Econometrics Project

Luis Green

2023-05-09

The 'credit' dataset consists of 1000 observations of 17 variables related to credit applications. The variables include information on the applicant's checking balance, loan duration, credit history, the purpose of the loan, amount requested, savings balance, employment duration, percent of income, years at residence, age, other credit, housing, existing loans count, job, dependents, phone, and whether the applicant defaulted on the loan or not.

The variables checking balance, credit history, purpose, savings balance, employment duration, other credit, housing, job, and phone are categorical variables, while the variables months_loan_duration, amount, percent_of_income, years_at_residence, age, existing_loans_count, and dependents are numerical variables. The variable default is the target variable, indicating whether the applicant defaulted on the loan or not.

```
library(readr)
credit <- read.csv("credit.csv", stringsAsFactors = TRUE)
str(credit)
```

```
## 'data.frame':    1000 obs. of  17 variables:
##  $ checking_balance    : Factor w/ 4 levels "< 0 DM","> 200 DM",..: 1 3 4
1 1 4 4 3 4 3 ...
##  $ months_loan_duration: int  6 48 12 42 24 36 24 36 12 30 ...
##  $ credit_history      : Factor w/ 5 levels "critical","good",..: 1 2 1 2
4 2 2 2 2 1 ...
##  $ purpose             : Factor w/ 6 levels "business","car",..: 5 5 4 5 2
4 5 2 5 2 ...
##  $ amount              : int  1169 5951 2096 7882 4870 9055 2835 6948 3059
5234 ...
##  $ savings_balance     : Factor w/ 5 levels "< 100 DM","> 1000 DM",..: 5 1
1 1 1 5 4 1 2 1 ...
##  $ employment_duration : Factor w/ 5 levels "< 1 year","> 7 years",..: 2 3
4 4 3 3 2 3 4 5 ...
##  $ percent_of_income   : int  4 2 2 2 3 2 3 2 2 4 ...
##  $ years_at_residence  : int  4 2 3 4 4 4 4 2 4 2 ...
##  $ age                 : int  67 22 49 45 53 35 53 35 61 28 ...
##  $ other_credit        : Factor w/ 3 levels "bank","none",..: 2 2 2 2 2 2
2 2 2 2 ...
##  $ housing             : Factor w/ 3 levels "other","own",..: 2 2 2 1 1 1
2 3 2 2 ...
##  $ existing_loans_count: int  2 1 1 1 2 1 1 1 1 2 ...
##  $ job                 : Factor w/ 4 levels "management","skilled",..: 2 2
4 2 2 4 2 1 4 1 ...
```

```
##  $ dependents          : int  1 1 2 2 2 2 1 1 1 1 ...
##  $ phone               : Factor w/ 2 levels "no","yes": 2 1 1 1 1 2 1 2 1
1 ...
##  $ default             : Factor w/ 2 levels "no","yes": 1 2 1 1 2 1 1 1 1
2 ...
```

```
unique(credit$credit_history)
```

```
## [1] critical  good      poor      perfect   very good
## Levels: critical good perfect poor very good
```

```
unique(credit$employment_duration)
```

```
## [1] > 7 years   1 - 4 years 4 - 7 years unemployed  < 1 year
## Levels: < 1 year > 7 years 1 - 4 years 4 - 7 years unemployed
```

```
unique(credit$existing_loans_count)
```

```
## [1] 2 1 3 4
```

```
unique(credit$job)
```

```
## [1] skilled    unskilled  management unemployed
## Levels: management skilled unemployed unskilled
```

I selected this dataset with the objective of developing a linear model that can effectively predict default rates. My decision to choose a financial dataset stems from my proficiency in this domain, which has enabled me to carefully select the variables that have the highest predictive potential.

Research Question

How well can the variables in the "credit" data set explain the probability of loan default among credit applicants? This question aims to investigate the relationship between the predictive variables and the dependent variable, which in this case is "default" in predicting the likelihood of loan default. I will be using a linear model to examine the strength and direction of the relationships between the variables and the extent to which they can explain the variance in default rates.

Part 1) of the analysis will be breaking the data down. Understanding the variables and creating visuals to better interpret the data.

```
colnames(credit)
```

```
##  [1] "checking_balance"     "months_loan_duration" "credit_history"
##  [4] "purpose"              "amount"               "savings_balance"
##  [7] "employment_duration"  "percent_of_income"    "years_at_residence"
## [10] "age"                  "other_credit"         "housing"
## [13] "existing_loans_count" "job"                  "dependents"
## [16] "phone"                "default"
```

Let's take a look at the two characteristics of the applicant. The results are a frequency table that shows the number of times each value appears in the credit checking_balance column of the credit data frame. The different values of credit$checking_balance are displayed as the row names of the table, which are < 0 DM, > 200 DM, 1 - 200 DM, and unknown.

The counts of each value appear in the corresponding columns of the table, where < 0 DM appears 274 times, > 200 DM appears 63 times, 1 - 200 DM appears 269 times, and unknown appears 394 times. We can see that the majority of the data falls into the unknown category, which suggests that there may be missing data or that this category was intentionally used to represent a certain group. We can also see that the number of values above 200 DM is relatively small compared to the other categories.

```
# Show the different levels, they add up to 1000, Dutch Marks
table(credit$checking_balance)

##
##     < 0 DM   > 200 DM 1 - 200 DM    unknown
##        274          63         269         394

table(credit$savings_balance)

##
##      < 100 DM     > 1000 DM  100 - 500 DM 500 - 1000 DM       unknown
##           603            48           103            63           183
```

Let's take a look at the two characteristics of the loan, which include "months_loan_duration" and "amount". The minimum value of months_loan_duration is 4, which is the smallest value in the dataset. The first quartile (25th percentile) is 12, which means that 25% of the loans have a duration of 12 months or less. The median (50th percentile) is 18, which means that 50% of the loans have a duration of 18 months or less. The mean duration of the loans is 20.9 months, which is the arithmetic average of all the durations. The third quartile (75th percentile) is 24, which means that 75% of the loans have a duration of 24 months or less. The maximum value of months_loan_duration is 72, which is the longest duration in the dataset.

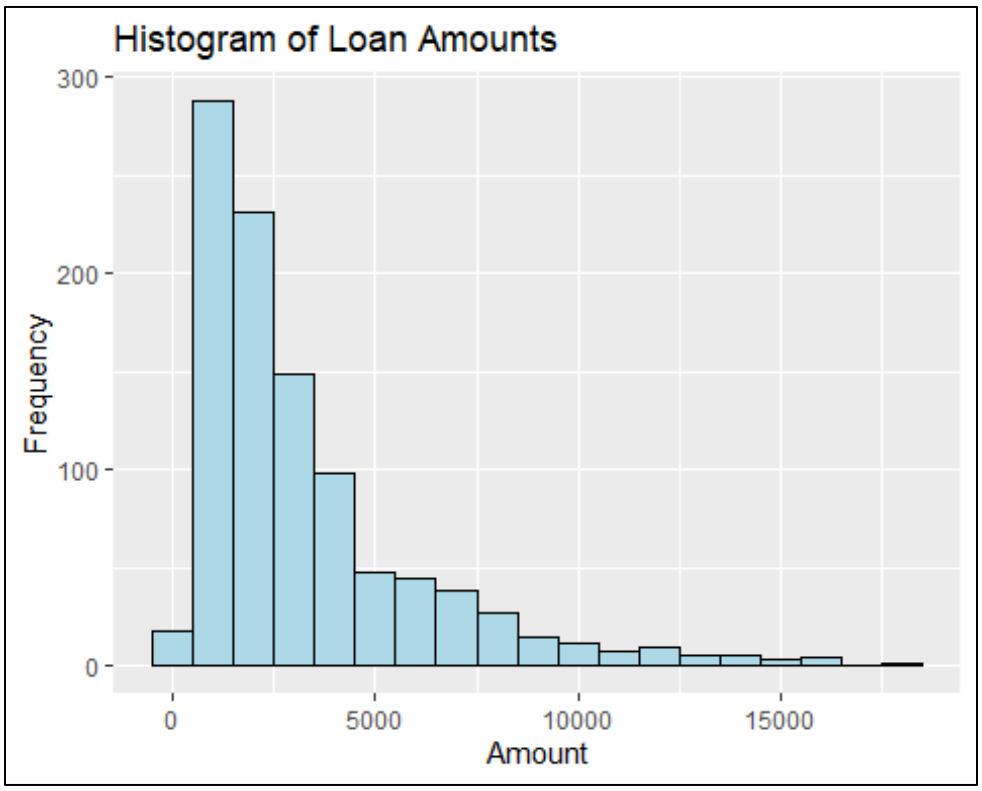
```
summary(credit$months_loan_duration)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     4.0    12.0    18.0    20.9    24.0    72.0
```

Look at the class variable where 1 is no and yes is 2. No means Good that they did not defaulted, and yes means Bad or that they defaulted on the loan. There appear to 700 that did not default out of 1000 or 70%, while the remaining 30% defaulted.

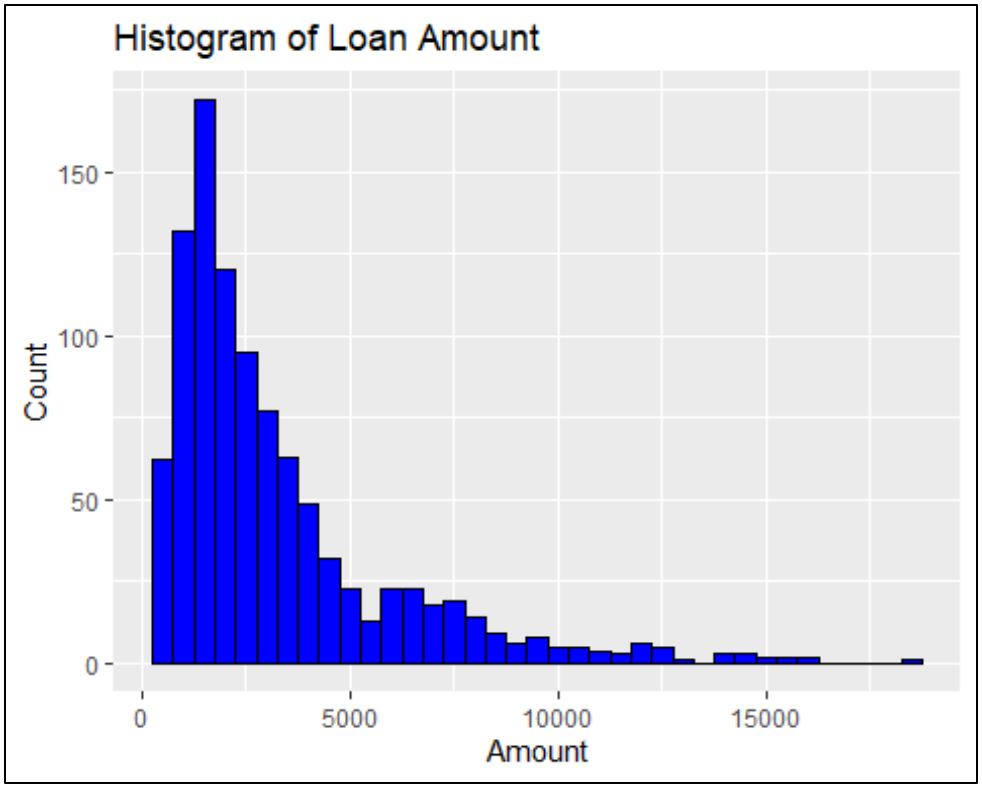
Part 2) Visual Portion of the Exploratory Analysis

```
# Loading the library to create the visuals
library(ggplot2)
ggplot(data = credit, aes(x = amount)) +
  geom_histogram(binwidth = 1000, color = "black", fill = "lightblue") +
  labs(title = "Histogram of Loan Amounts", x = "Amount", y = "Frequency")
```

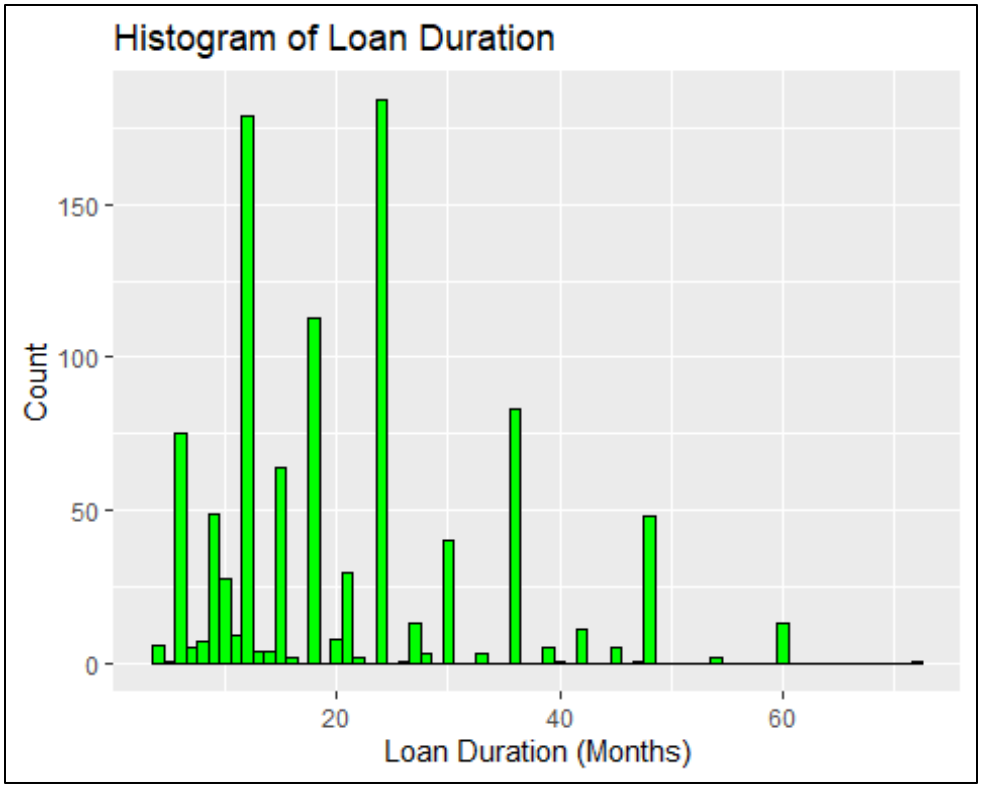


Data is skewed toward the left indicating that there is a higher frequency of loans that are in the range of $0 to $5,000.

```
library(ggplot2)
ggplot(data = credit, aes(x = amount)) +
  geom_histogram(binwidth = 500, fill = "blue", color = "black") +
  labs(x = "Amount", y = "Count", title = "Histogram of Loan Amount")
```

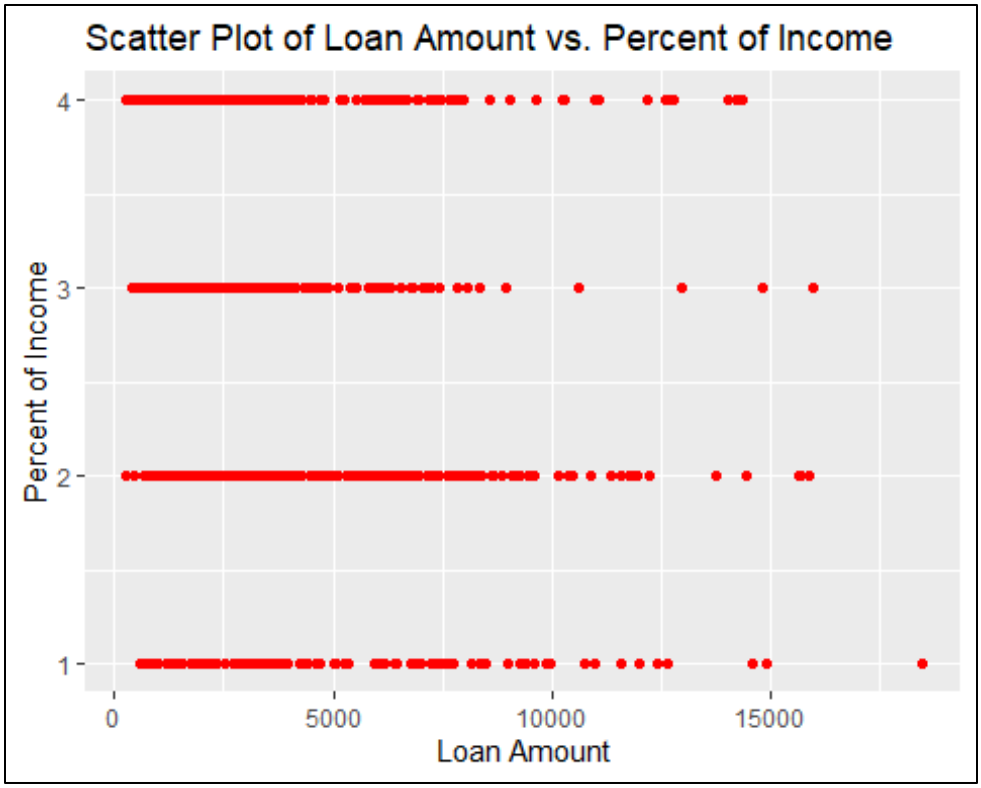



```
ggplot(data = credit, aes(x = months_loan_duration)) +
  geom_histogram(binwidth = 1, fill = "green", color = "black") +
  labs(x = "Loan Duration (Months)", y = "Count", title = "Histogram of Loan
Duration")
```



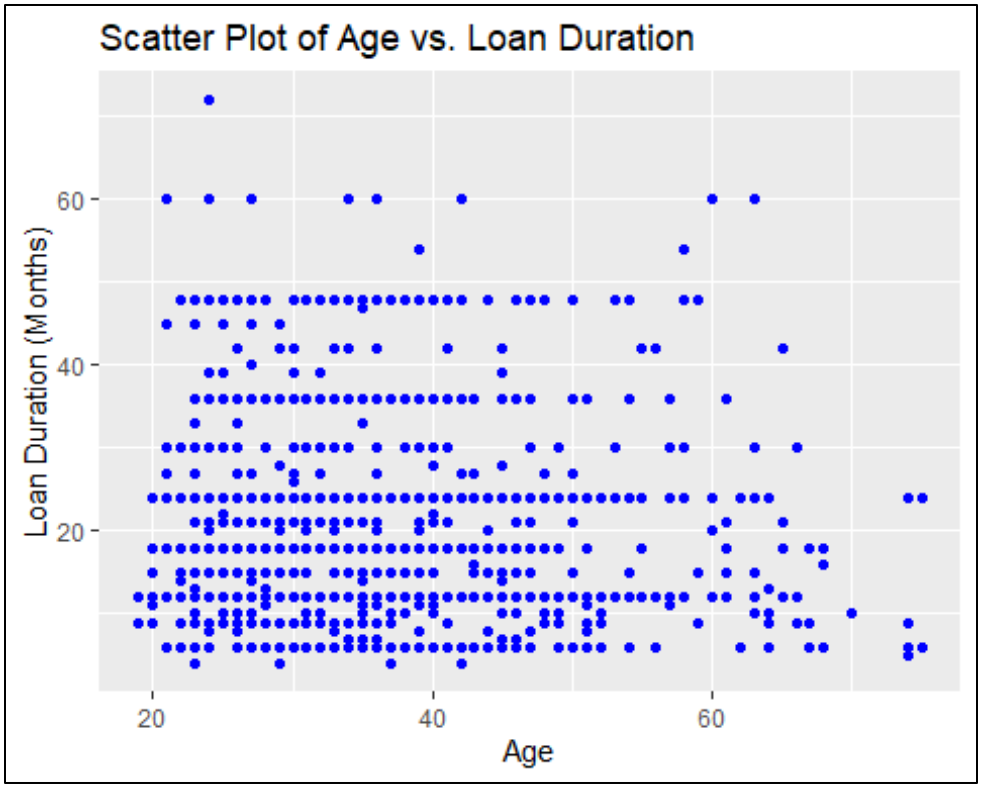
There appears to be a variation in the loan duration. The majority of loans last between 1 month and 30 months. This makes sense since most of the loans are less than $5,000 in value. Larger loans would require a longer time frame to repay it back.

```
ggplot(data = credit, aes(x = amount, y = percent_of_income)) +
  geom_point(color = "red") +
  labs(x = "Loan Amount", y = "Percent of Income", title = "Scatter Plot of
Loan Amount vs. Percent of Income")
```



This was surprising. Weather loan as a percent of income were between 1% and 4% the loan amount was relatively in the same range.

```
ggplot(data = credit, aes(x = age, y = months_loan_duration)) +
  geom_point(color = "blue") +
  labs(x = "Age", y = "Loan Duration (Months)", title = "Scatter Plot of Age
vs. Loan Duration")
```



This scatter plot was an interesting one given the relationship between age and loan amount. There does not appear to be a consistent pattern. However, the bulk of the data is between 20 months and age 50.

Part 3) I will be creating a linear model.

First, we need to identify the response or dependent variable and the predictor or independent variables. In this case, the dependent variable is the "default" variable, which indicates whether a person defaulted on a credit loan (yes or no).

We can use the other variables as predictors, but we need to make sure they are appropriate for a linear model. For example, the "checking_balance" and "savings_balance" variables are categorical, so we need to convert them into numerical variables using dummy encoding or one-hot encoding. Similarly, the "credit_history", "purpose", "employment_duration", "other_credit", "housing", and "job" variables are also categorical, so we need to convert them into numerical variables using the same method.

Once we have converted the categorical variables into numerical variables, we can use them as predictors in a linear model. We can use the "months_loan_duration", "amount", "percent_of_income", "years_at_residence", "age", "existing_loans_count", and "dependents" variables as numerical predictors directly.

This model will fit a linear regression line to predict the default variable based on the amount and months_loan_duration variables in the credit dataset. The lm() function specifies the formula for the linear model, and the data argument tells R to use the credit

dataset. The summary() function will display the summary of the linear model, including the coefficients and their corresponding p-values.

```
credit <- na.omit(credit)
# Turning the factor variables into numeric variables to use them for the
linear model
credit$amount <- as.numeric(credit$amount)
credit$months_loan_duration <- as.numeric(credit$months_loan_duration)
credit$default <- as.numeric(credit$default)
credit$credit_history <- as.numeric(credit$credit_history)
credit$job <- as.numeric(credit$job)
credit$percent_of_income <- as.numeric(credit$percent_of_income)
credit$employment_duration <- as.numeric(credit$employment_duration)
credit$years_at_residence <- as.numeric(credit$years_at_residence)

credit$credit_history <- as.integer(credit$credit_history)
credit$employment_duration <- as.integer(credit$employment_duration)
credit$existing_loans_count <- as.integer(credit$existing_loans_count)
credit$job <- as.integer(credit$job)
```

Correlation and Multicolliniarity

```
correlations <- cor(credit[c("default", "amount", "months_loan_duration",
"credit_history", "percent_of_income", "years_at_residence",
"credit_history", "employment_duration", "existing_loans_count", "job")])
print(correlations)

##                           default      amount months_loan_duration
## default                1.000000000  0.15473864           0.21492667
## amount                 0.154738641  1.00000000           0.62498420
## months_loan_duration   0.214926665  0.62498420           1.00000000
## credit_history         0.193729621  0.10959783           0.14823931
## percent_of_income      0.072403937 -0.27131570           0.07474882
## years_at_residence     0.002967159  0.02892632           0.03406720
## credit_history.1       0.193729621  0.10959783           0.14823931
## employment_duration   -0.056655440  0.09303002           0.05580028
## existing_loans_count  -0.045732489  0.02079455          -0.01128360
## job                   -0.032755605 -0.26113918          -0.21543812
##                       credit_history percent_of_income years_at_residence
## default                   0.193729621        0.07240394       0.0029671588
## amount                    0.109597829       -0.27131570       0.0289263231
## months_loan_duration      0.148239313        0.07474882       0.0340672016
## credit_history            1.000000000       -0.01698587      -0.0318050303
## percent_of_income        -0.016985871        1.00000000       0.0493023708
## years_at_residence       -0.031805030        0.04930237       1.0000000000
## credit_history.1          1.000000000       -0.01698587      -0.0318050303
## employment_duration       0.015005199       -0.04998331      -0.0018835869
## existing_loans_count     -0.177466609        0.02166874       0.0896252326
## job                      -0.009164894       -0.07809000       0.0004503643
##                       credit_history.1 employment_duration
existing_loans_count
```

```
## default                    0.193729621        -0.056655440         -
0.045732489
## amount                     0.109597829         0.093030016
0.020794552
## months_loan_duration       0.148239313         0.055800276         -
0.011283597
## credit_history             1.000000000         0.015005199         -
0.177466609
## percent_of_income         -0.016985871        -0.049983315
0.021668743
## years_at_residence        -0.031805030        -0.001883587
0.089625233
## credit_history.1           1.000000000         0.015005199         -
0.177466609
## employment_duration        0.015005199         1.000000000
0.034827691
## existing_loans_count      -0.177466609         0.034827691
1.000000000
## job                       -0.009164894        -0.081560347
0.004544204
##                                    job
## default                 -0.0327556046
## amount                  -0.2611391802
## months_loan_duration    -0.2154381167
## credit_history          -0.0091648936
## percent_of_income       -0.0780899970
## years_at_residence       0.0004503643
## credit_history.1        -0.0091648936
## employment_duration     -0.0815603467
## existing_loans_count     0.0045442040
## job                      1.0000000000
```

The variables with the highest correlation to the default variable in descending order are:

- months_loan_duration (0.214926665)
- credit_history (0.193729621)
- amount (0.154738641)
- credit_history.1 (0.193729621)
- percent_of_income (0.072403937)
- years_at_residence (0.002967159)
- employment_duration (-0.056655440)
- existing_loans_count (-0.045732489)
- job (-0.032755605)

```
credit_num <- credit[, c("default", "amount", "months_loan_duration",
"credit_history", "percent_of_income", "years_at_residence",
"credit_history", "employment_duration", "existing_loans_count", "job")]
cor(credit_num)
```

```
##                            default      amount months_loan_duration
## default                1.000000000  0.15473864           0.21492667
## amount                 0.154738641  1.00000000           0.62498420
## months_loan_duration   0.214926665  0.62498420           1.00000000
## credit_history         0.193729621  0.10959783           0.14823931
## percent_of_income      0.072403937 -0.27131570           0.07474882
## years_at_residence     0.002967159  0.02892632           0.03406720
## credit_history.1       0.193729621  0.10959783           0.14823931
## employment_duration   -0.056655440  0.09303002           0.05580028
## existing_loans_count  -0.045732489  0.02079455          -0.01128360
## job                   -0.032755605 -0.26113918          -0.21543812
##                       credit_history percent_of_income years_at_residence
## default                   0.193729621        0.07240394       0.0029671588
## amount                    0.109597829       -0.27131570       0.0289263231
## months_loan_duration      0.148239313        0.07474882       0.0340672016
## credit_history            1.000000000       -0.01698587      -0.0318050303
## percent_of_income        -0.016985871        1.00000000       0.0493023708
## years_at_residence       -0.031805030        0.04930237       1.0000000000
## credit_history.1          1.000000000       -0.01698587      -0.0318050303
## employment_duration       0.015005199       -0.04998331      -0.0018835869
## existing_loans_count     -0.177466609        0.02166874       0.0896252326
## job                      -0.009164894       -0.07809000       0.0004503643
##                       credit_history.1 employment_duration
existing_loans_count
## default                     0.193729621        -0.056655440         -
0.045732489
## amount                      0.109597829         0.093030016
0.020794552
## months_loan_duration        0.148239313         0.055800276         -
0.011283597
## credit_history              1.000000000         0.015005199         -
0.177466609
## percent_of_income          -0.016985871        -0.049983315
0.021668743
## years_at_residence         -0.031805030        -0.001883587
0.089625233
## credit_history.1            1.000000000         0.015005199         -
0.177466609
## employment_duration         0.015005199         1.000000000
0.034827691
## existing_loans_count       -0.177466609         0.034827691
1.000000000
## job                        -0.009164894        -0.081560347
0.004544204
##                                    job
## default                     -0.0327556046
## amount                      -0.2611391802
## months_loan_duration        -0.2154381167
## credit_history              -0.0091648936
## percent_of_income           -0.0780899970
```

```
## years_at_residence    0.0004503643
## credit_history.1     -0.0091648936
## employment_duration  -0.0815603467
## existing_loans_count  0.0045442040
## job                   1.0000000000
```

The diagonal elements are all equal to 1 because they represent the correlation between a variable and itself, which is always perfect. The non-diagonal elements show the correlation between pairs of variables. For example, the correlation between “default” and “amount” is -0.1547, indicating a weak negative correlation.

The strongest correlation is between “amount” and “months_loan_duration” with a value of 0.625, indicating a moderate positive correlation. There is also a weak negative correlation between “default” and “percent_of_income” (-0.0724) and a weak positive correlation between “percent_of_income” and “years_at_residence” (0.0493).

Overall, there is no significant multicollinearity issue, as all the correlation coefficients are below 0.7, and no two variables are highly correlated with each other.

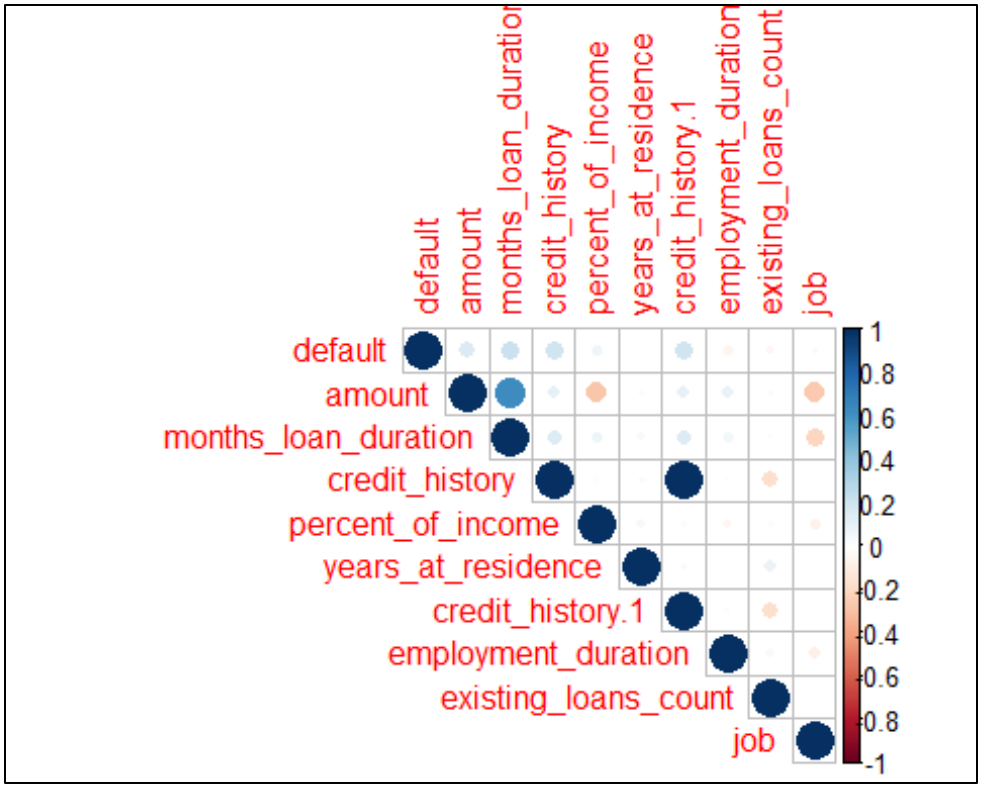
Next, I am displaying a circle plot of the correlation matrix, where the size and color of the circles indicate the strength and direction of the correlation between the variables.

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.2.3
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(credit_num), type = "upper", method = "circle")
```



The Variance Inflation Factor (VIF) measures the degree of multicollinearity among predictor variables in a regression model. Typically, a VIF of 1 indicates no correlation between the predictor variables and other variables, while a VIF greater than 1 indicates some degree of correlation. A VIF greater than 5 or 10 indicates high multicollinearity. If any of the VIF values are high, you may need to remove one or more of the correlated variables from the model to avoid issues with multicollinearity.

```
library(car)

## Loading required package: carData

## Warning: package 'carData' was built under R version 4.2.3

vif(lm(default ~ amount + months_loan_duration + credit_history +
percent_of_income + years_at_residence + credit_history + employment_duration
+ existing_loans_count + job, data = credit))

##               amount months_loan_duration       credit_history
##             2.073469             1.857116             1.058407
##    percent_of_income   years_at_residence  employment_duration
##             1.241022             1.012959             1.015087
## existing_loans_count                  job
##             1.044770             1.107998
```

In this case, the VIF values for the variables amount, months_loan_duration, percent_of_income, and years_at_residence are all below 2. This indicates that there is not a

significant degree of multicollinearity among these variables. Therefore, we can say that these variables are not highly correlated with each other in the model.

Fit the linear model

```
model <- lm(default ~ amount + months_loan_duration + credit_history +
percent_of_income + years_at_residence + credit_history + employment_duration
+ existing_loans_count + job, data = credit)

# view the summary of the model
summary(model)

##
## Call:
## lm(formula = default ~ amount + months_loan_duration + credit_history +
##     percent_of_income + years_at_residence + credit_history +
##     employment_duration + existing_loans_count + job, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7385 -0.2999 -0.1992  0.5014  0.9409
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.620e-01  9.230e-02  10.422  < 2e-16 ***
## amount                1.402e-05  7.107e-06   1.973 0.048799 *
## months_loan_duration  5.277e-03  1.574e-03   3.351 0.000834 ***
## credit_history        7.073e-02  1.353e-02   5.228 2.09e-07 ***
## percent_of_income     3.567e-02  1.387e-02   2.571 0.010283 *
## years_at_residence   -8.390e-04  1.270e-02  -0.066 0.947356
## employment_duration  -2.766e-02  1.243e-02  -2.225 0.026277 *
## existing_loans_count -1.301e-02  2.465e-02  -0.528 0.597764
## job                   1.090e-02  1.549e-02   0.703 0.482000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4403 on 991 degrees of freedom
## Multiple R-squared:  0.08501,    Adjusted R-squared:  0.07763
## F-statistic: 11.51 on 8 and 991 DF,  p-value: 1.024e-15
```

The model shows that the default variable is significantly associated with the amount, months_loan_duration, credit_history, and percent_of_income variables. The years_at_residence variable is not a significant predictor of the default variable. The model has a low R-squared value, indicating that the predictors explain only a small proportion of the variation in the default variable. The model's overall significance level is very low, indicating that the predictors as a group have a significant effect on the default variable.

I am going to use stepwise selection to select the best variables. This will perform forward and backward stepwise seleciton to find the best variables to include in the model.

```
model_1 <- lm(default ~ amount + months_loan_duration + credit_history +
percent_of_income + years_at_residence + credit_history + employment_duration
+ existing_loans_count + job, data = credit)
step(model_1)
```

```
## Start:  AIC=-1631.49
## default ~ amount + months_loan_duration + credit_history +
percent_of_income +
##     years_at_residence + credit_history + employment_duration +
##     existing_loans_count + job
##
##                         Df Sum of Sq    RSS     AIC
## - years_at_residence     1    0.0008 192.15 -1633.5
## - existing_loans_count   1    0.0540 192.20 -1633.2
## - job                    1    0.0959 192.24 -1633.0
## <none>                               192.15 -1631.5
## - amount                 1    0.7546 192.90 -1629.6
## - employment_duration    1    0.9603 193.11 -1628.5
## - percent_of_income      1    1.2817 193.43 -1626.8
## - months_loan_duration   1    2.1779 194.32 -1622.2
## - credit_history         1    5.2995 197.45 -1606.3
##
## Step:  AIC=-1633.49
## default ~ amount + months_loan_duration + credit_history +
percent_of_income +
##     employment_duration + existing_loans_count + job
##
##                         Df Sum of Sq    RSS     AIC
## - existing_loans_count   1    0.0555 192.20 -1635.2
## - job                    1    0.0956 192.24 -1635.0
## <none>                               192.15 -1633.5
## - amount                 1    0.7538 192.90 -1631.6
## - employment_duration    1    0.9600 193.11 -1630.5
## - percent_of_income      1    1.2819 193.43 -1628.8
## - months_loan_duration   1    2.1773 194.32 -1624.2
## - credit_history         1    5.3048 197.45 -1608.3
##
## Step:  AIC=-1635.2
## default ~ amount + months_loan_duration + credit_history +
percent_of_income +
##     employment_duration + job
##
##                         Df Sum of Sq    RSS     AIC
## - job                    1    0.0926 192.30 -1636.7
## <none>                               192.20 -1635.2
## - amount                 1    0.7342 192.94 -1633.4
## - employment_duration    1    0.9784 193.18 -1632.1
## - percent_of_income      1    1.2622 193.47 -1630.7
## - months_loan_duration   1    2.1954 194.40 -1625.8
## - credit_history         1    5.6823 197.89 -1608.1
```

```
##
## Step:  AIC=-1636.72
## default ~ amount + months_loan_duration + credit_history +
percent_of_income +
##     employment_duration
##
##                         Df Sum of Sq    RSS     AIC
## <none>                               192.30 -1636.7
## - amount                 1    0.6583 192.95 -1635.3
## - employment_duration    1    1.0219 193.32 -1633.4
## - percent_of_income      1    1.1892 193.49 -1632.5
## - months_loan_duration   1    2.1778 194.47 -1627.5
## - credit_history         1    5.7224 198.02 -1609.4

##
## Call:
## lm(formula = default ~ amount + months_loan_duration + credit_history +
##     percent_of_income + employment_duration, data = credit)
##
## Coefficients:
##          (Intercept)                amount  months_loan_duration
##            9.744e-01             1.279e-05             5.274e-03
##       credit_history     percent_of_income   employment_duration
##            7.228e-02             3.390e-02            -2.846e-02
```

The code is performing a stepwise regression analysis, which is a method of selecting the best subset of predictor variables for a linear regression model.

The initial model (model_1) includes 9 predictor variables (amount, months_loan_duration, credit_history, percent_of_income, years_at_residence, employment_duration, existing_loans_count, and job) to predict the response variable default.

The output shows the results of each step of the stepwise regression process. At each step, the algorithm removes one predictor variable to find the best subset of predictor variables that minimizes the Akaike information criterion (AIC).

The results show the change in AIC as each variable is removed and the resulting model with the lowest AIC. The final model has only five predictor variables (amount, months_loan_duration, credit_history, percent_of_income, and employment_duration) and a lower AIC than the initial model. The coefficients for each predictor variable in the final model are also shown.

The final model has the lowest AIC (Akaike Information Criterion) value of -1636.7, compared to the other models with higher AIC values. Therefore, this model is considered to be the best among the ones tested using stepwise selection.

```
model_2 <- lm(default ~ amount + months_loan_duration + credit_history +
percent_of_income + employment_duration, data = credit)
summary(model_2);summary(model_1)
```

```
##
## Call:
## lm(formula = default ~ amount + months_loan_duration + credit_history +
##     percent_of_income + employment_duration, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7378 -0.2982 -0.1982  0.4993  0.9299
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.744e-01  6.333e-02  15.385  < 2e-16 ***
## amount                1.279e-05  6.931e-06   1.845 0.065384 .
## months_loan_duration  5.274e-03  1.572e-03   3.355 0.000823 ***
## credit_history        7.228e-02  1.329e-02   5.439 6.76e-08 ***
## percent_of_income     3.390e-02  1.367e-02   2.479 0.013327 *
## employment_duration  -2.846e-02  1.238e-02  -2.298 0.021747 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4398 on 994 degrees of freedom
## Multiple R-squared:  0.0843, Adjusted R-squared:  0.0797
## F-statistic:  18.3 on 5 and 994 DF,  p-value: < 2.2e-16

##
## Call:
## lm(formula = default ~ amount + months_loan_duration + credit_history +
##     percent_of_income + years_at_residence + credit_history +
##     employment_duration + existing_loans_count + job, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7385 -0.2999 -0.1992  0.5014  0.9409
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.620e-01  9.230e-02  10.422  < 2e-16 ***
## amount                1.402e-05  7.107e-06   1.973 0.048799 *
## months_loan_duration  5.277e-03  1.574e-03   3.351 0.000834 ***
## credit_history        7.073e-02  1.353e-02   5.228 2.09e-07 ***
## percent_of_income     3.567e-02  1.387e-02   2.571 0.010283 *
## years_at_residence   -8.390e-04  1.270e-02  -0.066 0.947356
## employment_duration  -2.766e-02  1.243e-02  -2.225 0.026277 *
## existing_loans_count -1.301e-02  2.465e-02  -0.528 0.597764
## job                   1.090e-02  1.549e-02   0.703 0.482000
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4403 on 991 degrees of freedom
```

```
## Multiple R-squared:  0.08501,    Adjusted R-squared:  0.07763
## F-statistic: 11.51 on 8 and 991 DF,  p-value: 1.024e-15
```

Based on the summary statistics, it appears that model_2 is slightly better than model_1. The Adjusted R-squared value of model_2 is 0.0797, while that of model_1 is 0.07763. Additionally, model_2 has a higher F-statistic value than model_1, which is an indication of a better fit. However, it is important to note that the difference in performance between the two models is quite small and may not be practically significant.

Research Question

How well can the variables in the "credit" data set explain the probability of loan default among credit applicants?

```
summary(model_2)

##
## Call:
## lm(formula = default ~ amount + months_loan_duration + credit_history +
##     percent_of_income + employment_duration, data = credit)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.7378 -0.2982 -0.1982  0.4993  0.9299
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)           9.744e-01  6.333e-02  15.385  < 2e-16 ***
## amount                1.279e-05  6.931e-06   1.845 0.065384 .
## months_loan_duration  5.274e-03  1.572e-03   3.355 0.000823 ***
## credit_history        7.228e-02  1.329e-02   5.439 6.76e-08 ***
## percent_of_income     3.390e-02  1.367e-02   2.479 0.013327 *
## employment_duration  -2.846e-02  1.238e-02  -2.298 0.021747 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4398 on 994 degrees of freedom
## Multiple R-squared:  0.0843, Adjusted R-squared:  0.0797
## F-statistic:  18.3 on 5 and 994 DF,  p-value: < 2.2e-16
```

Based on the linear regression analysis provided, the variables in the "credit" data set can explain the probability of loan default among credit applicants to some extent, but not very well. The multiple R-squared value of 0.0843 indicates that only about 8.4% of the variability in the default variable can be explained by the predictor variables (amount, months_loan_duration, credit_history, percent_of_income, and employment_duration) included in the model. This means that there may be other important factors that are not captured by these variables that also affect the probability of loan default.

However, some of the predictor variables do show significant effects on the default variable. Specifically, credit_history, months_loan_duration, percent_of_income, and

employment_duration all have significant coefficients at the 5% level or better, suggesting that they may be important predictors of loan default. The amount variable, on the other hand, only has a marginally significant coefficient at the 10% level, indicating that it may not be as important a predictor as the other variables in the model.

Other Questions

1) Which variables have a statistically significant relationship with the dependent variable "default" at a certain level of significance?

The variables "months_loan_duration", "credit_history", "percent_of_income", and "employment_duration" have a statistically significant relationship with the dependent variable "default" at a certain level of significance, as their p-values are less than 0.05.

2) How much does a unit increase in the "amount" variable affect the probability of default, holding all other variables constant?

Holding all other variables constant, a unit increase in the "amount" variable results in a 1.279e-05 increase in the odds of defaulting on the loan. This effect is marginally statistically significant as its p-value is 0.065.

3) What is the estimated change in the probability of default associated with a one-month increase in the "months_loan_duration" variable, while controlling for other variables?

Controlling for other variables, a one-month increase in the "months_loan_duration" variable is associated with a 0.005274 increase in the odds of defaulting on the loan.

4) How does the level of "credit_history" affect the probability of defaulting on a loan, while holding other variables constant?

Holding other variables constant, a one-unit increase in the "credit_history" variable is associated with a 0.07228 decrease in the odds of defaulting on the loan. In other words, better credit history is associated with a lower probability of default.

5) What is the expected change in the probability of default for every 1% increase in "percent_of_income", holding other variables constant?

Keeping other variables constant, every 1% increase in "percent_of_income" is associated with a 0.0339 increase in the odds of defaulting on the loan.