**Student: Seif Kungulio**

**Date: 02/09/2025**

**Subject: Project 4**

**Class: DSCI 502**

**Section: 01W**

**Instructor: Sean Yang**

**File Name: Project4\_Kungulio\_Seif.docx**

1. Read the dataset in loan.csv into R. Call the loaded data, loan. Make sure that you have the directory set to the correct location for the data.

>

> ## 1. Read the dataset in loan.csv into R. Call the loaded data, loan.

> ### Make sure that you have the directory set to the correct location

> ### for the data.

>

> # Set the working directory to the correct location for the dataset.

> setwd("C:/PROJECTS/Maryville/DSCI 502/Week4")

>

> # Import necessary libraries

> # (Optional) Load any necessary libraries, e.g., dplyr, tidyr if needed.

>

> # Load the data from loan.csv

> loan <- read.csv("loan.csv", stringsAsFactors = TRUE)

>

> # Display the dimensions (rows and columns) of the dataframe

> dim(loan) # Shows the number of rows and columns in the dataset.

[1] 10000 11

>

A screenshot of a computer

Description automatically generated

1. Which variables (columns) are continuous/numerical variables? Which columns are factors (categorical variables)?

>

> ## 2. Which variables (columns) are continuous/numerical variables? Which

> ### columns are factors (categorical variables)?

>

> # Identify variable types

> str(loan) # Displays the structure of the dataset including variable types.

'data.frame': 10000 obs. of 11 variables:

$ id : int 1077501 1077430 1077175 1076863 1075358 1075269 1069639 1072053 1071795 1071570 ...

$ loan\_amnt : int 5000 2500 2400 10000 3000 5000 7000 3000 5600 5375 ...

$ term : Factor w/ 2 levels " 36 months"," 60 months": 1 2 1 1 2 1 2 1 2 2 ...

$ int\_rate : num 10.6 15.3 16 13.5 12.7 ...

$ installment : num 162.9 59.8 84.3 339.3 67.8 ...

$ grade : Factor w/ 7 levels "A","B","C","D",..: 2 3 3 3 2 1 3 5 6 2 ...

$ emp\_length : Factor w/ 12 levels "< 1 year","1 year",..: 3 1 3 3 2 5 10 11 6 1 ...

$ home\_ownership : Factor w/ 3 levels "MORTGAGE","OWN",..: 3 3 3 3 3 3 3 3 2 3 ...

$ annual\_inc : num 24000 30000 12252 49200 80000 ...

$ verification\_status: Factor w/ 3 levels "Not Verified",..: 3 2 1 2 2 2 1 2 2 3 ...

$ loan\_status : Factor w/ 7 levels "Charged Off",..: 4 1 4 4 2 4 2 4 1 1 ...

>

> # Identify continuous (numerical) and categorical (factor) variables

> # Checks which variables are numerical.

> numerical\_vars <- sapply(loan, is.numeric)

>

> # Checks which variables are categorical.

> categorical\_vars <- sapply(loan, is.factor)

>

> # Extract the names of numerical variables

> numerical\_columns <- names(numerical\_vars[numerical\_vars])

> cat("Numerical Variables:\n", numerical\_columns, "\n\n")

Numerical Variables:

id loan\_amnt int\_rate installment annual\_inc

>

> # Extract the names of categorical variables

> categorical\_columns <- names(categorical\_vars[categorical\_vars])

> cat("Categorical Variables:\n", categorical\_columns, "\n\n")

Categorical Variables:

term grade emp\_length home\_ownership verification\_status loan\_status

>

A computer screen shot of a program

Description automatically generated

1. Calculate the minimum, maximum, mean, median, standard deviation and three quartiles (25th, 50th and 75th percentiles) of loan\_amnt.

>

> ## 3. Calculate the minimum, maximum, mean, median, standard deviation and

> ### three quartiles (25th, 50th and 75th percentiles) of loan\_amnt.

>

> # Calculate and display the minimum value of loan\_amnt

> cat("Minimum of loan\_amnt:", min(loan$loan\_amnt, na.rm = TRUE), "\n")

Minimum of loan\_amnt: 1000

>

> # Calculate and display the maximum value of loan\_amnt

> cat("Maximum of loan\_amnt:", max(loan$loan\_amnt, na.rm = TRUE), "\n")

Maximum of loan\_amnt: 35000

>

> # Calculate and display the mean value of loan\_amnt

> cat("Mean of loan\_amnt:", mean(loan$loan\_amnt, na.rm = TRUE), "\n")

Mean of loan\_amnt: 12861.64

>

> # Calculate and display the median value of loan\_amnt

> cat("Median of loan\_amnt:", median(loan$loan\_amnt, na.rm = TRUE), "\n")

Median of loan\_amnt: 11200

>

> # Calculate and display the standard deviation of loan\_amnt

> cat("Standard deviation of loan\_amnt:", sd(loan$loan\_amnt, na.rm = TRUE), "\n")

Standard deviation of loan\_amnt: 8491.814

>

> # Calculate the quartiles of loan\_amnt

> percent <- quantile(loan$loan\_amnt, probs = c(0.25, 0.50, 0.75), na.rm = TRUE)

>

> # Display the 25th percentile of loan\_amnt

> cat("25% of loan\_amnt:", percent[1], "\n")

25% of loan\_amnt: 6000

>

> # Display the 50th percentile of loan\_amnt (median)

> cat("50% of loan\_amnt:", percent[2], "\n")

50% of loan\_amnt: 11200

>

> # Display the 75th percentile of loan\_amnt

> cat("75% of loan\_amnt:", percent[3], "\n")

75% of loan\_amnt: 17500

>

A screenshot of a computer program

Description automatically generated

1. Calculate the minimum, maximum, mean, median, standard deviation and three quartiles (25th, 50th and 75th percentiles) of int\_rate.

>

> ## 4. Calculate the minimum, maximum, mean, median, standard deviation and

> ### three quartiles (25th, 50th and 75th percentiles) of int\_rate.

>

> # Summary statistics for int\_rate

> summary(loan$int\_rate)

Min. 1st Qu. Median Mean 3rd Qu. Max.

5.42 8.90 12.42 12.43 15.27 24.11

>

> # Extract minimum value of int\_rate

> cat("Minimum of int\_rate:", summary(loan$int\_rate)["Min."], "\n")

Minimum of int\_rate: 5.42

>

> # Extract maximum value of int\_rate

> cat("Maximum of int\_rate:", summary(loan$int\_rate)["Max."], "\n")

Maximum of int\_rate: 24.11

>

> # Extract mean value of int\_rate

> cat("Mean of int\_rate:", summary(loan$int\_rate)["Mean"], "\n")

Mean of int\_rate: 12.42855

>

> # Extract median value of int\_rate

> cat("Median of int\_rate:", summary(loan$int\_rate)["Median"], "\n")

Median of int\_rate: 12.42

>

> # Calculate and display the standard deviation of int\_rate

> cat("Standard deviation of int\_rate:", sd(loan$int\_rate, na.rm = TRUE), "\n")

Standard deviation of int\_rate: 4.239117

>

> # Calculate the quartiles of int\_rate

> percentile <- quantile(loan$int\_rate, probs = c(0.25, 0.50, 0.75), na.rm = TRUE)

>

> # Display the 25th percentile of int\_rate

> cat("25% of int\_rate:", percentile["25%"], "\n")

25% of int\_rate: 8.9

>

> # Display the 50th percentile of int\_rate (median)

> cat("50% of int\_rate:", percentile["50%"], "\n")

50% of int\_rate: 12.42

>

> # Display the 75th percentile of int\_rate

> cat("75% of int\_rate:", percentile["75%"], "\n")

75% of int\_rate: 15.27

>

A computer screen shot of a black screen

Description automatically generated

1. Calculate the correlation coefficient of the two variables: int\_rate and installment. Do they have a strong relationship?

>

> ## 5. Calculate the correlation coefficient of the two variables: int\_rate

> ### and installment. Do they have a strong relationship?

>

> # Compute the correlation coefficient between int\_rate and installment

> correlation\_value <- cor(loan$int\_rate, loan$installment, use = "complete.obs")

>

> # Display the correlation coefficient

> cat("Correlation between int\_rate and installment:", correlation\_value, "\n")

Correlation between int\_rate and installment: 0.2819849

>

A screen shot of a computer code

Description automatically generated

The correlation coefficient between int\_rate and installment is 0.2819849, indicating a weak to moderate positive relationship. This suggests that while higher interest rates may lead to higher installment amounts, the connection is not strong. Since a strong correlation typically exceeds 0.7, this low value implies that other factors, such as loan amount and term length, likely have a greater influence on installment amounts.

1. Calculate the frequency table of term? What’s the mode of term variable?

>

> ## 6. Calculate the frequency table of term? What’s the mode of term variable?

>

> # Create a frequency table for the term variable

> term\_table <- table(loan$term)

>

> # Identify the mode of the term variable

> mode\_term <- names(term\_table[term\_table == max(term\_table)])

>

> # Print the frequency table

> print(term\_table)

36 months 60 months

6649 3351

>

> # Display the mode of the term variable

> cat("Mode of term:", mode\_term, "\n")

Mode of term: 36 months

>

A computer screen shot of a black screen

Description automatically generated

1. Calculate the proportion table of loan\_status? What’s the mode of loan\_status variable?

>

> ## 7. Calculate the proportion table of loan\_status? What’s the mode of

> ### loan\_status variable?

>

> # Compute the proportion table for loan\_status

> loan\_status\_table <- prop.table(table(loan$loan\_status))

>

> # Identify the mode of the loan\_status variable

> mode\_loan\_status <- names(loan\_status\_table[loan\_status\_table ==

+ max(loan\_status\_table)])

>

> # Print the proportion table

> print(loan\_status\_table)

Charged Off Current Default Fully Paid

0.1517 0.0956 0.0002 0.7487

In Grace Period Late (16-30 days) Late (31-120 days)

0.0008 0.0006 0.0024

>

> # Display the mode of the loan\_status variable

> cat("Mode of loan\_status:", mode\_loan\_status, "\n")

Mode of loan\_status: Fully Paid

>

A computer screen shot of a computer code

Description automatically generated

1. Calculate the cross table of term and loan\_status. Then produce proportions by row and column respectively.

>

> ## 8. Calculate the cross table of term and loan\_status. Then produce

> ### proportions by row and column respectively.

>

> # Compute the cross table of term and loan\_status

> table\_term\_status <- table(loan$term, loan$loan\_status)

>

> # Compute and print row proportions

> cross\_table\_row <- prop.table(table\_term\_status, margin = 1)

> cat("Row proportions of term and loan\_status:\n")

Row proportions of term and loan\_status:

> print(cross\_table\_row)

Charged Off Current Default Fully Paid In Grace Period

36 months 0.1134005114 0.0000000000 0.0000000000 0.8865994886 0.0000000000

60 months 0.2276932259 0.2852879737 0.0005968368 0.4750820651 0.0023873471

Late (16-30 days) Late (31-120 days)

36 months 0.0000000000 0.0000000000

60 months 0.0017905103 0.0071620412

>

> # Compute and print column proportions

> cross\_table\_col <- prop.table(table\_term\_status, margin = 2)

> cat("Column proportions of term and loan\_status:\n")

Column proportions of term and loan\_status:

> print(cross\_table\_col)

Charged Off Current Default Fully Paid In Grace Period

36 months 0.4970336 0.0000000 0.0000000 0.7873648 0.0000000

60 months 0.5029664 1.0000000 1.0000000 0.2126352 1.0000000

Late (16-30 days) Late (31-120 days)

36 months 0.0000000 0.0000000

60 months 1.0000000 1.0000000

>

A computer screen with text and numbers

Description automatically generated

1. The data is stored in the data frame, loan. Please summarize all the variables using one command.

>

> ## 9. The data is stored in the data frame, loan. Please summarize all the

> ### variables using one command.

>

> # Generate summary statistics for all variables in the dataset

> summary(loan)

id loan\_amnt term int\_rate

Min. : 458165 Min. : 1000 36 months:6649 Min. : 5.42

1st Qu.: 878178 1st Qu.: 6000 60 months:3351 1st Qu.: 8.90

Median : 987925 Median :11200 Median :12.42

Mean : 963545 Mean :12862 Mean :12.43

3rd Qu.:1033696 3rd Qu.:17500 3rd Qu.:15.27

Max. :1077501 Max. :35000 Max. :24.11

installment grade emp\_length home\_ownership annual\_inc

Min. : 22.24 A:2765 10+ years:2548 MORTGAGE:4612 Min. : 6000

1st Qu.: 193.58 B:3113 2 years : 987 OWN : 748 1st Qu.: 42000

Median : 322.25 C:1825 3 years : 904 RENT :4640 Median : 60000

Mean : 363.82 D:1220 < 1 year : 900 Mean : 70267

3rd Qu.: 480.33 E: 718 4 years : 861 3rd Qu.: 84500

Max. :1288.10 F: 292 5 years : 855 Max. :1782000

G: 67 (Other) :2945

verification\_status loan\_status

Not Verified :3050 Charged Off :1517

Source Verified:3069 Current : 956

Verified :3881 Default : 2

Fully Paid :7487

In Grace Period : 8

Late (16-30 days) : 6

Late (31-120 days): 24

>

A screenshot of a computer

Description automatically generated