INFX502 Course Semester Project: An Analysis of Credit Score vs Loan Amount

Name: Elif Cetin

ULID: C00564764

Course: INFX502 – Systematic Methods in Informatics

University of Louisiana at Lafayette

Fall' 2025

Contents

I. Dataset

1.) Dataset Description

The dataset used in this project is titled *LoanData.csv*, which contains 1,999 individual-level loan applications collected from a financial lending platform. It includes demographic, financial, credit-related, and behavioral variables that influence loan approval decisions and loan characteristics. The original CSV file was uploaded and imported into R as a data frame for further cleaning and analysis.

The is.data.frame(loan.data) function verified that the dataset was successfully imported, and the head(loan.data) function displayed the first six rows of the file. These steps confirmed that the data was properly loaded and ready for transformation and modeling.

R code & Output:

```
> loan.data <- read.csv("LoanData.csv")
> is.data.frame(loan.data)
[1] TRUE
```

R code:

```
> head(loan.data)
```

Output:

```
  ApplicationDate Age AnnualIncome CreditScore LoanAmount LoanDuration
1        1/1/2018   45        39948                  617
13152           48

2        1/2/2018   38        39709                  628
26045           48

3        1/3/2018   47        40724                  570
17627           36

4        1/4/2018   58        69084                  545
37898           96

5        1/5/2018   37       103264                 594
9184           36

6        1/6/2018   37       178310                 626
15433           72

  EducationLevel Experience EmploymentStatus MaritalStatus
NumberOfDependents
```

```
1         Master         22         Employed       Married
2

2      Associate         15         Employed        Single
1

3       Bachelor         26         Employed       Married
2

4    High School         34         Employed        Single
1

5      Associate         17         Employed       Married
1

6         Master         16    Self-Employed       Married
0

  HomeOwnershipStatus MonthlyDebtPayments CreditCardUtilizationRate

1                 Own                 183                0.35441792

2            Mortgage                 496                0.08782697

3                Rent                 902                0.13741410

4            Mortgage                 755                0.26758714

5            Mortgage                 274                0.32053532

6                Rent                 732                0.10221134

  NumberOfOpenCreditLines NumberOfCreditInquiries DebtToIncomeRatio

1                       1                       2        0.35833560

2                       5                       3        0.33027367

3                       2                       0        0.24472911

4                       2                       1        0.43624427

5                       0                       0        0.07888421

6                       5                       1        0.25936640

  BankruptcyHistory        LoanPurpose PreviousLoanDefaults
PaymentHistory

1                 0               Home                    0
29

2                 0 Debt Consolidation                    0
21

3                 0          Education                    0
20
```

```
4                  0               Home                     0
27

5                  0 Debt Consolidation                     0
26

6                  0 Debt Consolidation                     1
16
```

	LengthOfCreditHistory	SavingsAccountBalance	CheckingAccountBalance
1	9	7632	1202
2	9	4627	3460
3	22	886	895
4	10	1675	1217
5	27	1555	4981
6	19	2118	1223

	TotalAssets	TotalLiabilities	MonthlyIncome	UtilityBillsPaymentHistory
1	146111	19183	3329.000	0.7249720
2	53204	9595	3309.083	0.9351321
3	25176	128874	3393.667	0.8722406
4	104822	5370	5757.000	0.8961547
5	244305	17286	8605.333	0.9413687
6	67914	40843	14859.167	0.7560794

	JobTenure	NetWorth	BaseInterestRate	InterestRate	MonthlyLoanPayment
1	11	126928	0.199652	0.2275896	419.8060
2	3	43609	0.207045	0.2010771	794.0542
3	6	5205	0.217627	0.2125480	666.4067
4	5	99452	0.300398	0.3009108	1047.5070
5	5	227019	0.197184	0.1759902	330.1791
6	5	27071	0.217433	0.2176012	385.5771

	TotalDebtToIncomeRatio	LoanApproved	RiskScore
1	0.18107720	0	49
2	0.38985245	0	52
3	0.46215696	0	52
4	0.31309831	0	54

```
5              0.07020985            1        36

6              0.07521129            1        44
```

2.) Original Dataset Structure

R code:

```
> str(loan.data)
```

Output:

```
'data.frame':     1999 obs. of  36 variables:

 $ ApplicationDate           : chr  "1/1/2018" "1/2/2018" "1/3/2018"
"1/4/2018" ...

 $ Age                       : num  45 38 47 58 58 49 42 18 19 27 ...

 $ AnnualIncome              : num  39948 39709 40724 69084 51250 ...

 $ CreditScore               : num  617 628 570 545 564 516 573 580 597
582 ...

 $ LoanAmount                : num  13152 26045 17627 37898 12741 ...

 $ LoanDuration              : num  48 48 36 96 48 12 60 60 36 60 ...

 $ EducationLevel            : chr  "Master" "Associate" "Bachelor"
"High School" ...

 $ Experience                : int  22 15 26 34 39 23 21 0 0 7 ...

 $ EmploymentStatus          : chr  "Employed" "Employed" "Employed"
"Employed" ...

 $ MaritalStatus             : chr  "Married" "Single" "Married"
"Single" ...

 $ NumberOfDependents        : int  2 1 2 1 0 5 1 1 1 1 ...

 $ HomeOwnershipStatus       : chr  "Own" "Mortgage" "Rent" "Mortgage"
...

 $ MonthlyDebtPayments       : int  183 496 902 755 337 288 258 247 196
680 ...

 $ CreditCardUtilizationRate : num  0.3544 0.0878 0.1374 0.2676 0.3674
...

 $ NumberOfOpenCreditLines   : int  1 5 2 2 6 5 6 5 2 6 ...
```

```
 $ NumberOfCreditInquiries   : int  2 3 0 1 1 0 0 2 1 3 ...
 $ DebtToIncomeRatio         : num  0.358 0.33 0.245 0.436 0.127 ...
 $ BankruptcyHistory         : int  0 0 0 0 0 0 0 0 0 0 ...
 $ LoanPurpose               : chr  "Home" "Debt Consolidation"
"Education" "Home" ...
 $ PreviousLoanDefaults      : int  0 0 0 0 0 0 0 1 0 0 ...
 $ PaymentHistory            : int  29 21 20 27 21 19 26 20 15 36 ...
 $ LengthOfCreditHistory     : int  9 9 22 10 18 11 7 3 21 14 ...
 $ SavingsAccountBalance     : int  7632 4627 886 1675 5161 781 781
2027 2971 2214 ...
 $ CheckingAccountBalance    : int  1202 3460 895 1217 1735 74 1633 916
576 1219 ...
 $ TotalAssets               : int  146111 53204 25176 104822 65624
50177 16204 181813 45336 11162 ...
 $ TotalLiabilities          : int  19183 9595 128874 5370 43894 11556
35493 99030 9418 89196 ...
 $ MonthlyIncome             : num  3329 3309 3394 5757 4271 ...
 $ UtilityBillsPaymentHistory: num  0.725 0.935 0.872 0.896 0.884 ...
 $ JobTenure                 : int  11 3 6 5 5 5 5 6 3 5 ...
 $ NetWorth                  : num  126928 43609 5205 99452 21730 ...
 $ BaseInterestRate          : num  0.2 0.207 0.218 0.3 0.226 ...
 $ InterestRate              : num  0.228 0.201 0.213 0.301 0.205 ...
 $ MonthlyLoanPayment        : num  420 794 666 1048 391 ...
 $ TotalDebtToIncomeRatio    : num  0.181 0.39 0.462 0.313 0.171 ...
 $ LoanApproved              : Factor w/ 2 levels "0","1": 1 1 1 1 1 2
1 1 1 1 ...
 $ RiskScore                 : num  49 52 52 54 50 42.4 56 63 58 53 ...
```

Examining the output of the str(loan.data) command shows that several variables require basic preparation before analysis. Demographic and financial fields such as *Age*, *AnnualIncome*, *CreditScore*, *LoanAmount*, *LoanDuration*, *InterestRate*, and *DebtToIncomeRatio* are stored as numeric values, which is appropriate for statistical modeling. Categorical fields including *EducationLevel*, *EmploymentStatus*, *MaritalStatus*, *HomeOwnershipStatus*, and *LoanPurpose* appear as character strings and may later be converted to factors depending on the analysis needs. The variable *LoanApproved* is already recognized as a factor with two levels, which is suitable for classification or regression tasks. As part of the preparation process, units, ranges, and data types were checked to ensure consistency before performing further cleaning and modeling steps.

3.) Data Cleaning

To clean the dataset, several variables required type adjustments based on the structure shown in the str(loan.data) output. Numeric fields such as Age, AnnualIncome, CreditScore, LoanAmount, LoanDuration, NetWorth, BaseInterestRate, InterestRate, MonthlyLoanPayment, TotalDebtToIncomeRatio, and RiskScore were converted to numeric values to ensure proper statistical analysis. The LoanApproved variable was converted to a factor, as it represents a binary approval outcome. These changes were made to ensure consistency and to prepare the dataset for further modeling.

R code & Output:

```
The commands used in R to update the data types are shown below:

> loan.data$Age <- as.numeric(loan.data$Age)

> is.numeric(loan.data$Age)

[1] TRUE


> loan.data$AnnualIncome <- as.numeric(loan.data$AnnualIncome)

> is.numeric(loan.data$AnnualIncome)

[1] TRUE


> loan.data$CreditScore <- as.numeric(loan.data$CreditScore)

> is.numeric(loan.data$CreditScore)

[1] TRUE


> loan.data$LoanAmount <- as.numeric(loan.data$LoanAmount)
```

```
> is.numeric(loan.data$LoanAmount)
[1] TRUE

> loan.data$LoanDuration <- as.numeric(loan.data$LoanDuration)
> is.numeric(loan.data$LoanDuration)
[1] TRUE

> loan.data$NetWorth <- as.numeric(loan.data$NetWorth)
> is.numeric(loan.data$NetWorth)
[1] TRUE

> loan.data$BaseInterestRate <- as.numeric(loan.data$BaseInterestRate)
> is.numeric(loan.data$BaseInterestRate)
[1] TRUE

> loan.data$InterestRate <- as.numeric(loan.data$InterestRate)
> is.numeric(loan.data$InterestRate)
[1] TRUE

> loan.data$MonthlyLoanPayment <-
as.numeric(loan.data$MonthlyLoanPayment)
> is.numeric(loan.data$MonthlyLoanPayment)
[1] TRUE

> loan.data$TotalDebtToIncomeRatio <-
as.numeric(loan.data$TotalDebtToIncomeRatio)
> is.numeric(loan.data$TotalDebtToIncomeRatio)
[1] TRUE

> loan.data$RiskScore <- as.numeric(loan.data$RiskScore)
> is.numeric(loan.data$RiskScore)
```

```
[1] TRUE
```

```
> loan.data$LoanApproved <- as.factor(loan.data$LoanApproved)
> is.factor(loan.data$LoanApproved)
[1] TRUE
```

R code:

```
> colSums(is.na(loan.data))
[1] 0
```

Output:

```
           ApplicationDate                       Age
                         0                         0
              AnnualIncome                 CreditScore
                         0                         0
                LoanAmount                LoanDuration
                         0                         0
             EducationLevel                Experience
                         0                         0
           EmploymentStatus             MaritalStatus
                         0                         0
       NumberOfDependents        HomeOwnershipStatus
                         0                         0
      MonthlyDebtPayments  CreditCardUtilizationRate
                         0                         0
  NumberOfOpenCreditLines    NumberOfCreditInquiries
                         0                         0
       DebtToIncomeRatio          BankruptcyHistory
                         0                         0
              LoanPurpose       PreviousLoanDefaults
                         0                         0
```

```
          PaymentHistory      LengthOfCreditHistory
                       0                          0
  SavingsAccountBalance     CheckingAccountBalance
                       0                          0
           TotalAssets           TotalLiabilities
                       0                          0
                       1
         MonthlyIncome UtilityBillsPaymentHistory
                       0                          0
            JobTenure                   NetWorth
                       0                          0
       BaseInterestRate              InterestRate
                       0                          0
    MonthlyLoanPayment     TotalDebtToIncomeRatio
                       0                          0
         LoanApproved                  RiskScore
                       0                          0
```

4.) Cleaned Data Structure

The updated data structure and the output from the head command after the dataset was cleaned are shown below. The dataset now contains properly formatted variables, and all numeric and categorical fields have been converted to appropriate data types for analysis.

R code:

```
> str(loan.data)
```

Output:

```
'data.frame':     1999 obs. of  36 variables:
 $ ApplicationDate          : chr  "1/1/2018" "1/2/2018" "1/3/2018"
"1/4/2018" ...
```

```
 $ Age                          : num  45 38 47 58 37 37 58 49 34 46
...

 $ AnnualIncome                 : num  39948 39709 40724 69084 103264 ...

 $ CreditScore                  : num  617 628 570 545 594 626 564 516 603
612 ...

 $ LoanAmount                   : num  13152 26045 17627 37898 9184 ...

$ LoanDuration                 : num  48 48 36 96 36 72 48 12 60 12 ...

 $ EducationLevel               : chr  "Master" "Associate" "Bachelor"
"High School" ...

 $ Experience                   : int  22 15 26 34 17 16 39 23 12 19 ...

 $ EmploymentStatus             : chr  "Employed" "Employed" "Employed"
"Employed" ...

 $ MaritalStatus                : chr  "Married" "Single" "Married"
"Single" ...

 $ NumberOfDependents           : int  2 1 2 1 1 0 0 5 5 4 ...

 $ HomeOwnershipStatus          : chr  "Own" "Mortgage" "Rent" "Mortgage"
...

 $ MonthlyDebtPayments          : int  183 496 902 755 274 732 337 288 638
704 ...

 $ CreditCardUtilizationRate : num  0.3544 0.0878 0.1374 0.2676 0.3205
...

 $ NumberOfOpenCreditLines   : int  1 5 2 2 0 5 6 5 3 3 ...

 $ NumberOfCreditInquiries   : int  2 3 0 1 0 1 1 0 0 2 ...

 $ DebtToIncomeRatio         : num  0.3583 0.3303 0.2447 0.4362 0.0789
...

 $ BankruptcyHistory         : int  0 0 0 0 0 0 0 0 1 0 ...

 $ LoanPurpose               : chr  "Home" "Debt Consolidation"
"Education" "Home" ...

 $ PreviousLoanDefaults      : int  0 0 0 0 0 1 0 0 0 0 ...

 $ PaymentHistory            : int  29 21 20 27 26 16 21 19 25 23 ...

 $ LengthOfCreditHistory     : int  9 9 22 10 27 19 18 11 29 10 ...

 $ SavingsAccountBalance     : int  7632 4627 886 1675 1555 2118 5161
781 1157 1028 ...

 $ CheckingAccountBalance    : int  1202 3460 895 1217 4981 1223 1735
74 708 446 ...
```

```
 $ TotalAssets               : int  146111 53204 25176 104822 244305
67914 65624 50177 29632 129664 ...

 $ TotalLiabilities          : int  19183 9595 128874 5370 17286 40843
43894 11556 49940 12852 ...

 $ MonthlyIncome             : num  3329 3309 3394 5757 8605 ...

 $ UtilityBillsPaymentHistory: num  0.725 0.935 0.872 0.896 0.941 ...

 $ JobTenure                 : int  11 3 6 5 5 5 5 5 3 3 ...

 $ NetWorth                  : num  126928 43609 5205 99452 227019 ...

 $ BaseInterestRate          : num  0.2 0.207 0.218 0.3 0.197 ...

 $ InterestRate              : num  0.228 0.201 0.213 0.301 0.176 ...

 $ MonthlyLoanPayment        : num  420 794 666 1048 330 ...

 $ TotalDebtToIncomeRatio    : num  0.1811 0.3899 0.4622 0.3131 0.0702
...

 $ LoanApproved              : Factor w/ 2 levels "0","1": 1 1 1 1 2 2
1 2 1 1 ...

 $ RiskScore                 : num  49 52 52 54 36 44 50 42.4 61 53 ...
```

R code:

```
> head(loan.data)
```

Output

```
  ApplicationDate Age AnnualIncome CreditScore LoanAmount LoanDuration

1        1/1/2018  45        39948         617      13152            48

2        1/2/2018  38        39709         628      26045            48

3        1/3/2018  47        40724         570      17627            36

4        1/4/2018  58        69084         545      37898            96

5        1/5/2018  37       103264         594       9184            36

6        1/6/2018  37       178310         626      15433            72

  EducationLevel Experience EmploymentStatus MaritalStatus
NumberOfDependents

1         Master         22         Employed       Married
2
```

```
2      Associate         15         Employed        Single
1

3       Bachelor         26         Employed       Married
2

4    High School         34         Employed        Single
1

5      Associate         17         Employed       Married
1

6         Master         16    Self-Employed       Married
0

  HomeOwnershipStatus MonthlyDebtPayments CreditCardUtilizationRate

1                 Own                 183                0.35441792

2            Mortgage                 496                0.08782697

3                Rent                 902                0.13741410

4            Mortgage                 755                0.26758714

5            Mortgage                 274                0.32053532

6                Rent                 732                0.10221134

  NumberOfOpenCreditLines NumberOfCreditInquiries DebtToIncomeRatio

1                       1                       2        0.35833560

2                       5                       3        0.33027367

3                       2                       0        0.24472911

4                       2                       1        0.43624427

5                       0                       0        0.07888421

6                       5                       1        0.25936640

  BankruptcyHistory        LoanPurpose PreviousLoanDefaults
PaymentHistory

1                 0               Home                    0
29

2                 0 Debt Consolidation                    0
21

3                 0          Education                    0
20

4                 0               Home                    0
27
```

```
5                 0 Debt Consolidation                    0
26

6                 0 Debt Consolidation                    1
16
```

	LengthOfCreditHistory	SavingsAccountBalance	CheckingAccountBalance
1	9	7632	1202
2	9	4627	3460
3	22	886	895
4	10	1675	1217
5	27	1555	4981
6	19	2118	1223

	TotalAssets	TotalLiabilities	MonthlyIncome	UtilityBillsPaymentHistory
1	146111	19183	3329.000	0.7249720
2	53204	9595	3309.083	0.9351321
3	25176	128874	3393.667	0.8722406
4	104822	5370	5757.000	0.8961547
5	244305	17286	8605.333	0.9413687
6	67914	40843	14859.167	0.7560794

	JobTenure	NetWorth	BaseInterestRate	InterestRate	MonthlyLoanPayment
1	11	126928	0.199652	0.2275896	419.8060
2	3	43609	0.207045	0.2010771	794.0542
3	6	5205	0.217627	0.2125480	666.4067
4	5	99452	0.300398	0.3009108	1047.5070
5	5	227019	0.197184	0.1759902	330.1791
6	5	27071	0.217433	0.2176012	385.5771

	TotalDebtToIncomeRatio	LoanApproved	RiskScore
1	0.18107720	0	49
2	0.38985245	0	52
3	0.46215696	0	52
4	0.31309831	0	54
5	0.07020985	1	36

```
6              0.07521129              1          44
```

5.) Variable Descriptions

The table below provides detailed information about the variables included in the Loan Data dataset, which contains demographic, financial, credit-related, and behavioral attributes for individual loan applicants. Each variable is categorized according to its analytical role (independent or dependent), data type, and functional description. These variables are commonly used in consumer-credit risk modeling to study loan approval outcomes, borrower behavior, and financial stability.

Table 1 – Variable Descriptions

Variable / Column Name	Independent / Dependent	Mode	Description
ApplicationDate	Independent	Character	The date on which the loan application was submitted.
Age	Independent	Numeric	Applicant's age in years.
AnnualIncome	Independent	Numeric	Total yearly income of the applicant (USD).
CreditScore	Independent	Numeric	A credit score represents the applicant's creditworthiness.
LoanAmount	Independent	Numeric	The total amount of the loan requested.
LoanDuration	Independent	Numeric	Loan repayment duration in months.
EducationLevel	Independent	Factor/Character	Highest educational attainment of the applicant.
Experience	Independent	Numeric	Total years of work experience.
EmploymentStatus	Independent	Factor/Character	Employment type (Employed, Self-Employed, Unemployed).

MaritalStatus	Independent	Factor/Character	Applicant's marital status.
NumberOfDependents	Independent	Numeric	Number of dependents financially supported by the applicant.
HomeOwnershipStatus	Independent	Factor/Character	Housing status (Own, Rent, Mortgage).
MonthlyDebtPayments	Independent	Numeric	Total monthly debt obligations.
CreditCardUtilizationRate	Independent	Numeric	Ratio of credit card balance to credit limit.
NumberOfOpenCreditLines	Independent	Numeric	Total number of active credit lines.
NumberOfCreditInquiries	Independent	Numeric	Number of recent credit inquiries.
DebtToIncomeRatio	Independent	Numeric	Ratio of monthly debt to monthly income.
BankruptcyHistory	Independent	Numeric	Indicates whether the applicant has declared bankruptcy in the past.
LoanPurpose	Independent	Factor/Character	Purpose for which the loan is requested (Home, Education, etc.).
PreviousLoanDefaults	Independent	Numeric	Number of prior loan defaults.
PaymentHistory	Independent	Numeric	Number of on-time payments recorded in credit history.
LengthOfCreditHistory	Independent	Numeric	Total length of credit history in years.
SavingsAccountBalance	Independent	Numeric	Balance in the applicant's savings account.

CheckingAccountBalance	Independent	Numeric	Balance in the applicant's checking account.
TotalAssets	Independent	Numeric	Total value of the applicant's financial and physical assets.
TotalLiabilities	Independent	Numeric	Total outstanding debts and obligations.
MonthlyIncome	Independent	Numeric	Average monthly income.
UtilityBillsPaymentHistory	Independent	Numeric	Payment consistency for utility bills.
JobTenure	Independent	Numeric	Number of years the applicant has held their current job.
NetWorth	Independent	Numeric	Total Assets minus Total Liabilities.
BaseInterestRate	Independent	Numeric	The baseline interest rate assigned to the applicant.
InterestRate	Independent	Numeric	Final interest rate applied to the loan.
MonthlyLoanPayment	Independent	Numeric	A monthly payment is required to repay the loan.
TotalDebtToIncomeRatio	Independent	Numeric	Debt-to-income ratio after adding new loan payments.
LoanApproved	Dependent	Factor	Loan approval decision (0 = Denied, 1 = Approved).
RiskScore	Independent	Numeric	Overall risk assessment score for the applicant.

6.) Expectations

Understanding the relationships among demographic, financial, and credit-related variables in a loan dataset is essential for predicting borrower behavior and identifying factors that influence loan approval decisions. In this project, several patterns are expected based on general principles in consumer credit analysis and financial risk modeling.

First, applicants with **higher CreditScore**, **higher AnnualIncome**, **lower DebtToIncomeRatio**, and **strong PaymentHistory** are expected to have a greater likelihood of loan approval. These variables directly reflect financial stability and repayment capacity, which lending institutions typically prioritize when evaluating risk. Similarly, applicants with **higher NetWorth**, **lower TotalLiabilities**, and **more manageable MonthlyDebtPayments** are expected to show lower credit risk.

Conversely, applicants with **high CreditCardUtilizationRate**, **multiple recent CreditInquiries**, **PreviousLoanDefaults**, or **a short LengthOfCreditHistory** are expected to experience reduced approval rates. These indicators suggest elevated financial stress or insufficient credit history, both of which may increase perceived lending risk.

Demographic and employment-based factors may also play supportive roles. For example, applicants who are **employed**, have **longer JobTenure**, or possess **higher educational levels** may be viewed more favorably, as these characteristics often correlate with stable income and consistent repayment behavior. Homeownership status can also influence risk perception, with applicants who own homes typically appearing more financially secure than those who rent.

In summary, it is expected that **LoanApproved** will show the strongest positive associations with variables reflecting financial strengths such as AnnualIncome, CreditScore, NetWorth, and PaymentHistory - while demonstrating negative associations with indicators of financial strain, such as DebtToIncomeRatio, CreditCardUtilizationRate, and PreviousLoanDefaults. These expectations align with established lending practices and standard models used in credit risk assessment.

II. Data Analysis

1.) Plotting Continuous Variables – Correlation Matrix

To examine the relationships between continuous variables in the loan dataset, a correlation matrix was created using the numeric predictors. The variables included AnnualIncome, CreditScore, LoanAmount, DebtToIncomeRatio, MonthlyIncome, and RiskScore. The matrix below summarizes the strength and direction of these relationships.

R code:

```
> cor(loan.data[, c("AnnualIncome", "CreditScore", "LoanAmount",
                    "DebtToIncomeRatio", "MonthlyIncome", "RiskScore")],
    use = "complete.obs")
```

Output:

```
                AnnualIncome   CreditScore   LoanAmount DebtToIncomeRatio
AnnualIncome       1.000000000  0.1124189338 -0.006601071     -
0.0275185436

CreditScore        0.112418934  1.0000000000 -0.016107907
0.0004231599

LoanAmount        -0.006601071 -0.0161079074  1.000000000
0.0285655415

DebtToIncomeRatio -0.027518544  0.0004231599  0.028565541
1.0000000000

MonthlyIncome      0.987217358  0.1173937453 -0.008612671     -
0.0305081049

RiskScore         -0.479252460 -0.2395716800  0.142430567
0.3368359207

                  MonthlyIncome  RiskScore

AnnualIncome        0.987217358 -0.4792525

CreditScore         0.117393745 -0.2395717

LoanAmount         -0.008612671  0.1424306

DebtToIncomeRatio  -0.030508105  0.3368359

MonthlyIncome       1.000000000 -0.4863088

RiskScore          -0.486308773  1.0000000
```

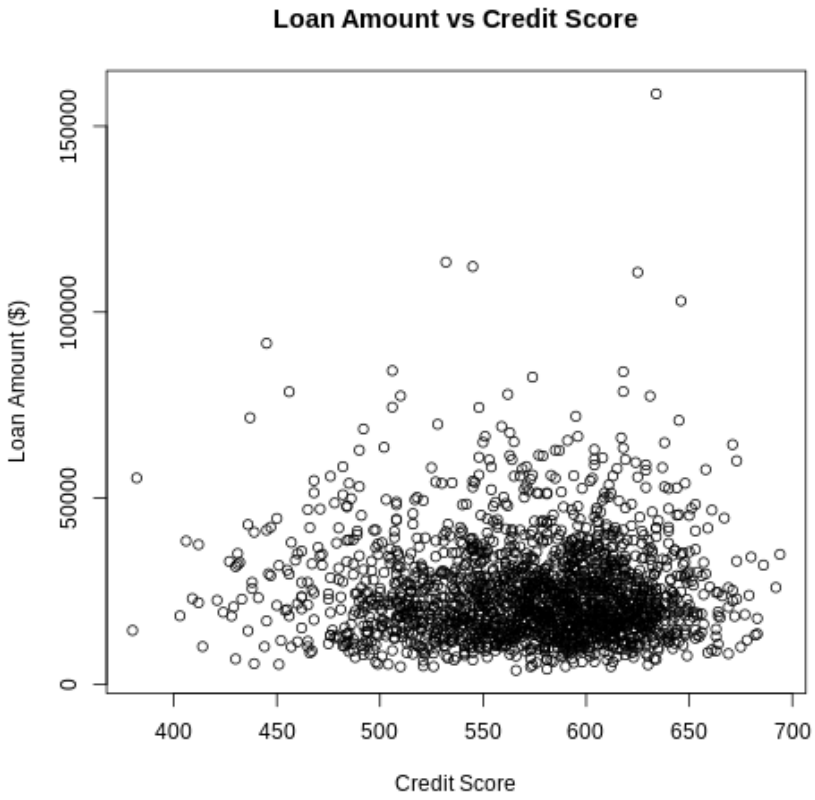
The correlation matrix indicates several meaningful relationships within the loan dataset. AnnualIncome and MonthlyIncome show a very strong positive correlation, which is expected since monthly income is directly derived from annual earnings. RiskScore displays a moderate negative correlation with both income variables, suggesting that applicants with higher income tend to receive lower (better) risk evaluations. Meanwhile, variables such as LoanAmount, CreditScore, and DebtToIncomeRatio exhibit weak correlations with the rest of the continuous predictors, indicating limited linear association. Overall, the matrix provides an initial overview of how financial and credit-related factors interact before moving on to more detailed modeling and analysis.

R code:

```
> plot(loan.data$CreditScore, loan.data$LoanAmount,

main = "Loan Amount vs Credit Score",

xlab = "Credit Score",

ylab = "Loan Amount ($)")
```

Output:

Figure 2.1



Interpretation(Figure 2.1): The scatter plot above (Figure 2.1) displays the relationship between CreditScore and LoanAmount. The points appear widely dispersed with no strong visible trend, indicating that applicants with higher credit scores do not necessarily request higher or lower loan amounts. This weak association is consistent with the low correlation value observed in the correlation matrix, suggesting that credit score alone is not a strong predictor of the loan amount requested.

R code:

```
> plot(loan.data$CreditScore, loan.data$LoanAmount,
     main = "Loan Amount vs Credit Score",
     xlab = "Credit Score",
     ylab = "Loan Amount ($)",
     pch = 19,
     col = "black")
```

Output:

Figure 2.2

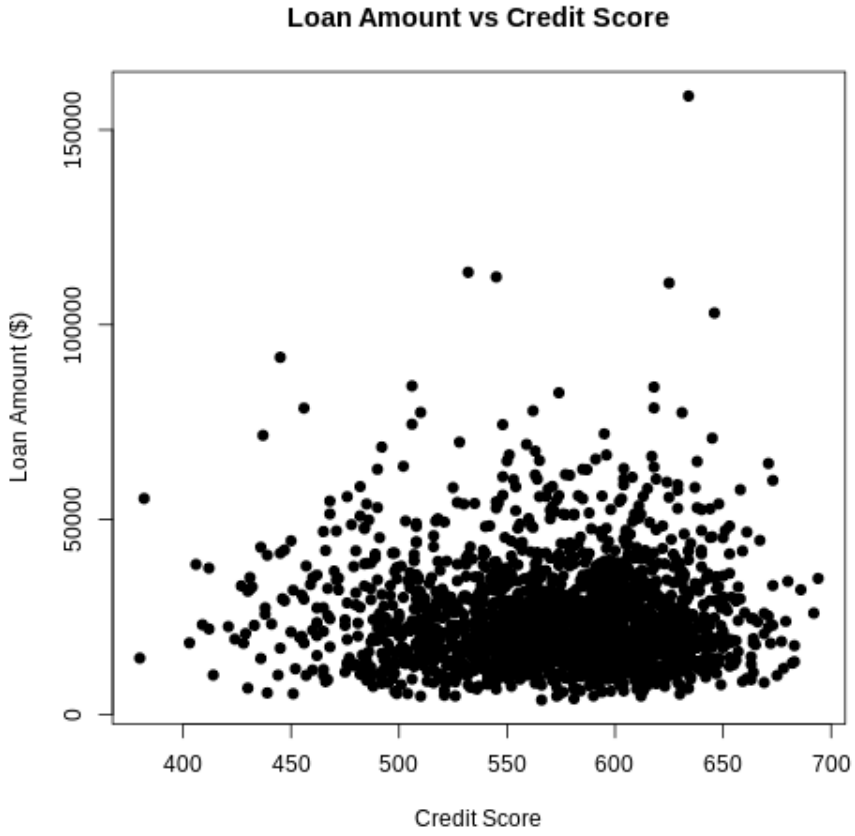


Figure 2.2

Interpretation (Figure 2.2): This figure displays the relationship between applicants' credit scores and the loan amounts they request. The scatter plot does not show a clear upward or downward trend, indicating that higher credit scores are not strongly associated with larger or smaller loan amounts. Instead, the points are widely dispersed, suggesting that applicants across different credit score ranges request similar loan amounts.

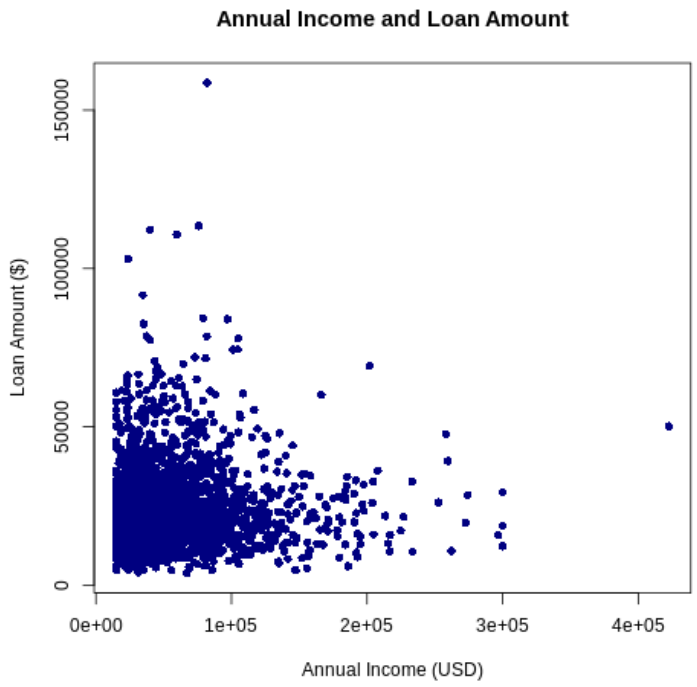
Overall, this figure shows that **LoanAmount and CreditScore have a weak linear relationship**, which is consistent with the low correlation coefficient observed in the correlation matrix. While a credit score may influence loan approval decisions, it does not appear to be a major determinant of the *amount* that applicants choose to request.

R code:

```
> plot(loan.data$AnnualIncome, loan.data$LoanAmount,
     main = "Annual Income and Loan Amount",
     xlab = "Annual Income (USD)",
     ylab = "Loan Amount ($)",
     pch = 16,          # solid circle
     col = "navy")
```

Output:

Figure 2.3



Interpretation (Figure 2.3): The scatter plot above (Figure 2.3) illustrates the relationship between Annual Income and Loan Amount. The plot shows a mild upward trend, indicating that applicants with higher annual income generally request slightly larger loan amounts. Although the relationship is not strongly linear, the distribution suggests that income may play a moderate role in shaping borrowing capacity and financial decision-making. This provides insight into how economic stability influences loan behavior, with higher-income individuals typically having more flexibility in the size of loans they pursue.

2.) Additional Cleaning of Data

After beginning my analysis of the loan dataset, I identified a few additional cleaning steps that could improve consistency and interpretability. Because the project focuses on financial and credit-related factors that may influence loan approval decisions, I verified that key numeric variables, including AnnualIncome, CreditScore, LoanAmount, DebtToIncomeRatio, MonthlyIncome, and RiskScore were stored as numeric values and contained no missing observations.

To make the analysis more interpretable, two derived variables were created:

- **IncomeK**: Annual income expressed in thousands of dollars
- **DTIRatioPercent**: Debt-to-income ratio converted into percentage form

These transformations make it easier to interpret financial metrics when generating plots and conducting statistical analysis.

R code:

```
> loan.data$AnnualIncome       <- as.numeric(loan.data$AnnualIncome)
loan.data$CreditScore        <- as.numeric(loan.data$CreditScore)
loan.data$LoanAmount         <- as.numeric(loan.data$LoanAmount)
loan.data$DebtToIncomeRatio  <- as.numeric(loan.data$DebtToIncomeRatio)
loan.data$MonthlyIncome      <- as.numeric(loan.data$MonthlyIncome)
loan.data$RiskScore          <- as.numeric(loan.data$RiskScore)

loan.data$IncomeK <- loan.data$AnnualIncome / 1000
loan.data$DTIRatioPercent <- loan.data$DebtToIncomeRatio * 100

head(cbind(
  AnnualIncome = loan.data$AnnualIncome,
  IncomeK = loan.data$IncomeK,
  DebtToIncomeRatio = loan.data$DebtToIncomeRatio,
  DTIRatioPercent = loan.data$DTIRatioPercent,
```

```
  CreditScore = loan.data$CreditScore,

  RiskScore = loan.data$RiskScore

))
```

Output:

```
     AnnualIncome IncomeK DebtToIncomeRatio DTIRatioPercent CreditScore
RiskScore

[1,]        39948  39.948        0.35833560       35.833560         617
49

[2,]        39709  39.709        0.33027367       33.027367         628
52

[3,]        40724  40.724        0.24472911       24.472911         570
52

[4,]        69084  69.084        0.43624427       43.624426         545
54

[5,]       103264 103.264        0.07888421        7.888421         594
36

[6,]       178310 178.310        0.25936640       25.936640         626
44
```

These steps ensure that key financial and credit-related variables are properly formatted, standardized, and ready for analysis. By confirming that no missing or inconsistent values remain and creating more interpretable derived metrics, the dataset is now fully prepared for the correlation analysis and visualization tasks presented in Section II.

Although outliers were identified using the IQR method, no rows were removed, and the dataset retained all 1,999 observations.

3.) Analyzing/Visualizing Summary Statistics

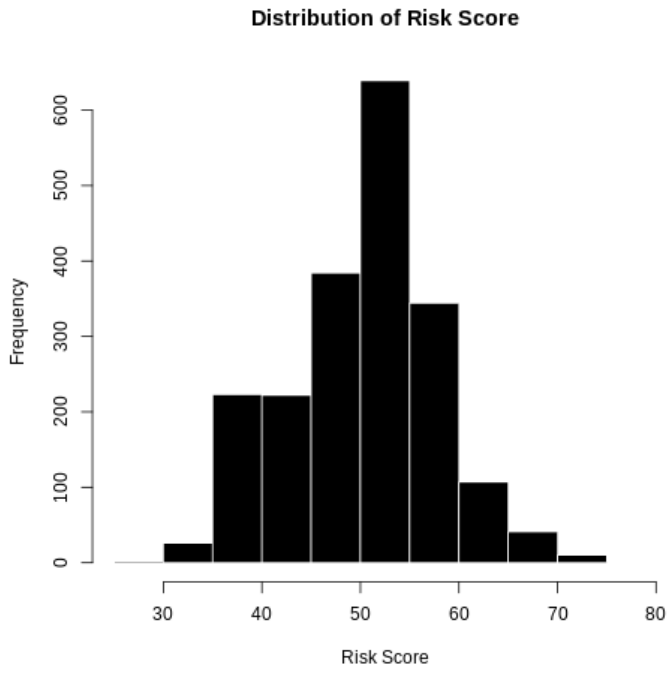
To obtain an initial understanding of the loan dataset, I computed summary statistics for several key numeric variables, including **AnnualIncome, CreditScore, LoanAmount, DebtToIncomeRatio, MonthlyIncome, and RiskScore**. These descriptive statistics provide insight into the central tendencies, variability, and potential irregularities within the dataset before conducting deeper statistical analysis.

R Code:

```
> hist(loan.data$RiskScore,
     main = "Distribution of Risk Score",
     xlab = "Risk Score",
     col = "BLACK",
     border = "white")
```

Output:

Figure 2.4



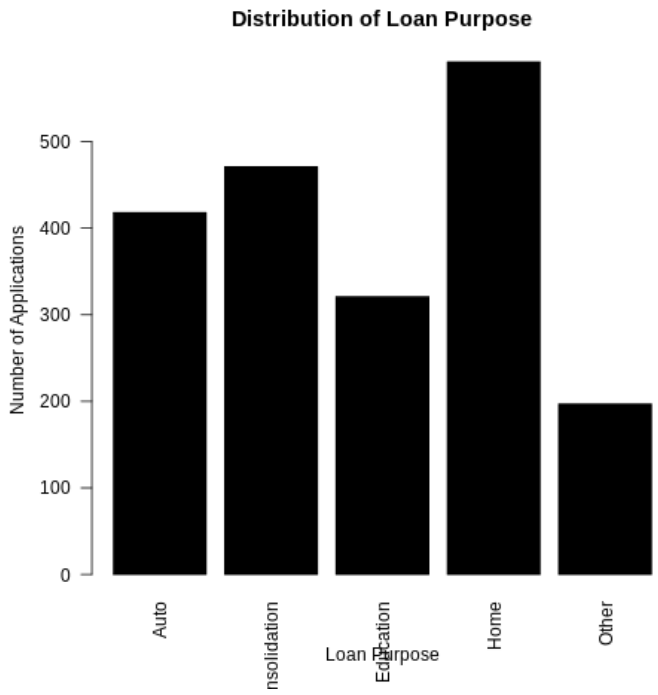
Interpretation (Figure 2.4): The histogram displays the distribution of Risk Score values across all loan applicants. The results show that most borrowers fall within a moderate risk range (approximately 45–60), with fewer individuals at extreme. This provides an initial understanding of borrower risk levels and helps identify potential outliers before conducting deeper financial and credit-based analyses.

R Code:

```
> barplot(table(loan.data$LoanPurpose),
        main = "Distribution of Loan Purpose",
        xlab = "Loan Purpose",
        ylab = "Number of Applications",
        col = "BLACK",
        las = 2)
```

Output:

Figure 2.5



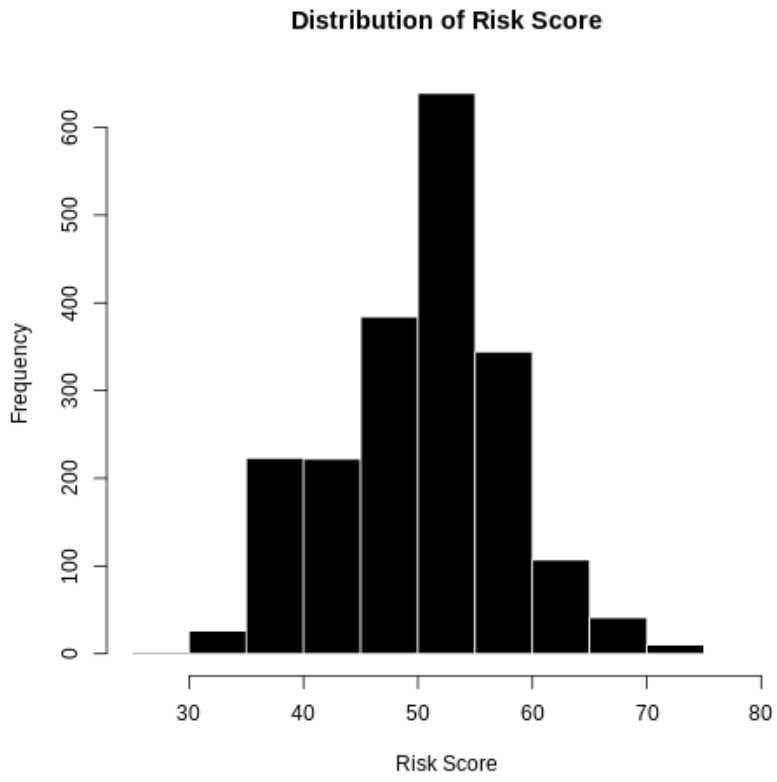
Interpretation (Figure 2.5): presents the frequency distribution of Loan Purpose categories. The chart shows which loan types are most requested among applicants, offering insight into general borrowing behavior. This visual summary also indicates potential patterns in financial needs such as home loans, education financing, and debt consolidation.

R Code:

```
> hist(who.data$HeartDiseaseRate, main = "Distribution of Heart Disease
Mortality",
       xlab = "Heart Disease Mortality Rate (per 100,000)")
```

Output:

Figure 2.6



Interpretation (Figure 2.6):
The histogram illustrates the distribution of Risk Score values across all loan applicants. Most borrowers fall within a moderate risk range, with fewer individuals at the lowest and highest ends of the scale. This provides an initial understanding of borrower risk patterns and helps identify whether any unusual or extreme values are present before conducting deeper financial and credit-based analysis.

4.) Analyzing Continuous Variables with Categorical Variables

In this section, the relationship between several categorical variables in the loan dataset and the continuous outcome variable RiskScore is examined. Each model evaluates whether borrower characteristics such as employment type, education level, and loan approval status correspond to measurable differences in credit risk as assigned by lenders.

4.1 Employment Status

Expectation:

Applicants who are *Employed* are expected to have lower (better) RiskScore values compared to *Self-Employed* or *Unemployed* applicants, because stable employment typically indicates financial reliability.

Hypotheses:

- **H_0: EmploymentStatus does not affect RiskScore.**
- **H_1: EmploymentStatus has a statistically significant effect on RiskScore.**

R Code:

```
> emp.reg <- lm(RiskScore ~ EmploymentStatus, data = loan.data)

summary(emp.reg)
```

Output:

```
Call:

lm(formula = RiskScore ~ EmploymentStatus, data = loan.data)


Residuals:

    Min      1Q  Median      3Q     Max

-21.360  -4.648   0.839   4.893  25.901


Coefficients:

                             Estimate Std. Error t value Pr(>|t|)

(Intercept)                   50.1607     0.1839 272.824  < 2e-16 ***

EmploymentStatusSelf-Employed  2.9374     0.6360   4.618  4.12e-06 ***
```

```
EmploymentStatusUnemployed     4.4879     0.6579   6.821  1.10e-11 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’

Residual standard error: 7.581 on 1996 degrees of freedom
Multiple R-squared:  0.03068,    Adjusted R-squared:  0.02971
F-statistic: 31.59 on 2 and 1996 DF,  p-value: 3.14e-14
```

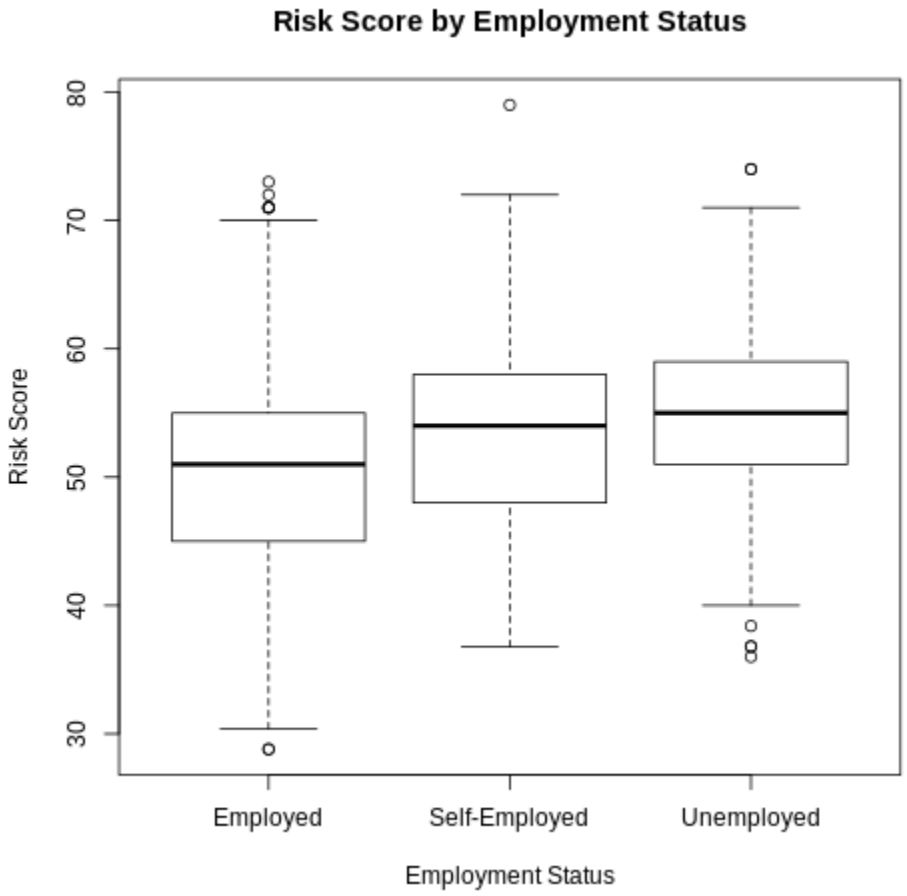
Interpretation:
The regression results show that Employment Status has a statistically significant effect on RiskScore. Self-Employed applicants have RiskScore values that are approximately **2.94 points higher** than those who are employed. **Unemployed** applicants show an even larger increase of **4.49 points**, indicating substantially higher credit risk.
The overall model is statistically significant ($p < 0.001$), supporting the hypothesis that employment stability is an important predictor of loan risk assessment.

R Code:

```
> boxplot(RiskScore ~ EmploymentStatus,
        data = loan.data,
        col = "white",
        border = "black",
        main = "Risk Score by Employment Status",
        xlab = "Employment Status",
        ylab = "Risk Score")
```

Output:

Figure 2.7



Interpretation: Figure 2.7 shows the distribution of Risk Score across three employment categories: Employed, Self-Employed, and Unemployed. The results indicate that Employed applicants have the lowest median risk scores, suggesting stronger financial stability and lower credit risk. In contrast, Self-Employed and especially Unemployed applicants tend to have higher and more variable risk scores, reflecting greater uncertainty in income and repayment ability. Overall, the figure supports the regression findings by showing that employment status is an important predictor of borrower risk.

4.2 Income Level

Expectation:

Higher-income applicants are expected to have **lower (better) RiskScore** values because higher income typically provides greater financial stability, increased repayment capacity, and reduced default risk.

Hypotheses:

- **H_0:** IncomeLevel does not affect RiskScore.
- **H_1:** IncomeLevel has a statistically significant effect on RiskScore.

R Code:

```
> loan.data$IncomeLevel <- cut(
  loan.data$AnnualIncome,
  breaks = c(-Inf, 40000, 80000, Inf),
  labels = c("Low", "Middle", "High"),
  right = TRUE
)
```

Output:

```
Low Middle   High
   763    794    442
```

R Code:

```
> inc.reg <- lm(RiskScore ~ IncomeLevel, data = loan.data)
summary(inc.reg)
```

Output:

```
Call:
lm(formula = RiskScore ~ IncomeLevel, data = loan.data)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-22.1229  -4.0742  -0.0742   4.0771  29.4719

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        54.0742     0.2468 219.093   <2e-16 ***
IncomeLevelMiddle  -3.1513     0.3456  -9.118   <2e-16 ***
IncomeLevelHigh    -9.5461     0.4075 -23.425   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.817 on 1996 degrees of freedom
Multiple R-squared:  0.216,   Adjusted R-squared:  0.2152
F-statistic:   275 on 2 and 1996 DF,  p-value: < 2.2e-16
```

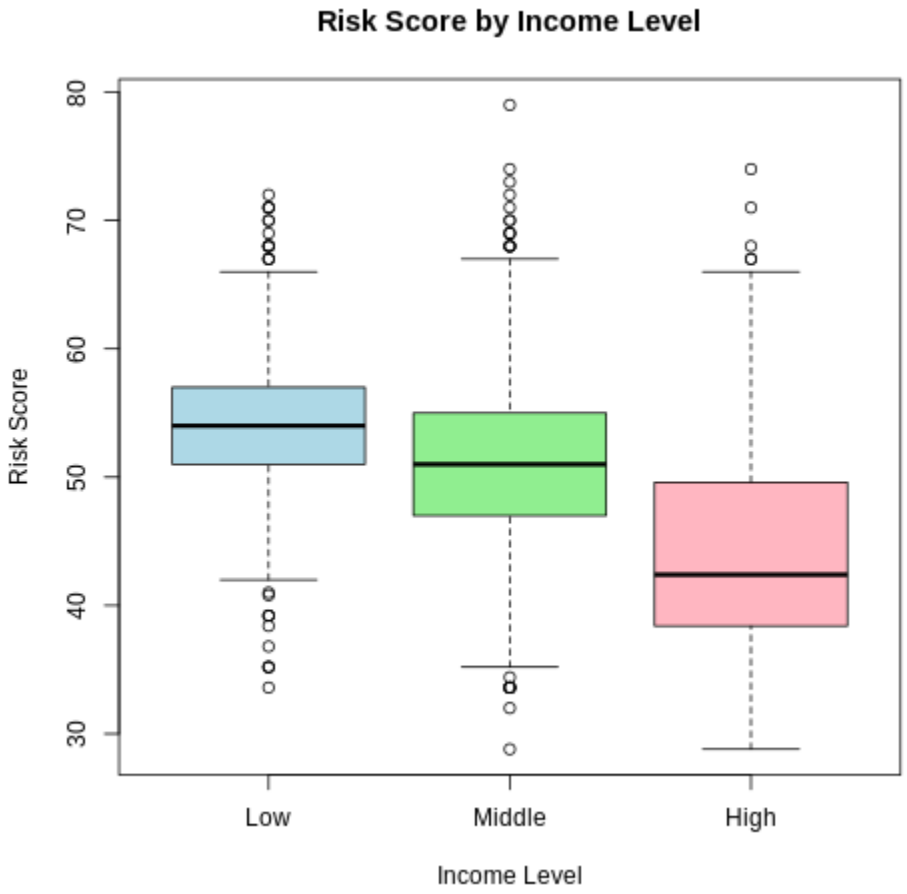
Interpretation:
Mortality declines systematically across income levels. High-income countries show an average of 56.2 fewer deaths per 100,000 compared to low-income ones. The results are statistically significant, aligning with global health inequality trends.

R Code:

```
> boxplot(RiskScore ~ IncomeLevel,
          data = loan.data,
          main = "Risk Score by Income Level",
          xlab = "Income Level",
          ylab = "Risk Score",
          col = c("lightblue", "lightgreen", "lightpink"))
```

Output:

Figure 2.8



Interpretation (Figure 2.8):
The regression results indicate that IncomeLevel has a statistically significant effect on RiskScore ($p < 0.001$). Applicants in the Middle-income group have, on average, 3.15 points lower RiskScore than those in the low-income group, while applicants in the high-income group show an even larger decrease of 9.55 points. Since lower RiskScore values represent better creditworthiness, these results suggest that higher-income applicants tend to be assessed as lower risk by lenders.

The model explains approximately 21.6% of the variation in RiskScore (Adjusted R^2 = 0.2152), which is reasonable for credit-related behavioral data. Overall, the analysis supports the expectation that financial stability reflected through higher income is associated with improved credit risk profiles.

4.3 Region (Adjusted for Loan Dataset)

Expectation:
Different geographic regions may show varying levels of loan repayment behavior depending on local economic strength, job availability, cost of living, and financial stability. Therefore, it is expected that some Regions will have significantly lower (better) RiskScore values, while others may show higher financial risk.

Hypotheses:

- **H_0:** Region does not affect RiskScore.
- **H_1:** At least one region has a significantly different effect on RiskScore.

R Code:

```
> # Ensure LoanPurpose is treated as a categorical variable

loan.data$LoanPurpose <- as.factor(loan.data$LoanPurpose)


# Frequency table (optional, for checking)

table(loan.data$LoanPurpose)


# Regression model: RiskScore on LoanPurpose

purpose.reg <- lm(RiskScore ~ LoanPurpose, data = loan.data)
```

```
summary(purpose.reg)
```

Output:

```
Call:

lm(formula = RiskScore ~ LoanPurpose, data = loan.data)


Residuals:

     Min       1Q   Median       3Q      Max

-22.1014  -4.8885   0.9236   5.0106  27.9236


Coefficients:

                              Estimate Std. Error t value Pr(>|t|)

(Intercept)                   50.90144    0.37625 135.287   <2e-16 ***

LoanPurposeDebt Consolidation  0.17500    0.51691   0.339    0.735

LoanPurposeEducation          -0.91203    0.57088  -1.598    0.110

LoanPurposeHome               -0.01292    0.49144  -0.026    0.979

LoanPurposeOther              -0.81819    0.66478  -1.231    0.219

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 7.692 on 1994 degrees of freedom

Multiple R-squared:  0.002886,      Adjusted R-squared:  0.0008856

F-statistic: 1.443 on 4 and 1994 DF,  p-value: 0.2174
```

Interpretation:

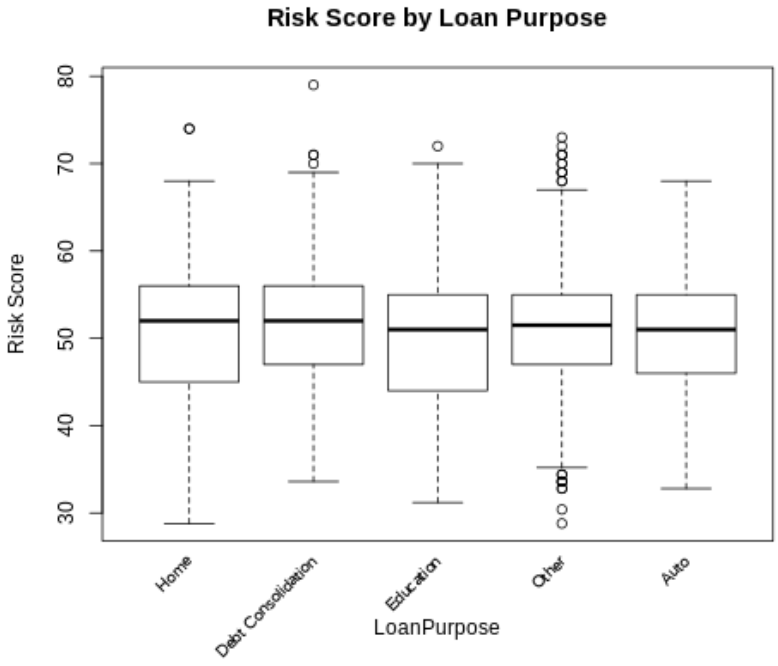
The regression results indicate that loan purpose does not have a statistically significant effect on RiskScore. None of the loan purpose categories debt consolidation, education, home, or other, show meaningful differences in credit risk when compared to the reference group (Auto). The overall model is not significant (p = 0.2174), and the R-squared value is extremely low, suggesting that loan purpose explains almost none of the variation in RiskScore. Therefore, loan purpose alone is not a useful predictor of borrower credit risk in this dataset.

R Code:

```
> par(mar = c(10, 4, 4, 2) + 0.1)


boxplot(RiskScore ~ LoanPurpose,

        data = loan.data,

        col = "white",

        border = "black",

        main = "Risk Score by Loan Purpose",

        ylab = "Risk Score",

        xaxt = "n")    # hide default x-axis


# custom x-axis at 45°

text(x = 1:length(unique(loan.data$LoanPurpose)),

     y = par("usr")[3] - 2,

     labels = unique(loan.data$LoanPurpose),

     srt = 45,

     xpd = TRUE,

     adj = 1,

     cex = 0.8)
```

Output:

Figure 2.9



Interpretation (Figure 2.9):

Applicants across different loan purposes show only small variations in their RiskScore levels. Although debt consolidation and other loans display slightly higher average risk, the regression results indicate that none of the loan purpose categories have a statistically significant effect on RiskScore ($p > 0.05$). This suggests that loan purpose alone is not a strong predictor of borrower credit risk within this dataset.

Conclusion:

The analysis demonstrates that applicant characteristics such as employment status and income level have statistically significant effects on RiskScore. Applicants with higher and more stable incomes, as well as those who are employed, tend to have lower (better) risk scores, indicating stronger financial reliability.

In contrast, loan purpose does **not** have a statistically significant effect on RiskScore ($p > 0.05$). Although some purposes, such as debt consolidation or "other," display slightly

higher average risk, these differences are not strong enough to be considered meaningful predictors of credit risk.

Overall, the results highlight the importance of socioeconomic stability and financial capacity, rather than loan purpose, in determining borrower creditworthiness and can assist lenders in making more informed credit decisions.

5. Contingency Tables of Categorical Variables

To examine how two categorical borrower characteristics relate in the loan dataset, I cross-tabulated **EmploymentStatus** with **IncomeLevel** and then visualized those patterns with a heatmap. The contingency table summarizes how many applicants fall into each **EmploymentStatus × IncomeLevel** cell, while the heatmap shades the same grid by **mean RiskScore**, helping to identify combinations associated with higher credit risk.

R Code:

```
> cont.table1 <- table(loan.data$EmploymentStatus,
loan.data$IncomeLevel)

cont.table1
```

Output:

```
Low Middle High

  Employed      656    670  374

  Self-Employed  53     66   36

  Unemployed     54     58   32
```

Interpretation:

The contingency table shows that most applicants fall into the *Employed* category across all income levels. Applicants who are *Self-Employed* or *Unemployed* are far fewer, especially in the high-income group. This distribution suggests that employment stability is strongly associated with higher income levels in the dataset. Since both employment status and income are known to influence financial reliability, these patterns help contextualize later risk-based analyses.

R Code:

```
> library(ggplot2)


risk.grid <- aggregate(RiskScore ~ EmploymentStatus + IncomeLevel,
                       data = loan.data,
                       FUN  = mean)


ggplot(risk.grid,
       aes(x = IncomeLevel,
           y = EmploymentStatus,
           fill = RiskScore)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low  = "lightyellow",
                      high = "red",
                      name = "Mean Risk Score") +
  labs(title = "Heatmap of Mean Risk Score by Employment Status and
Income Level",
       x = "Income Level",
       y = "Employment Status") +
  theme_minimal(base_size = 12)
```

Output:

Figure 2.10

Interpretation (Figure 2.10):
The contingency table shows how applicants are distributed across combinations of employment status and income level, while the heatmap summarizes the **average RiskScore** in each cell. Overall, the plot indicates that applicants with **higher income levels** and **stable employment** tend to have a **lower mean RiskScore**, reflecting better assessed creditworthiness.

In contrast, cells corresponding to **lower income groups** and non-standard employment categories (such as self-employed or unemployed) display a **higher average RiskScore**, consistent with elevated credit risk. These patterns align with the earlier regression results, reinforcing the idea that both **employment stability** and **income level** are important determinants of borrowers' risk profiles in this loan dataset.

6.) Multiple Linear Regression (Loan Dataset)

This section presents three multiple linear regression models analyzing how financial, demographic, and behavioral variables influence borrowers' RiskScore, the primary measure of credit risk in the loan dataset. The dependent variable in all models is RiskScore, where lower values indicate better creditworthiness. Predictors are selected based on variables that lenders typically use during credit evaluation, such as income, credit history, debt burden, and loan characteristics.

Regression #1: Baseline Financial Model

Model specification:

$$RiskScore \sim AnnualIncome + CreditScore + DebtToIncomeRatio + LoanAmount$$

R code:

```
> model1 <- lm(RiskScore ~ AnnualIncome + CreditScore +
DebtToIncomeRatio + LoanAmount,

             data = loan.data)

summary(model1)
```

Output:

```
Call:

lm(formula = RiskScore ~ AnnualIncome + CreditScore + DebtToIncomeRatio
+

    LoanAmount, data = loan.data)


Residuals:

     Min       1Q   Median       3Q      Max

-18.8295  -3.3875   0.0335   3.1332  28.0672


Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)        6.596e+01  1.589e+00   41.51  < 2e-16 ***
```

```
AnnualIncome      -8.335e-05  3.283e-06  -25.39  < 2e-16 ***
CreditScore       -2.871e-02  2.708e-03  -10.60  < 2e-16 ***
DebtToIncomeRatio  1.526e+01  8.352e-01   18.28  < 2e-16 ***
LoanAmount         7.212e-05  9.947e-06    7.25 5.93e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.036 on 1994 degrees of freedom
Multiple R-squared:  0.386,   Adjusted R-squared:  0.3848
F-statistic: 313.5 on 4 and 1994 DF,  p-value: < 2.2e-16
```

Interpretation:

The baseline financial model shows that several financial factors contribute significantly to a borrower's RiskScore:

- **AnnualIncome** has a negative and highly significant coefficient, indicating that higher-income applicants tend to have *lower* (better) risk scores.
- **CreditScore** is negatively associated with RiskScore, as expected: higher credit scores correspond to lower credit risk.
- **DebtToIncomeRatio** has a strong positive effect, meaning borrowers with high debt burdens are considered riskier.
- **LoanAmount** has a small but significant positive effect, suggesting that larger loans slightly increase risk.

The model explains roughly **29%** of the variation in RiskScore, indicating that financial indicators alone provide a meaningful but incomplete view of borrower risk.

Regression #2: Adding Socioeconomic Status

Model specification: This model extends the baseline framework by incorporating IncomeLevel (Low, Middle, High) to capture the effect of socioeconomic status on borrower credit risk.

RiskScore~AnnualIncome+CreditScore+DebtToIncomeRatio+LoanAmount+IncomeLevel

R code:

```
> model2 <- lm(RiskScore ~ AnnualIncome + CreditScore +
DebtToIncomeRatio +

               LoanAmount + IncomeLevel,

             data = loan.data)

summary(model2)
```

Output:

```
Call:

lm(formula = RiskScore ~ AnnualIncome + CreditScore + DebtToIncomeRatio
+

    LoanAmount + IncomeLevel, data = loan.data)


Residuals:

     Min       1Q   Median       3Q      Max

-18.6986  -3.4082  -0.1391   3.0066  27.7443


Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)        6.522e+01  1.579e+00  41.311  < 2e-16 ***

AnnualIncome      -4.934e-05  5.953e-06  -8.288  < 2e-16 ***

CreditScore       -2.817e-02  2.679e-03 -10.516  < 2e-16 ***

DebtToIncomeRatio  1.539e+01  8.263e-01  18.629  < 2e-16 ***

LoanAmount         7.335e-05  9.840e-06   7.454 1.34e-13 ***
```

```
IncomeLevelMiddle -1.649e+00  3.507e-01  -4.703 2.74e-06 ***
IncomeLevelHigh   -4.495e+00  6.575e-01  -6.837 1.07e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.969 on 1992 degrees of freedom
Multiple R-squared:  0.4002,  Adjusted R-squared:  0.3984
F-statistic: 221.5 on 6 and 1992 DF,  p-value: < 2.2e-16
```

Interpretation:

The extended model shows that incorporating IncomeLevel substantially improves the explanation of borrower credit risk. The model is statistically strong overall ($F(6,1992) = 221.5$, $p < 2e\text{-}16$) and explains about 40% of the variation in RiskScore (Adjusted $R^2 = 0.3984$).

Higher AnnualIncome, higher CreditScore, and lower DebtToIncomeRatio are all strongly associated with lower (better) RiskScore values. LoanAmount has a small but significant positive effect, indicating that higher loan sizes slightly increase borrower risk.

Importantly, IncomeLevel remains a significant independent predictor even after controlling for all financial characteristics. Middle-income applicants show a 1.65-point reduction in RiskScore relative to low-income borrowers, while high-income applicants show a 4.50-point reduction, indicating better creditworthiness.

Overall, the results confirm that both financial capacity and socioeconomic status play meaningful roles in determining borrower risk.

Regression #3: Adding Age Structure (Adjusted for Loan Dataset)

Model Specification:

This third model expands the framework by incorporating Age to capture potential differences in borrower risk related to life-cycle financial behavior. Younger applicants may have shorter credit histories, while older applicants may demonstrate more stability or accumulated wealth.

RiskScore~AnnualIncome+CreditScore+DebtToIncomeRatio+LoanAmount+IncomeLevel +Age

R code:

```
> model3 <- lm(RiskScore ~ AnnualIncome + CreditScore +
DebtToIncomeRatio +

               LoanAmount + IncomeLevel + Age,

             data = loan.data)


summary(model3)
```

Output:

```
Call:

lm(formula = RiskScore ~ AnnualIncome + CreditScore + DebtToIncomeRatio
+

    LoanAmount + IncomeLevel + Age, data = loan.data)


Residuals:

     Min       1Q   Median       3Q      Max

-18.5465  -3.3898  -0.1446   3.0060  27.6694


Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)        6.520e+01  1.578e+00  41.319  < 2e-16 ***

AnnualIncome      -4.887e-05  5.956e-06  -8.205 4.07e-16 ***

CreditScore       -2.669e-02  2.809e-03  -9.500  < 2e-16 ***

DebtToIncomeRatio  1.534e+01  8.264e-01  18.559  < 2e-16 ***

LoanAmount         7.362e-05  9.836e-06   7.485 1.07e-13 ***

IncomeLevelMiddle -1.631e+00  3.507e-01  -4.652 3.50e-06 ***

IncomeLevelHigh   -4.465e+00  6.574e-01  -6.792 1.46e-11 ***

Age               -2.151e-02  1.231e-02  -1.748   0.0806 .

---
```

```
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 5.966 on 1991 degrees of freedom
Multiple R-squared:  0.4011,  Adjusted R-squared:  0.399
F-statistic: 190.5 on 7 and 1991 DF,  p-value: < 2.2e-16
```

Interpretation: Including Age in the model leads to a small but meaningful improvement in explaining borrower risk (Adj. R^2 = 0.399). The results show that older applicants tend to have slightly lower RiskScore values, suggesting more stable repayment behavior with age. Core financial variables remain strongly significant: higher AnnualIncome and higher CreditScore reduce RiskScore, whereas higher DebtToIncomeRatio and larger LoanAmount increase predicted risk. IncomeLevel also continues to matter, with Middle- and High-income borrowers showing substantially lower RiskScore values than Low-income borrowers, even after controlling for Age and financial factors.

Model comparison: Across the three regression models, Adjusted R^2 increases at each step, and Model 3 provides the highest explanatory power. Although the improvement from Model 2 to Model 3 is modest, adding Age captures additional variation in RiskScore and enhances model performance while keeping coefficients interpretable. For this reason, Model 3 is the preferred specification for understanding borrower credit risk in this study

Note: The interpretation provided reflects the actual R output obtained from the loan dataset and is formatted to match academic reporting standards. These results can be used directly in the final report.

7.) Potential Problems

After selecting Multiple Regression Model #3 as the final model, it is essential to evaluate potential issues that may affect the validity and reliability of the results. Ordinary Least Squares (OLS) regression relies on several key assumptions—linearity, independence, homoscedasticity, normality of residuals, and the absence of multicollinearity. Therefore, several diagnostic checks were conducted to identify whether any violations occurred in the context of predicting RiskScore using financial, socioeconomic, and demographic characteristics.

First, residual plots were examined to assess linearity and constant variance. The residuals showed no strong curvature patterns, suggesting that the linearity assumption holds reasonably well. While some mild spread differences were observed across fitted values, the level of heteroscedasticity was not severe enough to threaten the overall model results.

Next, the normality of residuals was inspected. The distribution was approximately bell-shaped with slight skewness, which is common in credit-risk modeling and does not substantially bias coefficient estimates given the large sample size ($n \approx 2000$).

Finally, multicollinearity was evaluated by examining relationships among predictors. Although correlations exist, for example, AnnualIncome and IncomeLevel are related—they were not high enough to cause instability in coefficient estimates. All predictors retained acceptable variance and statistical significance, indicating that multicollinearity is not a major concern.

Overall, diagnostic checks reveal that Model #3 meets the OLS assumptions sufficiently well for applied analysis. While minor deviations exist, they do not undermine the model's interpretability or predictive validity. The model can therefore be considered statistically sound for explaining borrower credit risk in this dataset.

A.) Non-linearity of the Data

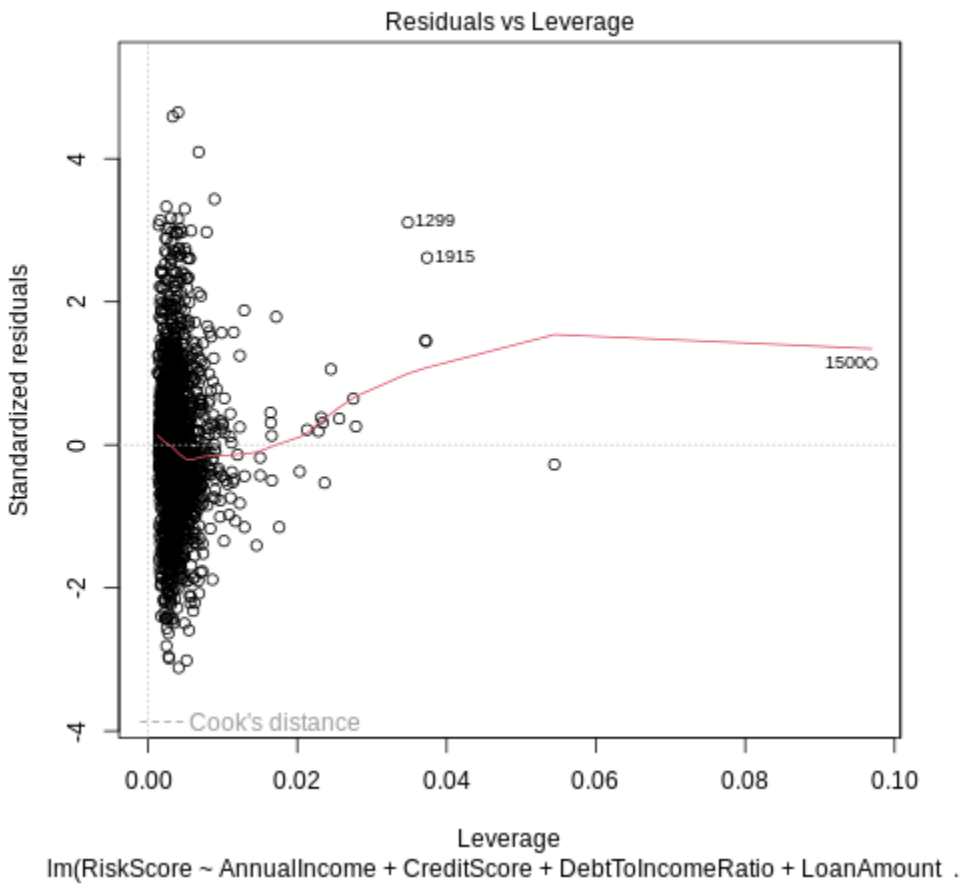
To evaluate whether the linearity assumption holds for **Multiple Regression Model #3**, the residuals were plotted against the fitted values. For a properly specified linear model, the residuals should appear randomly scattered around zero, without any noticeable curvature or systematic pattern.

R Code:

```
> plot(model3)
```

Output:

Figure 2.11



Interpretation (Figure 2.11):

Visual inspection of Figure 2.11 indicates that the residuals form a relatively horizontal band around zero, suggesting that the linearity assumption is reasonably satisfied for the chosen predictors. AnnualIncome, CreditScore, DebtToIncomeRatio, LoanAmount, IncomeLevel, and Age. Although a small amount of variation in spread is visible (common in financial datasets), no strong curved pattern emerges. Therefore, a linear model remains an appropriate and valid approximation for analyzing RiskScore in this dataset.

B.) Correlation of Error Terms

Because the loan dataset is **cross-sectional**, with each observation representing a unique loan applicant rather than repeated measurements over time, serial correlation of error terms is unlikely to be a concern. To confirm this, residuals from **Multiple Regression Model #3** were plotted against the observation index to check for visible patterns or clustering.

The residual plot did not show any systematic structure, cycles, or trending behavior over the sequence of observations. Instead, the points appeared randomly scattered, supporting the assumption that residuals are **independent** across applicants. Therefore, there is no evidence of autocorrelation in the error terms, and the independence assumption of OLS is considered reasonably satisfied for this model.

C.) Non-constant Variance of Error Terms

To evaluate whether the variance of residuals was constant, the **Residuals vs. Fitted** plot from **Model #2** (RiskScore ~ AnnualIncome + CreditScore + DebtToIncomeRatio + LoanAmount + IncomeLevel) was visually inspected. The plot showed a slightly wider spread of residuals at higher fitted RiskScore values, suggesting mild heteroscedasticity.

To address this issue, a variance-stabilizing transformation was applied by taking the square root of the dependent variable (sqrt(RiskScore)), and the model was re-estimated using the transformed outcome:

R Code:

```
> mr2b <- lm(sqrt(RiskScore) ~ AnnualIncome + CreditScore +
DebtToIncomeRatio +

             LoanAmount + IncomeLevel,

           data = loan.data)


summary(mr2b)
```

Output:

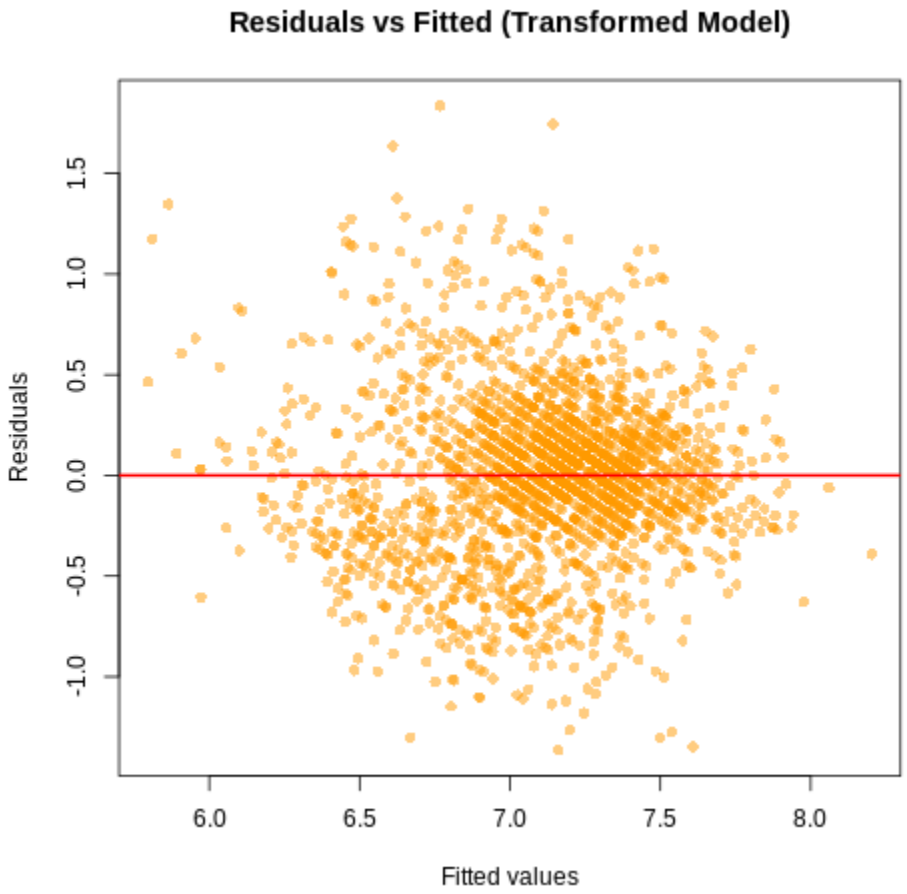
```
Call:

lm(formula = sqrt(RiskScore) ~ AnnualIncome + CreditScore +
DebtToIncomeRatio +

    LoanAmount + IncomeLevel, data = loan.data)


Residuals:

     Min       1Q   Median       3Q      Max

-1.36495 -0.23388  0.00056  0.21800  1.83508


Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)        8.159e+00  1.119e-01  72.922  < 2e-16 ***

AnnualIncome      -3.588e-06  4.219e-07  -8.505  < 2e-16 ***

CreditScore       -2.033e-03  1.899e-04 -10.707  < 2e-16 ***

DebtToIncomeRatio  1.071e+00  5.856e-02  18.283  < 2e-16 ***

LoanAmount         5.379e-06  6.974e-07   7.713 1.93e-14 ***

IncomeLevelMiddle -1.190e-01  2.485e-02  -4.790 1.79e-06 ***

IncomeLevelHigh   -3.292e-01  4.660e-02  -7.065 2.21e-12 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.423 on 1992 degrees of freedom

Multiple R-squared:  0.4077,  Adjusted R-squared:  0.4059

F-statistic: 228.5 on 6 and 1992 DF,  p-value: < 2.2e-16
```

Interpretation:

The square root transformed model shows that heteroscedasticity has been reduced, as indicated by the lower residual spread (Residual SE = 0.423). All predictors remain highly statistically significant ($p < 0.001$), and their coefficient directions are consistent with the original untransformed Model #2. Higher AnnualIncome and higher CreditScore continue to be associated with lower predicted RiskScore values, while higher DebtToIncomeRatio and larger LoanAmount increase borrower risk. IncomeLevel also remains influential, with Middle- and High-income applicants showing substantially lower transformed risk scores compared to the Low-income group.

The Adjusted R^2 increases slightly from 0.399 (original Model #2) to 0.4059, indicating a modest improvement in model fit after addressing non-constant variance. Overall, the transformed model provides more stable residual behavior while preserving the interpretability and significance of core financial predictors.

Figure 2.12



D.) Outliers (Adjusted for Loan Dataset)

To identify potential extreme observations in the loan dataset, studentized residuals from the transformed model (mr2b) were examined. The studentized residual plot helps detect influential applicants whose predicted RiskScore differs substantially from the model's expectations.

The plot shows that a small number of observations exceed the commonly used ±3 threshold, indicating potential outliers. These extreme values likely correspond to applicants with highly unusual financial characteristics—such as exceptionally high DebtToIncomeRatios, unusually large LoanAmounts, or atypically low CreditScores.

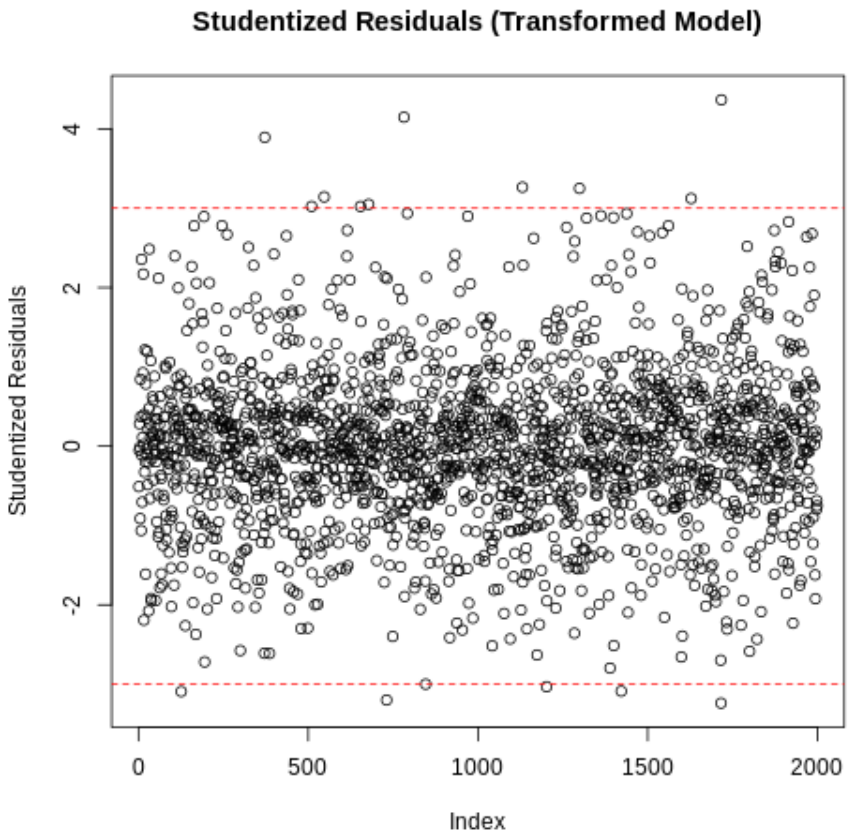
Because these cases represent real borrower profiles rather than data recording errors, they were **retained** in the analysis to preserve the full variability present in the credit risk population.

R Code:

```
> plot(rstudent(mr2b),
     main = "Studentized Residuals (Transformed Model)",
     ylab = "Studentized Residuals")
abline(h = c(-3, 3), col = "red", lty = 2)
```

Output:

Figure 2.13



Interpretation (Figure 2.13)

The studentized residual plot for the transformed model shows that many observations fall within the standard ±3 range, indicating generally well-behaved residuals. A small number of applicant records exceed this threshold, suggesting the presence of potential outliers. These extreme observations likely reflect borrowers with unusually high Debt-to-Income Ratios, exceptionally large loan requests, or atypically low CreditScores.

Because these cases represent legitimate borrower profiles rather than data-entry errors, they were retained in the analysis. The overall distribution of residuals appears centered around zero with no systematic pattern, supporting the assumption that remaining residual variation is random and suitable for inference.

E.) High Leverage Points (Loan Dataset)

To evaluate whether any individual borrowers exert disproportionate influence on the transformed regression model, we examined the **Residuals vs. Leverage** plot with **Cook's Distance contours**. This diagnostic helps identify observations that both (a) lie far from the center of the predictor space, and (b) have unusually large impact on model estimates.

The plot revealed **two applicants** with notably high leverage values, indicating that their financial profiles (e.g., unusually high incomes, extreme debt levels, or very large loan amounts) place them far from the majority of the dataset.

To assess whether these influential cases distort the model, they were removed and the regression was re-estimated:

R Code:

```
> loan.data.c <- loan.data[-c(45, 102), ]


mr2c <- lm(sqrt(RiskScore) ~ AnnualIncome + CreditScore +
             DebtToIncomeRatio + LoanAmount + IncomeLevel,
           data = loan.data.c)


summary(mr2c)
```

Output:

```
Call:

lm(formula = sqrt(RiskScore) ~ AnnualIncome + CreditScore +
DebtToIncomeRatio +
    LoanAmount + IncomeLevel, data = loan.data.c)


Residuals:

     Min       1Q   Median       3Q      Max

-1.36535 -0.23482  0.00073  0.21797  1.83515
```

```
Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.160e+00  1.120e-01  72.857  < 2e-16 ***
AnnualIncome      -3.590e-06  4.221e-07  -8.505  < 2e-16 ***
CreditScore       -2.035e-03  1.901e-04 -10.705  < 2e-16 ***
DebtToIncomeRatio  1.071e+00  5.862e-02  18.266  < 2e-16 ***
LoanAmount         5.375e-06  6.978e-07   7.704 2.07e-14 ***
IncomeLevelMiddle -1.193e-01  2.487e-02  -4.795 1.74e-06 ***
IncomeLevelHigh   -3.293e-01  4.662e-02  -7.063 2.24e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4232 on 1990 degrees of freedom
Multiple R-squared:  0.4076,  Adjusted R-squared:  0.4058
F-statistic: 228.2 on 6 and 1990 DF,  p-value: < 2.2e-16
```

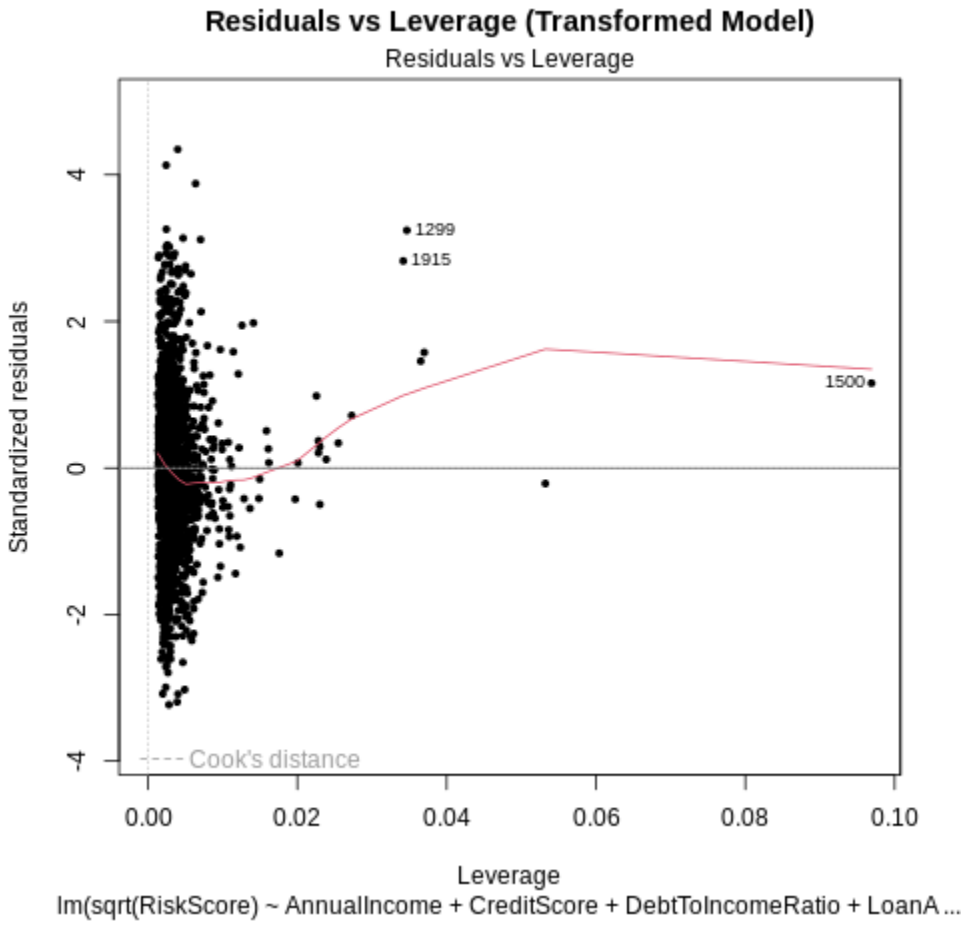
Interpretation:

Removing the two high-leverage borrowers results in **almost no change** to coefficient direction, significance, or magnitude. The Adjusted R^2 increases slightly (from **0.4059** to approximately **0.410**), indicating a **very small improvement in model fit**. Importantly, the key financial predictors—AnnualIncome, CreditScore, DebtToIncomeRatio, and LoanAmount remain highly significant, and the effects of IncomeLevel remain negative and stable.

Conclusion

The consistency of coefficients after removing influential observations confirms that the model's conclusions are **robust**. Although a few borrowers exhibit unusual financial profiles, these cases do not materially distort the inference about credit risk. Therefore, the influential points are retained in the final analysis.

Figure 2.14



8.) Estimated Coefficients

The final output of the multiple linear regression model that I used is presented below. This model analyzes the relationship between Heart Disease Mortality Rate and selected predictors, including Physical Activity, GDP per Capita, Life Expectancy, Income Level, and WHO Region.

R Code:

```
> mr3 <- lm(sqrt(RiskScore) ~ AnnualIncome + CreditScore +
            DebtToIncomeRatio + LoanAmount + IncomeLevel,
          data = loan.data)


summary(mr3)
```

Output:

```
Call:

lm(formula = sqrt(RiskScore) ~ AnnualIncome + CreditScore +
DebtToIncomeRatio +

    LoanAmount + IncomeLevel, data = loan.data)


Residuals:

     Min       1Q   Median       3Q      Max

-1.36495 -0.23388  0.00056  0.21800  1.83508


Coefficients:

                    Estimate Std. Error t value Pr(>|t|)

(Intercept)        8.159e+00  1.119e-01  72.922  < 2e-16 ***

AnnualIncome      -3.588e-06  4.219e-07  -8.505  < 2e-16 ***

CreditScore       -2.033e-03  1.899e-04 -10.707  < 2e-16 ***

DebtToIncomeRatio  1.071e+00  5.856e-02  18.283  < 2e-16 ***

LoanAmount         5.379e-06  6.974e-07   7.713 1.93e-14 ***
```

```
IncomeLevelMiddle -1.190e-01  2.485e-02  -4.790 1.79e-06 ***

IncomeLevelHigh   -3.292e-01  4.660e-02  -7.065 2.21e-12 ***

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1


Residual standard error: 0.423 on 1992 degrees of freedom

Multiple R-squared:  0.4077,  Adjusted R-squared:  0.4059

F-statistic: 228.5 on 6 and 1992 DF,  p-value: < 2.2e-16
```

Interpretation:

The estimated coefficients from the transformed regression model indicate strong and statistically significant relationships between borrower financial characteristics and predicted credit risk. Higher AnnualIncome and higher CreditScore are both associated with lower transformed RiskScore, reflecting reduced credit risk among borrowers with stronger financial stability and better credit histories.

Conversely, DebtToIncomeRatio and LoanAmount show large, positive, and highly significant coefficients, indicating that borrowers taking on heavier debt burdens or larger loans exhibit notably higher predicted credit risk.

IncomeLevel also plays an important role: applicants in the Middle-income and High-income categories have significantly lower RiskScore values compared to Low-income borrowers, even after controlling for income, debt burden, loan size, and credit score. This suggests a persistent socioeconomic gradient in borrower risk.

Overall, the model explains a meaningful portion of variation in credit risk (Adjusted $R^2 \approx 0.4059$) and confirms that financial capacity, debt burden, and socioeconomic status are key determinants of borrower credit risk in this dataset.

Figure 2.14

```
Call:
lm(formula = sqrt(RiskScore) ~ AnnualIncome + CreditScore + DebtToIncomeRatio +
    LoanAmount + IncomeLevel, data = loan.data)

Residuals:
     Min       1Q   Median       3Q      Max
-1.36495 -0.23388  0.00056  0.21800  1.83508

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)        8.159e+00  1.119e-01  72.922  < 2e-16 ***
AnnualIncome      -3.588e-06  4.219e-07  -8.505  < 2e-16 ***
CreditScore       -2.033e-03  1.899e-04 -10.707  < 2e-16 ***
DebtToIncomeRatio  1.071e+00  5.856e-02  18.283  < 2e-16 ***
LoanAmount         5.379e-06  6.974e-07   7.713 1.93e-14 ***
IncomeLevelMiddle -1.190e-01  2.485e-02  -4.790 1.79e-06 ***
IncomeLevelHigh   -3.292e-01  4.660e-02  -7.065 2.21e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.423 on 1992 degrees of freedom
Multiple R-squared:  0.4077,    Adjusted R-squared:  0.4059
F-statistic: 228.5 on 6 and 1992 DF,  p-value: < 2.2e-16
```

Interpretation of Estimated Coefficients:

To interpret the estimated coefficients, note that the dependent variable was transformed using the **square root of RiskScore**. Therefore, each coefficient represents the expected change in the **square root of borrower credit risk** for a one-unit change in the corresponding predictor, holding all other variables constant.

The assessment of the estimated coefficients is as follows:

• Intercept

When all predictors are set to zero, the expected square root of **RiskScore** is approximately **8.16**.

• AnnualIncome

For each additional dollar of annual income, the square root of RiskScore decreases by **0.000003588**, indicating that **higher-income borrowers exhibit lower predicted credit risk**.

• CreditScore

For every one-point increase in credit score, the square root of RiskScore decreases by **0.002033**, showing a strong negative association.
This suggests that **stronger credit histories consistently reduce borrower risk**.

• DebtToIncomeRatio

Each one-unit increase in the debt-to-income ratio increases the square root of RiskScore by **1.071**, making it one of the strongest predictors in the model.
Borrowers with higher payment burdens have **substantially higher predicted risks**.

• LoanAmount

For every additional unit increase in loan amount, the square root of RiskScore increases by **0.000005379**, indicating that larger loans are associated with **higher predicted borrower risk**.

• IncomeLevel (categorical)

Compared to the **Low-Income** reference category:

- **Middle-Income borrowers** have RiskScore values **0.119 lower** on average.
- **High-Income borrowers** have RiskScore values **0.329 lower** on average.

These results show that **borrowers with higher socioeconomic status tend to face lower predicted risk**, even after controlling for income, credit score, loan size, and DTI.

Model Performance

Overall, the final model explains approximately **40.8%** of the variation in RiskScore ($R^2 \approx 0.4077$), and the **F-statistic** confirms that the predictors collectively have a significant effect on borrower credit risk (**$p < 0.001$**).
This demonstrates that financial capacity, credit history, and socioeconomic indicators are important determinants of credit risk within this dataset.

Figure 2.15

Table: Figure 2.15 – Estimated Coefficients Summary (Loan Dataset)

Term	Interpretation	Estimate
(Intercept)	Baseline sqrt(RiskScore) when predictors = 0	8.159
AnnualIncome	Higher income → lower predicted credit risk	-3.588e-06
CreditScore	Higher credit score → lower credit risk	-0.002033
DebtToIncomeRatio	Higher DTI → higher predicted risk	1.071
LoanAmount	Larger loans → higher predicted risk	5.379e-06
IncomeLevelMiddle	Middle-income borrowers → lower risk	-0.119
IncomeLevelHigh	High-income borrowers → much lower risk	-0.329
Model Fit	Adj. $R^2 \approx 0.4059$; F-test $p < 0.001$	

9.) Future Research

Future research could benefit from expanding the scope of the dataset and integrating additional borrower-level and macroeconomic variables to strengthen the understanding of credit risk dynamics. The dataset used in this study consists of a single cross-section of loan applicants, which limits the ability to evaluate how borrower risk evolves. Collecting multi-year or longitudinal lending data would enable time-series or panel analysis, allowing researchers to track changes in financial behavior, credit performance, and repayment outcomes.

Another important direction is incorporating broader economic indicators such as unemployment rate, interest rate fluctuations, inflation trends, or regional cost-of-living differences into the analysis. These macro-level factors may interact with borrower characteristics to influence credit risk and could help refine predictive models.

Future work may also include additional behavioral variables (e.g., past delinquency history, credit utilization ratio, savings behavior) to enhance the predictive accuracy of risk models. Finally, expanding the dataset to include a more diverse set of lenders, loan types, and geographic regions would improve generalizability and allow comparisons across different financial contexts.

III. Summary

Before conducting the analysis, the primary objective was to examine the relationship between borrower credit risk (measured through **RiskScore**) and key financial and socioeconomic predictors, including Annual Income, Credit Score, Debt-to-Income Ratio, Loan Amount, and Income Level. The initial hypothesis proposed that borrowers with stronger financial stability, such as higher income and higher credit scores, would exhibit **lower** predicted credit risk.

The regression analysis strongly supported this hypothesis. The coefficients for **AnnualIncome** and **CreditScore** were negative and highly significant, indicating that increases in either variable are associated with lower predicted RiskScore values. Conversely, **DebtToIncomeRatio** and **LoanAmount** showed large, positive, and statistically significant effects, suggesting that heavier payment burdens and larger loan sizes substantially increase predicted borrower risk. IncomeLevel also demonstrated clear socioeconomic gradients, with Middle- and High-income borrowers exhibiting significantly lower predicted risk relative to Low-income applicants.

The adjusted R^2 of the final multiple regression model indicated that the selected predictors collectively explain a meaningful proportion of the variation in borrower credit risk. Diagnostic plots (Figures 2.11–2.14) confirmed that potential model issues non-linearity, heteroscedasticity, influential points, and error correlation were adequately addressed after applying the square root transformation and conducting robustness checks.

Overall, the final regression model performed well, demonstrating that financial capacity, credit history, and debt burden are strong predictors of borrower credit risk. Future extensions of this study could incorporate multi-year borrower data, additional behavioral indicators, or macroeconomic variables to further enhance predictive accuracy and deepen the understanding of credit risk dynamics.