        97          360
## 57         2132           1591        96          360
## 58         3366           2200       135          360
```

```
## 59             8080              2250        180              360
## 60             3357              2859        144              360
## 61             2500              3796        120              360
## 62             3029                 0         99              360
## 63             2609              3449        165              180
## 64             4945                 0         NA              360
## 65             4166                 0        116              360
## 66             5726              4595        258              360
## 67             3200              2254        126              180
## 68            10750                 0        312              360
## 69             7100                 0        125               60
## 70             4300                 0        136              360
## 71             3208              3066        172              360
## 72             1875              1875         97              360
## 73             3500                 0         81              300
## 74             4755                 0         95               NA
## 75             5266              1774        187              360
## 76             3750                 0        113              480
## 77             3750              4750        176              360
## 78             1000              3022        110              360
## 79             3167              4000        180              300
## 80             3333              2166        130              360
## 81             3846                 0        111              360
## 82             2395                 0         NA              360
## 83             1378              1881        167              360
## 84             6000              2250        265              360
## 85             3988                 0         50              240
## 86             2366              2531        136              360
## 87             3333              2000         99              360
## 88             2500              2118        104              360
## 89             8566                 0        210              360
## 90             5695              4167        175              360
## 91             2958              2900        131              360
## 92             6250              5654        188              180
## 93             3273              1820         81              360
## 94             4133                 0        122              360
## 95             3620                 0         25              120
## 96             6782                 0         NA              360
## 97             2484              2302        137              360
## 98             1977               997         50              360
## 99             4188                 0        115              180
## 100            1759              3541        131              360
## 101            4288              3263        133              180
## 102            4843              3806        151              360
## 103           13650                 0         NA              360
## 104            4652              3583         NA              360
## 105            3816               754        160              360
## 106            3052              1030        100              360
## 107           11417              1126        225              360
##     Credit_History Property_Area Loan_Status
## 1                1         Urban           Y
## 2                1         Rural           N
## 3                1         Urban           Y
## 4                1         Urban           Y
## 5                1         Urban           Y
## 6                1         Urban           Y
```

```
## 7                1         Urban           Y
## 8                0     Semiurban           N
## 9                1         Urban           Y
## 10               1     Semiurban           N
## 11               1         Urban           Y
## 12               1         Urban           Y
## 13               1         Urban           Y
## 14               1         Rural           N
## 15               1         Urban           Y
## 16               1         Urban           Y
## 17              NA         Urban           Y
## 18               0         Urban           N
## 19               1         Rural           N
## 20               1         Urban           Y
## 21               0         Urban           N
## 22               1         Urban           Y
## 23               0     Semiurban           N
## 24               0         Rural           N
## 25              NA     Semiurban           N
## 26               1     Semiurban           Y
## 27               1     Semiurban           Y
## 28               1         Urban           Y
## 29               1         Urban           N
## 30               1     Semiurban           Y
## 31              NA         Urban           N
## 32               1         Urban           N
## 33               1         Rural           N
## 34               1     Semiurban           Y
## 35               1         Rural           N
## 36               1         Urban           Y
## 37               0         Urban           N
## 38               1     Semiurban           Y
## 39               1         Urban           Y
## 40               1     Semiurban           Y
## 41               1         Urban           N
## 42               1         Urban           Y
## 43              NA         Urban           Y
## 44               1     Semiurban           Y
## 45               1         Urban           Y
## 46               1         Urban           Y
## 47               1         Urban           Y
## 48               1         Urban           Y
## 49               0         Urban           N
## 50               1     Semiurban           Y
## 51               1     Semiurban           Y
## 52               1     Semiurban           Y
## 53               1     Semiurban           N
## 54               1         Urban           N
## 55               0         Urban           N
## 56               1     Semiurban           Y
## 57               1     Semiurban           Y
## 58               1         Rural           N
## 59               1         Urban           Y
## 60               1         Urban           Y
## 61               1         Urban           Y
## 62               1         Urban           Y
```

```
## 63               0         Rural           N
## 64               0         Rural           N
## 65               0     Semiurban           N
## 66               1     Semiurban           N
## 67               0         Urban           N
## 68               1         Urban           Y
## 69               1         Urban           Y
## 70               0     Semiurban           N
## 71               1         Urban           Y
## 72               1     Semiurban           Y
## 73               1     Semiurban           Y
## 74               0     Semiurban           N
## 75               1     Semiurban           Y
## 76               1         Urban           N
## 77               1         Urban           N
## 78               1         Urban           N
## 79               0     Semiurban           N
## 80              NA     Semiurban           Y
## 81               1     Semiurban           Y
## 82               1     Semiurban           Y
## 83               1         Urban           N
## 84              NA     Semiurban           N
## 85               1         Urban           Y
## 86               1     Semiurban           Y
## 87              NA     Semiurban           Y
## 88               1     Semiurban           Y
## 89               1         Urban           Y
## 90               1     Semiurban           Y
## 91               1     Semiurban           Y
## 92               1     Semiurban           Y
## 93               1         Urban           Y
## 94               1     Semiurban           Y
## 95               1     Semiurban           Y
## 96              NA         Urban           N
## 97               1     Semiurban           Y
## 98               1     Semiurban           Y
## 99               1     Semiurban           Y
## 100              1     Semiurban           Y
## 101              1         Urban           Y
## 102              1     Semiurban           Y
## 103              1         Urban           Y
## 104              1     Semiurban           Y
## 105              1         Urban           Y
## 106              1         Urban           Y
## 107              1         Urban           Y
```

```
# Load necessary libraries
library(dplyr)      # For data manipulation & Filtering, summarizing and manipulating data.
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(ggplot2)    # For visualizations & creating plots and graphs.
library(corrplot)   # For correlation matrix & visualizing correlation matrices.
```

```
## corrplot 0.95 loaded
```

```
library(GGally)     # For pairwise scatter plot and used for creating pairwise scatter plot m
atrices.
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
# These are like "extra tools" that help R perform tasks like organizing data and making char
ts in the output.
# Ensure that you have dplyr installed
# install.packages("dplyr")  # Uncomment if you haven't installed it

# Ensure ggplot2, corrplot, and GGally are installed
# install.packages("ggplot2")
# install.packages("corrplot")
# install.packages("GGally")

# Your existing data preprocessing steps here...


# Inspect the dataset structure
# This command shows us what kind of data each column contains (numbers, text, etc in your da
ta).
str(my_data)   # Displays structure of the dataset
```

```
## 'data.frame':    107 obs. of  13 variables:
##  $ Loan_ID          : chr  "LP001002" "LP001003" "LP001005" "LP001006" ...
##  $ Gender           : chr  "Male" "Male" "Male" "Male" ...
##  $ Married          : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Dependents       : chr  "0" "1" "0" "0" ...
##  $ Education        : chr  "Graduate" "Graduate" "Graduate" "Not Graduate" ...
##  $ Self_Employed    : chr  "No" "No" "Yes" "No" ...
##  $ ApplicantIncome  : int  5849 4583 3000 2583 6000 5417 2333 3036 4006 12841 ...
##  $ CoapplicantIncome: int  0 1508 0 2358 0 4196 1516 2504 1526 10968 ...
##  $ LoanAmount       : int  NA 128 66 120 141 267 95 158 168 349 ...
##  $ Loan_Amount_Term : int  360 360 360 360 360 360 360 360 360 360 ...
##  $ Credit_History   : int  1 1 1 1 1 1 1 0 1 1 ...
##  $ Property_Area    : chr  "Urban" "Rural" "Urban" "Urban" ...
##  $ Loan_Status      : chr  "Y" "N" "Y" "Y" ...
```

```
head(my_data)  # Shows first few rows
```

```
##    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
```

```
# Displays the first few rows so we can quickly see what the data looks like.

# Identify missing values
# Finds all the missing values & Counts how many missing values are in each column.
missing_values <- colSums(is.na(my_data))  # Count missing values per column
print(missing_values)
```

```
##           Loan_ID            Gender           Married        Dependents
##                 0                 0                 0                 0
##         Education     Self_Employed   ApplicantIncome CoapplicantIncome
##                 0                 0                 0                 0
##        LoanAmount  Loan_Amount_Term    Credit_History     Property_Area
##                 7                 5                 8                 0
##       Loan_Status
##                 0
```

```
# Handling missing values
# Drop columns with too many missing values (more than 50% missing)
# This line shows us If more than half the rows in a column are missing, remove that column.
threshold <- 0.5 * nrow(my_data)
my_data <- my_data[, colSums(is.na(my_data)) < threshold]

# Impute missing values
# For numerical columns: Replace missing values with the median
# This line shows us Finds all the numeric columns (numbers)    Replaces missing values in th
ose columns with the median (middle value).
num_cols <- sapply(my_data, is.numeric)
my_data[num_cols] <- lapply(my_data[num_cols], function(x) ifelse(is.na(x), median(x, na.rm =
TRUE), x))

# For categorical columns: Replace missing values with the mode & get unique non-NA values fr
equent values.
fill_mode <- function(x) {
  unique_x <- unique(x[!is.na(x)])
  mode_val <- unique_x[which.max(tabulate(match(x, unique_x)))]
  return(ifelse(is.na(x), mode_val, x))
}
# apply mode replacement to the categorical columns
cat_cols <- sapply(my_data, is.factor)
my_data[cat_cols] <- lapply(my_data[cat_cols], fill_mode)

# Check for duplicates rows from this line.
duplicates <- my_data[duplicated(my_data), ]
print(duplicates)  # Display duplicate rows if any
```

```
##  [1] Loan_ID           Gender            Married           Dependents
##  [5] Education         Self_Employed     ApplicantIncome   CoapplicantIncome
##  [9] LoanAmount        Loan_Amount_Term  Credit_History    Property_Area
## [13] Loan_Status
## <0 rows> (or 0-length row.names)
```

```
# Remove duplicates rows from this line.
my_data <- my_data[!duplicated(my_data), ]

# Identify and handle outliers using IQR method and replace outliers with NA & identify numer
ic columns and apply outlier removal to numerical columns.
remove_outliers <- function(x) {
  Q1 <- quantile(x, 0.25, na.rm = TRUE)
  Q3 <- quantile(x, 0.75, na.rm = TRUE)
  IQR_value <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR_value
  upper_bound <- Q3 + 1.5 * IQR_value
  x[x < lower_bound | x > upper_bound] <- NA
  return(x)
}

# Apply outlier removal to numerical columns with median values.
my_data[num_cols] <- lapply(my_data[num_cols], remove_outliers)

# Impute outliers with median (same as missing values handling)
my_data[num_cols] <- lapply(my_data[num_cols], function(x) ifelse(is.na(x), median(x, na.rm =
TRUE), x))

# Final dataset summary
# This line shows us command gives a quick overview of the dataset & It shows minimum, maximu
m, median, mean, and quartiles for numeric columns.
summary(my_data)
```

```
##    Loan_ID             Gender            Married           Dependents
##  Length:107         Length:107         Length:107         Length:107
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##   Education         Self_Employed      ApplicantIncome CoapplicantIncome
##  Length:107         Length:107         Min.   :1000    Min.   :   0
##  Class :character   Class :character   1st Qu.:2754    1st Qu.:   0
##  Mode  :character   Mode  :character   Median :3505    Median :1512
##                                        Mean   :3651    Mean   :1513
##                                        3rd Qu.:4228    3rd Qu.:2319
##                                        Max.   :7660    Max.   :5654
##    LoanAmount    Loan_Amount_Term Credit_History Property_Area
##  Min.   : 25.0   Min.   :360      Min.   :1      Length:107
##  1st Qu.:102.0   1st Qu.:360      1st Qu.:1      Class :character
##  Median :120.5   Median :360      Median :1      Mode  :character
##  Mean   :122.1   Mean   :360      Mean   :1
##  3rd Qu.:136.0   3rd Qu.:360      3rd Qu.:1
##  Max.   :225.0   Max.   :360      Max.   :1
##  Loan_Status
##  Length:107
##  Class :character
##  Mode  :character
##
##
##
```

```
# Save the cleaned dataset (optional)
# This line shows us that "cleaned_data.csv in the working directory & row.names = FALSE remo
ves unnecessary row numbers #print("Data preprocessing complete!") just confirms that everyth
ing is done.
write.csv(my_data, "cleaned_data.csv", row.names = FALSE)

# Print success message
# This line tells that "data preprocessing is complete"
print("Data preprocessing complete!")
```

```
## [1] "Data preprocessing complete!"
```

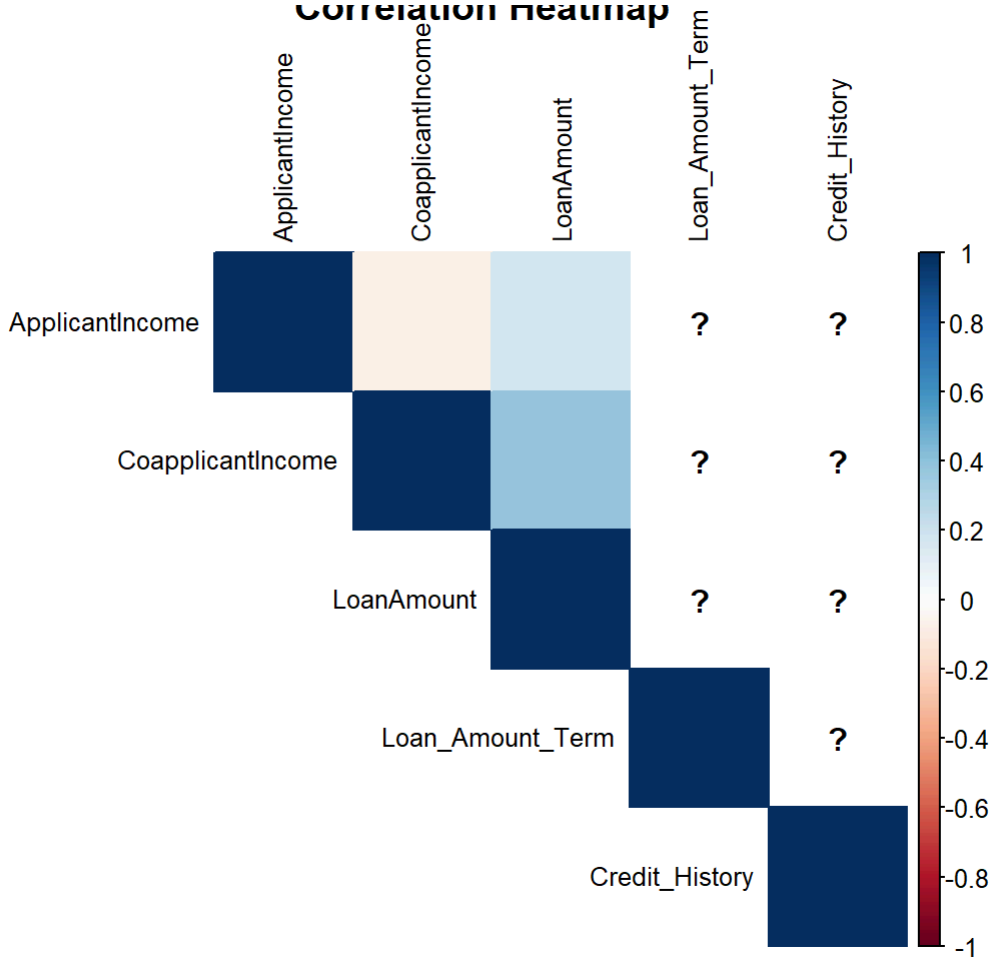
```
# ============================================
# 1. SUMMARY STATISTICS
# ============================================
# this code shows us that ExtRACTED ONLY NUMERIC Columns from the dataset calculates through
median & mean and min max.
summary_stats <- my_data[, num_cols] %>% summarise_all(list(
  mean = mean, median = median, sd = sd,
  min = min, max = max, range = ~ max(.) - min(.),
  Q1 = ~ quantile(., 0.25), Q3 = ~ quantile(., 0.75)
))
print(summary_stats)
```

```
##   ApplicantIncome_mean CoapplicantIncome_mean LoanAmount_mean
## 1             3651.234               1512.794        122.0794
##   Loan_Amount_Term_mean Credit_History_mean ApplicantIncome_median
## 1                   360                   1                   3505
##   CoapplicantIncome_median LoanAmount_median Loan_Amount_Term_median
## 1                     1512             120.5                      360
##   Credit_History_median ApplicantIncome_sd CoapplicantIncome_sd LoanAmount_sd
## 1                     1           1293.575             1471.742      36.88963
##   Loan_Amount_Term_sd Credit_History_sd ApplicantIncome_min
## 1                   0                 0                1000
##   CoapplicantIncome_min LoanAmount_min Loan_Amount_Term_min Credit_History_min
## 1                     0             25                  360                   1
##   ApplicantIncome_max CoapplicantIncome_max LoanAmount_max Loan_Amount_Term_max
## 1                7660                  5654            225                  360
##   Credit_History_max ApplicantIncome_range CoapplicantIncome_range
## 1                  1                  6660                    5654
##   LoanAmount_range Loan_Amount_Term_range Credit_History_range
## 1              200                     0                    0
##   ApplicantIncome_Q1 CoapplicantIncome_Q1 LoanAmount_Q1 Loan_Amount_Term_Q1
## 1             2753.5                    0           102                 360
##   Credit_History_Q1 ApplicantIncome_Q3 CoapplicantIncome_Q3 LoanAmount_Q3
## 1                 1               4228                 2319           136
##   Loan_Amount_Term_Q3 Credit_History_Q3
## 1                 360                 1
```

```
# ============================================
# 2. CORRELATION ANALYSIS
# ============================================

library(corrplot)
# This code shows us Calculates the correlation between all numeric columns & ingnore missing
valuews abd cretae a heatmap in the realtionships.
# Compute correlation matrix
correlation_matrix <- cor(my_data[, num_cols], use = "complete.obs")
```

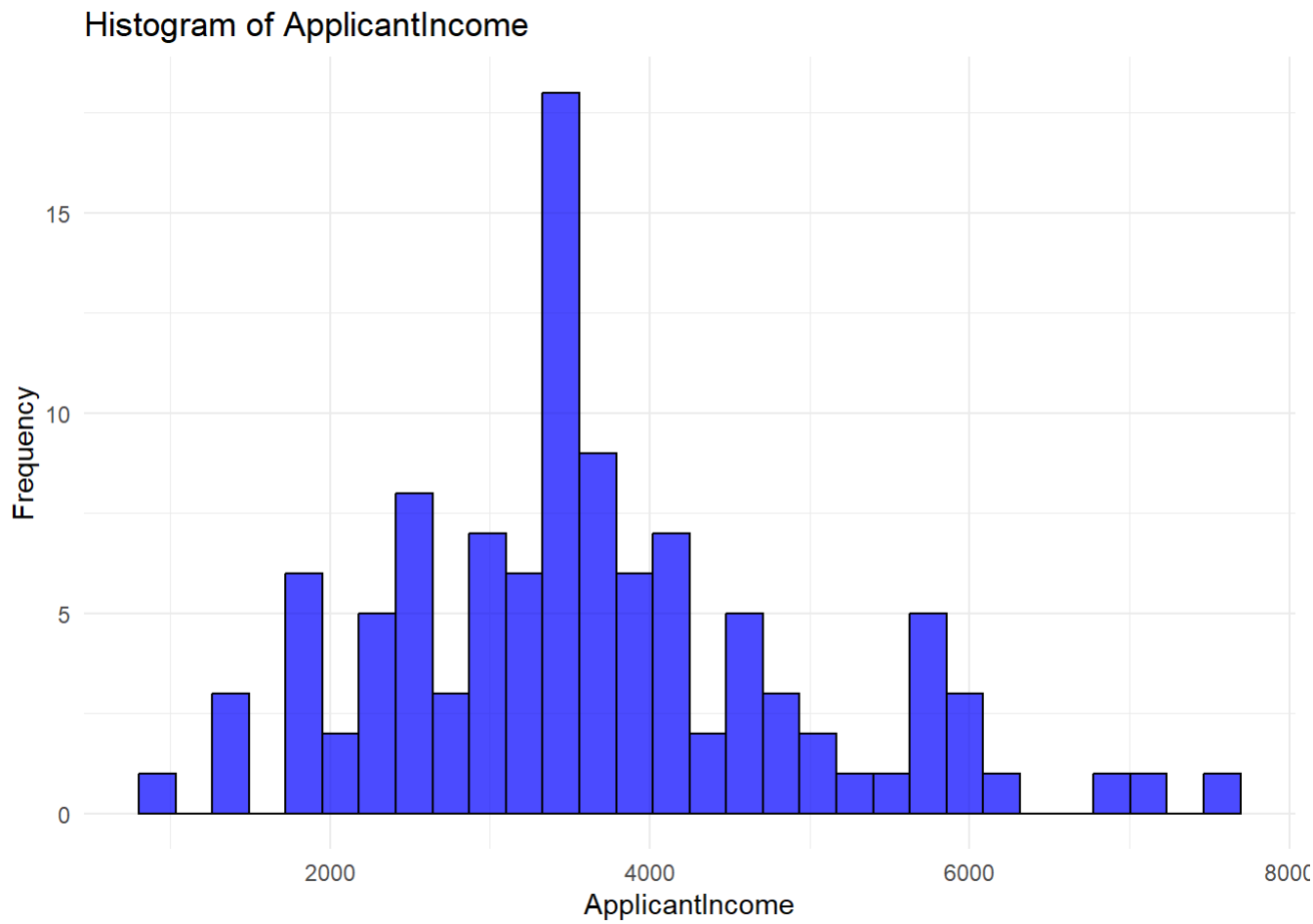
```
## Warning in cor(my_data[, num_cols], use = "complete.obs"): the standard
## deviation is zero
```

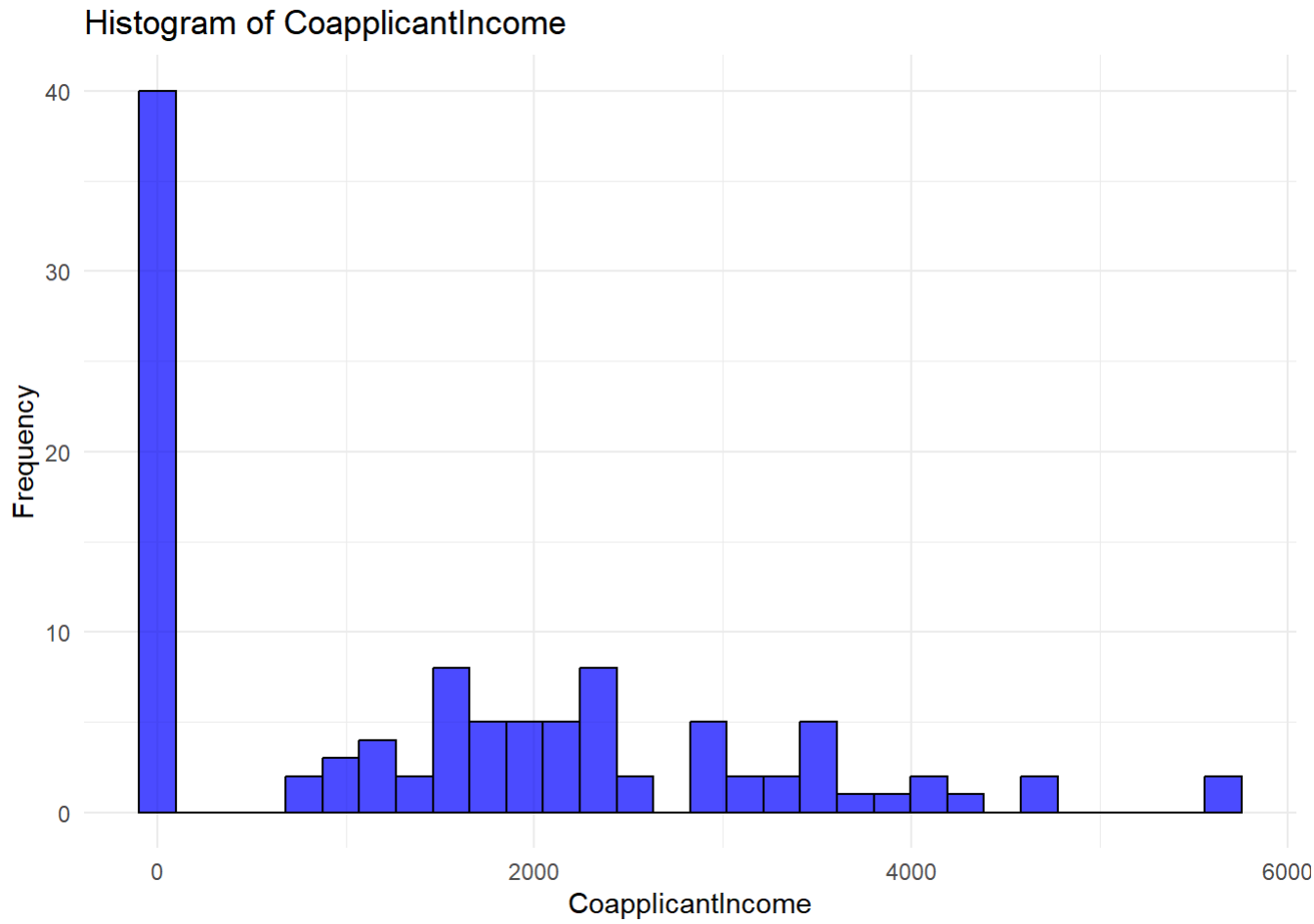
```
# Plot correlation heatmap
corrplot(correlation_matrix, method = "color", type = "upper",
         tl.cex = 0.8, tl.col = "black", title = "Correlation Heatmap")
```

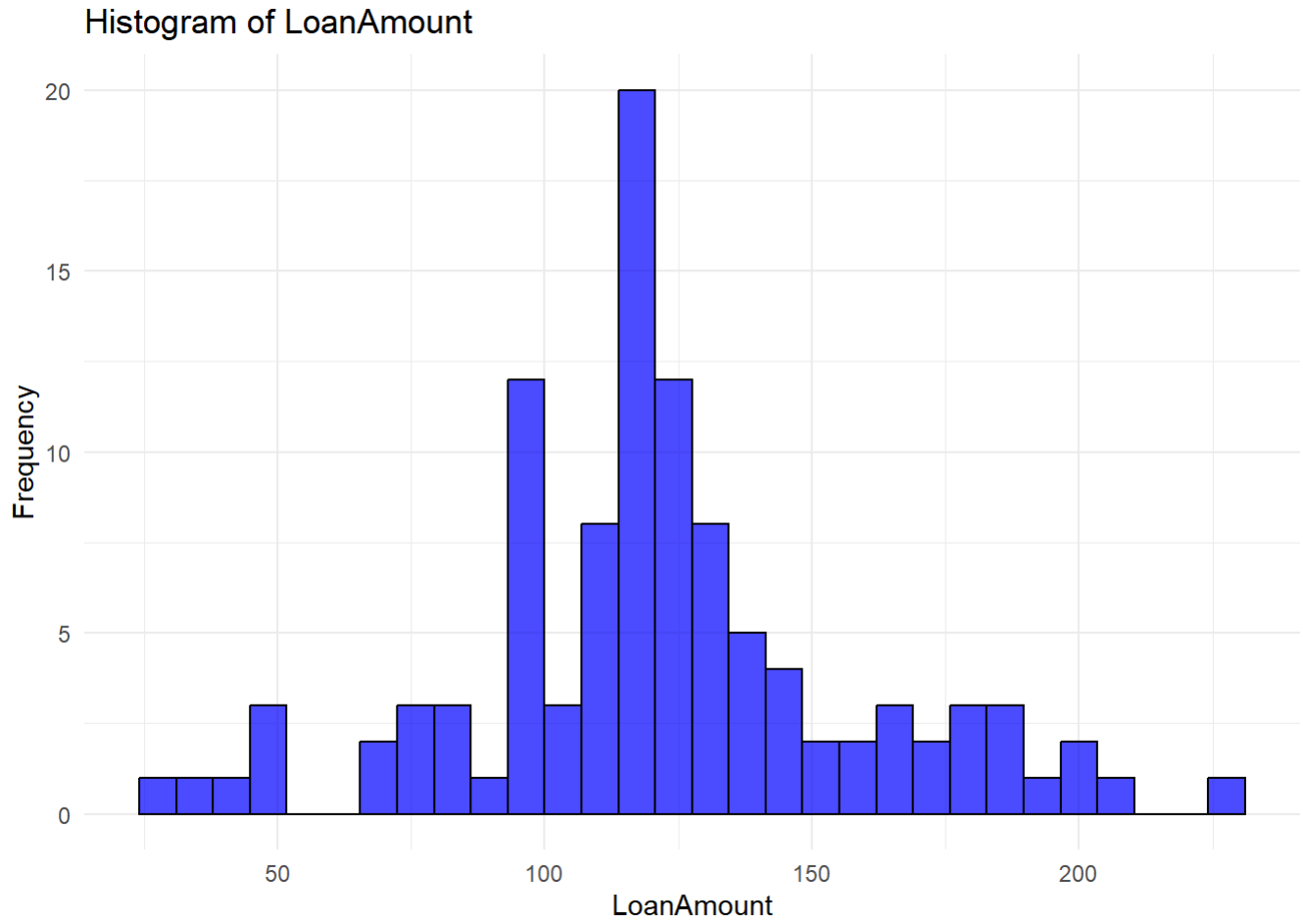


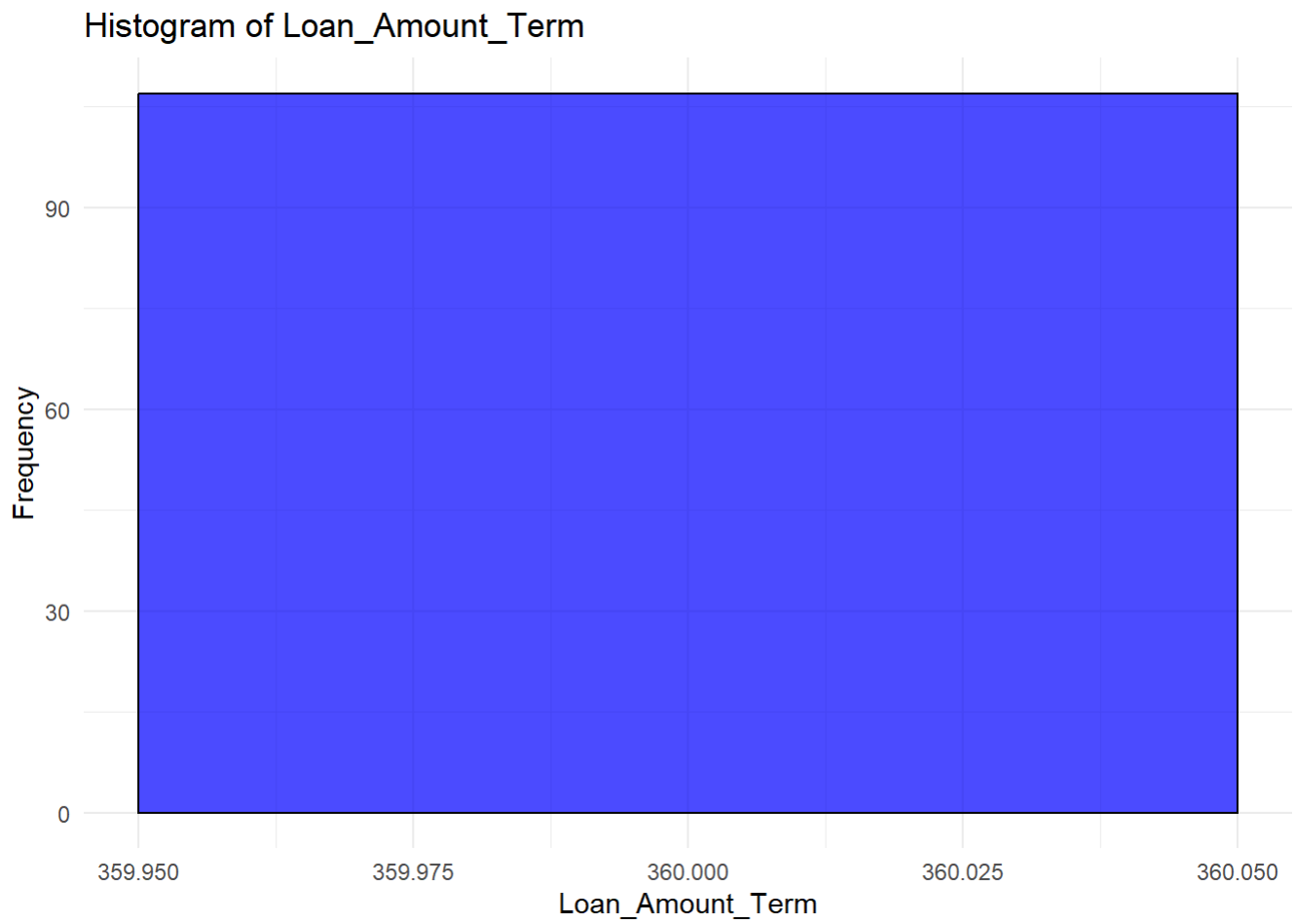

```
# ============================================
# 3. DATA VISUALIZATION
# ============================================

library(ggplot2)
library(GGally)

# Histograms for numerical variables
# The code shows us Loops through all numeric columns and plots a histogram & Divides data in
to bins and shows frequency & sets bar clours to identify the data.
for (col in names(my_data[, num_cols])) {
  print(ggplot(my_data, aes_string(x = col)) +
          geom_histogram(bins = 30, fill = "blue", color = "black", alpha = 0.7) +
          labs(title = paste("Histogram of", col), x = col, y = "Frequency") +
          theme_minimal())
}
```

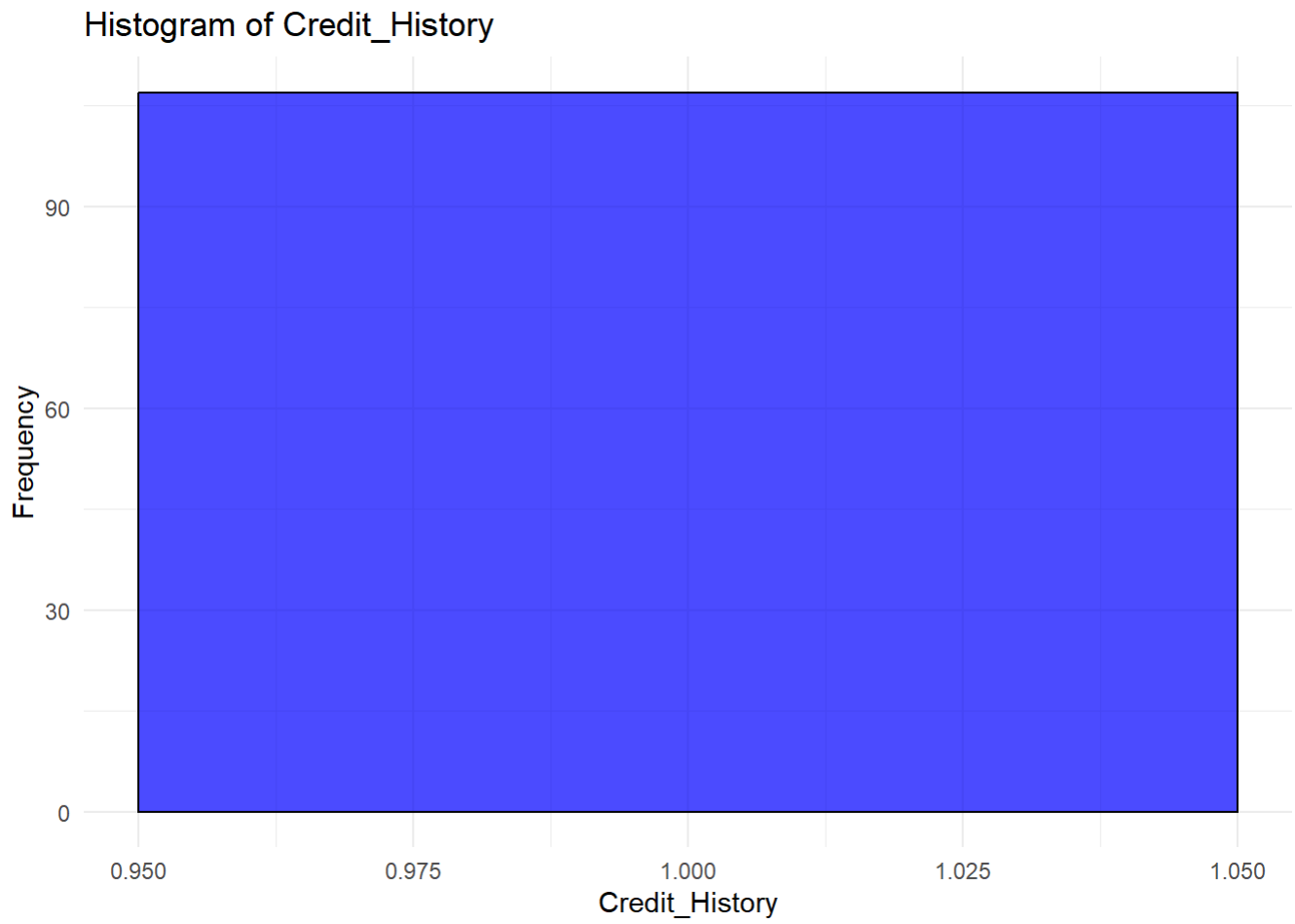
```
## Warning: `aes_string()` was deprecated in ggplot2 3.0.0.
## ℹ Please use tidy evaluation idioms with `aes()`.
## ℹ See also `vignette("ggplot2-in-packages")` for more information.
## This warning is displayed once every 8 hours.
## Call `lifecycle::last_lifecycle_warnings()` to see where this warning was
## generated.
```


Histogram of ApplicantIncome
Frequency
0
5
10
15
2000
4000
6000
8000
ApplicantIncome

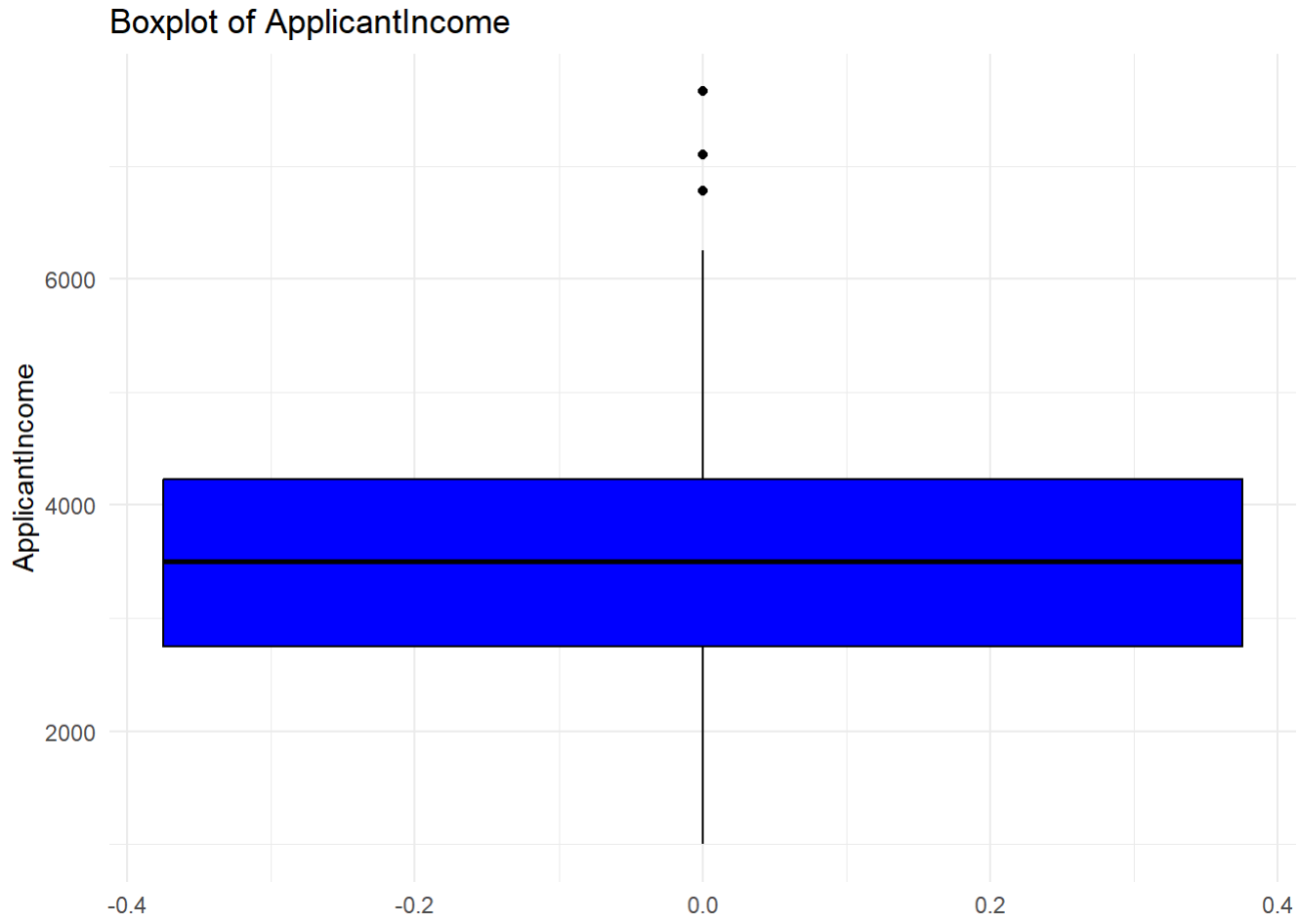

Histogram of CoapplicantIncome
Frequency
0
10
20
30
40
0
2000
4000
6000
CoapplicantIncome

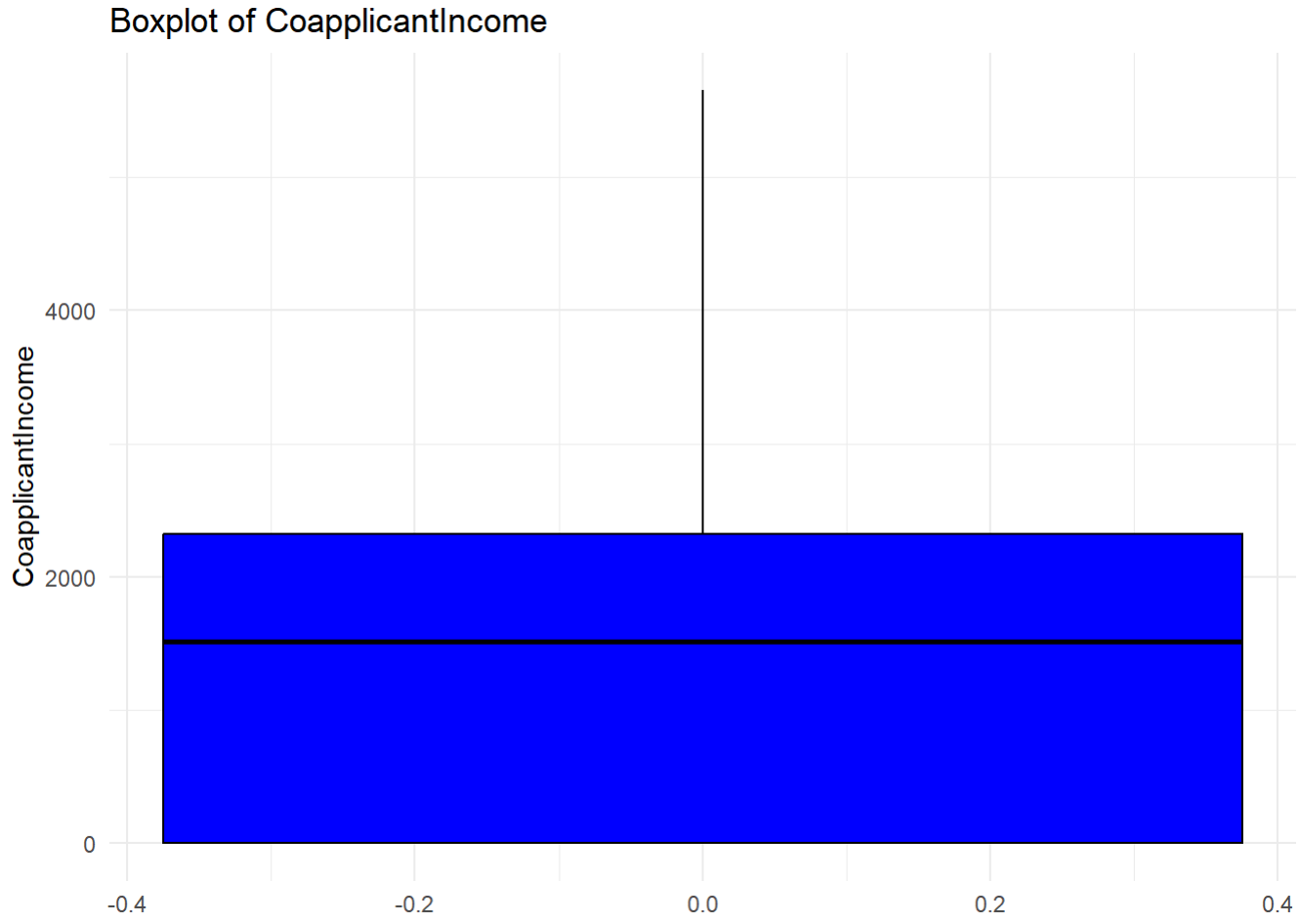

Histogram of LoanAmount
Frequency
0
5
10
15
20
50
100
150
200
LoanAmount

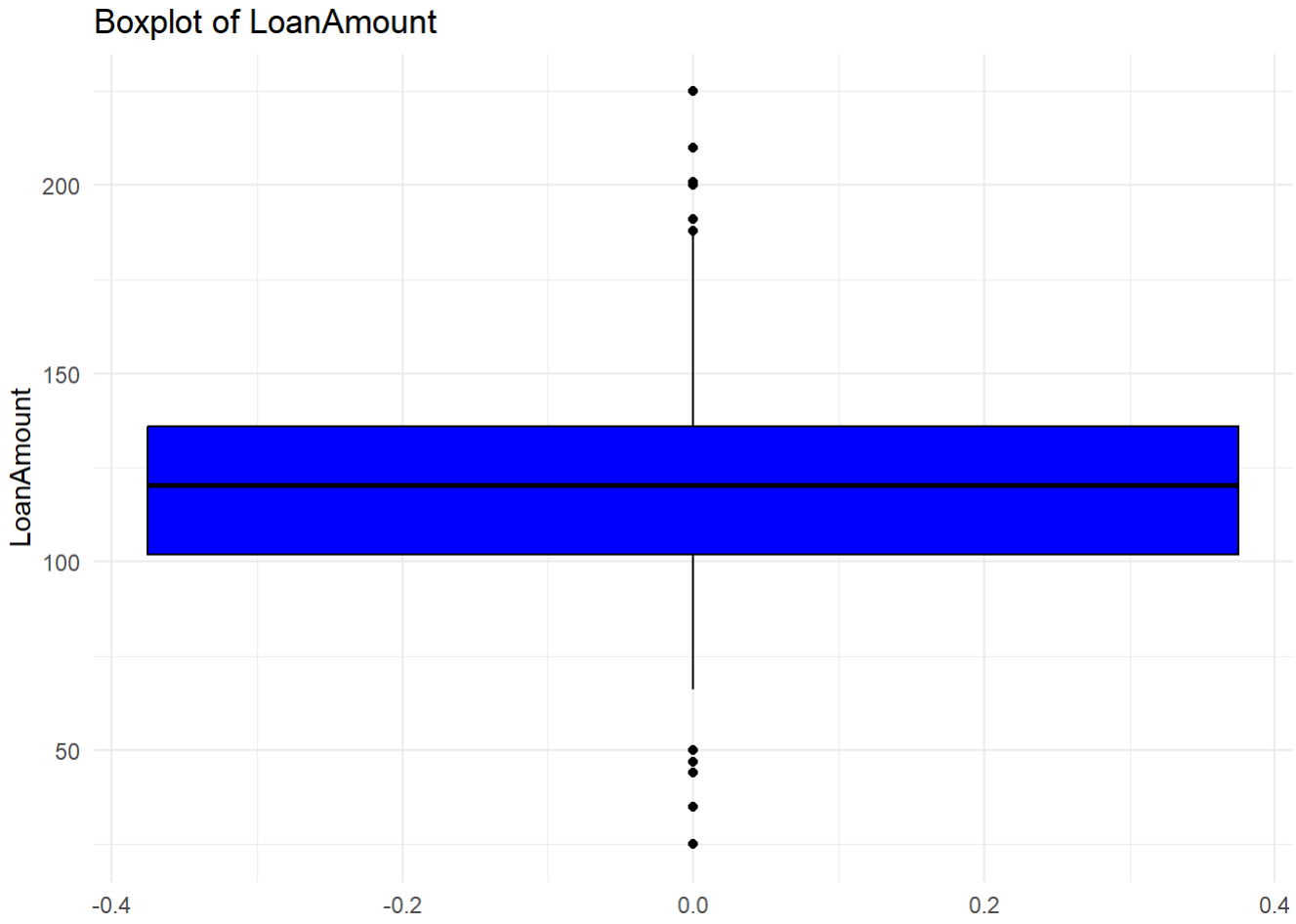

Histogram of Loan_Amount_Term
Frequency
0
30
60
90
359.950
359.975
360.000
360.025
360.050
Loan_Amount_Term

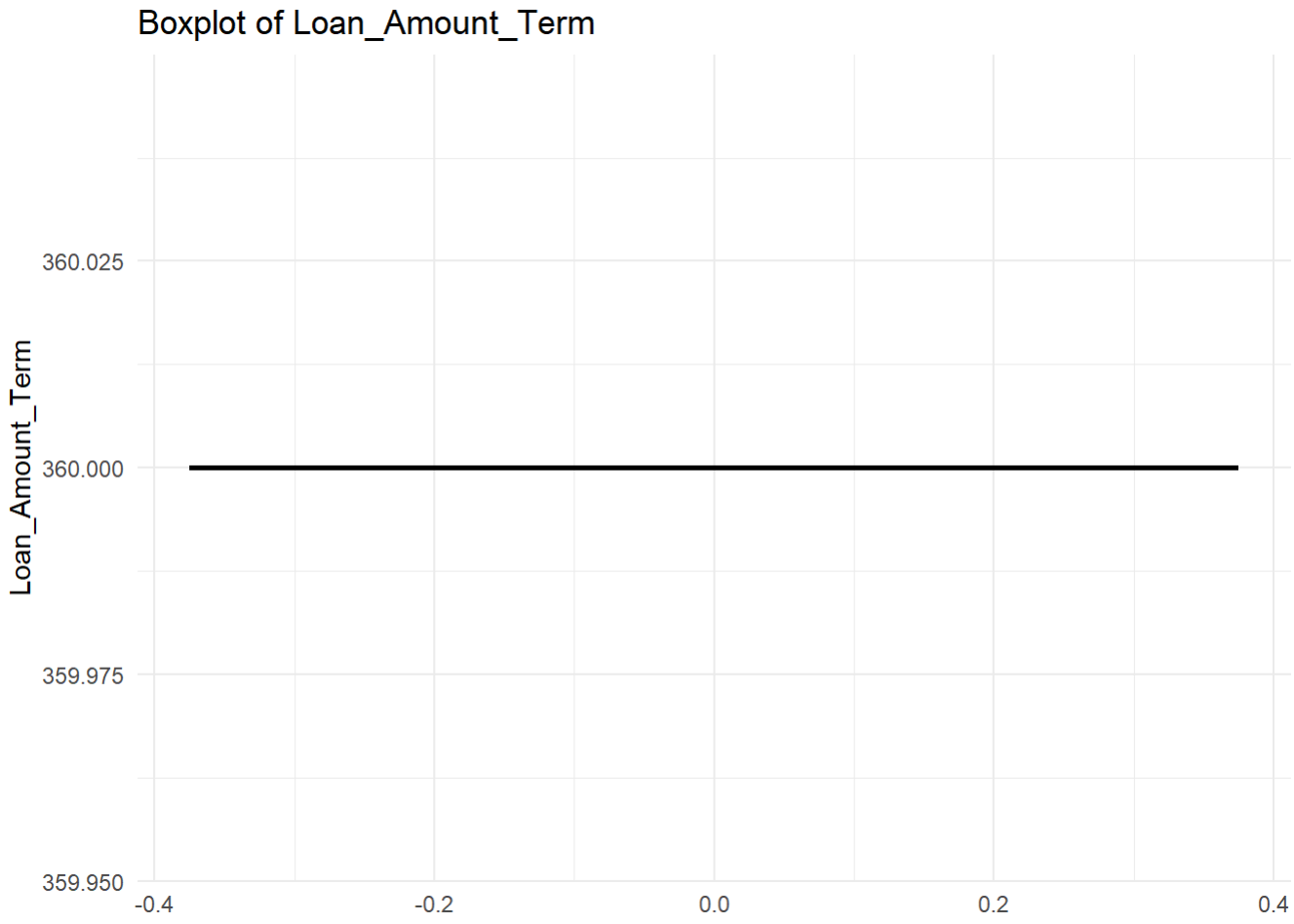


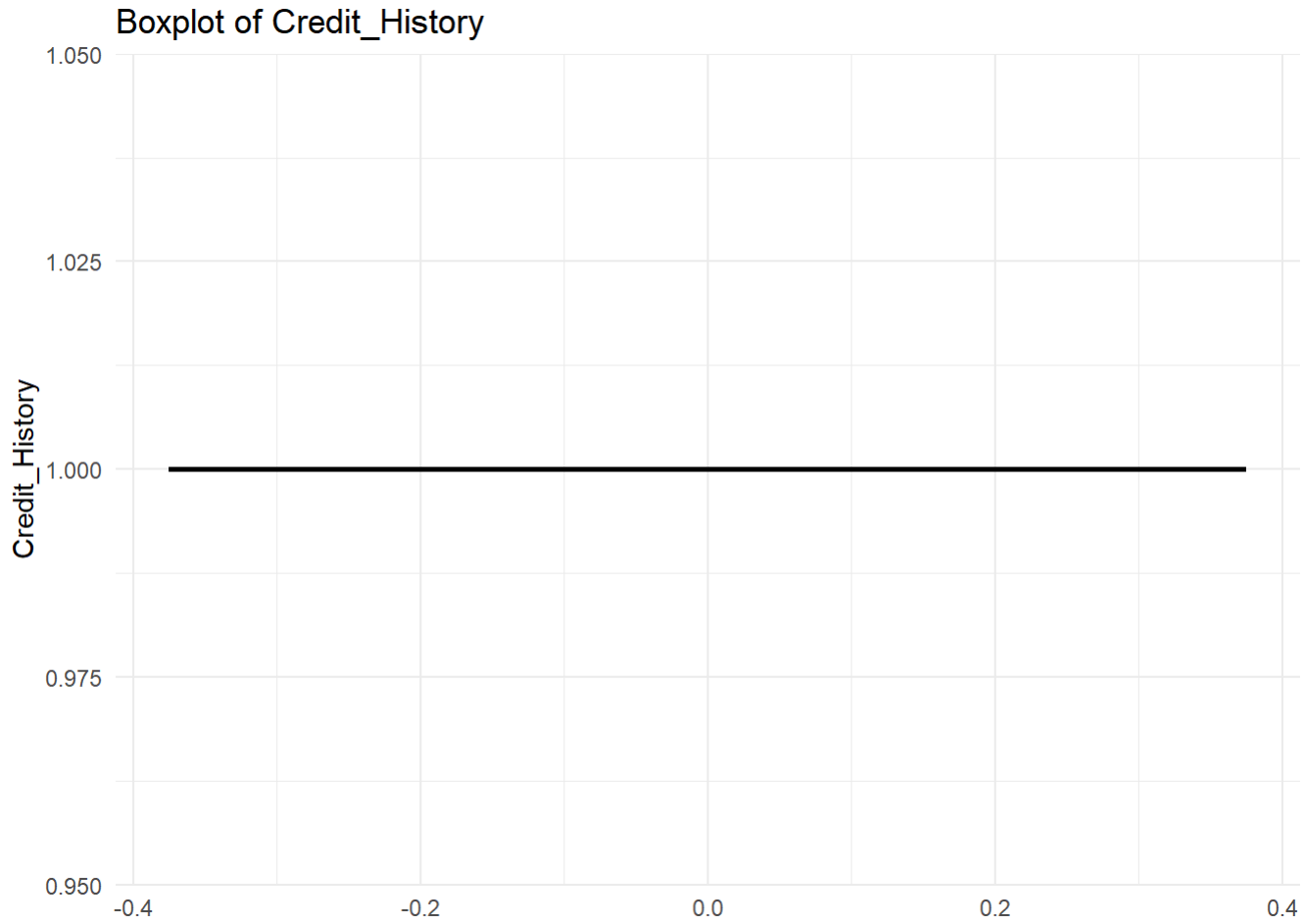

```
# This code shows us Loops through all numeric columns and creates a boxplot & Draws a boxplo
t.

# Boxplots for numerical variables
for (col in names(my_data[, num_cols])) {
  print(ggplot(my_data, aes_string(y = col)) +
          geom_boxplot(fill = "blue", color = "black") +
          labs(title = paste("Boxplot of", col), y = col) +
          theme_minimal())
}
```


Boxplot of ApplicantIncome
ApplicantIncome
6000
4000
2000
-0.4
-0.2
0.0
0.2
0.4


Boxplot of CoapplicantIncome
CoapplicantIncome
4000
2000
0
-0.4
-0.2
0.0
0.2
0.4


Boxplot of LoanAmount
LoanAmount
200
150
100
50
-0.4
-0.2
0.0
0.2
0.4

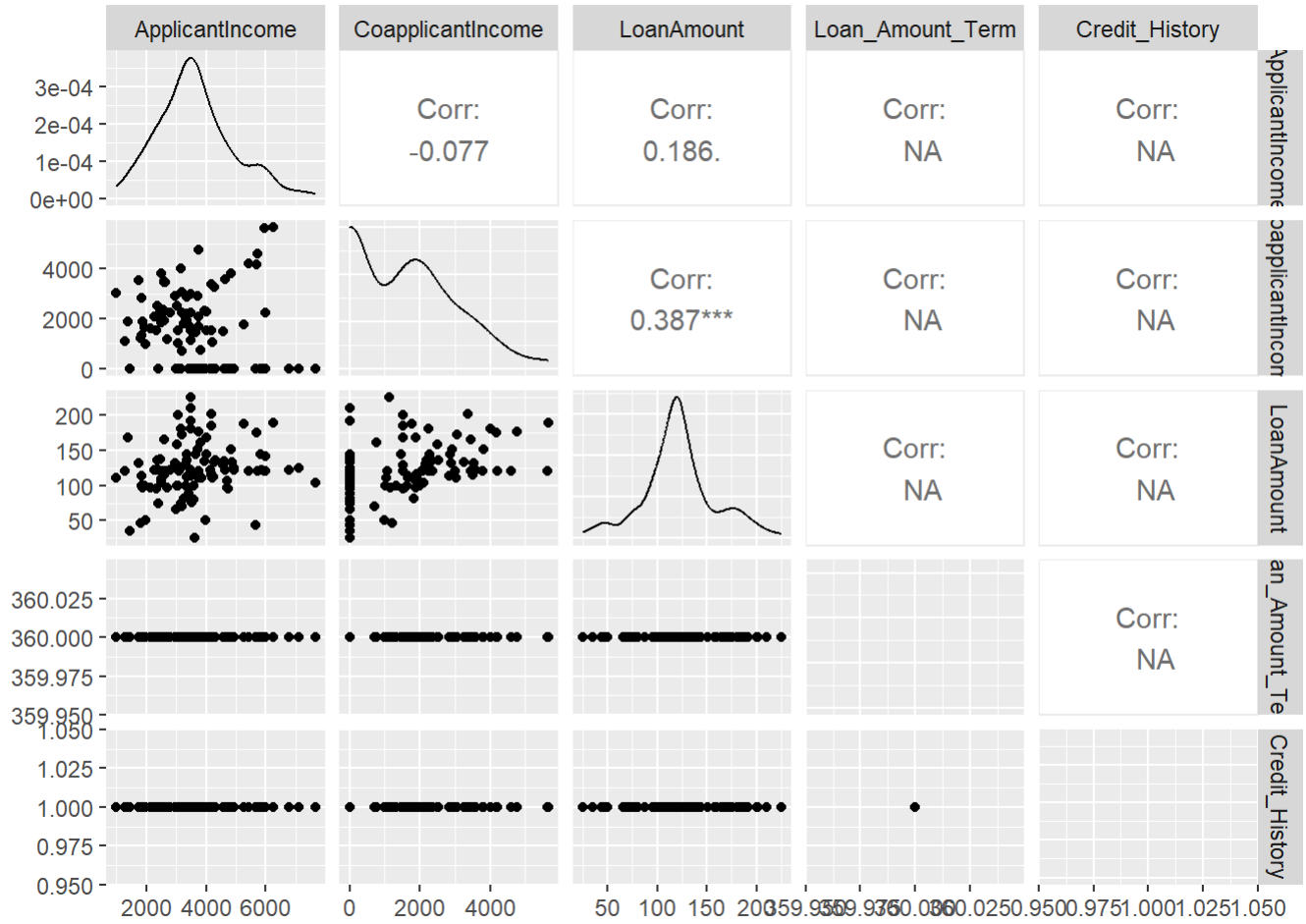

Boxplot of Loan_Amount_Term
Loan_Amount_Term
360.025
360.000
359.975
359.950
-0.4
-0.2
0.0
0.2
0.4



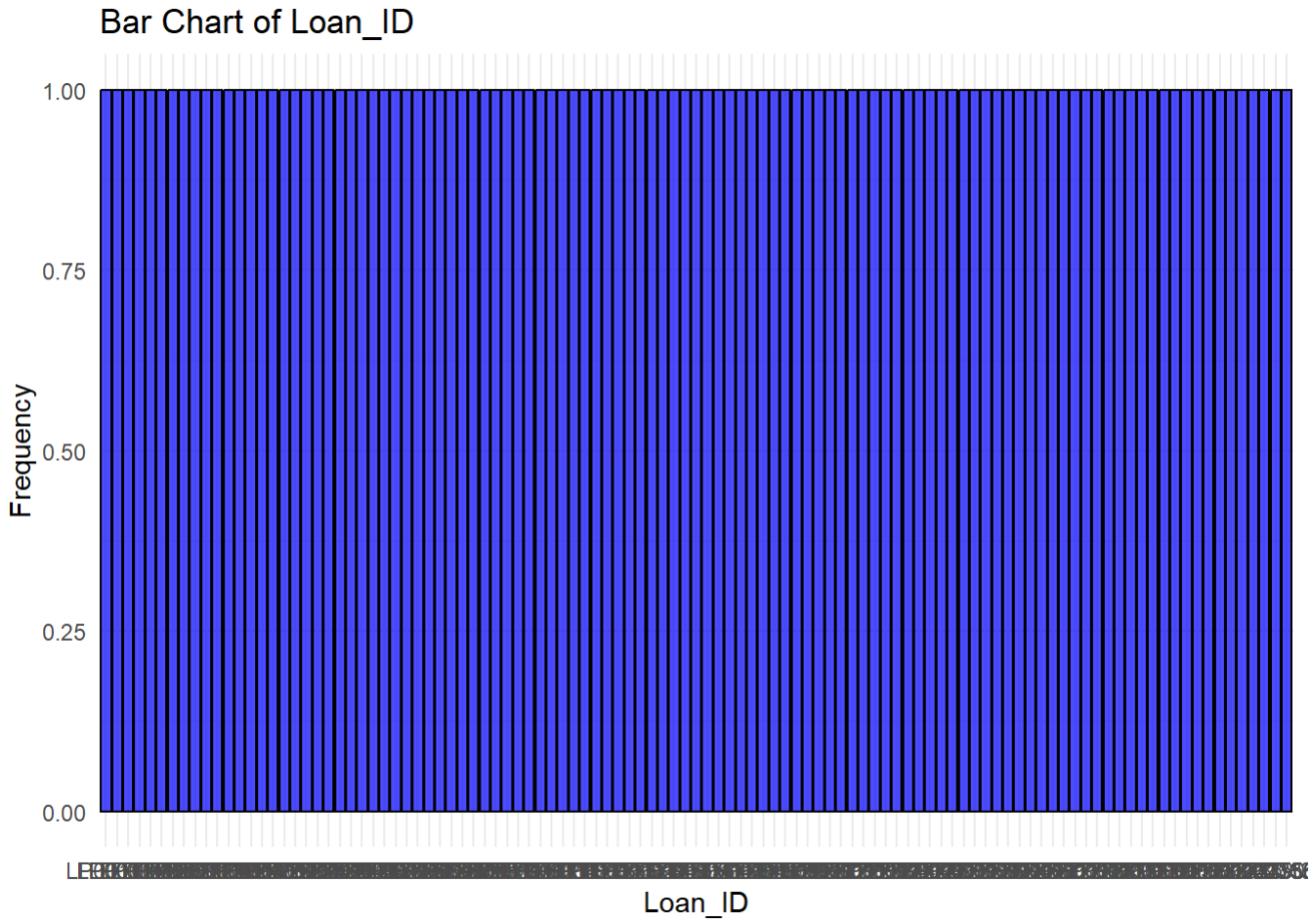

```
# Pairwise scatter plot matrix (correlation visualization)
# Creates scatter plots for all numeric variables.
ggpairs(my_data[, num_cols])
```

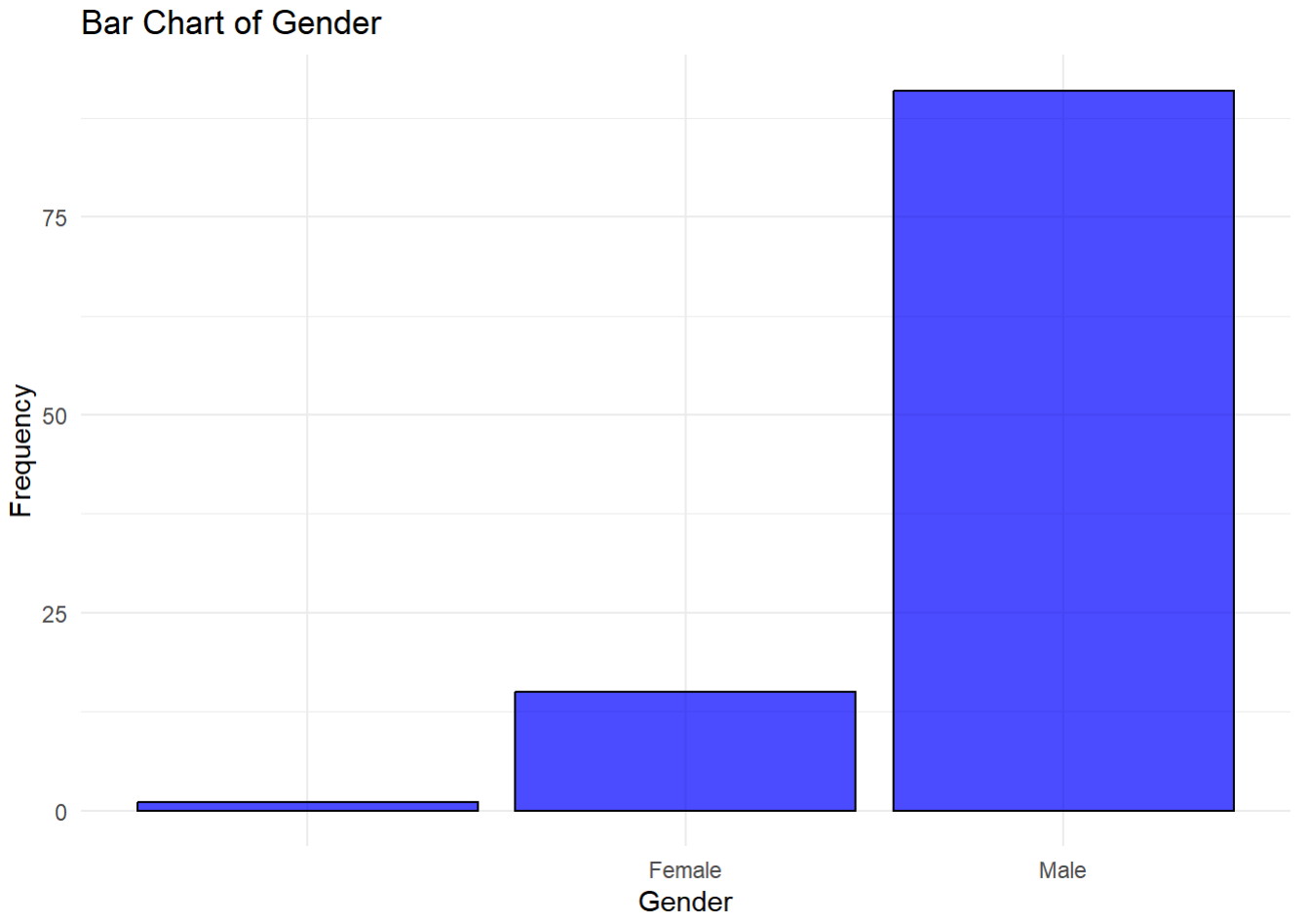
```
## Warning in cor(x, y): the standard deviation is zero
```

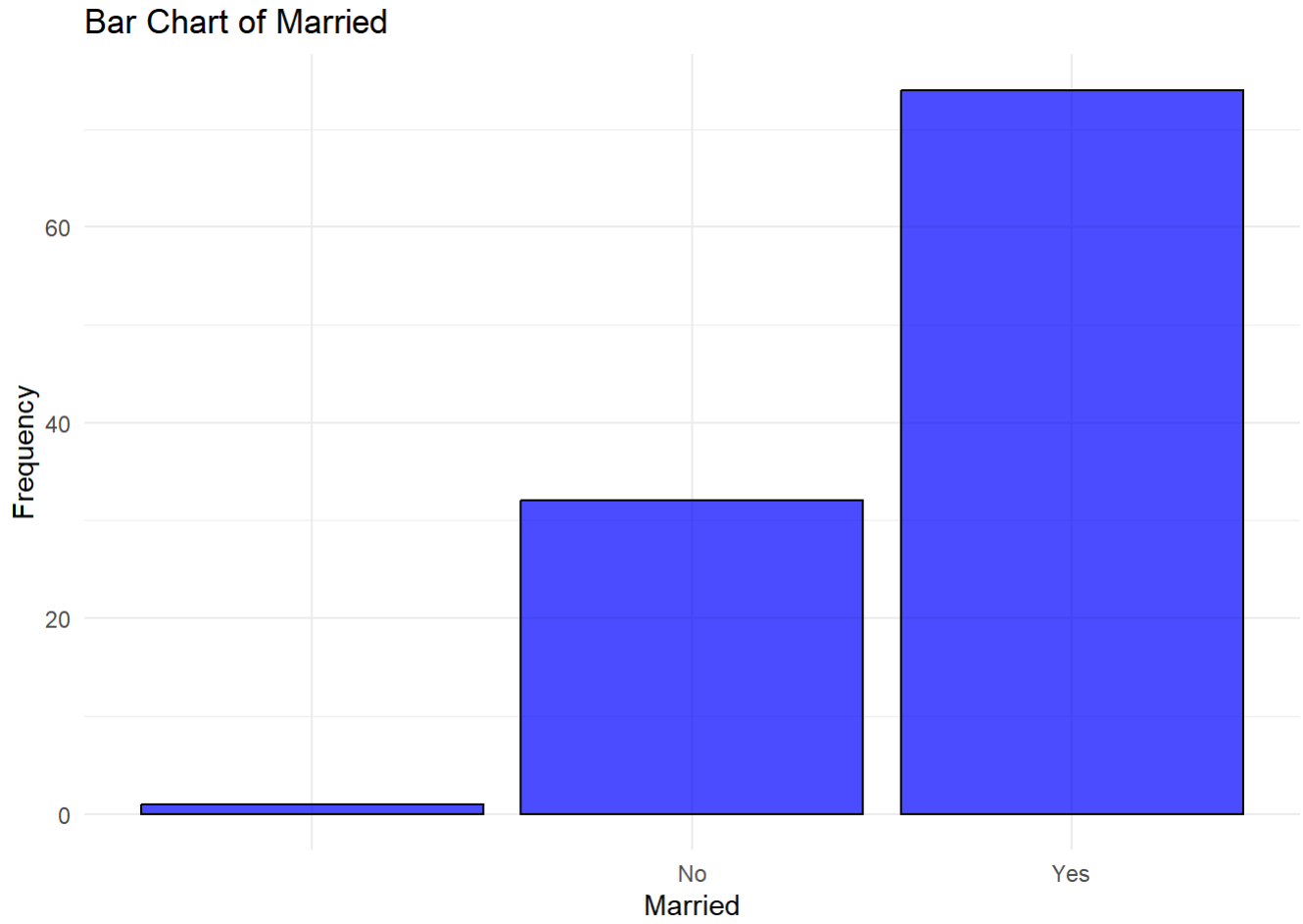
```
## Warning in cor(x, y): the standard deviation is zero
## Warning in cor(x, y): the standard deviation is zero
## Warning in cor(x, y): the standard deviation is zero
## Warning in cor(x, y): the standard deviation is zero
## Warning in cor(x, y): the standard deviation is zero
## Warning in cor(x, y): the standard deviation is zero
```

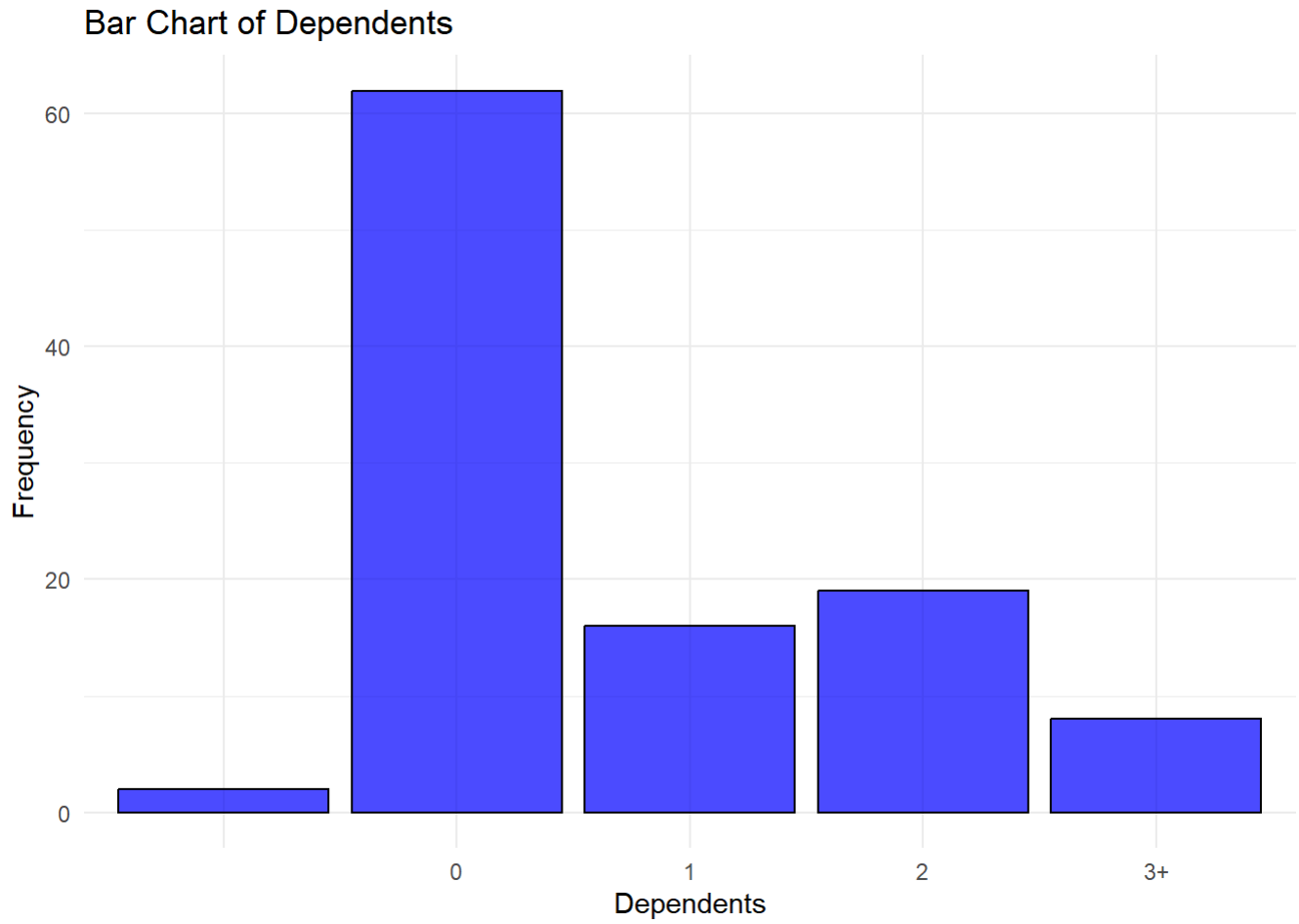


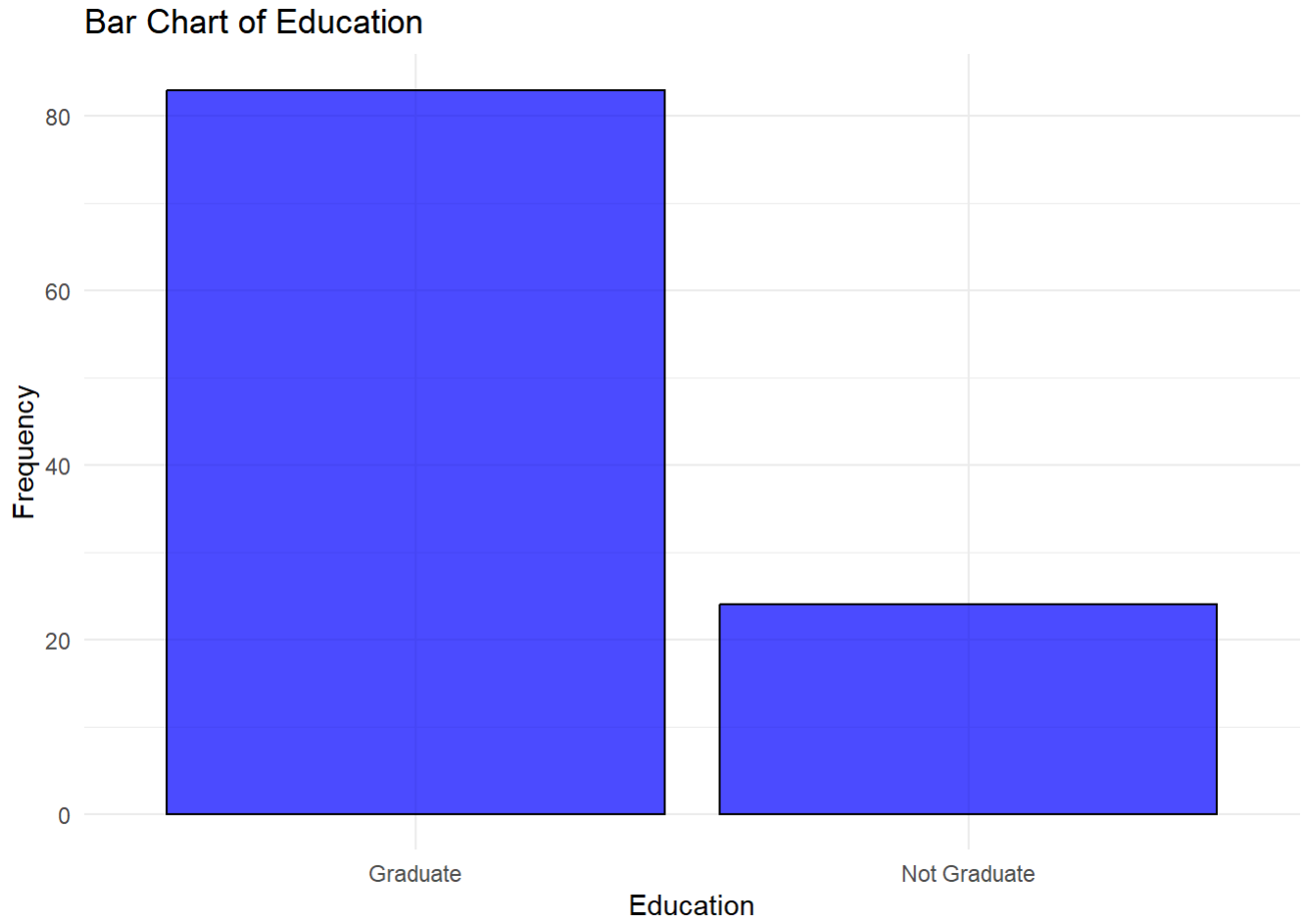

```
# Bar charts for categorical variables
# this code shows us Finds categorical (non-numeric) columns & Creates a bar chart showing co
unts of different categories.
cat_cols <- sapply(my_data, function(x) is.character(x) | is.factor(x))
for (col in names(my_data[, cat_cols])) {
  print(ggplot(my_data, aes_string(x = col)) +
          geom_bar(fill = "blue", color = "black", alpha = 0.7) +
          labs(title = paste("Bar Chart of", col), x = col, y = "Frequency") +
          theme_minimal())
}
```

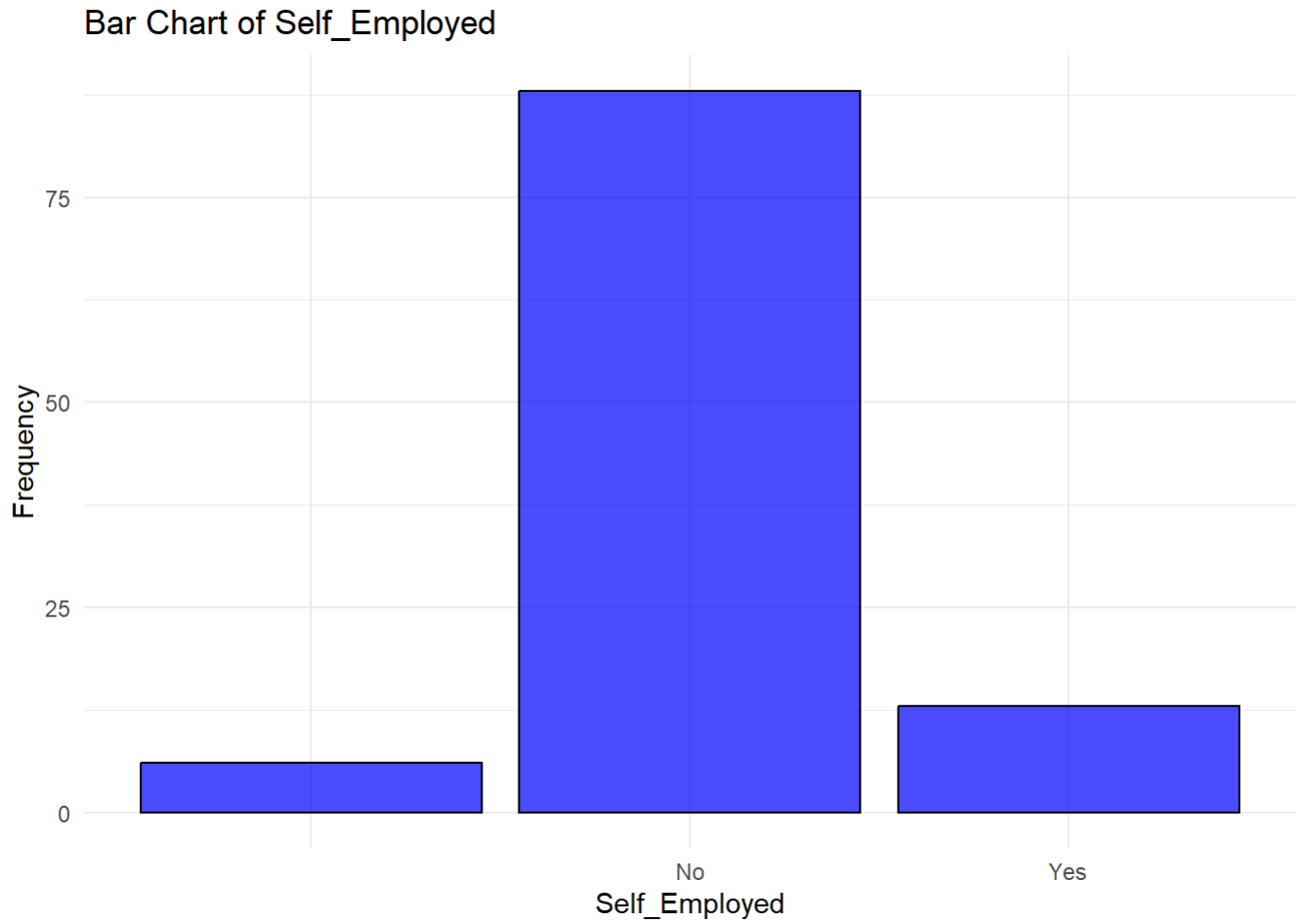

Bar Chart of Loan_ID
1.00
0.75
0.50
0.25
0.00
Frequency
Loan_ID

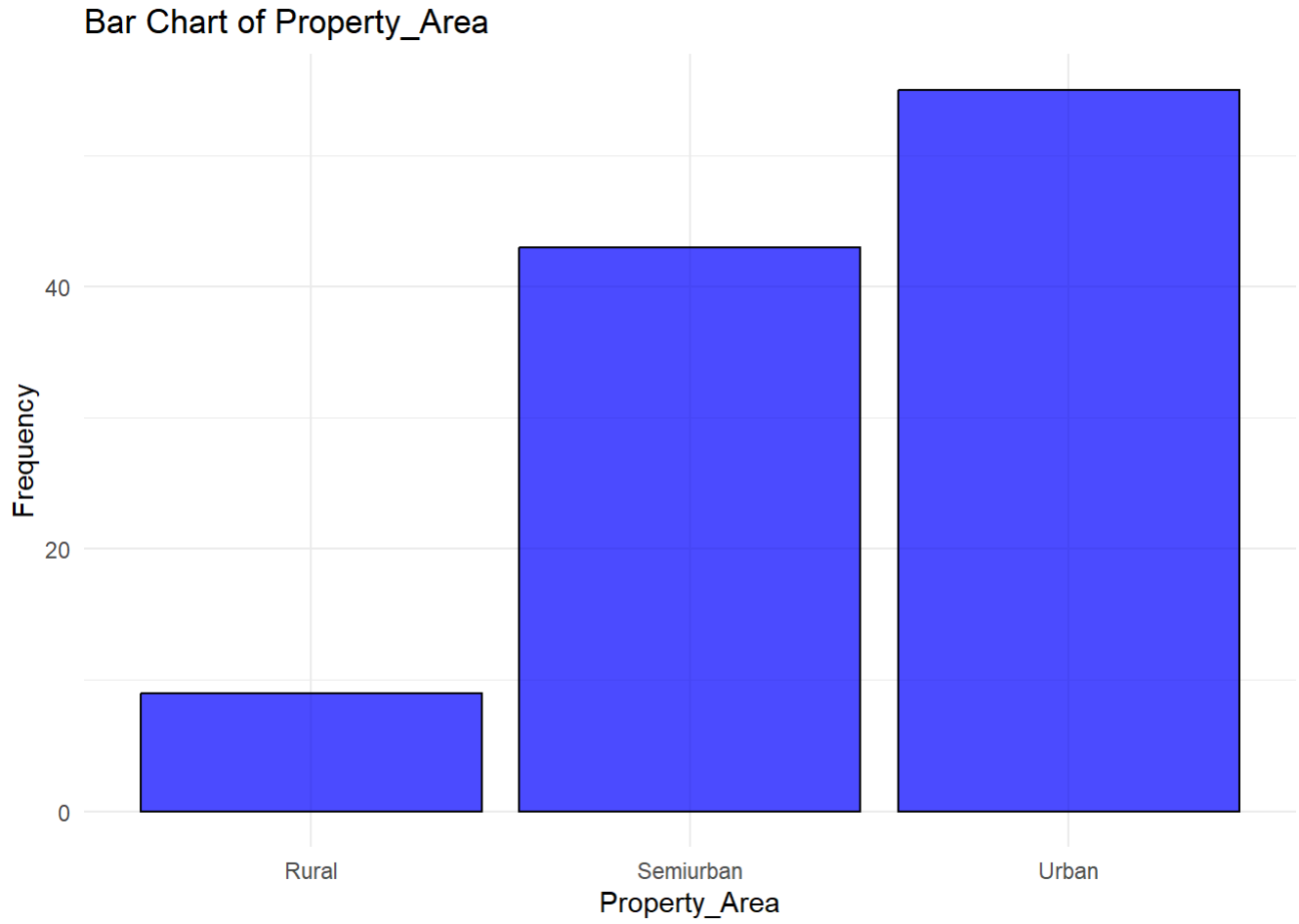

Bar Chart of Gender
75
50
25
0
Frequency
Female
Male
Gender

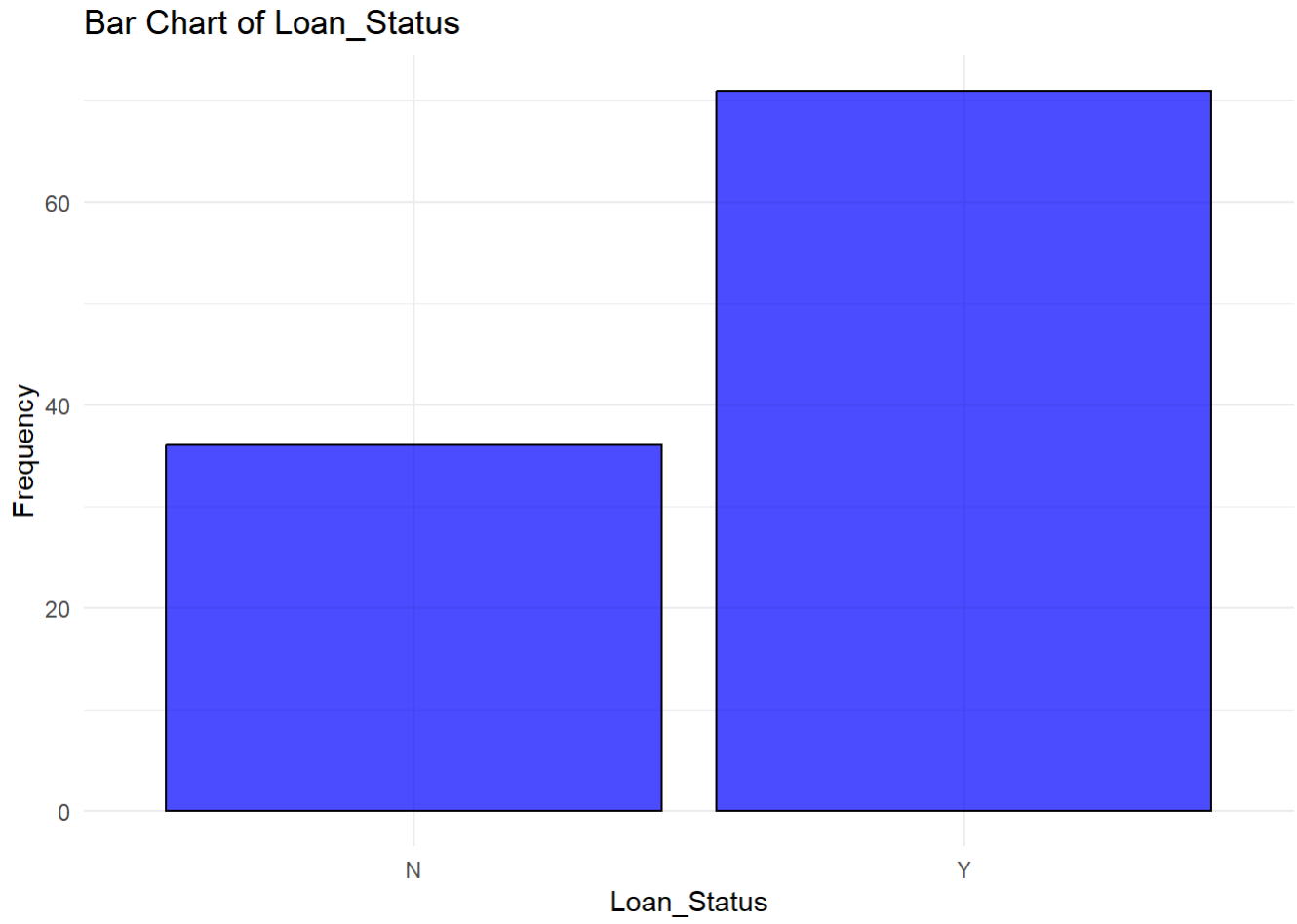

Bar Chart of Married
Frequency
0
20
40
60
No
Yes
Married


Bar Chart of Dependents
Frequency
0
20
40
60
0
1
2
3+
Dependents


Bar Chart of Education
Frequency
80
60
40
20
0
Graduate
Not Graduate
Education


Bar Chart of Self_Employed
Frequency
75
50
25
0
No
Yes
Self_Employed


Bar Chart of Property_Area
Frequency
0
20
40
Rural
Semiurban
Urban
Property_Area


Bar Chart of Loan_Status
Frequency
0
20
40
60
N
Y
Loan_Status

```
# this code shows us hows the first six rows of the dataset& str(my_data) → Displays column n
ames, data types, and first few values.Helps verify that the data looks as expected before fu
rther processing.

# ============================================
# 4. FINAL CLEANED DATA CHECK
# ============================================

head(my_data)
```

```
##    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      121.0              360              1         Urban
## 2              1508      128.0              360              1         Rural
## 3                 0       66.0              360              1         Urban
## 4              2358      120.0              360              1         Urban
## 5                 0      141.0              360              1         Urban
## 6              4196      120.5              360              1         Urban
##   Loan_Status
## 1           Y
## 2           N
## 3           Y
## 4           Y
## 5           Y
## 6           Y
```

```
str(my_data)
```

```
## 'data.frame':    107 obs. of  13 variables:
##  $ Loan_ID          : chr  "LP001002" "LP001003" "LP001005" "LP001006" ...
##  $ Gender           : chr  "Male" "Male" "Male" "Male" ...
##  $ Married          : chr  "No" "Yes" "Yes" "Yes" ...
##  $ Dependents       : chr  "0" "1" "0" "0" ...
##  $ Education        : chr  "Graduate" "Graduate" "Graduate" "Not Graduate" ...
##  $ Self_Employed    : chr  "No" "No" "Yes" "No" ...
##  $ ApplicantIncome  : num  5849 4583 3000 2583 6000 ...
##  $ CoapplicantIncome: num  0 1508 0 2358 0 ...
##  $ LoanAmount       : num  121 128 66 120 141 ...
##  $ Loan_Amount_Term : num  360 360 360 360 360 360 360 360 360 360 ...
##  $ Credit_History   : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ Property_Area    : chr  "Urban" "Rural" "Urban" "Urban" ...
##  $ Loan_Status      : chr  "Y" "N" "Y" "Y" ...
```

```
# Print success message
print("Data analysis and visualization complete!")
```

```
## [1] "Data analysis and visualization complete!"
```