Logistic Regression Analysis

Nur Ahmed

2024-03-18

Introduction (4 points)

In this analysis, we look at data gathered from loan borrowers to assess their risk of default and seeking valuable insight on these risk assessments, whether or not they seem reasonable or make sense. The general purpose of my analysis is to see how big of an influence these variables factored in loan default.

```
loanDefaultData<-read.csv("LoanDefaultData.csv")
```

The Variables (8 points)

The variables available for this analysis are:

Response:

- Default (Binary/Categorical): This records whether the loan defaulted or not. This is the variable we seek to predict with our logistic regression model.

```
table(loanDefaultData$Default)
```

```
##
##    0    1
## 3091  409
```

Predictors:

```
names(loanDefaultData)
```

```
##  [1] "Age"            "Income"         "LoanAmount"     "CreditScore"
##  [5] "MonthsEmployed" "NumCreditLines" "InterestRate"   "LoanTerm"
##  [9] "Education"      "EmploymentType" "MaritalStatus"  "HasCoSigner"
## [13] "Default"
```

- Age (Numeric): This records the age of the borrower in years. The standard deviation is 14.96 years and the mean is 43.74 years.

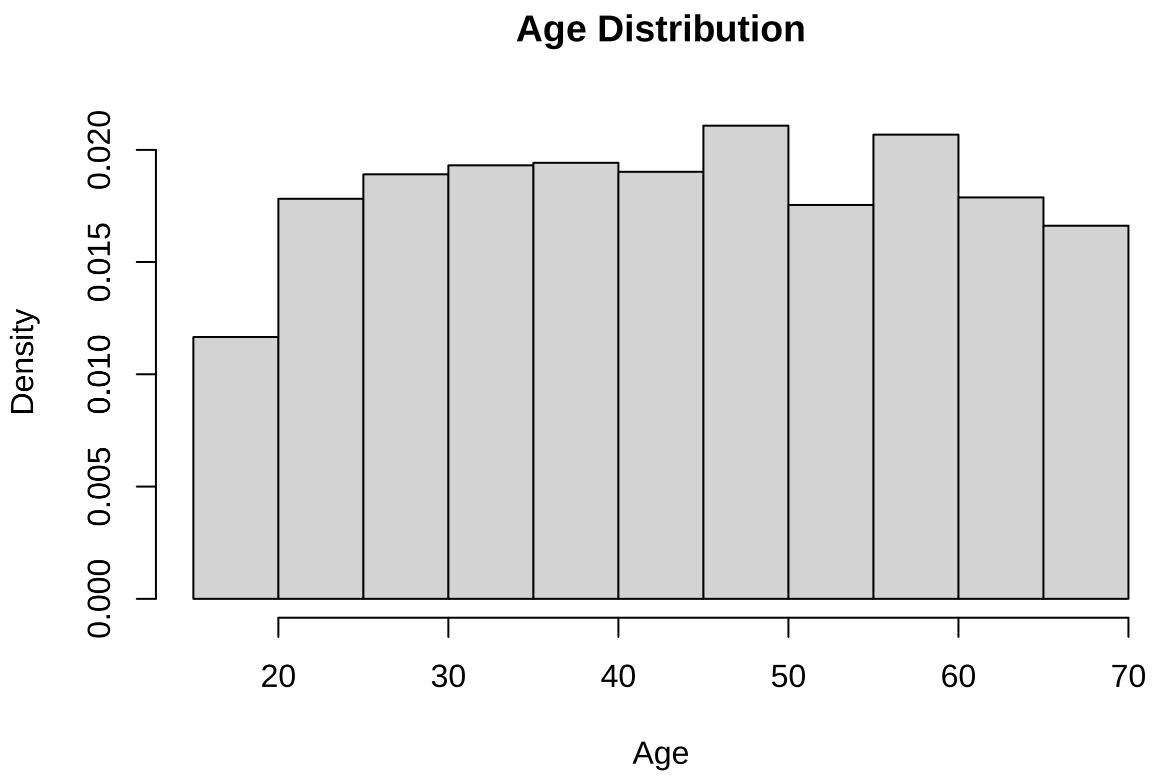
```
sd(loanDefaultData$Age)
```

```
## [1] 14.95721
```

```
mean(loanDefaultData$Age)
```

```
## [1] 43.74314
```

```
hist(loanDefaultData$Age, main="Age Distribution", xlab="Age", freq = FALSE)
```



- Income (Numeric): This records the yearly income of the borrower in dollars ($). The standard deviation is 39177.07 and the mean is 82351.62

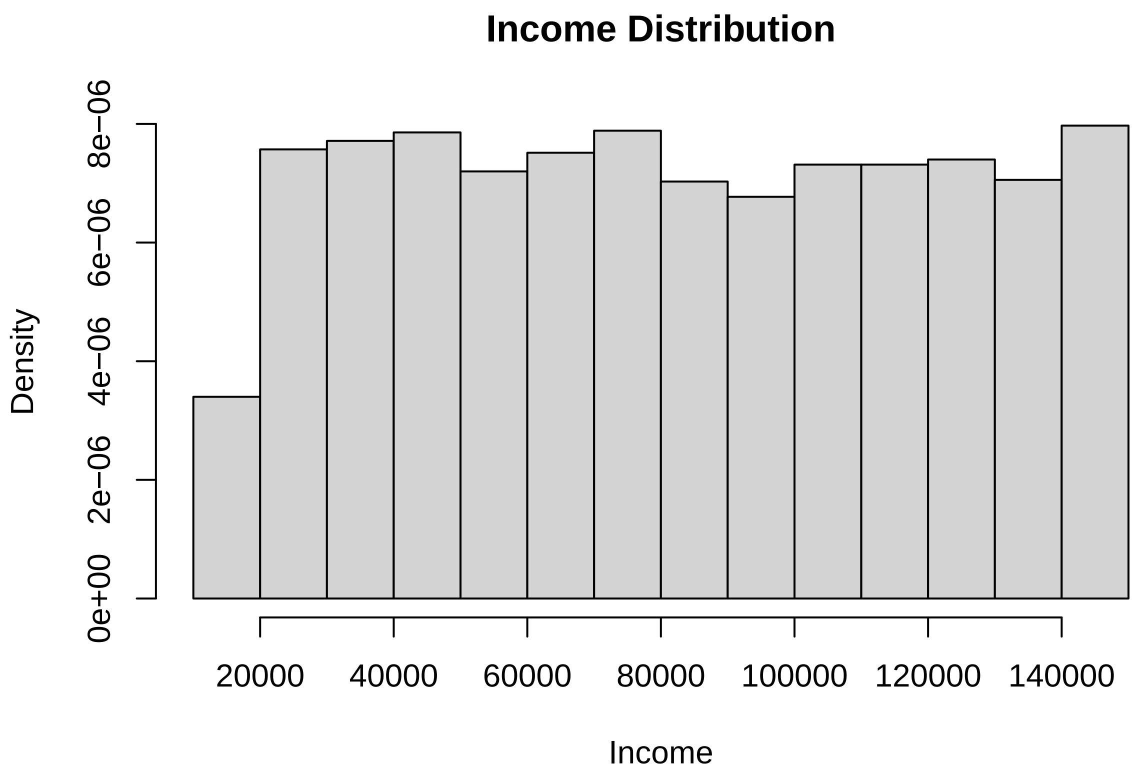
```
sd(loanDefaultData$Income)
```

```
## [1] 39177.07
```

```
mean(loanDefaultData$Income)
```

```
## [1] 82351.62
```

```
hist(loanDefaultData$Income, main="Income Distribution", xlab="Income", freq = FALSE)
```



- LoanAmount (Numeric): The amount of money being borrowed in dollars ($). The standard deviation is 70830.42 and the mean is 128733.50

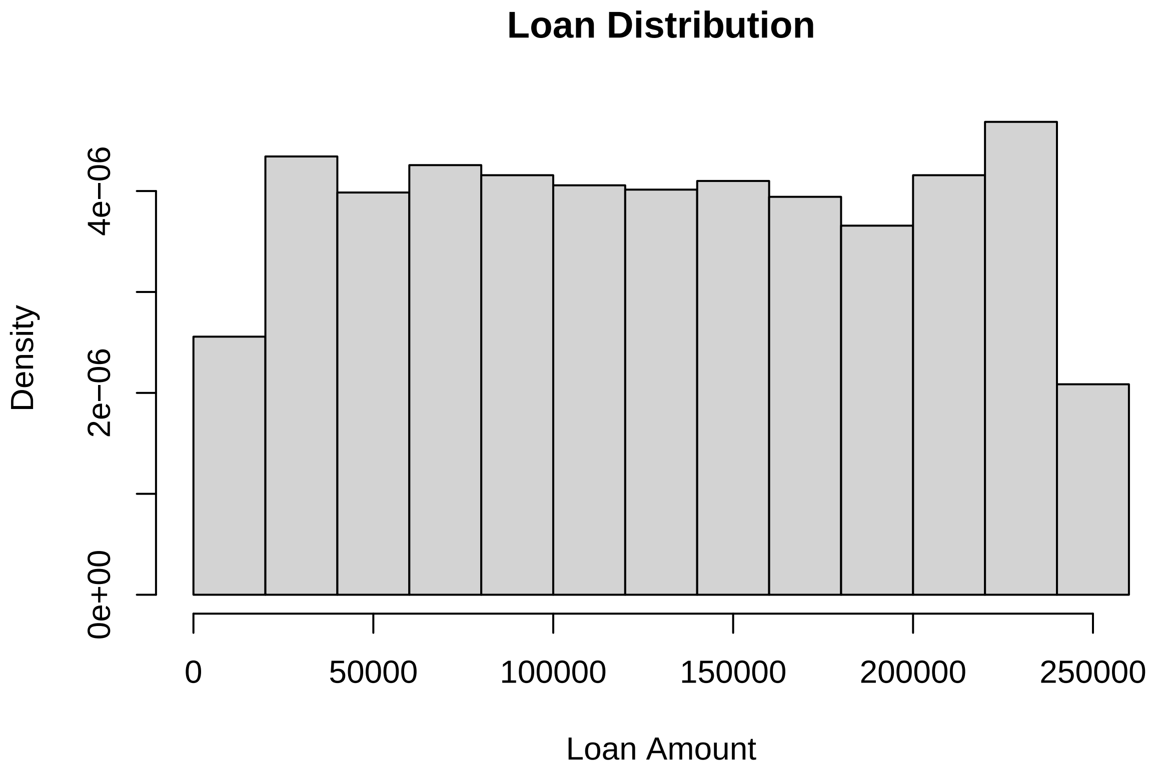
```
sd(loanDefaultData$LoanAmount)
```

```
## [1] 70830.42
```

```
mean(loanDefaultData$LoanAmount)
```

```
## [1] 128733.5
```

```
hist(loanDefaultData$LoanAmount, main="Loan Distribution", xlab="Loan Amount", freq = FALSE)
```



- CreditScore (Numeric): The credit score of the borrower. The standard deviation is 159.42 and the mean is 575.07

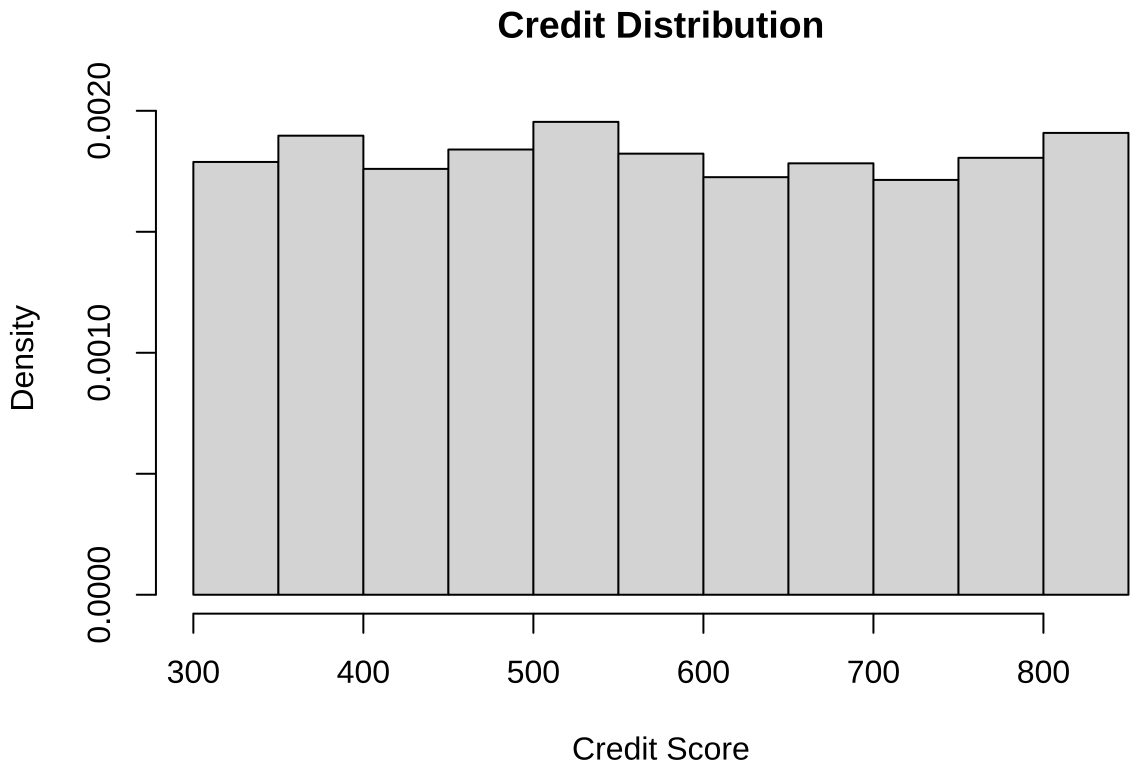
```r
sd(loanDefaultData$CreditScore)
```

```
## [1] 159.4175
```

```r
mean(loanDefaultData$CreditScore)
```

```
## [1] 575.0686
```

```r
hist(loanDefaultData$CreditScore, main="Credit Distribution", xlab="Credit Score", freq = FALSE)
```



- MonthsEmployed (Numeric): The number of months the borrower has been employed. The standard deviation is 34.1 months and the mean is 60.16 months

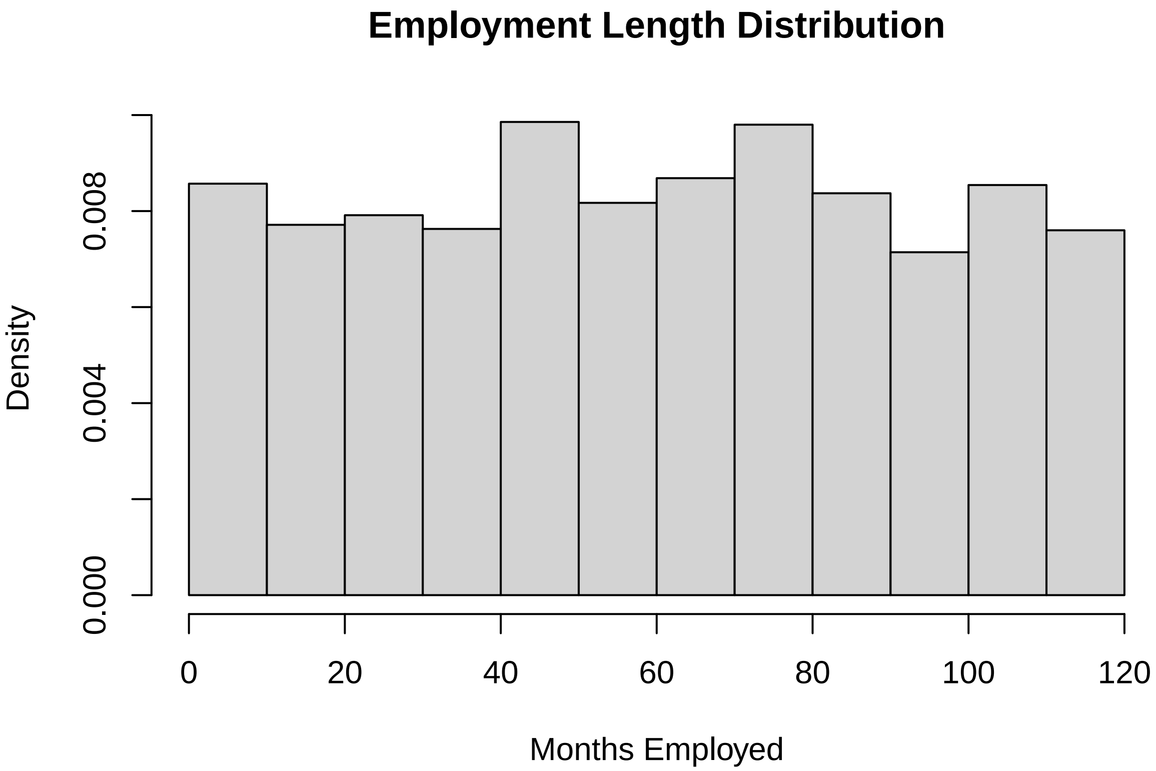
```
sd(loanDefaultData$MonthsEmployed)
```

```
## [1] 34.10338
```

```
mean(loanDefaultData$MonthsEmployed)
```

```
## [1] 60.15971
```

```
hist(loanDefaultData$MonthsEmployed, main="Employment Length Distribution", xlab="Months Employed", frec
```



- NumCreditLines (Numeric): The number of credit lines the borrower has open. The standard deviation is 1.12 and the mean is 2.47

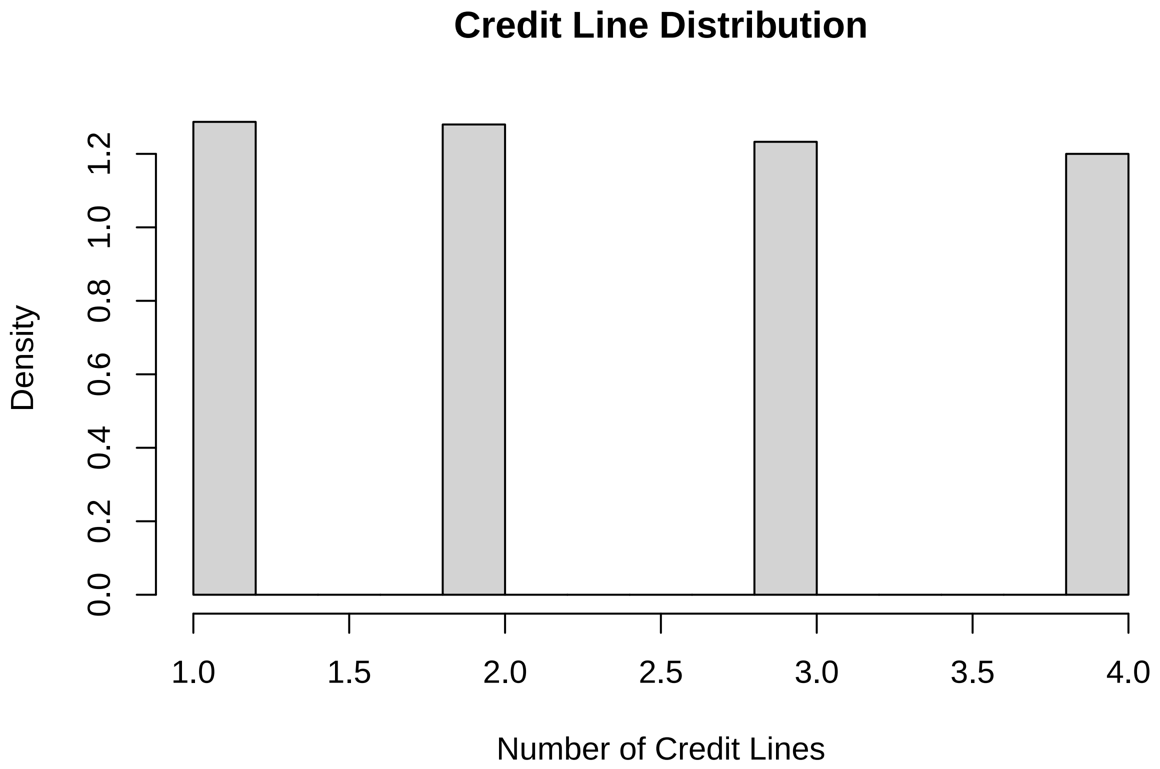
```
sd(loanDefaultData$NumCreditLines)
```

```
## [1] 1.115464
```

```
mean(loanDefaultData$NumCreditLines)
```

```
## [1] 2.469143
```

```
hist(loanDefaultData$NumCreditLines, main="Credit Line Distribution", xlab="Number of Credit Lines", fr
```



- InterestRate (Numeric): The interest rate for the loan. The standard deviation is 6.66% and the mean is 13.61%

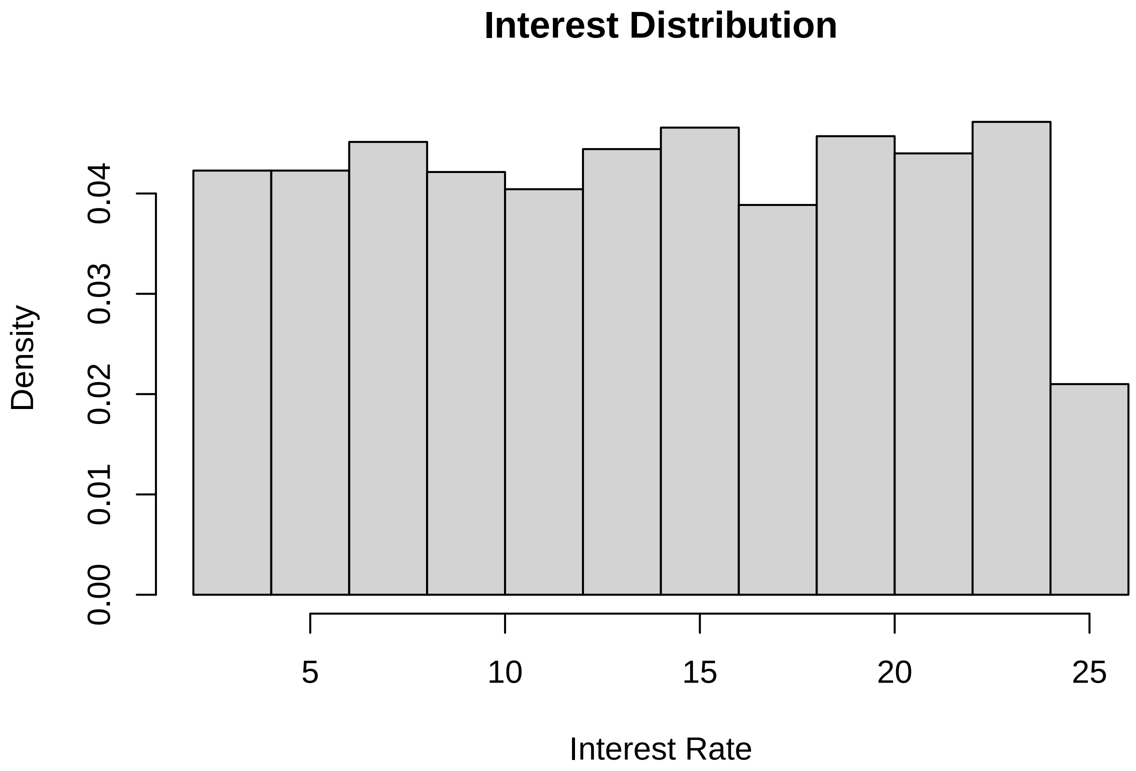
```
sd(loanDefaultData$InterestRate)
```

```
## [1] 6.661623
```

```
mean(loanDefaultData$InterestRate)
```

```
## [1] 13.61331
```

```
hist(loanDefaultData$InterestRate, main="Interest Distribution", xlab="Interest Rate", freq = FALSE)
```



- LoanTerm (Numeric): The term length of the loan in months. The standard deviation is 16.94 months and the mean is 36.04 months

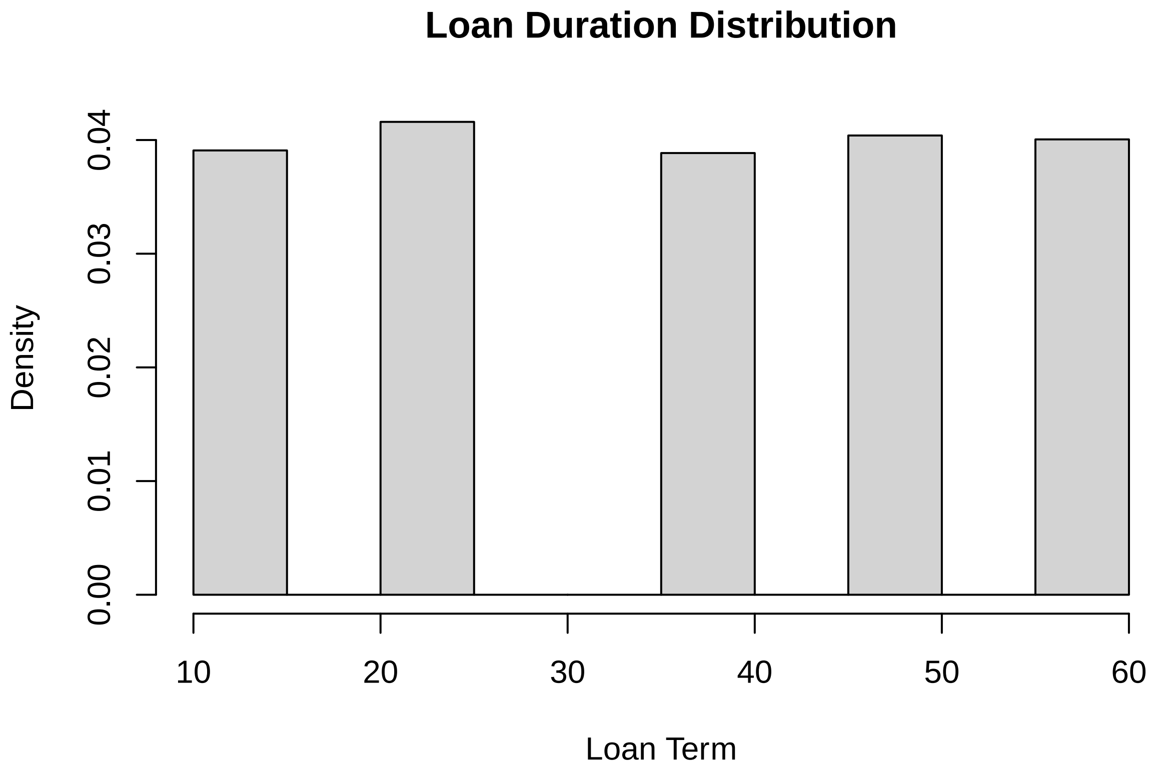
```
sd(loanDefaultData$LoanTerm)
```

```
## [1] 16.94259
```

```
mean(loanDefaultData$LoanTerm)
```

```
## [1] 36.04457
```

```
hist(loanDefaultData$LoanTerm, main="Loan Duration Distribution", xlab="Loan Term", freq = FALSE)
```



- Education (Categorical): The highest level of education attained by the borrower.

```r
table(loanDefaultData$Education)
```

```
##
##  Bachelor's High School    Master's         PhD
##         902         849         893         856
```

- EmploymentType (Categorical): The type of employment status of the borrower.

```r
table(loanDefaultData$EmploymentType)
```

```
##
##     Full-time     Part-time Self-employed    Unemployed
##           872           873           908           847
```

- MaritalStatus (Categorical): The marital status of the borrower.

```r
table(loanDefaultData$MaritalStatus)
```

```
##
## Divorced  Married   Single
##     1176     1126     1198
```

- HasCoSigner (Binary/Categorical): Whether the loan has a co-signer.

```r
table(loanDefaultData$HasCoSigner)
```

```
##
##   No  Yes
## 1744 1756
```

Model Fitting (12 points)

Here I utilize AIC scoring to determine the best model. I chose CreditScore and Age as my specfic interest due to my guess being those are the major influence predictors. I chose to step back with my model under the assumption that the model would incorporate most values in the data.

```r
fullmodel=glm(Default ~ ., data=loanDefaultData, family = "binomial")

basicmodel=glm(Default ~ CreditScore+Age, data=loanDefaultData, family = "binomial")
```

```r
step(fullmodel,
     scope=list(lower=basicmodel, upper=fullmodel),
     direction='backward', steps=25)
```

```
## Start:  AIC=2295.69
## Default ~ Age + Income + LoanAmount + CreditScore + MonthsEmployed +
##     NumCreditLines + InterestRate + LoanTerm + Education + EmploymentType +
##     MaritalStatus + HasCoSigner
##
##                  Df Deviance    AIC
## - Education       3   2262.9 2292.9
## - LoanTerm        1   2260.5 2294.5
## <none>                2259.7 2295.7
## - MaritalStatus   2   2265.3 2297.3
## - HasCoSigner     1   2267.8 2301.8
## - NumCreditLines  1   2268.3 2302.3
## - EmploymentType  3   2273.2 2303.2
## - LoanAmount      1   2275.1 2309.1
## - Income          1   2281.1 2315.1
## - MonthsEmployed  1   2292.4 2326.4
## - InterestRate    1   2317.2 2351.2
##
## Step:  AIC=2292.93
## Default ~ Age + Income + LoanAmount + CreditScore + MonthsEmployed +
##     NumCreditLines + InterestRate + LoanTerm + EmploymentType +
##     MaritalStatus + HasCoSigner
##
##                  Df Deviance    AIC
## - LoanTerm        1   2263.9 2291.9
## <none>                2262.9 2292.9
## - MaritalStatus   2   2268.7 2294.7
## - HasCoSigner     1   2271.3 2299.3
## - NumCreditLines  1   2271.4 2299.4
## - EmploymentType  3   2276.3 2300.3
## - LoanAmount      1   2278.3 2306.3
## - Income          1   2284.8 2312.8
## - MonthsEmployed  1   2296.3 2324.3
## - InterestRate    1   2319.7 2347.7
##
## Step:  AIC=2291.94
## Default ~ Age + Income + LoanAmount + CreditScore + MonthsEmployed +
##     NumCreditLines + InterestRate + EmploymentType + MaritalStatus +
##     HasCoSigner
##
##                  Df Deviance    AIC
## <none>                2263.9 2291.9
## - MaritalStatus   2   2269.5 2293.5
## - HasCoSigner     1   2272.2 2298.2
## - NumCreditLines  1   2272.5 2298.5
## - EmploymentType  3   2277.2 2299.2
## - LoanAmount      1   2279.4 2305.4
## - Income          1   2285.9 2311.9
## - MonthsEmployed  1   2297.3 2323.3
## - InterestRate    1   2320.7 2346.7
```

```
##
## Call:  glm(formula = Default ~ Age + Income + LoanAmount + CreditScore +
##     MonthsEmployed + NumCreditLines + InterestRate + EmploymentType +
##     MaritalStatus + HasCoSigner, family = "binomial", data = loanDefaultData)
##
## Coefficients:
##                 (Intercept)                          Age
##                  -6.007e-01                   -3.647e-02
##                      Income                   LoanAmount
##                  -6.639e-06                    3.090e-06
##                 CreditScore               MonthsEmployed
##                  -1.152e-03                   -9.383e-03
##              NumCreditLines                 InterestRate
##                   1.439e-01                    6.328e-02
##     EmploymentTypePart-time  EmploymentTypeSelf-employed
##                   2.677e-01                    4.337e-01
##    EmploymentTypeUnemployed         MaritalStatusMarried
##                   5.475e-01                   -3.047e-01
##         MaritalStatusSingle               HasCoSignerYes
##                  -4.632e-02                   -3.186e-01
##
## Degrees of Freedom: 3499 Total (i.e. Null);  3486 Residual
## Null Deviance:       2524
## Residual Deviance: 2264  AIC: 2292
```

Here, I decided to explore interaction terms with Age and CreditScore.

```r
current.model=glm(Default ~ Age + Income + LoanAmount + CreditScore + MonthsEmployed + NumCreditLines +

interactions.model=glm(Default ~ (CreditScore+Age)*(Income + LoanAmount + MonthsEmployed + NumCreditLin
```

```r
step(current.model,
     scope=list(lower=current.model, upper=interactions.model),
     direction='forward', steps=25)
```

```
## Start:  AIC=2291.94
## Default ~ Age + Income + LoanAmount + CreditScore + MonthsEmployed + 
##     NumCreditLines + InterestRate + EmploymentType + MaritalStatus + 
##     HasCoSigner
## 
##                               Df Deviance    AIC
## + Age:EmploymentType          3   2250.6 2284.6
## <none>                            2263.9 2291.9
## + Age:InterestRate            1   2262.6 2292.6
## + CreditScore:EmploymentType  3   2258.8 2292.8
## + Age:LoanAmount              1   2263.2 2293.2
## + Age:Income                  1   2263.5 2293.5
## + CreditScore:Income          1   2263.7 2293.7
## + CreditScore:HasCoSigner     1   2263.7 2293.7
## + CreditScore:NumCreditLines  1   2263.8 2293.8
## + CreditScore:MonthsEmployed  1   2263.8 2293.8
## + CreditScore:InterestRate    1   2263.8 2293.8
## + Age:NumCreditLines          1   2263.8 2293.8
## + Age:MonthsEmployed          1   2263.9 2293.9
## + Age:HasCoSigner             1   2263.9 2293.9
## + CreditScore:LoanAmount      1   2263.9 2293.9
## + CreditScore:MaritalStatus   2   2262.9 2294.9
## + Age:MaritalStatus           2   2263.5 2295.5
## 
## Step:  AIC=2284.61
## Default ~ Age + Income + LoanAmount + CreditScore + MonthsEmployed + 
##     NumCreditLines + InterestRate + EmploymentType + MaritalStatus + 
##     HasCoSigner + Age:EmploymentType
## 
##                               Df Deviance    AIC
## <none>                            2250.6 2284.6
## + CreditScore:EmploymentType  3   2245.3 2285.3
## + Age:LoanAmount              1   2249.7 2285.7
## + Age:InterestRate            1   2249.7 2285.7
## + Age:Income                  1   2250.1 2286.1
## + CreditScore:HasCoSigner     1   2250.4 2286.4
## + Age:HasCoSigner             1   2250.4 2286.4
## + CreditScore:NumCreditLines  1   2250.4 2286.4
## + CreditScore:Income          1   2250.4 2286.4
## + CreditScore:InterestRate    1   2250.5 2286.5
## + Age:MonthsEmployed          1   2250.5 2286.5
## + Age:NumCreditLines          1   2250.5 2286.5
## + CreditScore:MonthsEmployed  1   2250.6 2286.6
## + CreditScore:LoanAmount      1   2250.6 2286.6
## + CreditScore:MaritalStatus   2   2249.6 2287.6
## + Age:MaritalStatus           2   2250.2 2288.2
```

```
## 
## Call:  glm(formula = Default ~ Age + Income + LoanAmount + CreditScore + 
##     MonthsEmployed + NumCreditLines + InterestRate + EmploymentType + 
##     MaritalStatus + HasCoSigner + Age:EmploymentType, family = "binomial", 
##     data = loanDefaultData)
## 
## Coefficients:
```

```
##                      (Intercept)                              Age
##                       -1.569e+00                       -1.164e-02
##                           Income                       LoanAmount
##                       -6.725e-06                        3.101e-06
##                      CreditScore                   MonthsEmployed
##                       -1.206e-03                       -9.319e-03
##                   NumCreditLines                     InterestRate
##                        1.436e-01                        6.382e-02
##          EmploymentTypePart-time      EmploymentTypeSelf-employed
##                        1.223e+00                        1.878e+00
##         EmploymentTypeUnemployed             MaritalStatusMarried
##                        1.926e+00                       -3.012e-01
##              MaritalStatusSingle                   HasCoSignerYes
##                       -5.061e-02                       -3.322e-01
##      Age:EmploymentTypePart-time  Age:EmploymentTypeSelf-employed
##                       -2.371e-02                       -3.678e-02
##     Age:EmploymentTypeUnemployed
##                       -3.467e-02
##
## Degrees of Freedom: 3499 Total (i.e. Null);  3483 Residual
## Null Deviance:       2524
## Residual Deviance: 2251  AIC: 2285
```

I assessed that moving forward would be more efficient thinking much interactions wouldn't be made which concluded with an interaction between Age and EmploymentType.

```
current.model=glm(formula = Default ~ Age + Income + LoanAmount + CreditScore +
    MonthsEmployed + NumCreditLines + InterestRate + EmploymentType +
    MaritalStatus + HasCoSigner + Age:EmploymentType, family = "binomial",
    data = loanDefaultData)
```

I assess the current model with the significance of the terms

```
summary(current.model)
```

```
##
## Call:
## glm(formula = Default ~ Age + Income + LoanAmount + CreditScore +
##     MonthsEmployed + NumCreditLines + InterestRate + EmploymentType +
##     MaritalStatus + HasCoSigner + Age:EmploymentType, family = "binomial",
##     data = loanDefaultData)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.569e+00  4.862e-01  -3.226 0.001256 **
## Age                             -1.164e-02  8.058e-03  -1.445 0.148559
## Income                          -6.725e-06  1.436e-06  -4.682 2.85e-06 ***
## LoanAmount                       3.101e-06  7.924e-07   3.914 9.09e-05 ***
## CreditScore                     -1.206e-03  3.521e-04  -3.426 0.000613 ***
## MonthsEmployed                  -9.319e-03  1.652e-03  -5.643 1.67e-08 ***
## NumCreditLines                   1.436e-01  4.955e-02   2.898 0.003756 **
## InterestRate                     6.382e-02  8.637e-03   7.390 1.47e-13 ***
## EmploymentTypePart-time          1.223e+00  4.806e-01   2.544 0.010955 *
## EmploymentTypeSelf-employed      1.878e+00  4.728e-01   3.971 7.15e-05 ***
## EmploymentTypeUnemployed         1.926e+00  4.755e-01   4.051 5.10e-05 ***
## MaritalStatusMarried            -3.012e-01  1.382e-01  -2.179 0.029320 *
## MaritalStatusSingle             -5.061e-02  1.308e-01  -0.387 0.698832
## HasCoSignerYes                  -3.322e-01  1.116e-01  -2.977 0.002910 **
## Age:EmploymentTypePart-time     -2.371e-02  1.120e-02  -2.117 0.034235 *
## Age:EmploymentTypeSelf-employed -3.678e-02  1.123e-02  -3.276 0.001054 **
## Age:EmploymentTypeUnemployed    -3.467e-02  1.110e-02  -3.124 0.001783 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2524.3  on 3499  degrees of freedom
## Residual deviance: 2250.6  on 3483  degrees of freedom
## AIC: 2284.6
##
## Number of Fisher Scoring iterations: 5
```

I conclude that Age and MaritalStatusSingle had no significance, but have decided to keep them in my final model due to my Age:EmploymentType interaction and the significance for MaritalStatusMarried in the MaritalStatus variable.

```
final.model = glm(formula = Default ~ Age + Income + LoanAmount + CreditScore +
    MonthsEmployed + NumCreditLines + InterestRate + EmploymentType +
    MaritalStatus + HasCoSigner + Age:EmploymentType, family = "binomial",
    data = loanDefaultData)

summary(final.model)
```

```
##
## Call:
## glm(formula = Default ~ Age + Income + LoanAmount + CreditScore +
##     MonthsEmployed + NumCreditLines + InterestRate + EmploymentType +
##     MaritalStatus + HasCoSigner + Age:EmploymentType, family = "binomial",
##     data = loanDefaultData)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     -1.569e+00  4.862e-01  -3.226 0.001256 **
## Age                             -1.164e-02  8.058e-03  -1.445 0.148559
## Income                          -6.725e-06  1.436e-06  -4.682 2.85e-06 ***
## LoanAmount                       3.101e-06  7.924e-07   3.914 9.09e-05 ***
## CreditScore                     -1.206e-03  3.521e-04  -3.426 0.000613 ***
## MonthsEmployed                  -9.319e-03  1.652e-03  -5.643 1.67e-08 ***
## NumCreditLines                   1.436e-01  4.955e-02   2.898 0.003756 **
## InterestRate                     6.382e-02  8.637e-03   7.390 1.47e-13 ***
## EmploymentTypePart-time          1.223e+00  4.806e-01   2.544 0.010955 *
## EmploymentTypeSelf-employed      1.878e+00  4.728e-01   3.971 7.15e-05 ***
## EmploymentTypeUnemployed         1.926e+00  4.755e-01   4.051 5.10e-05 ***
## MaritalStatusMarried            -3.012e-01  1.382e-01  -2.179 0.029320 *
## MaritalStatusSingle             -5.061e-02  1.308e-01  -0.387 0.698832
## HasCoSignerYes                  -3.322e-01  1.116e-01  -2.977 0.002910 **
## Age:EmploymentTypePart-time     -2.371e-02  1.120e-02  -2.117 0.034235 *
## Age:EmploymentTypeSelf-employed -3.678e-02  1.123e-02  -3.276 0.001054 **
## Age:EmploymentTypeUnemployed    -3.467e-02  1.110e-02  -3.124 0.001783 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 2524.3  on 3499  degrees of freedom
## Residual deviance: 2250.6  on 3483  degrees of freedom
## AIC: 2284.6
##
## Number of Fisher Scoring iterations: 5
```

Model summary (8 points)

log-odds(Default)= 0.2083477 +0.9884259*(Age) +0.9999933*(Income) +1.0000031*(LoanAmount) +0.9987945*(CreditScore) +0.9907240*(MonthsEmployed) +1.1544144*(NumCreditLines) +1.0659012*(InterestRate) +3.3967974*(EmploymentTypePart-time) +6.5377859*(EmploymentTypeSelf-employed) +6.8644685*(EmploymentTypeUnemployed) +0.7399581*(MaritalStatusMarried) +0.9506527*(MaritalStatusSingle) +0.7173221*(HasCoSignerYes) +0.9765642*(Age:EmploymentTypePart-time) +0.9638929*(Age:EmploymentTypeSelf-employed) +0.9659194*(Age:EmploymentTypeUnemployed)

The final resulting model found was:

The effect on the odds of each of the terms are listed below.

```
summary
```

```
## function (object, ...)
## UseMethod("summary")
## <bytecode: 0x000002e04b2098f8>
## <environment: namespace:base>
```

```
oddsimpact = exp(final.model$coefficients)
oddsimpact
```

```
##                     (Intercept)                             Age
##                       0.2083477                       0.9884259
##                          Income                      LoanAmount
##                       0.9999933                       1.0000031
##                     CreditScore                  MonthsEmployed
##                       0.9987945                       0.9907240
##                  NumCreditLines                    InterestRate
##                       1.1544144                       1.0659012
##         EmploymentTypePart-time     EmploymentTypeSelf-employed
##                       3.3967974                       6.5377859
##        EmploymentTypeUnemployed            MaritalStatusMarried
##                       6.8644685                       0.7399581
##             MaritalStatusSingle                  HasCoSignerYes
##                       0.9506527                       0.7173221
##     Age:EmploymentTypePart-time Age:EmploymentTypeSelf-employed
##                       0.9765642                       0.9638929
##    Age:EmploymentTypeUnemployed
##                       0.9659194
```

Note: The baseline individual in this study is Employed Full-time and Divorced

Conclusion (8 points)

Key Predictors

- Age: Every year increase in age decreases the odds by 1.2%
- Employment Type: Being unemployed, you are 6.9 times more likely to default compared to Full-time. Part-time and Self-Employed are 3.4 and 6.5 times more likely to default than Full-time respectively. This makes sense since Full-time employed borrowers have an active source income meaning less risk compared to others.
- CreditScore: Each point increase in credit decreases the odds by 0.12%. This makes sense because the purpose of credit is to assess creditworthiness. However, this change isn't super significant which goes to show that this is just one of the many factors that influences loan default.
- MonthsEmployed: Each additional month of employment decreases the odds by 0.93%. This also makes sense because the longer you had been employed shows the financial stability and consistent earned income.
- NumCreditLines: Each additional credit lines increases the odds by 15.4%. This makes sense because multiple credit lines can lead to more debt.
- InterestRate: Each percentage increase in the interest rate increases the odds by 6.6%. This checks out because higher interest rates leads to higher payments making loans harder to manage.
- MaritalStatus: Being married decreases the odds of default by 26% and being single decreases it by 4.9% compared to being divorced. This makes sense because married couples can support each others finances. My reasoning for why a single person may have a decreased odd compared to a divorced person would be because of divorce settlements and significant changes when it comes to finances but then again, this is a less significant decrease than it is being married.
- HasCoSigner: Having a cosigner decreases the odds of default by 28.3%. This makes sense in why the odds decrease similar to marital status due to having another person financially support you.
- Age:EmploymentType: This interaction indicates for every unit increase for age as Part-time is a 2.3% decrease, 3.6% for Self-employed, and 3.4% for unemployed in odds. I found this to be quite confusing but I think there seems to be more of an impact in age than the baseline because it is more acceptable accruing finances at an older age while not working.

All in all, the predictors that positively impacted the default odds were the number of credit lines open, interest rates, being self-employed, being unemployed, and being part-time. Predictors that negatively impacted the default odds were your age, credit score, months employed, being married, having a cosigner, and the interactions between age and employment status (part-time, self-employed, and unemployed).