Module Code	:	CT127-3-2-PFDA (Programming for Data Analysis)
Coursework Title	:	Credit Risk Classification
Lecturer Name	:	Farhana Illiani Binti Hassan
Hand Out Date	:	6 October 2024
Hand In Date	:	9 December 2024
Intake	:	APD2F2409IT(IOT), APU2F2409IT(IOT), APU2F2409IT(BIS)
Group No.	:	13
Name, TP Number	:	1. Chong Zhi Yue (TP067869) 2. Eleanor Permata Fry (TP072606) 3. Lee Jun Keat (TP067856) 4. Tham Wing Hein (TP67080)

Table of Contents

1.0 Introduction

The bank sector act as an important role in assessing credit risks to ensure financial stability and minimize losses. The case study of this assignment mainly focuses on implementing the application of data analytics techniques to identify the credit risks based on customer demographics and their credit status. Inside this dataset, it contains various attributes such as gender, marital status, employment status, loan purpose, and instalment's duration, providing a comprehensive view of customer profiles and their credit-related activities.

The objective of this assignment is to conduct an analysis of the dataset to identify the factors between high-risk and low-risk individuals. This will involve applying data exploration, manipulation, transformation, and visualization techniques to pre-process the data, and reflect our important insights to develop models for credit risk classification. Advanced analytical concepts will also be incorporated to enhance the effectiveness of the analysis.

1.1 Data Description

In this assignment, the team conducted data analysis on the "3. credit_risk_classification.csv," which contains approximately 6000 records, where each row corresponds to a unique customer's record. The dataset includes various key variables:

A. Customer Demographics:
    - Gender
    - Marital Status
    - Employment Status
    - Age
B. Credit Behaviour:
    - Credit Amount
    - Duration of Instalment Commitment
    - Purpose of Loan
    - Credit History
    - Account Type
C. Credit Risk:

- Credit Class

To determine the main variables impacting credit risk, a dataset with numerous numerical and categorical elements will be analysed. Before doing the analysis of each objective, the data will be pre-processed, cleaned, and explored to find patterns and relationships that can help with predictive modelling.

1.2 Assumptions

For this assignment, each group member chose one goal to watch and evaluate to discover how it affected the classification of credit risk. We have done analyse the missing information during data cleaning and handle categorical and continuous variable properly. We make our predictions that the features featured are relevant to the categorization of credit risk and that it can predict the connections between customer demographics, credit behaviour, and credit risk classification.

1.3 Hypothesis & Objectives

Hypothesis:

1. Classification of credit risk is influenced by customer demographics and credit behaviours (e.g., credit amount, loan term, and job status).
2. Higher credit amounts and longer loan durations are associated with a higher probability of being categorized as a “Bad” credit risk.

Objectives:

1. To investigate the relationship between customer demographics and credit risk classification.
2. To analyse the effect of loan-related factors on credit risk classification.
3. To assess the role of financial commitments on credit risk classification.
4. To explore the impact of customer credit history on credit risk classification.

2.0 Data Preparation

2.1 Install Required Packages and Load The Libraries, Then Import Data

```
#Install packages and load libraries needed
install.packages("dplyr")
library(dplyr)
install.packages("ggplot2")
library(ggplot2)
install.packages("ggridges")
library(ggridges)
install.packages("igraph")
library(igraph)
install.packages("plotrix")
library(plotrix)
install.packages("vcd")
library(vcd)
```

Figure: Install Packages and Load Libraries

The code loads several packages that are commonly used for data manipulation, visualization and analysis. 'dplyr' was being used to perform data manipulations such as filtering, selecting and summarizing data frames. 'ggplot2' was being used to create visualizations such as plots. 'ggridges' was being used to create ridge plots. 'igraph' was being used for network analysis and graph manipulation which can be used to create, manipulate, and visualize graphs. 'plotrix' was being used to do various plotting functions like 3d plots and others. 'vcd' is being used for visual display of categorical data.

```
AData <- read.csv("C:\\Users\\Itsuki\\Documents\\Degree Year 2 Semester 1\\Programming For Data Analysis\\Assignment\\3. credit_risk_classification.csv", sep=",", header=TRUE)
AData
#Table View
View(AData)
```

Figure 1: Import Dataset

The code above reads the .csv file named "3. Credit_risk_classification.csv" from the mentioned directory into an R data frame called 'AData' using 'read.csv' function. Then the data frame is being displayed into table form using 'View' function.

2.2 Data Pre-processing and Cleaning

```
#View Missing Data
colSums(is.na(AData))
sum(complete.cases(AData))
which(is.na(AData))
sum(is.na(AData))
```

Figure 2: View Missing Data in Dataset

- colsums(is.na(AData)) was used to showing how many missing values exist in each columns.
- sum(complete.cases(AData)) will tell how many rows in AData have no missing data.

- which(is.na(AData)) gives the exact positions of the missing values in rows and columns.
- sum(is.na(AData)) show how many total missing values that are across the entire dataset.

```
#Remove Unused Column
AData<- AData %>% select(-X)

#Remove Duplicated Data
AData <- AData %>% distinct()
sum(duplicated(AData))
```

Figure 3: Remove Unused Column & Duplicated Data

- "-X" here means unnecessary or irrelevant columns. "-X" here will be removed.
- distinct() is a function from "dplyr" package that removes duplicated rows based on all columns.
- sum() will then check for any remaining duplicates and returns the total count of duplicated rows.

```
#Remove Error In age
AData <- AData[AData$age >= 18 & AData$age <= 100, ]
AData$age <- as.integer(AData$age)


#Remove Error In duration
AData$duration <- round(AData$duration)
AData <- AData[AData$duration >= 4 & AData$duration <= 72, ]


#Remove Error In installment commitment
AData$installment_commitment <- round(AData$installment_commitment)
AData <- AData[AData$installment_commitment >= 1 & AData$installment_commitment <= 4, ]

#Remove Error In residence since
AData$residence_since <- round(AData$residence_since)
AData <- AData[AData$residence_since >= 1 & AData$residence_since <= 4, ]


#Remove Error In existing credits
AData$existing_credits <- round(AData$existing_credits)
AData <- AData[AData$existing_credits >= 1 & AData$existing_credits <= 4, ]


#Remove Error In number dependents
AData$num_dependents <- round(AData$num_dependents)
AData <- AData[AData$num_dependents >= 1 & AData$num_dependents <= 2, ]
```

Figure 4: Remove Errors in Specified Column

The codes above filters out invalid or out-of-range values for "age", "duration", "installment_commitment", "residence_since", "existing_credits", and "num_dependents". It then rounds continuous numeric variables to the nearest appropriate integer ensuring the dataset only contains valid and logical entries for the columns.

```r
#Filling other_payment_plans
AData$other_payment_plans[AData$other_payment_plans == ""] <- NA
mode_value <- names(sort(table(AData$other_payment_plans), decreasing = TRUE))[1]
print(mode_value) # shows stores
AData$other_payment_plans[is.na(AData$other_payment_plans)] <- mode_value
```

Figure 5: Filling Blank Cell and Replace with Mode Value

This code was being used to find and fill in the missing values (NA) in "other_payment_plans" with the most frequent value (mode).

```r
#Change Label Name
AData <- AData%>%
  mutate(
    personal_status = case_when(
      grepl("male", personal_status) & grepl("single", personal_status) ~ "male single",
      grepl("female", personal_status) & grepl("div|dep|mar", personal_status) ~ "female divorced/dependent/married",
      grepl("male", personal_status) & grepl("div|sep", personal_status) ~ "male divorced/separated",
      grepl("male", personal_status) & grepl("mar|wid", personal_status) ~ "male married/widowed",
      TRUE ~ personal_status
    )
  )

AData <- AData %>%
  mutate(
    job = case_when(
      job == "skilled" ~ "skilled",
      job == "unskilled resident" ~ "unskilled, resident",
      job == "high qualif/self emp/mgmt" ~ "high qualification, self-employed, management",
      job == "unemp/unskilled non res" ~ "unemployed, unskilled, non-resident",
      TRUE ~ job
    )
  )

AData <- AData %>%
  mutate(
    employment = case_when(
      employment == ">=7" ~ "7 or more years",
      employment == "1<=X<4" ~ "1 to 3 years",
      employment == "4<=X<7" ~ "4 to 6 years",
      employment == "<1" ~ "less than 1 year",
      employment == "unemployed" ~ "unemployed",
      TRUE ~ "other"
    )
  )
```

Figure 6: Change the Data Label Names in The Columns

```r
#add column "gender"
AData$gender <- ifelse(grepl("female", AData$personal_status, ignore.case =TRUE), "female",
                       ifelse(grepl("male", AData$personal_status, ignore.case = TRUE), "male", NA))
```

Figure 7: Adding extra column called "gender"

2.3 Data Validation

```r
#Conversion
AData$age <- as.numeric(AData$age)
AData$credit_amount <- as.numeric(AData$credit_amount)
AData$installment_commitment <- as.numeric(AData$installment_commitment)
AData$duration <- as.numeric(AData$duration)
AData$residence_since <- as.numeric(AData$residence_since)
AData$existing_credits <- as.numeric(AData$existing_credits)
AData$num_dependents <- as.numeric(AData$num_dependents)

#credit history conversion
AData$credit_history <- gsub("delayed previously", "delays", AData$credit_history)
AData$credit_history <- gsub("critical/other existing credit", "critical accounts", AData$credit_history)
AData$credit_history <- gsub("no credits/all paid", "all paid",AData$credit_history)
```

Figure 8: Converting Data Types

```r
#Validate Dataset
#Recheck structure and summary
str(AData)
summary(AData)
```

```
'data.frame':   1400 obs. of  21 variables:
 $ checking_status     : chr  "<0" "0<=X<200" "no checking" "<0" ...
 $ duration            : num  6 48 12 42 24 36 24 36 12 30 ...
 $ credit_history      : chr  "critical accounts" "existing paid" "critical accounts" "existing paid" ...
 $ purpose             : chr  "radio/tv" "radio/tv" "education" "furniture/equipment" ...
 $ credit_amount       : num  1169 5951 2096 7882 4870 ...
 $ savings_status      : chr  "no known savings" "<100" "<100" "<100" ...
 $ employment          : chr  ">=7" "1<=X<4" "4<=X<7" "4<=X<7" ...
 $ installment_commitment: num  4 2 2 2 3 2 3 2 2 4 ...
 $ personal_status     : chr  "male single" "female divorced/dependent/married" "male single" "male single" ...
 $ other_parties       : chr  "none" "none" "none" "guarantor" ...
 $ residence_since     : num  4 2 3 4 4 4 4 2 4 2 ...
 $ property_magnitude  : chr  "real estate" "real estate" "real estate" "life insurance" ...
 $ age                 : num  67 22 49 45 53 35 53 35 61 28 ...
 $ other_payment_plans : chr  "stores" "stores" "stores" "stores" ...
 $ housing             : chr  "own" "own" "own" "for free" ...
 $ existing_credits    : num  2 1 1 1 2 1 1 1 1 2 ...
 $ job                 : chr  "skilled" "skilled" "unskilled, resident" "skilled" ...
 $ num_dependents      : num  1 1 2 2 2 2 1 1 1 1 ...
 $ own_telephone       : chr  "yes" "none" "none" "none" ...
 $ foreign_worker      : chr  "yes" "yes" "yes" "yes" ...
 $ class               : chr  "good" "bad" "good" "good" ...
```

Figures 9 & 10: Output of Corrected Data Types

```
#Save Cleaned Dataset
write.csv(AData, "C:\\Users\\Dev\\Downloads\\Assignment\\Assignment\\3. cleaned_credit_risk_classification.csv", row.names = FALSE)

#Read Cleaned Dataset
AData <- read.csv("C:\\Users\\Dev\\Downloads\\Assignment\\Assignment\\3. cleaned_credit_risk_classification.csv")
View(AData)
```

Figure 11: Save the processed data and read again (refresh)

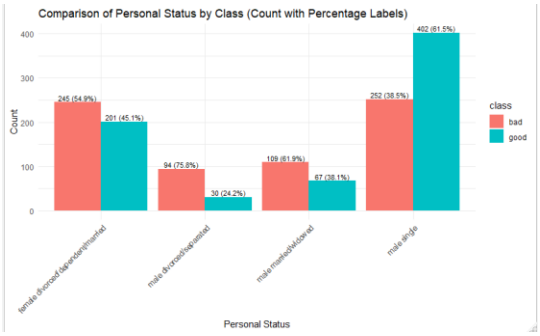
3.0 Data Analysis

Objective 3.1: To Investigate The Relationship Between Customer Demographics And Credit Risk Classification. Lee Jun Keat (TP067856)

3.1.1: Is There A Significant Relationship Between Personal Status And Credit Risk Classification?

The first step is to combine all the count within the personal status and classify it with the class to determine the credit risk score.

```
#Analysis 1
personal_status_frequency <- AData %>%
  count(personal_status) %>%
  as_tibble()
personal_status_frequency
# Display the tibble
personal_status_frequency_tibble
status_class_counts <- as.data.frame(table(AData$personal_status, AData$class))
colnames(status_class_counts) <- c("personal_status", "class", "count")

status_class_counts <- status_class_counts %>%
  group_by(personal_status) %>%
  mutate(percentage = count / sum(count) * 100)

ggplot(status_class_counts, aes(x = personal_status, y = count, fill = class)) +
  geom_bar(stat = "identity", position = "dodge") +
  geom_text(aes(label = paste0(count, " (", round(percentage, 1), "%)")),
            position = position_dodge(width = 0.9),
            vjust = -0.3,
            size = 3) +
  labs(title = "Comparison of Personal Status by Class (Count with Percentage Labels)",
       x = "Personal Status", y = "Count") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Figures 12 & 13: Bar Chart of Analysis 1

```
> personal_status_frequency
# A tibble: 4 × 2
  personal_status                 n
  <chr>                       <int>
1 female divorced/dependent/married   446
2 male divorced/separated             124
3 male married/widowed                176
4 male single                         654
```

Figure 14: Tibble table for Analysis 1

The analysis took place with 1400 individuals in the credit risk classification with the current personal status. Each bar displays the count of “bad” and “good” credit class for every category

Category:

Female divorced/dependent/married: There are 446 individuals inside this category, there are 201 (45.1%) individuals got a “good” class for the credit risk and 245 (54.9%) individuals are classified as “bad” class.

Male divorced/separated: Among 124 individuals, it shows that only 30 (24.2%) individuals got a "good" class and most of them which are 94 (75.8%) individuals got a "bad" class. The significance shows that it skews into the bad credit class in this category.

Male married/widowed: The result returns that only 67 (38.1%) individuals got "good" credit class during the analysis. While 109 of out 176 individuals is got a "bad" credit score. Same as the previous category, it shows that majority of them are towards higher financial risk.

Male singles: This category got the most count of individual among 4 of the categories. Which are 654 out of 1400 individuals. It shows that majority of them shows that they had a "good" credit class which are 402 (61.5%) individuals, and the rest 252 (38.5%) individual are "bad" credit class.

```
> chi_table <- table(AData$personal_status, AData$class)
>
> cstest <- chisq.test(chi_table)
>
> cstest

        Pearson's Chi-squared test

data:  chi_table
X-squared = 81.799, df = 3, p-value < 2.2e-16
```

Figure 15: Chi Squared Statistic for Analysis 1

In the chi square table, it shows that the chi squared statistic is on the value of 81.799 with the degree of freedom = 3 and a p value <2.2e-16.

3.1.2: Do Job Skills Significantly Influence Credit Risk Classification?

```
#Analysis 2
library(dplyr)
job_frequency <- AData %>%
  count(job,class) %>%
  as_tibble()
job_frequency

library(plotrix)

skilled_data <- job_frequency %>% filter(job == "skilled")
labels <- paste0(skilled_data$class, " (", skilled_data$n, ") ", "(",round(skilled_data$n / sum(skilled_data$n) * 100, 1), "%)")
pie3D(skilled_data$n,
      labels = labels,
      main = "Credit Class Distribution for Job Skills: skilled",
      explode = 0.15,
      col = c("blue", "red"),
      labelcex = 1)


unskilled_data <- job_frequency %>% filter(job == "unskilled, resident")
labels <- paste0(unskilled_data$class, " (", unskilled_data$n, ") ", "(",round(unskilled_data$n / sum(unskilled_data$n) * 100, 1), "%)")
pie3D(unskilled_data$n,
      labels = labels,
      main = "Credit Class Distribution for Job Skills: Unskilled/resident",
      explode = 0.15,
      col = c("blue", "red"),
      labelcex = 1)

highskilled_data <- job_frequency %>% filter(job == "high qualification, self-employed, management")
labels <- paste0(highskilled_data$class, " (", highskilled_data$n, ") ", "(",round(highskilled_data$n / sum(highskilled_data$n) * 100, 1), "%)")
pie3D(skilled_data$n,
      labels = labels,
      main = "Credit Class Distribution for Job Skills: High qualification/ self-employed/ management",
      explode = 0.15,
      col = c("blue", "red"),
      labelcex = 1)
```

```
# A tibble: 8 × 3
  job                                          class     n
  <chr>                                        <chr> <int>
1 high qualification, self-employed, management bad     148
2 high qualification, self-employed, management good     97
3 skilled                                       bad     411
4 skilled                                       good    444
5 unemployed, unskilled, non-resident           bad      60
6 unemployed, unskilled, non-resident           good     15
7 unskilled, resident                           bad      81
8 unskilled, resident                           good    144
>
```

Figure 16 & 17: Code Sampling and Tibble Table for Analysis 2

An analysis about a credit risk score of an individual within a dataset about the job skills with its credit risk score. 3 categories which are "skilled", "unskilled/resident" and "high qualification/self-employed/management".

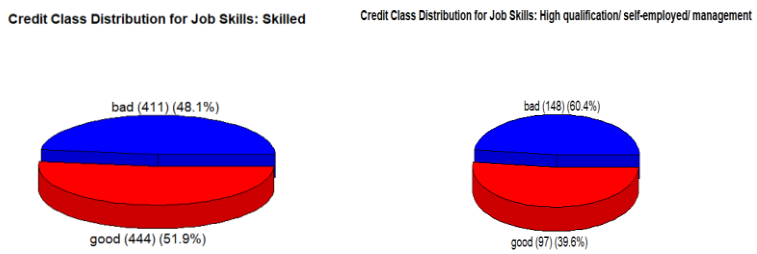


Figure 18 & 19: 3D Pie Chart- Job Skills:Skilled / Job Skills-High Qualifications/Self-Employed/Management

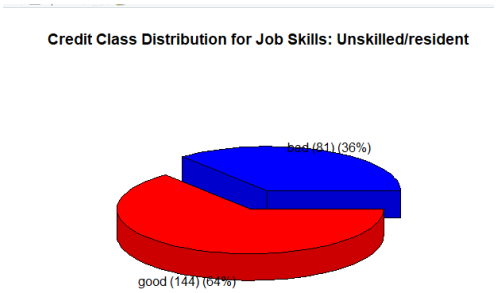


Figure 20 & 21: 3D Pie Chart- Unskilled/Resident

1. Skilled Workers:

Good Credit Class: 444 individuals (51.9%)

Bad Credit Class: 411 individuals (48.1%)

In this category, both sides are nearly split and there is less difference between two sides, with a slight majority in the good credit class. It means that most of them is able to handle their financial status, but it is close that they might fall into bad credit class.

The analysis provided explores the impact of job skills on credit risk classification, examining three job categories: "skilled," "unskilled, resident," and "high qualification, self-employed, management." For each job category, the credit class distribution is divided into "good" and "bad" classifications and visualized using pie charts. Here's a breakdown:

2. Unskilled, Resident Workers:

Good Credit Class: 144 individuals (16.8%)

Bad Credit Class: 81 individuals (9.5%)

Unskilled/resident workers show a larger proportion in the good credit class (64%) compared to the bad (36%). It shows that unskilled workers have a higher chance of being classified as good credit than bad as they don’t have financial issues, but their overall distribution still tell us that unskilled status alone does not prevent a good credit classification.

3. High Qualification, Self-Employed, Management

- **Good Credit Class**: 97 individuals (39.6%)
- **Bad Credit Class**: 148 individuals (60.4%)

In this category, individuals with high qualification or management roles show a larger proportion in the bad credit category 148 individuals (60.4%), with only 97 individuals (39.6%) in the good credit class. This reflects that even with high qualifications or self-employment status, there is a risk of getting bad credit classification due to income instability or other financial risks associated with self-employment or high-level management roles.

In conclusion, job skills do influence credit classification, but other underlying factors likely contribute to credit risk. High-skill or high-management roles do not guarantee a good credit classification, and some unskilled individuals can still achieve favourable credit ratings.

3.1.3: Foreign Workers Got A Better Credit Risk Class?

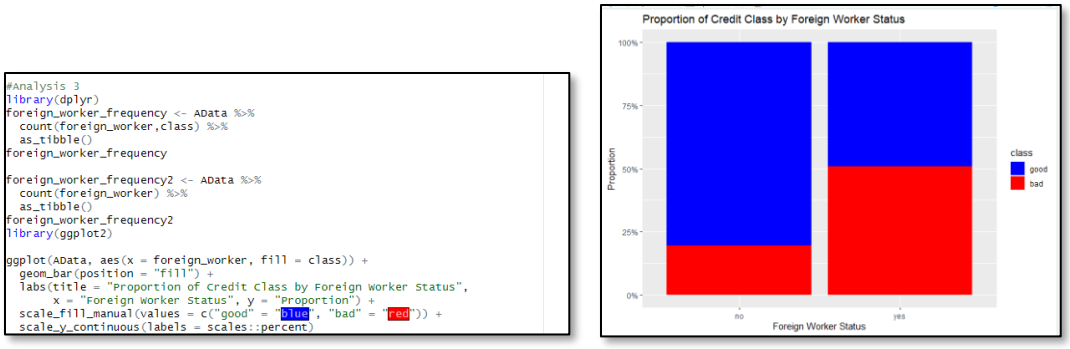


Figure 22 & 23: Sample code and Stacked Bar Graph (Proportion) for Analysis 3

This graph shows the proportion of credit class based on the foreign worker status. For the category who are not foreign workers, it shows that a large proportion falls into the good credit class and the bad credit class only takes a small proportion, which indicates that most of the individuals who are not foreign workers have a good credit rating.

Meanwhile, for the individuals who are foreign workers, the proportion of bad is higher than good class. We can see that there are barely any small differences between bad credit class and good credit class. We can understand that individuals who are from other countries are more likely to have a bad credit class.

In conclusion, individuals who are not foreign workers got a better credit class, while individuals who are foreign workers mostly got a bad credit class. This may relate that underlying factors associated with foreign worker status that impact creditworthiness, potentially due to economic, employment, or stability differences.

3.1.4: Is There A Significant Relationship Between Customer Gender And Credit Risk Classification?

```
#Analysis 4
ADataSgender <- ifelse(grepl("female", ADataSpersonal_status, ignore.case = TRUE), "female",
                       ifelse(grepl("male", ADataSpersonal_status, ignore.case = TRUE), "male", NA))

data_summary <- AData %>%
  count(gender, class) %>%
  group_by(gender) %>%
  mutate(percentage = n / sum(n) * 100)

ggplot(data_summary, aes(x = gender, y = n, group = class, color = class)) +
  geom_line(size = 1) +
  geom_point(size = 3) +
  geom_text(aes(label = paste0(n, " (", round(percentage, 1), "%)")),
            vjust = -0.5, size = 3.5) +
  labs(title = "Line Graph of Gender and Credit Class",
       x = "Gender", y = "Count") +
  scale_color_manual(values = c("good" = "blue", "bad" = "red")) +
  theme_minimal()
```

gender
male
female
male
male
male
male
male
male
male
female
female
female
male
female
female
male
male
female
male
male
male

Figure 24 & 25: Sampling Code and New Column for Analysis 4

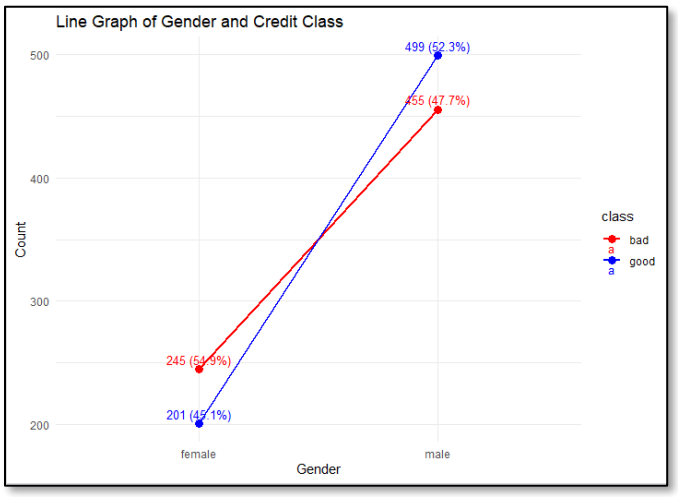


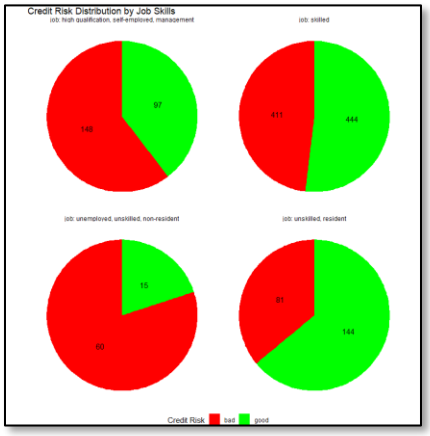
Figure 26: Line Graph for Analysis 4

The line graph above shows the relationship between the credit risk classification and their gender. Among the 1400 individuals, we found that there are 446 females and 954 males in our research. In 446 individuals who are female, it indicates that the one who get a bad credit score in credit risk classification is higher than the good one. Where 245 (54.9%) individuals got a “bad” credit score and the rest 201 (45.1%) got a “good” credit score. In contrast, males show a slightly better credit classification, with 499 (52.3%) individuals were classified as having "good" credit class, and the rest of 455(47.7%) individuals having "bad" credit class. These figures show a higher prevalence of good credit among males compared to females in this dataset. And there is potential influence of gender on credit risk classification, with males generally having a better credit profile than females.

3.1.5: What Are The Combined Effects Of Gender, Job Skills, On Credit Risk Classification?

```
Analysis 5
library(dplyr)
library(ggplot2)

ggplot(AData, aes(x = "", fill = class)) +
 geom_bar(width = 1, stat = "count") +
 coord_polar("y") +
 facet_wrap(~gender, scales = "free_y") +
 scale_fill_manual(values = c("bad" = "red", "good" = "green")) +
 labs(title = "Credit Risk Distribution by Gender",
      x = "",
      y = "",
      fill = "Credit Risk") +
 geom_text(aes(label = ..count..), stat = "count", position = position_stack(vjust = 0.5)) +
 theme_void() +
 theme(legend.position = "bottom")`

ggplot(AData, aes(x = factor(1), fill = class)) +
 geom_bar(width = 1, stat = "count") +
 coord_polar(theta = "y") +
 facet_wrap(~job, labeller = label_both, scales = "free_y") +
 scale_fill_manual(values = c("bad" = "red", "good" = "green")) +
 geom_text(aes(label = ..count..), stat = "count", position = position_stack(vjust = 0.5)) +
 labs(title = "Credit Risk Distribution by Job Skills",
      x = "",
      y = "",
      fill = "Credit Risk") +
 theme_void() +
 theme(legend.position = "bottom")

ggplot(AData, aes(x = factor(1), fill = class)) +
 geom_bar(width = 1, stat = "count") +
 coord_polar(theta = "y") +
 facet_wrap(~foreign_worker, labeller = label_both,scales = "free_y") +
 scale_fill_manual(values = c("bad" = "red", "good" = "green")) +
 geom_text(aes(label = ..count..), stat = "count", position = position_stack(vjust = 0.5)) +
 labs(title = "Credit Risk Distribution by Foreign Status",
      x = "",
      y = "",
      fill = "Credit Risk") +
 theme_void() +
 theme(legend.position = "bottom")
```

Figure 27: Code Sampling for Analysis 5



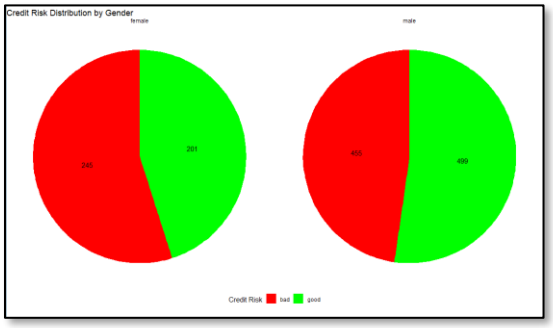


Figure 28 & 29: Pie Chart Distribution Based on Job Skills / Pie Chart Distribution Based on Gender

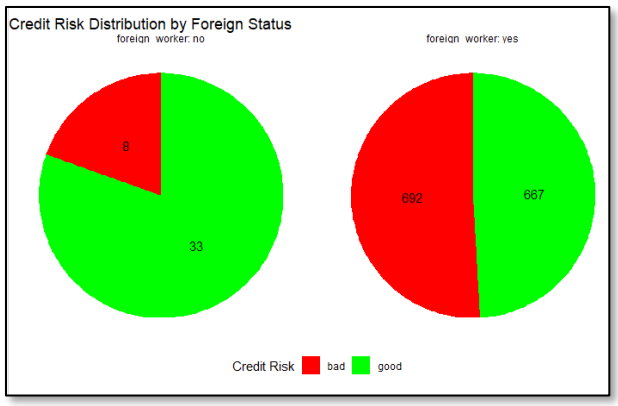


Figure 7: Pie Chart Distribution Based on Foreign Worker Status

The analysis of credit risk distribution reveals significant effects of both gender, foreign status and job skills on creditworthiness. Individuals who got higher qualifications or managerial roles show a higher proportion of bad credit risk compared to good, and skilled workers able to maintain balanced distribution with a little toward good credit risk. Unskilled and unemployed individuals, particularly non-residents, show the highest proportion of bad credit risk, which indicates that they might have financial vulnerability. Considering on gender factors, males got a more balanced credit risk profile, with a higher number of good credit cases compared to females, who show a greater proportion of bad credit risk. The combined effects suggest that skilled males are more likely to have favourable credit outcomes, while females, particularly in unskilled or unemployed categories, face greater credit challenges. These trends highlight the need for targeted interventions, such as job skill development and financial literacy programs, especially for vulnerable groups.

3.1.6 Additional Features

- Tibble functions

```
# Display the tibble
personal_status_frequency_tibble
status_class_counts <- as.data.frame(table(AData$personal_status, AData$class))
colnames(status_class_counts) <- c("personal_status", "class", "count")
```

- Chi-squared test

```
> chi_table <- table(AData$personal_status, AData$class)
>
> cstest <- chisq.test(chi_table)
>
> cstest

        Pearson's Chi-squared test

data:  chi_table
X-squared = 81.799, df = 3, p-value < 2.2e-16
```

3.2 Objective 3-2: To Analyse The Effect Of Loan Related Factors On Credit Risk Classification – Tham Wing Hein (TP067080)

3.2.1 Is there any relationship between loan purpose and credit risk classification?

```
> #Objective 2: To analyze the effect of loan-related factors on credit risk classification.
> #analysis 2-1: Is there any relationship between loan purpose and credit risk classification? (THAM WING HEIN TP067080)
> #view frequency of values for 'purpose' and 'class' separately
> table(AData$purpose)

          business  domestic appliance           education furniture/equipment            new car              other
               122                  36                  89                 268                344                 70
         radio/tv             repairs          retraining           used car
               313                  35                  18                 105
> table(AData$class)

 bad good
 700  700
```

```
> #create contingency table for loan purpose and credit risk classification
> PurposeClass_table <- table(AData$purpose,AData$class)
> print(PurposeClass_table)

                     bad good
  business            59   63
  domestic appliance  28    8
  education           61   28
  furniture/equipment 145  123
  new car             199  145
  other               63    7
  radio/tv            95   218
  repairs             21   14
  retraining          10    8
  used car            19   86
```

```
> prop.table(PurposeClass_table,1)

                              bad      good
business             0.4836066 0.5163934
domestic appliance   0.7777778 0.2222222
education            0.6853933 0.3146067
furniture/equipment  0.5410448 0.4589552
new car              0.5784884 0.4215116
other                0.9000000 0.1000000
radio/tv             0.3035144 0.6964856
repairs              0.6000000 0.4000000
retraining           0.5555556 0.4444444
used car             0.1809524 0.8190476
```

Figure 8: Output

```
#generate bar plot of the contingency table
ggplot(AData, aes(x = purpose, fill = class)) +
  geom_bar(position = "dodge") +
  labs(title = "Loan Purpose vs Credit Risk Classification", x = "Loan Purpose", y = "Count") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
  scale_fill_manual(values = c("pink","lightgreen"),name = "Credit Risk") +
  geom_text(stat = "count", aes(label = ..count..), position =position_dodge(width = 0.8), vjust = -0.5)
```

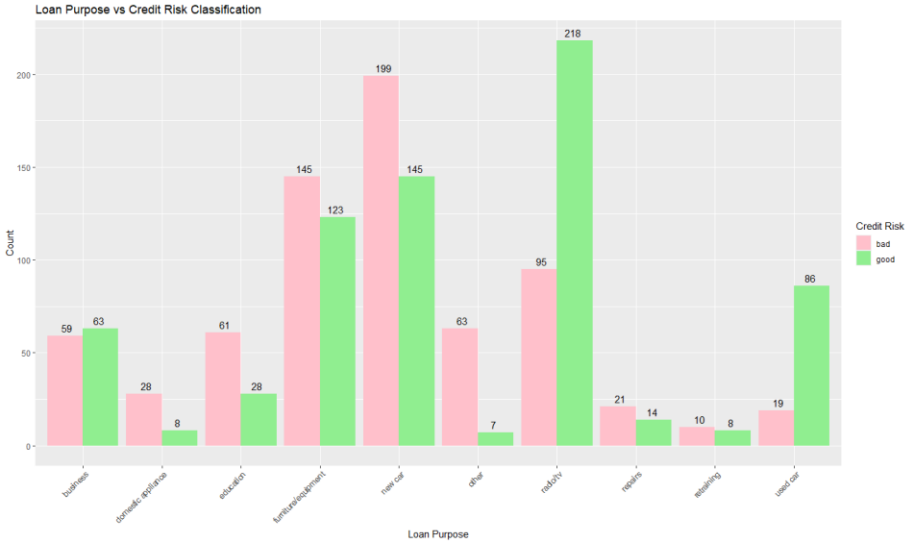


Figure 32 & 33: Bar Graph & its Code Snippet

In an analysis of the credit risk classification dataset comprising 1400 individuals, it was found that there are 700 people, or 50% of the total customers are having bad credit class. Among the 1400 individuals, the top 3 most common purposes of taking the credit are for “New Car” (344 cases), “Radio/TV” (313 cases) and “Furniture/Equipment” (268 cases) while the least common purposes include “Retraining” at 18 cases and “Repairs” at 35 cases.

Based on the contingency table and the bar plot generated, they revealed varying levels of risk associated with different loan purposes. The highest credit risk among the purposes is “Other” which 90% of it is classified as Bad, this marks the riskiest loan purpose in the dataset. “Domestic Appliance” comes along and becomes the second highest credit risk at around 77.78% classified as Bad, indicating a high likelihood of credit risk. On the other hand, “Used Car” scores 81.90% Good classification making it the least risky credit while “Radio/TV” contains 69.65% Good classification, suggesting a stronger association with good credit risk as well. At moderate risk categories, we found out that “Furniture/Equipment” is having 54.10% Bad classification showing a relatively balanced risk level. “Business” on the other hand indicates 51.64% Good classification showing a moderate risk.

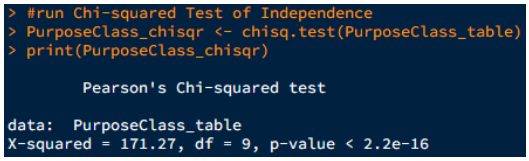


Figure 34： Chi-squared relationship

The Chi-squared test result reveals Chi-squared statistic of 171.27 with 9 degree of freedom and a p-value significantly less than 0.05 (at 2.2e-16 or 0.00000000000000022). This result indicates that the loan purpose is significantly associated with credit risk classification. In other words, it means that there is strong evidence of a relationship between loan purpose and credit risk classification in the dataset. For example, some certain loan purposes might be more likely to have a “bad” classification, while others might be correlate with a “good” classification.

3.2.2 Does the duration of the loan (installment duration) influence the credit risk classification?

```
#analysis 2-2: Does the duration of the loan (installment duration) influence the credit risk classification?
str(AData)    #check structure of data
unique(AData$class)    #check unique values in 'class' column
summary(AData$duration)    #statistics of 'duration' column

#generate density plot of the loan duration across credit risk classifications
ggplot(AData, aes(x = duration, fill = class)) +
  geom_density(alpha = 0.5) +
  labs(title = "Density of Loan Duration by Credit Risk Classification",
       x = "Loan Duration (months)",
       y = "Density") +
  scale_fill_manual(values = c("magenta","green"),name = "Credit Risk")
```

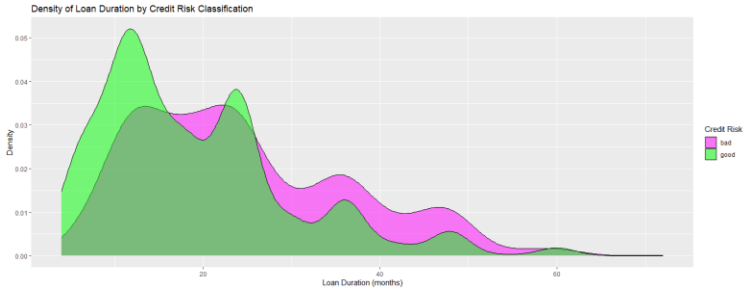


Figure 35 & 36: Density plot and its Code Snippet

```
> DurationClass_table <- table(AData$duration,AData$class)
> print(DurationClass_table)

    bad good
4     0    6
5     0    1
6    12   66
7     4    5
8     3    6
9    17   35
10   12   25
11   10    9
12   88  130
13   10    4
14   15    3
15   24   52
16   13    1
17    7    0
18   56   71
19   16    0
20   12    7
21   27   21
22   12    2
23   10    0
24   93  128
25    9    0
26    5    1
27    9    8
28    9    2
29   10    0
30   17   27
31    4    0
32    7    0
33    8    2
34    5    0
35   11    0
36   56   46
37    2    0
38    4    0
39    5    4
40    9    0
```

41	6	0
42	4	8
43	7	0
44	4	0
45	6	1
46	6	0
47	4	1
48	36	20
49	1	0
50	1	0
52	2	0
54	1	1
55	1	0
57	2	0
59	1	0
60	6	7
72	1	0

Figure 37 & 38: Output Data

Based on the density plot and data generated, they revealed that the loan durations of 12 months have the highest count for both Good (at 130 records) and Bad (at 88 records) classification. This suggests that short-term loans are common across both categories, but Good is slightly more than Bad classifications at this duration. While at longer durations of 36 to 60 months, there are more Bad classifications (56) than Good (46) at 36 months, then this trend continues at 48 months with 36 Bad and 20 Good classifications. When it reaches beyond 60 months, although loans are rare, but the majority are classified as bad credit risks. Shorter durations which are below 12 months on the other hand have very few Bad classifications and significantly more good classifications. For instance, at 6 months, there are 12 Bad and a whopping of 66 Good classifications, showing a strong tendency for short-term loans to be associated with good credit risks.

3.2.3 Are loan amounts a strong predictor for credit risk classification?

```
#analysis 2-3: Are loan amounts a strong predictor for credit risk classification?
#summarize credit amount by credit risk classification
summary(AData$credit_amount)
summary(AData$class)
str(AData$credit_amount)
str(AData$class)

Class_summary <- AData %>%
  group_by(class) %>%
  summarize(CreditAmount_mean = mean(credit_amount,na.rm = TRUE),
            CreditAmount_median = median(credit_amount,na.rm = TRUE),
            CreditAmount_sd = sd(credit_amount,na.rm = TRUE))
print(Class_summary)
```

```
> #analysis 2-3: Are loan amounts a strong predictor for credit risk classification?
> #summarize credit amount by credit risk classification
> summary(AData$credit_amount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    250    1355    2326    3383    4226   18424
> summary(AData$class)
   Length     Class      Mode
     1400 character character
> str(AData$credit_amount)
 num [1:1400] 1169 5951 2096 7882 4870 ...
> str(AData$class)
 chr [1:1400] "good" "bad" "good" "good" "bad" "good" "good" "good" "good" "bad" "bad" "bad" "good" "bad" "good" ...
>
> Class_summary <- AData %>%
+   group_by(class) %>%
+   summarize(CreditAmount_mean = mean(credit_amount,na.rm = TRUE),
+             CreditAmount_median = median(credit_amount,na.rm = TRUE),
+             CreditAmount_sd = sd(credit_amount,na.rm = TRUE))
> print(Class_summary)
# A tibble: 2 × 4
  class CreditAmount_mean CreditAmount_median CreditAmount_sd
  <chr>            <dbl>              <dbl>           <dbl>
1 bad              3788.              2450.           3405.
2 good             2985.              2244            2401.
```

Figure 39 & 40: Summary and Statistics

```
#generate box plot of the loan amount vs credit risk classification
ggplot(AData, aes(x = class, y = credit_amount, fill = class)) +
  geom_boxplot() +
  labs(title = "Loan Amount vs Credit Risk Classification",
       x = "Credit Risk Classification",
       y = "Loan Amount") +
  scale_fill_manual(values = c("pink","lightgreen"),name = "Credit Risk")
```

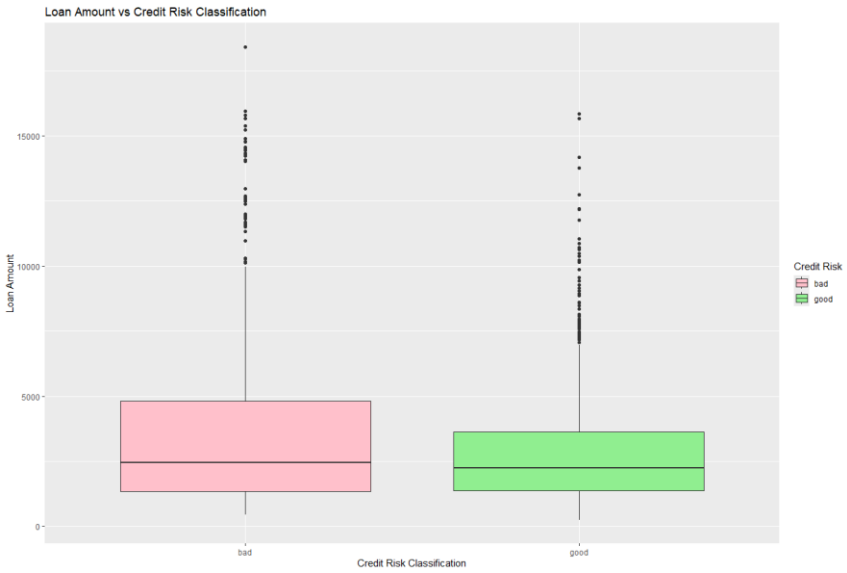


Figure 41 & 42: Boxplot and its Code Snippet

Based on the statistics and boxplot generated, we found out that individuals classified as “Bad” credit risk tend to have slightly higher loan amounts compared to individuals classified as “Good” credit risk. This can be proved with the median loan amount which Bad credit risk is 2450 while Good credit risk is slightly lower at 2244. Besides that, the loan amounts for “Bad” credit risk group show greater variability at 3405 comparing with “Good” credit risk group at 2401. This indicates that while “Bad” borrowers generally have higher loan amounts, there is

more inconsistency in the loan amounts within this group. On the other hand, we can see from the boxplot that both groups have outliers, but the “Bad” group have more extreme values which means higher loan amounts. These outliers indicate a tendency for some “Bad” credit risk borrowers to take out unusually large loans.

```
> #Run Welch Two Sample t-test (Extra Feature)
> CreditAmountClass_ttest <- t.test(credit_amount ~ class, data = AData)
> print(CreditAmountClass_ttest)

        Welch Two Sample t-test

data:  credit_amount by class
t = 5.0472, df = 1256.4, p-value = 5.144e-07
alternative hypothesis: true difference in means between group bad and group good is not equal to 0
95 percent confidence interval:
  485.9122 1103.8680
sample estimates:
 mean in group bad mean in group good
          3780.347          2985.457
```

Figure 43: T-test results

The T-test result reveals T-value of 5.0472 with 1256.4 degree of freedom and a p-value of 5.144e-7 or 0.0000005144 which is much smaller than common significant level of 0.05. This result indicates that the loan amounts is statistically significant difference between the credit risk classification. This means that the T-test confirms that loan amounts are significantly higher for “Bad” credit risk borrowers compared to “Good” one. However, this variable alone might not be able to fully explain credit risk classification due to the overlap in the distributions as seen in the boxplot.

3.2.4 Does employment status interacts with loan duration to influence credit risk classification?

```
#analysis 2-4: Does employment status interacts with loan duration to influence credit risk classification?
#convert employment status and credit risk classification columns to factors
AData$employment <- as.factor(AData$employment)
AData$class <- as.factor(AData$class)
AData$duration <- as.numeric(AData$duration)
str(AData)    #check the structure

#create and output summarized dataset for interaction plotting
summary2_4 <- AData %>%
  group_by(employment, duration, class) %>%
  summarize(count = n()) %>%
  ungroup()
print(summary2_4, n = 267)
```

Figure 44: Code snippet

```
#create interaction plot of employment status and loan duration by credit risk classification (Extra Feature)
ggplot(summary2_4, aes(x = duration, y = count, color = employment, group = employment)) +
  geom_line(size = 1) +
  geom_point(size = 2) +
  facet_wrap(~ class) +
  labs(
    title = "Interaction of Employment Status and Loan Duration on Credit Risk Classification",
    x = "Loan Duration (Months)",
    y = "Count of Classification",
    color = "Employment Status"
  )
```

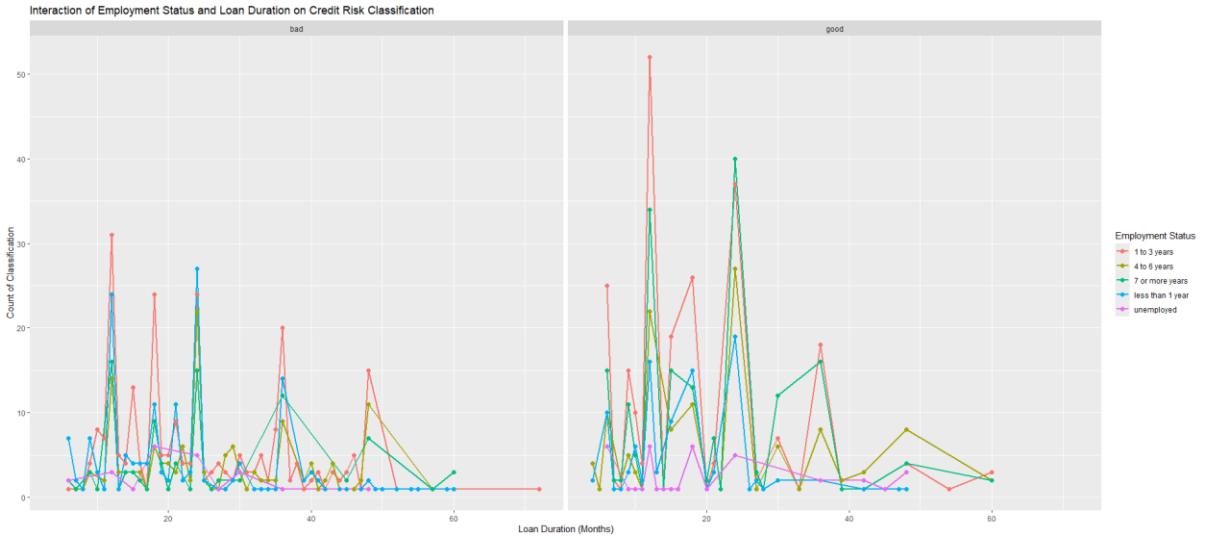


Figure 45 & 46: Interaction Plot and its Code Snippet (Extra Feature)

According to the interaction plot generated, we can identify that unemployed individuals have a significantly higher bad credit rate comparing with others who are employed, although there is some good credit, but it's much less frequent. For people in the "1 to 3 years" employment group, good credit risk is relatively more common across different loan durations, although there are instances where bad credit risk appears especially for longer durations at around 12 to 24 months.

Meanwhile based on the loan duration's role in credit risk, shorter loan durations like 6 to 12 months seems to have more balanced or better credit outcomes across most employment status. Then as loan duration increases, the tendency for bad credit increases in certain employment group, especially for the "unemployed" and those with "1 to 3 years" of employment.

Based on the visualization, good credit tends to be more prevalent in individuals with longer employment histories like "7 or more years" especially in loan duration of 12-24 months. Bad credit classification on the other hand is more frequent among unemployed individuals across all loan durations, indicating that employment status plays a major role in an individual's financial stability and ability to manage debt.

```
> #Run Logistic Regression Model with interaction term
> model2_4 <- glm(class ~ employment * duration, data =AData, family = "binomial")
> summary(model2_4)

Call:
glm(formula = class ~ employment * duration, family = "binomial",
    data = AData)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                      1.0380192  0.2022975   5.131 2.88e-07 ***
employment4 to 6 years          -0.2049345  0.3390466  -0.604   0.5455
employment7 or more years        0.2913978  0.3361316   0.867   0.3860
employmentless than 1 year      -0.5038148  0.3497896  -1.440   0.1498
employmentunemployed            -0.3379752  0.5696981  -0.593   0.5530
duration                        -0.0530114  0.0086444  -6.132 8.65e-10 ***
employment4 to 6 years:duration  0.0140047  0.0134849   1.039   0.2990
employment7 or more years:duration  0.0219562  0.0136659   1.607   0.1081
employmentless than 1 year:duration -0.0001525  0.0161580  -0.009   0.9925
employmentunemployed:duration    0.0447443  0.0237436   1.884   0.0595 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1940.8  on 1399  degrees of freedom
Residual deviance: 1803.3  on 1390  degrees of freedom
AIC: 1823.3

Number of Fisher Scoring iterations: 4
```

Figure 47: Logistic Regression Model

The Logistic Regression model shows an intercept of 1.038, meaning that when all factors are at their baseline, the likelihood of being classified as bad credit is higher. Employment status doesn't strongly affect the chances of being bad credit, except for those with "less than 1 year" of employment. Most employment categories on the other hand don't have a significant impact when compared to the "1 to 3 years" group. However, loan duration is a key factor. The negative coefficient for loan duration (-0.053) indicates that longer loan durations are linked to a lower chance of being classified as bad credit, suggesting that longer loans are associated with lower credit risk.

3.2.5 Does the proportion of income committed to loan payments (installment commitment) influence the likelihood of higher credit risk classification?

```
> summary(AData$installment_commitment)   #statistics of 'installment_commitment' column
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.000   3.026   4.000   4.000
>
> #summarize installment commitment and credit risk classification
> summary2_5 <- AData %>%
+   count(installment_commitment, class)
> print(summary2_5)
  installment_commitment class   n
1                      1   bad  55
2                      1  good 102
3                      2   bad 145
4                      2  good 169
5                      3   bad 153
6                      3  good 112
7                      4   bad 347
8                      4  good 317
```

Figure 48: Summary

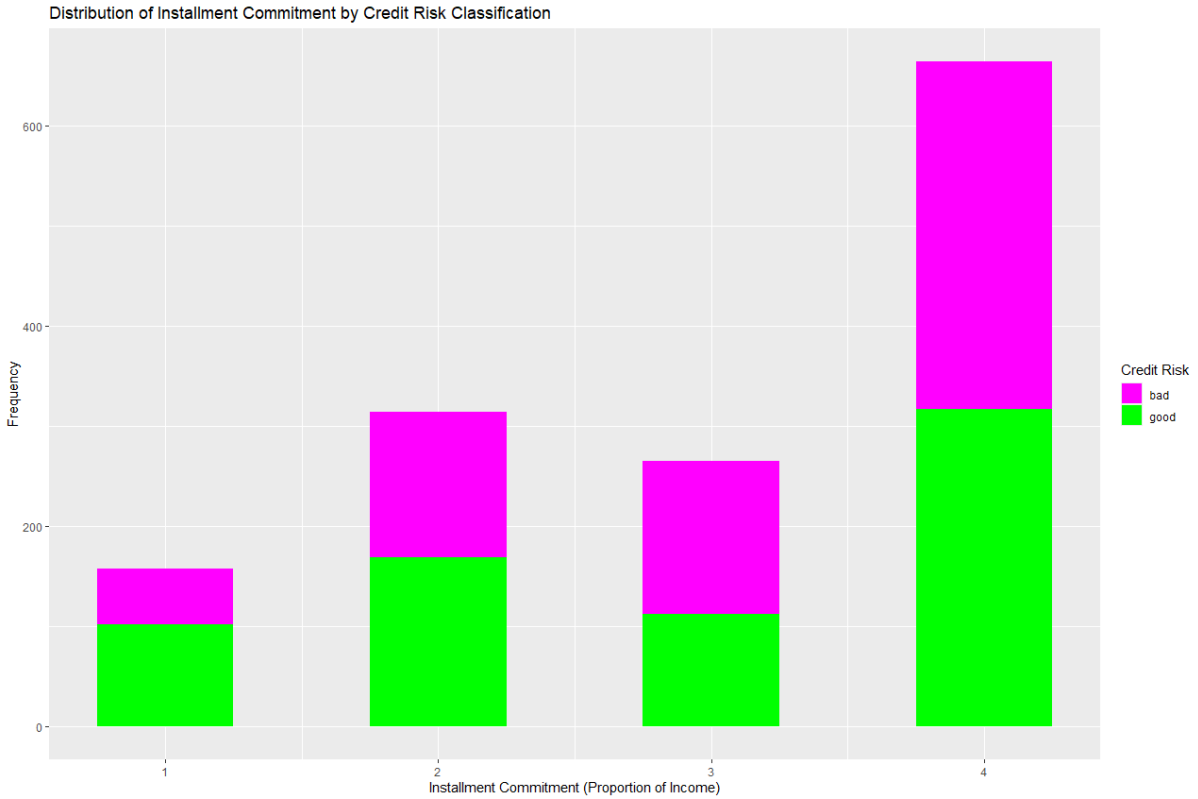


Figure 49: Histogram and its Code Snippet

Based on the output and histogram generated, as instalment commitment increases, the number of applicants classified as “bad credit risk” increases, particularly at higher level of commitment (3 and 4). This shows that the individuals are using more of their income to loan payment, which lead to the high-risk classification. On the other hand, the number of “good credit risk”

applicants generally decrease as instalment commitment increases, which declines at higher commitment levels. This can be seen at the moderate commitment levels (1 and 2) has a higher proportion of good credit risk especially at lowest commitment level (1).

3.2.6 Conclusion of Objective 3-2

The analysis showed how different loan-related factors (like loan purpose, duration, amount, and income commitment) affect credit risk classification.

- **Analysis 1** showed that while personal loans were more likely to be associated with bad credit risk, the loan purpose alone didn't significantly affect credit risk.
- **Analysis 2** found that longer loan durations were linked to higher chances of being classified as bad credit risk, but this effect wasn't as strong for shorter loans.
- **Analysis 3** looked at all factors together and found that applicants with high income commitment and longer loan durations were more likely to be high-risk. However, only a small percentage of applicants fit this description.
- **Analysis 5** showed that applicants who committed more of their income to loan repayments were more likely to be classified as bad credit risk, especially those with high commitments.

As conclusion, the findings suggest that loan-related factors alone don't significantly determine credit risk. Only a small group of applicants with high loan amounts, long durations, and high commitment were considered high-risk. Other factors, like employment status, income, and financial behavior, likely play a bigger role in determining credit risk.

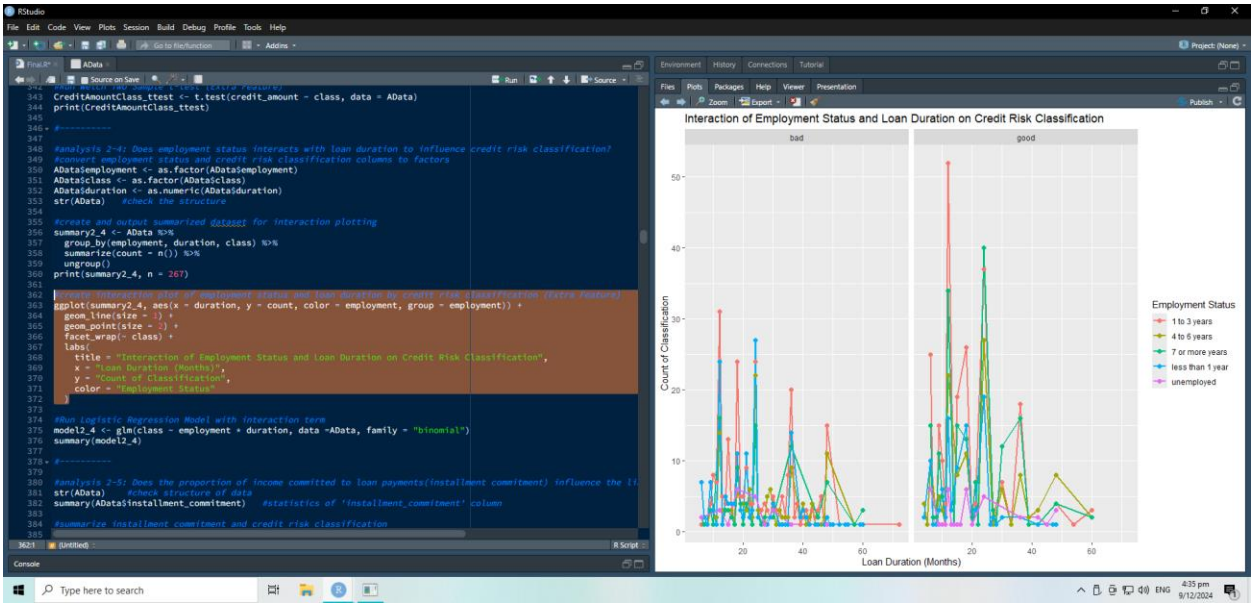
3.2.7 Additional Features

Welch Two Sample T-test

```
#Run Welch Two Sample t-test (Extra Feature)
CreditAmountClass_ttest <- t.test(credit_amount ~ class, data = AData)
print(CreditAmountClass_ttest)
```

This can be found at analysis 3. It was being used to statistically assess whether there is a significant difference in the mean loan amounts between the two credit risk groups (“good” and “bad”). It is used because it accounts for differences in variance between the two groups.

Interaction Plot



This can be found at analysis 4. An interaction plot is a helpful visualization tool when we want to examine whether the effect of one predictor (independent variable) on the dependent variable changes depending on the level of another predictor. In the context of my analysis, the interaction plot is used to explore whether the effect of loan duration on credit risk classification (good vs. bad) depends on the employment status of the individual.

3.3 Objective 3-3: To Assess the Role of Financial Commitments on Credit Risk Classification. Eleanor Permata Fry (TP072606)

3.3.1 Is There A Relationship Between The Instalment Commitments And Credit Risk Classification?

Step 1: Data Overview

```
> # Step 1: Data Overview
> # Check Structure of Data
> str(AData)
'data.frame':   1400 obs. of  22 variables:
 $ checking_status       : chr  "<0" "0<=X<200" "no checking" "<0" ...
 $ duration              : num  6 48 12 42 24 36 24 36 12 30 ...
 $ credit_history        : chr  "critical accounts" "existing paid" "critical accounts" "existing paid" ...
 $ purpose               : chr  "radio/tv" "radio/tv" "education" "furniture/equipment" ...
 $ credit_amount         : num  1169 5951 2096 7882 4870 ...
 $ savings_status        : chr  "no known savings" "<100" "<100" "<100" ...
 $ employment            : chr  ">=7" "1<=X<4" "4<=X<7" "4<=X<7" ...
 $ installment_commitment: num  4 2 2 2 3 2 3 2 2 4 ...
 $ personal_status       : chr  "male single" "female divorced/dependent/married" "male single" "male single" ...
 $ other_parties         : chr  "none" "none" "none" "guarantor" ...
 $ residence_since       : num  4 2 3 4 4 4 4 2 4 2 ...
 $ property_magnitude    : chr  "real estate" "real estate" "real estate" "life insurance" ...
 $ age                   : num  67 22 49 45 53 35 53 35 61 28 ...
 $ other_payment_plans   : chr  "stores" "stores" "stores" "stores" ...
 $ housing               : chr  "own" "own" "own" "for free" ...
 $ existing_credits      : num  2 1 1 1 2 1 1 1 1 2 ...
 $ job                   : chr  "skilled" "skilled" "unskilled, resident" "skilled" ...
 $ num_dependents        : num  1 1 2 2 2 2 1 1 1 1 ...
 $ own_telephone         : chr  "yes" "none" "none" "none" ...
 $ foreign_worker        : chr  "yes" "yes" "yes" "yes" ...
 $ class                 : chr  "good" "bad" "good" "good" ...
 $ income_level          : chr  "no known savings" "<100" "<100" "<100" ...
```

Figure 50: Data Structure

Step 2: Descriptive Statistics

```
> # Step 2: Descriptive Statistics
> # Calculate Mean, Median, and Standard Deviation of Installment Commitments by Credit Risk Class
> installment_summary <- AData %>%
+   group_by(class) %>%
+   summarise(
+     avg_installment = mean(installment_commitment, na.rm = TRUE),
+     median_installment = median(installment_commitment, na.rm = TRUE),
+     sd_installment = sd(installment_commitment, na.rm = TRUE)
+   )
> print(installment_summary)
# A tibble: 2 × 4
  class avg_installment median_installment sd_installment
  <chr>           <dbl>              <dbl>          <dbl>
1 bad              3.13                  3           1.00
2 good             2.92                  3           1.13
```

Figure 51: Summarise

The descriptive statistics above provides a summary of the central tendency (mean and median) and variability (standard deviation) of instalment commitments for each credit risk class.

Step 3: Data Visualization

```
# Step 2: Data Visualization
# 1.1 Density Plot: Counts of Credit Risk by Other Payment Plans

library(ggplot2) # Library for Density Plot

ggplot(AData, aes(x = installment_commitment, fill = class)) +
  geom_density(alpha = 0.5, color = "black") + # Add Border
  scale_fill_manual(values = c("good" = "green", "bad" = "red")) +
  labs(
    title = "Density Plot of Installment Commitments by Credit Risk Classification",
    x = "Number of Installment Commitments",
    y = "Density",
    fill = "Credit Risk Class"
    ) +
  theme_minimal(base_size = 12) +  # Increase Font Size
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    panel.grid.major = element_line(color = "lightgray"),  # Add Grid Lines
    panel.grid.minor = element_blank()
    )
```

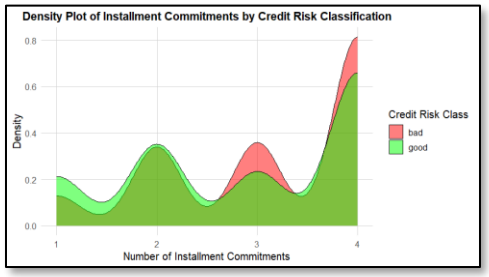
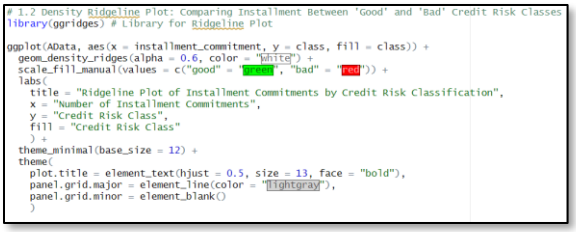


Figure 52 & 53: Density Plot



```
# 1.2 Density Ridgeline Plot: Comparing Installment Between "Good" and "Bad" Credit Risk Classes
library(ggridges) # Library for Ridgeline Plot

ggplot(AData, aes(x = installment_commitment, y = class, fill = class)) +
  geom_density_ridges(alpha = 0.6, color = "white") +
  scale_fill_manual(values = c("good" = "#00FF00", "bad" = "#FF0000")) +
  labs(
    title = "Ridgeline Plot of Installment Commitments by Credit Risk Classification",
    x = "Number of Installment Commitments",
    y = "Credit Risk Class",
    fill = "Credit Risk Class"
    ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank()
    )
```

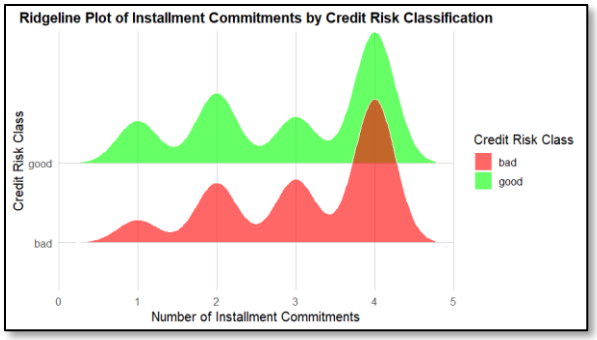


Figure 54 & 55: Ridgeline Plot

From the two figures above, both the density and ridgeline plot shares identical shapes. The distribution of "good" credit risk class peaks around the 2 instalment commitments, while the "bad" credit risk class shows peak around 3-4 instalment commitments, with a wider spread towards higher values than the "good" class. This indicates that the "good" credit risks have typically less instalment commitments, while the "bad" has more, suggesting higher financial exposure.

Step 4: Statistical Test

```
> # Step 4: Statistical Test
> # T-Test (Evaluate Differences in Mean Installment Commitments Between Credit Risk Classes)
> good_class <- AData %>%
+   filter(class == "good") %>%
+   pull(installment_commitment)
>
> bad_class <- AData %>%
+   filter(class == "bad") %>%
+   pull(installment_commitment)
>
> t_test_class_result <- t.test(good_class, bad_class, var.equal = FALSE)
> print(t_test_class_result)

        Welch Two Sample t-test

data:  good_class and bad_class
t = -3.7096, df = 1378.4, p-value = 0.0002158
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.32323390 -0.09962324
sample estimates:
mean of x mean of y
 2.920000  3.131429
```

Figure 56: Welch's T-Test

The statistical test was done using the Welch's t-test, to compare instalment commitments between two credit risk classes. The results showed a significant difference, with a t-statistic value of -3.71, indicating a strong difference. The p-value was extremely low, indicating the difference was unlikely due to coincidence. The "good" credit class had a mean value of 2.92, while the "bad" credit class had a mean value of 3.13. This suggests that individuals with "bad" credit risk tend to have higher average instalment commitments.

Conclusion

Based on the analysis done above, we can conclude that **there is a significantly strong relationship between instalment commitments and credit risk classification**. The "bad" credit risk has a statistically significantly higher mean of instalment commitment than the "good" credit risk class. The higher number of instalment commitments in the "bad" class could suggest an increase in financial strain. On the contrary, the "good" credit class has lesser commitments, suggesting better financial responsibility especially in managing loans.

3.3.2 Does Income Level Predict Credit Risk Classification When Considered Alongside Instalment Commitment?

Step 1: Data Visualization

```
# Step 2: Data Visualization
# 2.1 Bar Chart: Proportion of Credit Risk Classifications by Income Level and Installment Commitment

# Calculate Proportions
bar_data <- AData %>%
  group_by(income_level, installment_commitment, class) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count))

# Visualize
ggplot(bar_data, aes(x = income_level, y = proportion, fill = class)) +
  geom_bar(stat = "identity", position = "fill", alpha = 0.7) +
  geom_text(aes(label = scales::percent(proportion, accuracy = 0.1)),
            position = position_fill(vjust = 0.5), size = 3) +
  facet_wrap(~ installment_commitment) +
  scale_fill_manual(values = c("good" = "darkgreen", "bad" = "red")) +
  labs(title = "Credit Risk Classifications by Income Level and Installment Commitment",
       x = "Income Level",
       y = "Proportion",
       fill = "Credit Risk Class") +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 15, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  )
```

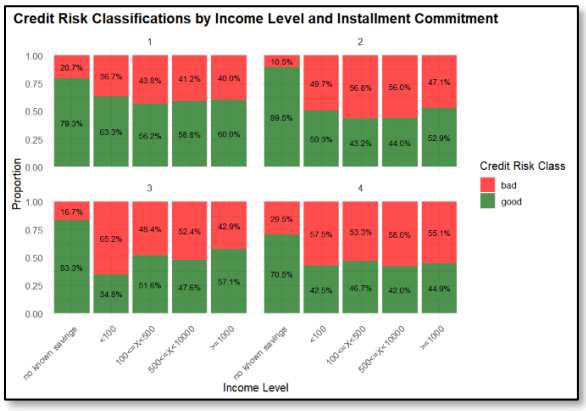


Figure 57 & 58: Stacked Bar Chart

Before the visualization, the '*savings_status'* column is used as a representative for the '*income_level'*, suggesting more savings means a higher income. From the stacked bar chart above, we can analyse that individuals with lower income levels (<100) have a higher proportion of "bad" credit risk classification, especially when the instalment commitment increases. For higher income levels (>=1000), the proportion of "bad" credit risk tends to be lower, even with the increase of instalment commitments. Suggesting that, a high income helps individuals to better manage their instalment commitments, despite the amount of isntalmetns.

```
# 2.2 Heat Map: Bad Credit Risk Proportions by Income Level and Installment Commitment
heatmap_data_bad <- AData %>%
  group_by(income_level, installment_commitment, class) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count)) %>%
  filter(class == "bad")

ggplot(heatmap_data_bad, aes(x = income_level, y = installment_commitment, fill = proportion)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightyellow", high = "red", name = "Proportion") +
  labs(title = "Proportion of Bad Credit Risks by Income Level and Installment Commitment",
       x = "Income Level",
       y = "Number of Installment Commitment"
       ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
    ) +
  geom_text(aes(label = scales::percent(proportion, accuracy = 0.1)), color = "black", size = 4)
```

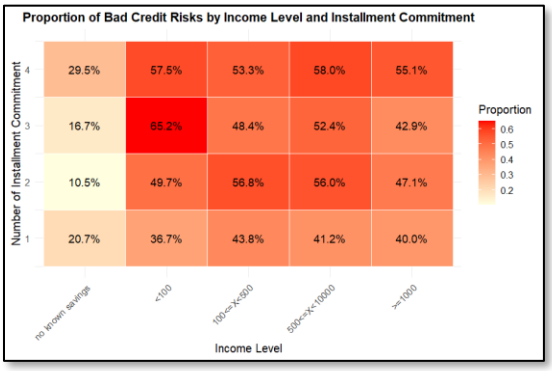


Figure 59 & 60: "Bad" Heatmap Snippet

In the heatmap above, a large proportion of "bad" credit risk class can be seen in the low income (<100) and 3 instalment commitments (65.2%). Individuals with no known savings / income and 2 instalments commitments has the lowest proportion of "bad" credit risks (10.5%).

The share of "bad" credit risks increases with the amount of instalment commitments. Meanwhile, the increase of income level indicates a decrease in "bad" credit risks for any given instalments.

```
# 2.3 Heat Map: Good Credit Risk Proportions by Income Level and Installment Commitment
heatmap_data_good <- AData %>%
  group_by(income_level, installment_commitment, class) %>%
  summarise(count = n()) %>%
  mutate(proportion = count / sum(count)) %>%
  filter(class == "good")

ggplot(heatmap_data_good, aes(x = income_level, y = installment_commitment, fill = proportion)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "lightyellow", high = "darkgreen", name = "Proportion") +
  labs(title = "Proportion of Good Credit Risks by Income Level and Installment Commitment",
       x = "Income Level",
       y = "Number of Installment Commitment"
       ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
    ) +
  geom_text(aes(label = scales::percent(proportion, accuracy = 0.1)), color = "black", size = 4)
```

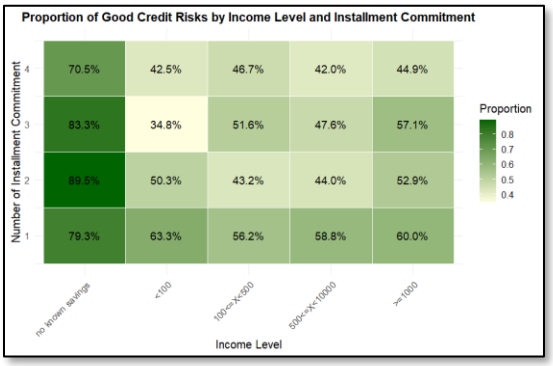


Figure 61 & 62: "Good" Heatmap Snippet

In the heatmap above, a large proportion of "good" credit risk class can be seen in the low income (<100) and with one instalment commitments (63.3%). Individuals with no know income, regardless of amount of instalments, has the highest share of "good" credit risks. Similar to the "bad" credit risk heatmap, the more the instalment commitment an individual has, despite the income level, the lower the "good" credit risks become, suggesting a notice of financial burden.

Step 2: Statistical Tests

```
# Step 3: Statistical Tests
# 3.2 Fisher's Exact Test

# Association Between Income Level and Credit Risk Classification
fisher_income_class <- fisher.test(
  table(AData$income_level,
        AData$class),
  simulate.p.value = TRUE
  )

# Association Between Installment Commitments and Credit Risk Classification
fisher_installment_class <- fisher.test(
  table(AData$installment_commitment,
        AData$class),
  simulate.p.value = TRUE
  )

# Association Between Installment Commitments and Income Level
fisher_income_installment <- fisher.test(
  table(AData$income_level,
        AData$installment_commitment),
  simulate.p.value = TRUE
  )

print(fisher_income_class)
print(fisher_installment_class)
print(fisher_income_installment)
```

```
> print(fisher_income_class)

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  table(AData$income_level, AData$class)
p-value = 0.0004998
alternative hypothesis: two.sided

> print(fisher_installment_class)

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  table(AData$installment_commitment, AData$class)
p-value = 0.0004998
alternative hypothesis: two.sided

> print(fisher_income_installment)

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  table(AData$income_level, AData$installment_commitment)
p-value = 0.001499
alternative hypothesis: two.sided
```

Figure 63 & 64: Fisher's Exact Test

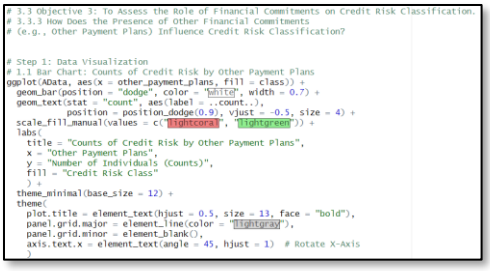
A Fisher's Exact Test was used for the statistical analysis between income level, instalment commitments, and credit risk classification. Based on all three's result, a p-value of 0.0005 was received from the fisher's test between income level, instalment commitments, and credit risk classification. Meanwhile, a p-value of 0.0015 was received between income level and instalment commitments. All three p-value are below the value of 0.005, which indicate that all three relationships are statistically significant to one another.

Conclusion

Based on the analysis done above, we can conclude that **income level does predict credit risk classification when considered alongside instalment commitments**. Instalment commitments has a big impact on credit risk, but if alone, it doesn't give a full picture. The descriptive statistics, visualizations and statistical tests show how important income level is in determining creditworthiness. Income levels provide essential context, as they show a person's ability to manage financial commitments, without it can lead to overestimating or underestimating credit risk classification.

3.3.3 How Does The Presence Of the Financial Commitments (e.g., Other Payment Plans) Influence Credit Risk Classification?

Step 1: Data Visualization



```
# 3.3 Objective 3: To Assess the Role of Financial Commitments on Credit Risk Classification.
# 3.3.3 How Does the Presence of Other Financial Commitments
# (e.g., Other Payment Plans) Influence Credit Risk Classification?


# Step 1: Data Visualization
# 1.1 Bar Chart: Counts of Credit Risk by Other Payment Plans
ggplot(AData, aes(x = other_payment_plans, fill = class)) +
  geom_bar(position = "dodge", color = "white", width = 0.7) +
  geom_text(stat = "count", aes(label = ..count..),
            position = position_dodge(0.9), vjust = -0.5, size = 4) +
  scale_fill_manual(values = c("lightcoral", "lightgreen")) +
  labs(
    title = "Counts of Credit Risk by Other Payment Plans",
    x = "Other Payment Plans",
    y = "Number of Individuals (Counts)",
    fill = "Credit Risk Class"
    ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_blank(),
    axis.text.x = element_text(angle = 45, hjust = 1)  # Rotate X-Axis
    )
```

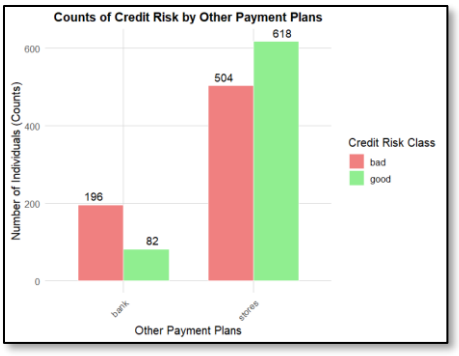
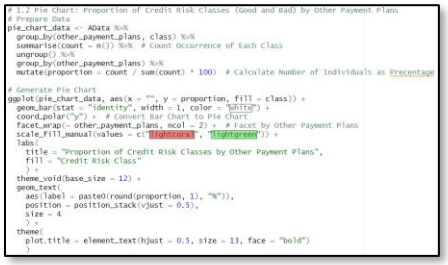


Figure 65 & 66: Bar Chart Snippet

The bar chart shows that a significant number of individuals prefer store payment plans, with 1122 individuals in total. Store-based plans are primarily used by individuals with "good" credit risks (618), while bank-based plans are more popular with "bad" credit risks (196). This preference is due to easier accessibility, flexible terms, and less demanding qualifying conditions. Bank payment plans are more likely to be linked to "bad" credit risks, suggesting that individuals with financial struggles may face more difficulties under these plans.



```
# 1.2 Pie Chart: Proportion of Credit Risk Classes (Good and Bad) by Other Payment Plans
# Prepare Data
pie_chart_data <- AData %>%
  group_by(other_payment_plans, class) %>%
  summarise(count = n()) %>%  # Count Occurrence of Each Class
  ungroup() %>%
  group_by(other_payment_plans) %>%
  mutate(proportion = count / sum(count) * 100)  # Calculate Number of Individuals as Precentage

# Generate Pie Chart
ggplot(pie_chart_data, aes(x = "", y = proportion, fill = class)) +
  geom_bar(stat = "identity", width = 1, color = "white") +
  coord_polar("y") +  # Convert Bar Chart to Pie Chart
  facet_wrap(~ other_payment_plans, ncol = 2) +  # Facet by Other Payment Plans
  scale_fill_manual(values = c("lightcoral", "lightgreen")) +
  labs(
    title = "Proportion of Credit Risk Classes by Other Payment Plans",
    fill = "Credit Risk Class"
    ) +
  theme_void(base_size = 12) +
  geom_text(
    aes(label = paste0(round(proportion, 1), "%")),
    position = position_stack(vjust = 0.5),
    size = 4
    ) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold")
    )
```

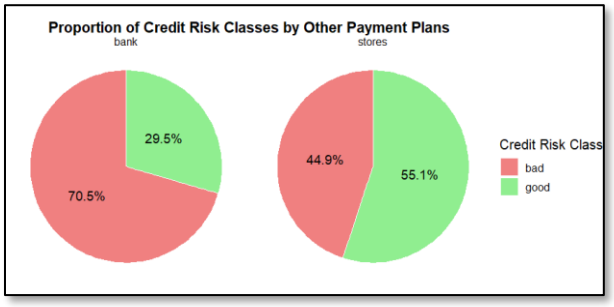


Figure 67 & 68: Pie Chart Snippet

From the pie charts above, we can evaluate that bank payment plans are mostly associated by individuals with “bad” credit risks (70.5%), while store payment plans are evenly split though slightly skewed to the “good” credit risks (55.1%). The result of these pie charts complements the bar chart as it emphasizes the differences in risk class distribution between the two payment types.

Step 2: Statistical Test

```
> # Step 2: Statistical Test
> # Fisher's Exact Test (Association Between Other Payment Plans and Credit Risk Classification)
> fisher_plans_class <- fisher.test(
+   table(AData$other_payment_plans,
+         AData$class)
+   )
> print(fisher_plans_class)

        Fisher's Exact Test for Count Data

data:  table(AData$other_payment_plans, AData$class)
p-value = 1.87e-14
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 2.190406 3.941332
sample estimates:
odds ratio
  2.928662
```

Figure 69: Fisher's Exact Test

The Fisher's Test revealed a statistically significant relationship between payment plan type and credit risk classification. The p-value is extremely low, confirming the significant relationship. The odds ratio indicates a strong association between bank payment plans and a higher chance of "bad" credit risk classification compared to store payment plans.

Conclusion

The analysis reveals a significant relationship between payment plan types and credit risk classification. Store payment plans are more suited to "good" credit risk individuals, indicating lower financial risk. However, a significant number of these individuals have "bad" credit, suggesting they may be financially secure but burdened due to their accessibility and less-strict repayment rules. Bank payment plans are more suited to "bad" credit risk individuals, making them more difficult to use. This relationship could help financial institutions analyse risks associated with specific payment plans and adjust their credit regulations accordingly.

3.3.4 What Demographic Factors (e.g., Age, Marital Status) Moderate The Relationship Between Instalment Commitments and Credit Risk Classification?

Step1 1: Data Visualization

```
# 3.3 Objective 3: To Assess the Role of Financial Commitments on Credit Risk Classification.
# 3.3.4 What Demographic Factors (e.g., Age, Marital Status) Moderate The Relationship
# Between Installment Commitments and Credit Risk Classification?

# Step 1: Data Visualization
# 1.1 Jittered Scatter Plot: Visualize Age by Installment Commitments and Credit Risk
ggplot(AData, aes(x = age, y = installment_commitment, color = class)) +
  geom_jitter(alpha = 0.6, width = 0.2, height = 0.2) +
  labs(
    title = "Relationship between Age, Installment Commitments, and Credit Risk Classification",
    x = "Age (years)",
    y = "Number of Installment Commitments",
    color = "Credit Risk Class"
    ) +
  theme_minimal() +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_line(color = "lightgray")
    ) +
  scale_color_manual(name = "Credit Risk Class", values = c("red", "darkgreen"))
```

Figure 70 & 71: Jittered Scatter Plot Snippet

The jittered plot above shows that aged 20-40 of individuals with high instalments fall under "bad" credit risk, indicating financial vulnerabilities. This highlights the need for financial literacy and supervision of high-commitment loans. Those above 40 have less commitments and are more stable.

```
# 1.2 Dot Plot: Visualize Marital Status by Installment Commitments and Credit Risk
ggplot(AData, aes(x = personal_status, y = installment_commitment, color = class)) +
  geom_jitter(width = 0.2, alpha = 0.6, height = 0.2) +
  labs(
    title = "Relationship between Marital Status, Installment Commitments, and Credit Risk Classification by Marital Status",
    x = "Marital Status",
    y = "Number of Installment Commitments",
    color = "Credit Risk Class"
    ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold"),
    panel.grid.major = element_line(color = "lightgray"),
    panel.grid.minor = element_line(color = "lightgray")
    ) +
  scale_color_manual(name = "Credit Risk Class", values = c("red", "darkgreen"))
```

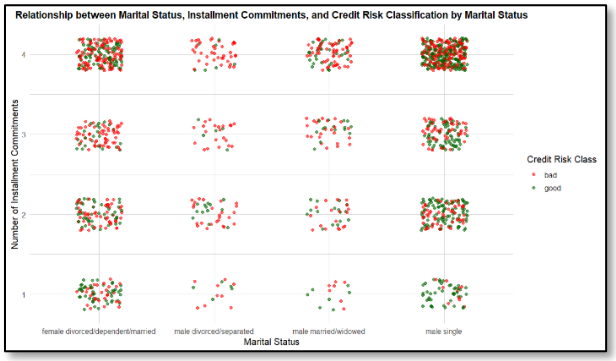
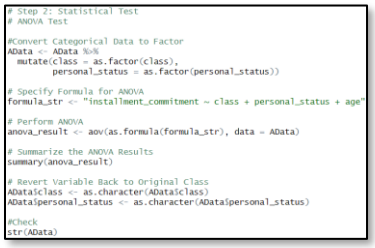


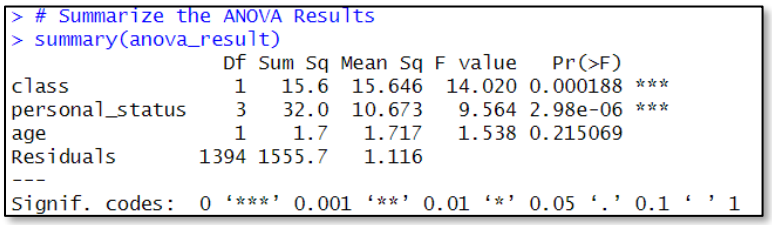
Figure 72 & 73: Dot Plot Snippet

The dot plot shows that males who are divorced or separated are more susceptible to "bad" credit risk classifications, while females exhibit better credit behaviour at lower commitments, and single males display balanced credit behaviour, but "bad" risks spike at the fourth instalment.

Step 2: Statistical Test



```
# Step 2: Statistical Test
# ANOVA Test

#Convert Categorical Data to Factor
AData <- AData %>%
  mutate(class = as.factor(class),
         personal_status = as.factor(personal_status))

# Specify Formula for ANOVA
formula_str <- "installment_commitment ~ class + personal_status + age"

# Perform ANOVA
anova_result <- aov(as.formula(formula_str), data = AData)

# Summarize the ANOVA Results
summary(anova_result)

# Revert Variable Back to Original Class
AData$class <- as.character(AData$class)
AData$personal_status <- as.character(AData$personal_status)

#Check
str(AData)
```



```
> # Summarize the ANOVA Results
> summary(anova_result)
                Df Sum Sq Mean Sq F value   Pr(>F)
class            1   15.6  15.646  14.020 0.000188 ***
personal_status  3   32.0  10.673   9.564 2.98e-06 ***
age              1    1.7   1.717   1.538 0.215069
Residuals     1394 1555.7   1.116
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Figure 74 & 75: ANOVA Test

The ANOVA test revealed that class significantly influences instalment commitments, while marital/personal status has a low impact. Age does not directly moderate the relationship between instalment commitments and credit risk classification, with a p-value of 0.215069 over 0.05. These findings suggest that age does not directly influence instalment commitments.

Conclusion

The study reveals that **personal/marital status moderates the relationship** between instalment commitments and credit risk classification. Groups like divorced/separated males face the highest risk, while single males, females, and married men vary in risk patterns. **Age doesn't directly moderate** this relationship but can influence risk classifications if it interacts with instalment commitments and marital status. Younger individuals are more vulnerable to "bad" credit risks with higher commitments, while older individuals show better financial stability.

3.3.5 How Does The Duration Of Residence Impact The Relationship Between Financial Commitments And Credit Risk Classification?

Step 1: Data Visualization

```
# 3.3 Objective 3: To Assess the Role of Financial Commitments on Credit Risk Classification.
# 3.3.5 How Does The Duration Of Residence Impact The Relationship Between Financial Commitments And Credit Risk Classification?

library(dplyr)

# Step 1: Data Visualization
# 1.1 Tile Plot: Relationship Between Residence Duration, Installment Commitments, and Credit Risk Class

# Summarize Data
tile_data <- AData %>%
  group_by(residence_since, class) %>%
  summarise(mean_commitment = mean(installment_commitment, na.rm = TRUE)) %>%
  ungroup()

# Generate Tile Plot
ggplot(tile_data, aes(x = factor(residence_since), y = class, fill = mean_commitment)) +
  geom_tile(color = "white", size = 0.5) +
  scale_fill_gradient(low = "lightblue", high = "darkblue", name = "Mean Commitment") +
  labs(
    title = "Tile Plot of Mean Financial Commitments by Residence Duration and Credit Risk Class",
    x = "Duration of Residence (years)",
    y = "Credit Risk Class"
    ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold")
    ) +
geom_text(aes(label = round(mean_commitment, 1)), color = "white", size = 4)  # Mean Commitment Labels
```

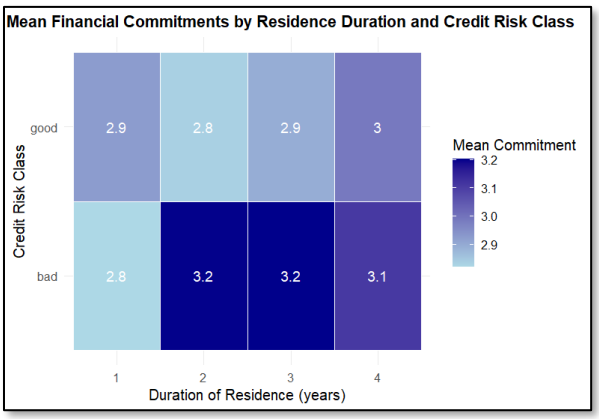


Figure 76 & 77: Tile Plot Snippet

The tile plot above summarizes the mean instalment commitments between residence duration and credit risk class. The tile plot shows that individuals with "bad" credit risk tends to have a higher mean of instalment commitments, specifically with the residence duration of 2 or 3 years.

```
# 1.2 Violin Plot: Distribution of Installment Commitments by Residence Duration and Credit Risk Class
ggplot(AData, aes(x = factor(residence_since), y = installment_commitment, fill = class)) +
  geom_violin(trim = FALSE, alpha = 0.6) +
  geom_boxplot(width = 0.1, position = position_dodge(0.9), outlier.shape = NA, color = "black") +
  scale_fill_manual(values = c("good" = "darkblue", "bad" = "red")) +
  labs(
    title = "Violin Plot of Financial Commitments by Residence Duration and Credit Risk Class",
    x = "Duration of Residence (years)",
    y = "Number of Installment Commitment",
    fill = "Credit Risk Class"
    ) +
  theme_minimal(base_size = 12) +
  theme(
    plot.title = element_text(hjust = 0.5, size = 13, face = "bold")
    )
```

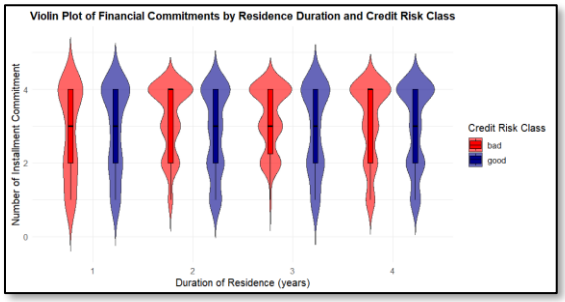


Figure 78 & 79: Violin Plot Snippet

In the violin plot above, individuals that falls under the "bad" credit risk class tend to have a wider and higher range of instalment commitments than "good" credit risk individuals throughout all residence durations years. A prominent difference can be seen clearly in year 2 and 3 of the residence duration. A nearly identical shape of violin can be seen in duration 1 and 4 years.

Step 2: Statistical Test

```
> # Step 2: Statistical Test
> # Kruskal-Wallis Test ( Compare Distributions Among Multiple Groups )
> kruskal_result <- kruskal.test(
+   installment_commitment ~ interaction(residence_since, class),
+   data = AData
+ )
> print(kruskal_result)

        Kruskal-Wallis rank sum test

data:  installment_commitment by interaction(residence_since, class)
Kruskal-Wallis chi-squared = 18.126, df = 7, p-value = 0.01141
```

Figure 80: Kruskal-Wallis Test

The Kruskal-Wallis's test was used to compare the distributions of residence duration, instalment commitments, and credit risk classification across multiple groups. The results showed a significant difference (p-value = 0.01141) at a 5% significance level, indicating that the combination of residence duration and credit risk classification significantly influences the distribution of instalment commitments.

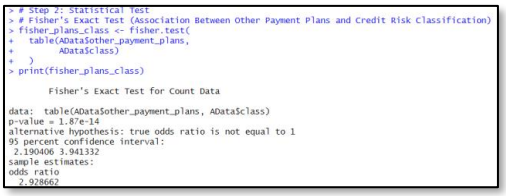
Conclusion

The study reveals that individuals with short residence durations maintain small instalment commitments, regardless of their credit risk class. Residence duration significantly influences instalment commitments, with the "bad" class showing higher commitments at 2 and 3 years. In the 4 years duration category, commitments slightly decrease, suggesting individuals in the "bad" credit risk class are trying to stabilize their instalment commitment amount. This analysis can help financial institutions design targeted interventions based on residence duration, making it a crucial factor in credit risk classification analysis.

3.3.6 Additional Features

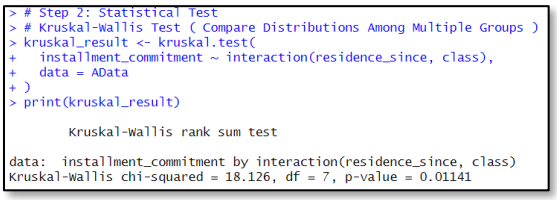
From the analysis done above, several additional features were used and added in order to support the analysis process. Here are some of the additional features used:

3.3.6.1 Fisher's Exact Test



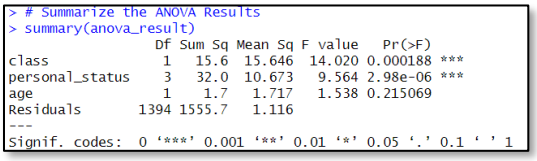
A Fisher's Exact test was used to determine if there is a significant relationship between two categorical variables, especially for small sample size. In the figure above, is used to determine the relationship between other payment plans and credit risk classes as they both are categorical (chr) variables.

3.3.6.2 Kruskal-Wallis Test



A Kruskal-Wallis test was used for the statistical test in the fifth analysis. This test is a non-parametric test for determining whether two or more groups' medians are from the same distribution. In the figure above, the test was used to compare the instalment commitments across different credit risk groups while also taking residence duration as one of the influences.

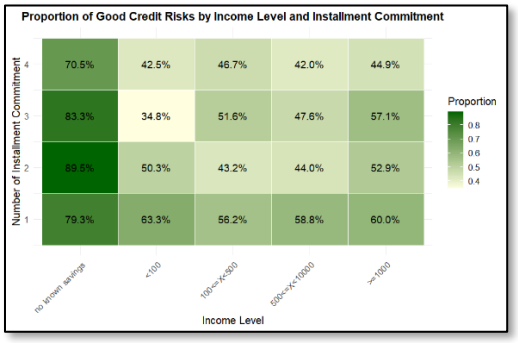
3.3.6.3 ANOVA (Analysis of Variance) Test



An ANOVA test was used to compare the means of three or more groups, assuming identical variances and normal distribution. To avoid incorrect results, variables must be converted

to categorical grouping variables. The test was used to assess differences between demographic factors like age and personal status across credit risk classes.

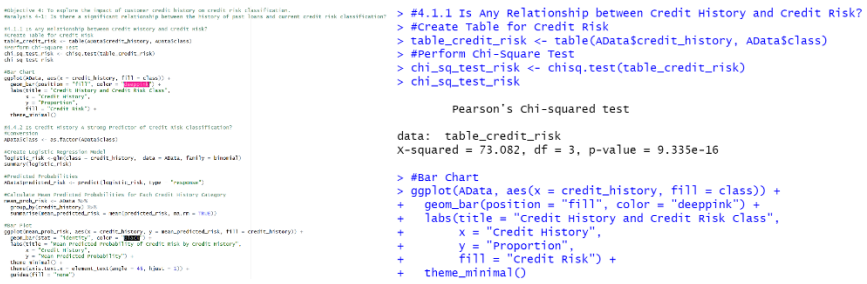
3.3.6.4 Heatmap



A heatmap is used to serve as a graphical representation of data in which the intensity of each hue represents a distinct value. A heatmap is often used to visualize correlations, associations, or patterns in complicated datasets. The key feature of a heatmap is the hue indicator where the value is differentiated by the intensity of that hue.

3.4 Objective 4: To explore the impact of customer credit history on credit risk classification. - CHONG ZHI YUE (TP067869)

3.4.1 Is there a significant relationship between the history of past loans and current credit risk classification?



```
> #4.1.1 Is Any Relationship between Credit History and Credit Risk?
> #Create Table for Credit Risk
> table_credit_risk <- table(ADataScredit_history, ADataSclass)
> #Perform Chi-Square Test
> chi_sq_test_risk <- chisq.test(table_credit_risk)
> chi_sq_test_risk

        Pearson's Chi-squared test

data:  table_credit_risk
X-squared = 73.082, df = 3, p-value = 9.335e-16

> #Bar Chart
> ggplot(AData, aes(x = credit_history, fill = class)) +
+   geom_bar(position = "fill", color = "deeppink") +
+   labs(title = "Credit History and Credit Risk Class",
+        x = "Credit History",
+        y = "Proportion",
+        fill = "Credit Risk") +
+   theme_minimal()
```

Figure 81 & 82: Analysis 4.1

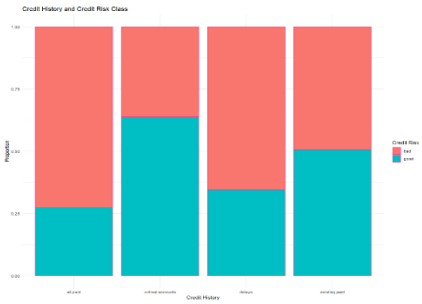


Figure 83: Create Bar Chart

Based on the chi-squared test of independence between the credit history and class variables in the contingency table, it shows that a p-value of 9.335e – 16 greatly below the significance level of 0.05. As a result, there is a highly significant relationship between credit history and credit class, we can reject the null hypothesis.

Categories for all paid and existing paid have a higher proportion of individuals with good credit risk; for critical accounts and delays have a higher proportion of individuals with bad credit risk. Customers who have always paid their bills on time and have a history of paying off existing accounts are good credit risk. Those with critical accounts, signifying overdue or missed payments, are strongly likely to be high-risk borrowers.

```
> #4.4.2 Is Credit History A Strong Predictor of Credit Risk Classification?
> #Conversion
> AData$class <- as.factor(AData$class)
>
> #Create Logistic Regression Model
> logistic_risk <-glm(class ~ credit_history,  data = AData, family = binomial)
> summary(logistic_risk)

Call:
glm(formula = class ~ credit_history, family = binomial, data = AData)

Coefficients:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                  -0.9808     0.1954  -5.019 5.20e-07 ***
credit_historycritical accounts   1.5466     0.2226   6.948 3.71e-12 ***
credit_historydelays          0.3390     0.2523   1.344    0.179
credit_historyexisting paid       1.0061     0.2093   4.807 1.53e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1940.8  on 1399  degrees of freedom
Residual deviance: 1866.0  on 1396  degrees of freedom
AIC: 1874

Number of Fisher Scoring iterations: 4
```

Figure 84: Testing Credit History Is a Strong Predictor or Not

The p-values for critical accounts, delays, and existing paid are all relevant to the 0.01 level. This shows that we can be certain that these credit history categories do affect the probability of being in the bad credit. When credit history information for a new customer is provided, the logistic model can be estimated the probability of the customer being a high-risk borrower. The probability can be used to inform decisions like approving or rejecting a loan application.

Low p-values and high coefficients show that critical accounts and already paid accounts have a large and meaningful impact on predicting credit risk. Although delays have some effect, they are not statistically significant, which suggests that they may not be a reliable indicator of credit risk on their own.

```
> #Predicted Probabilities
> AData$predicted_risk <- predict(logistic_risk, type = "response")
>
> #Calculate Mean Predicted Probabilities for Each Credit History Category
> mean_prob_risk <- AData %>%
+   group_by(credit_history) %>%
+   summarise(mean_predicted_risk = mean(predicted_risk, na.rm = TRUE))
>
> #Bar Plot
> ggplot(mean_prob_risk, aes(x = credit_history, y = mean_predicted_risk, fill = credit_history)) +
+   geom_bar(stat = "identity", color = "black") +
+   labs(title = "Mean Predicted Probability of Credit Risk by Credit History",
+        x = "Credit History",
+        y = "Mean Predicted Probability") +
+   theme_minimal() +
+   theme(axis.text.x = element_text(angle = 45, hjust = 1)) +
+   guides(fill = "none")
```

Figure 85: Calculation of Mean Predicted Probabilities & Create Bar Plot

predicted_risk
0.6377953
0.5063114
0.6377953
0.5063114
0.3448276
0.5063114
0.5063114
0.5063114
0.5063114
0.5063114
0.6377953
0.5063114
0.5063114
0.5063114
0.6377953
0.5063114
0.5063114
0.6377953

Figure 86: New Column for Predicted Probabilities of Credit Risk

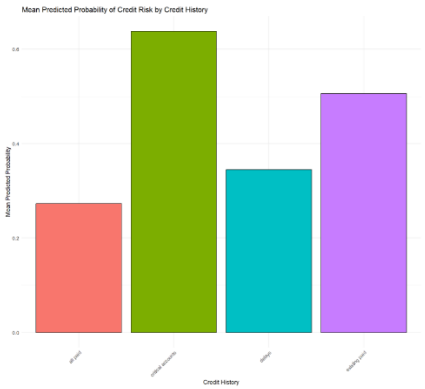


Figure 87: Bar Plot

From the bar chart above, the critical accounts categories have the highest predicted probability of credit risk. All paid categories have the lowest predicted probability, indicating that individuals with no prior credit issues are the least likely to default. They would require consideration of additional factors that could affect reputation such as the duration of payment history.

3.4.2 Does the number of previous loan defaults predict the likelihood of being classified as a high-risk customer?

```
#analysis 4.2: Does the number of previous loan defaults predict the likelihood of being classified as a high-risk customer?
#Conversion
ADataSclass <-  as.factor(ADataSclass)

#Logistic Regression Model
logistic_defaults <- glm(Class ~ credit_history, data = ADataa, family = binomial)
summary(logistic_defaults)

#Predicted Probabilities
ADataSpredicted_risk_defaults <- predict(logistic_defaults, type = "response")

#Calculate Mean Predicted Probabilities for Each Number of Defaults
mean_prob_defaults <- ADataa %>%
  group_by(credit_history) %>%
  summarise(mean_predicted_risk = mean(predicted_risk_defaults, na.rm = TRUE))

#Line Plot for Predicted Probabilities
ggplot(mean_prob_defaults, aes(x = credit_history, y = mean_predicted_risk, group = 1)) +
  geom_line(color = "blue", linewidth = 1) +
  geom_point(color = "red", size = 2) +
  labs(title = "Mean Predicted Probability of High Risk By Credit History",
       x = "Number of Previous Defaults",
       y = "Mean Predicted Probability") +
  theme_minimal()

#Violin Plot
ggplot(ADataa, aes(x = as.factor(credit_history), y = predicted_risk_defaults, fill = as.factor(credit_history))) +
  geom_violin(trim = FALSE) +
  labs(title = "Violin Plot of Predicted Probability by Credit History",
       x = "Number of Previous Defaults",
       y = "Predicted Probability") +
  theme_minimal()
```

Figure 88: Analysis 4.2

predicted_risk_defaults
0.6377953
0.5063114
0.6377953
0.5063114
0.3448276
0.5063114
0.5063114
0.5063114
0.5063114
0.6377953
0.5063114
0.5063114
0.5063114
0.6377953
0.5063114
0.5063114
0.6377953

Figure 89: New Column for Predicted Risk Defaults

```
> #Predicted Probabilities
> AData$predicted_risk_defaults <- predict(logistic_defaults, type = "response")
>
> #Calculate Mean Predicted Probabilities for Each Number of Defaults
> mean_prob_defaults <- AData %>%
+   group_by(credit_history) %>%
+   summarise(mean_predicted_risk = mean(predicted_risk_defaults, na.rm = TRUE))
> #Line Plot for Predicted Probabilities
> ggplot(mean_prob_defaults, aes(x = credit_history, y = mean_predicted_risk, group = 1)) +
+   geom_line(color = "blue", linewidth = 1) +
+   geom_point(color = "red", size = 2) +
+   labs(title = "Mean Predicted Probability of High Risk By Credit History",
+        x = "Number of Previous Defaults",
+        y = "Mean Predicted Probabilit") +
+   theme_minimal()
```

Figure 90: Calculation of Mean Predicted Probabilities & Create Point Plot

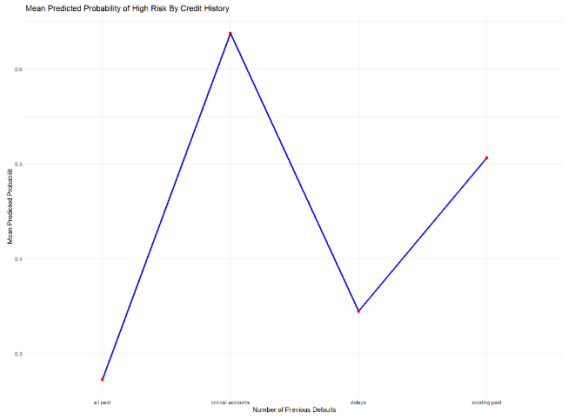


Figure 91: Point Plot

For the line plot graph, this shows the average predicted probability of high risk for each credit history group. Individuals with a history of critical accounts have the highest mean predicted probability of credit risk. In contrast, all paid accounts have the lowest risk among the other 3 categories.

```
> #Violin Plot
> ggplot(AData, aes(x = as.factor(credit_history), y = predicted_risk_defaults, fill = as.factor(credit_history))) +
+   geom_violin(trim = FALSE) +
+   labs(title = "Violin Plot of Predicted Probability by Credit History",
+        x = "Number of Previous Defaults",
+        y = "Predicted Probability") +
+   theme_minimal()
```

Figure 92: Create Violin Plot

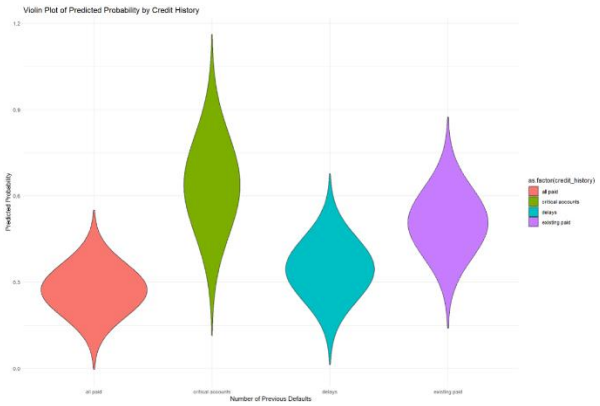


Figure 93: Violin Plot

In the violin plot, it shows that all paid group has a wider spread, with many individuals are predicted to have a low risk of bad credit and high predicted probabilities. The critical accounts have a narrow range centred around a more concentrated distribution of predicted probability, which means that they have higher predicted probabilities of bad risk.

3.4.3 How does the length of credit history impact the credit risk classification?

```
> #4.3 Length of Credit History and Credit Risk Classification
> #Conversion
> AData$duration <- as.numeric(as.character(AData$duration))
> #Convert Duration into Broader Categorical Group
> AData$duration <- cut(AData$duration,
+                       breaks = c(-Inf, 12, 24, 36, Inf), # Adjust ranges as needed
+                       labels = c("0-12", "13-24", "25-36", "36+"),
+                       right = TRUE)
> #Create Tables for Duration and Credit Class
> duration_class_table <- table(AData$duration, AData$class)
> duration_class_table

          0   1
  0-12  146 283
  13-24 295 289
  25-36 150  86
  36+   109  42
> #Chi-Square Test
> duration_chisq <- chisq.test(duration_class_table)
> duration_chisq

        Pearson's Chi-squared test

data:  duration_class_table
X-squared = 90.897, df = 3, p-value < 2.2e-16
```

Figure 94: Length of Credit History Impact Credit Risk Classification

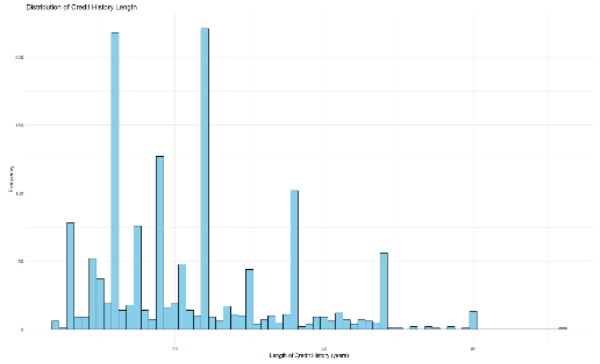


Figure 95: Histogram

The histogram clearly shows that most people in this dataset have a credit history between 20 and 25 years. On the histogram, we can observe a long tail, showing that some individuals have credit histories that are 40 years or more lengthy. A modest right-skewedness in the distribution shows that there are more people with shorter credit histories than those with longer ones.

```
> #Boxplot
> ggplot(AData, aes(x = class, y = duration, fill = class)) +
+   geom_boxplot() +
+   labs(title = "Credit History Length and Credit Risk Classification",
+        x = "Credit Class",
+        y = "Duration (Credit History Length)") +
+   theme_minimal()
```

Figure 96: Creating Boxplot

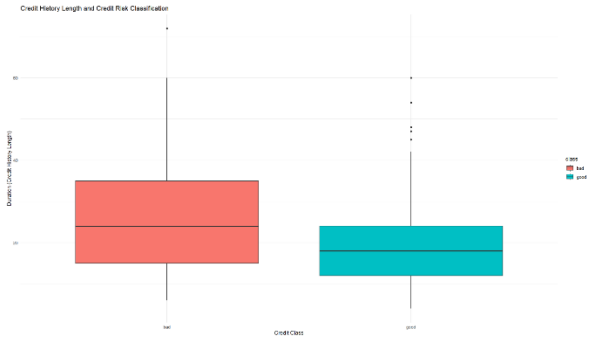


Figure 97: Boxplot

In comparison to people with good credit risk, those with bad credit risk typically have shorter median credit history lengths. Longer credit history is frequently linked to greater reliability. The boxplots for both good and negative credit risks show that the lengths of credit histories in each group range widely. The boxplot for bad credit risks is broader, showing that there is greater variance in the lengths of credit histories among those who have bad credit.

```
> #Convert duration into broader categorical group
> Adata$duration <- cut(Adata$duration,
+                       breaks = c(-Inf, 12, 24, 36, Inf), # Adjust ranges as needed
+                       labels = c("0-12", "13-24", "25-36", "36+"),
+                       right = TRUE)
> #Create tables for duration and credit class
> duration_class_table <- table(Adata$duration, Adata$class)
> duration_class_table

      bad good
  0-12  146  283
  13-24 295  286
  25-36 150   86
  36+   109   82
> #Fisher's Exact Test
> fisher_test_duration_sim <- fisher.test(duration_class_table, simulate.p.value = TRUE)
> fisher_test_duration_sim

        Fisher's Exact Test for Count Data with simulated p-value (based on 2000 replicates)

data:  duration_class_table
p-value = 0.0004998
alternative hypothesis: two.sided

> #Chi-Squared Test
> duration_chisq_sim <- chisq.test(duration_class_table, simulate.p.value = TRUE)
> duration_chisq_sim

        Pearson's Chi-squared test with simulated p-value (based on 2000 replicates)

data:  duration_class_table
X-squared = 90.867, df = NA, p-value = 0.0004998
```

Figure 98: Conversion, New Table, Fisher's Exact Test & Chi-Squared Test

After converting the variable into different categories, we would need to create a contingency table which can show the frequency counts of each combination of duration category and credit risk class. When sample sizes are limited, the fisher test is utilized to determine if two categorical variables are independent. In the chi-squared test, it reveals a strong relationship between credit risk classification and credit history duration, with both tests producing a p-value of 0.004998.

```
> #Conversion
> AData$duration <- as.factor(AData$duration)
> AData$class <- as.factor(AData$class)
> #Bar Chart
> ggplot(AData, aes(x = duration, fill = class)) +
+   geom_bar(position = "dodge") +
+   labs(title = "Credit History Length and Credit Risk",
+        x = "Duration",
+        y = "Count",
+        fill = "Credit Class") +
+   theme_minimal()
```

Figure 99: Conversion & Create Bar Chart

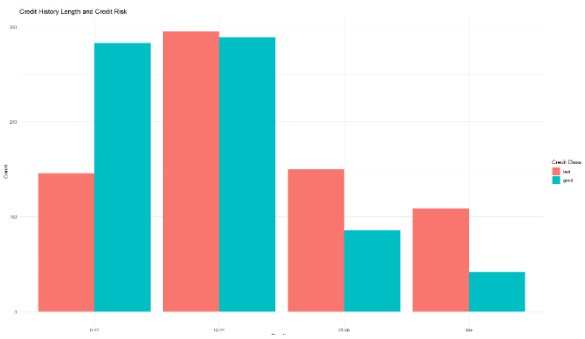


Figure 100: Bar Chart

This bar chart represents the distribution of good and bad credit classes across different duration categories. The result shows that the longer the credit histories, the lower the credit risk that might be associated with. From the graph for the category of 25 – 36 months and 36+ months, the proportion of bad classes decrease, the longer credit histories will associate with lower risk as the duration increases.

Additional Features – Showing Percentage in Graph

```
> #Calculate Percentages for Duration and Class
> duration_class_df <- as.data.frame(duration_class_table)
> colnames(duration_class_df) <- c("Duration", "Class", "Count")
>
> duration_class_df <- duration_class_df %>%
+   group_by(Duration) %>%
+   mutate(Percentage = Count / sum(Count) * 100)
> #Bar Chart with Percentages
> ggplot(duration_class_df, aes(x = Duration, y = Count, fill = Class)) +
+   geom_bar(stat = "identity") +
+   geom_text(aes(label = paste0(round(Percentage, 1), "%")),
+             position = position_stack(vjust = 0.5))+
+   labs(title = "Credit History Length and Risk Classification",
+        x = "Credit History Length (Duration)",
+        y = "Count",
+        fill = "Credit  Class") +
+   theme_minimal()
```

Figure 101: Calculation Percentages & Create Bar Chart with Percentages

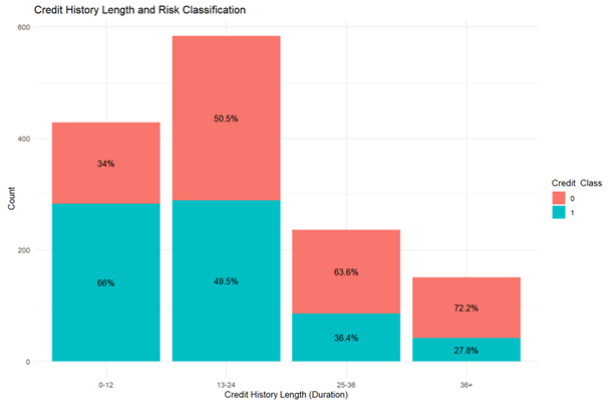
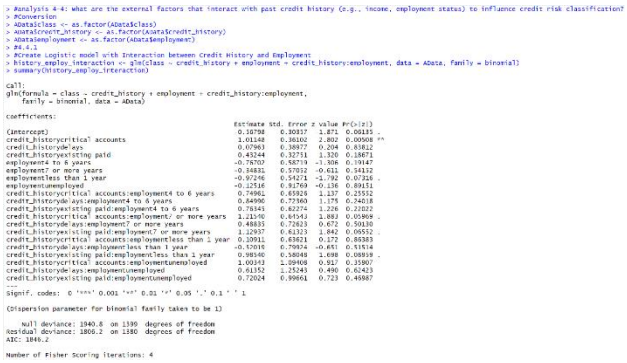


Figure 102: Bar Chart with Percentages

Based on the graph, it shows that:

- For 0-12 months duration, most individuals are classified as high-risk, making up 66% of the total for becoming risky.
- For 36+ months duration, the longest credit histories have a higher percentage of non-risky loans, with 72.2% of individuals classified as “Class 0” (non-risky).

3.4.4 What are the external factors that interact with past credit history (e.g., income, employment status) to influence credit risk classification?

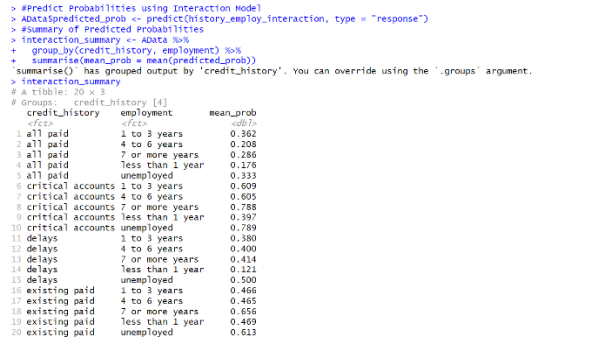


```
> #analysis 4-4: what are the external factors that interact with past credit history (e.g., income, employment status) to influence credit risk classification?
> #Conversion
> AData$class <- as.factor(AData$class)
> AData$credit_history <- as.factor(AData$credit_history)
> AData$employment <- as.factor(AData$employment)
> ##.4.1
> #Create Logistic model with Interaction between Credit History and Employment
> history_employ_interaction <- glm(class ~ credit_history + employment + credit_history:employment, data = ADataa, family = binomial)
> summary(history_employ_interaction)

Call:
glm(formula = class ~ credit_history + employment + credit_history:employment,
    family = binomial, data = ADataa)

Coefficients:
                                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                                        0.10798    0.30037   1.871  0.06185 .
credit_historycritical accounts                    1.01148    0.36102   2.802  0.00508 **
credit_historydelays                               0.07963    0.38977   0.204  0.83812
credit_historyexisting paid                        0.43244    0.32711   1.326  0.18471
employment4 to 6 years                            -0.76707    0.58719  -1.306  0.19147
employment7 or more years                         -0.34831    0.57032  -0.611  0.54152
employmentless than 1 year                        -0.97246    0.54271  -1.792  0.07316 .
employmentunemployed                              -0.12516    0.91769  -0.136  0.89151
credit_historycritical accounts:employment4 to 6 years   0.74961    0.65926   1.137  0.25552
credit_historydelays:employment4 to 6 years       0.84990    0.72360   1.175  0.24018
credit_historyexisting paid:employment4 to 6 years   0.76543    0.60274   1.226  0.22022
credit_historycritical accounts:employment7 or more years   1.21590    0.65543   1.882  0.05969 .
credit_historydelays:employment7 or more years    0.48835    0.72823   0.672  0.50130
credit_historyexisting paid:employment7 or more years   1.12937    0.61323   1.842  0.06552 .
credit_historycritical accounts:employmentless than 1 year  0.10911    0.63621   0.172  0.86383
credit_historydelays:employmentless than 1 year   -0.52018    0.79924  -0.651  0.51514
credit_historyexisting paid:employmentless than 1 year   0.98540    0.58048   1.698  0.08939 .
credit_historycritical accounts:employmentunemployed   1.00343    1.09408   0.917  0.35907
credit_historydelays:employmentunemployed          0.61152    1.25243   0.490  0.62423
credit_historyexisting paid:employmentunemployed   0.72024    0.99661   0.723  0.46987
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1940.8  on 1399  degrees of freedom
Residual deviance: 1806.2  on 1380  degrees of freedom
AIC: 1846.2

Number of Fisher Scoring iterations: 4
```

Figure 103: Conversion & Analysis 4.4.1

The probability of default compared to the reference category to the reference category, which is most likely to all paid, is displayed by the coefficients for the different credit history categories. For instance, if employment status maintains the same, a positive coefficient for critical accounts indicates that those with critical accounts are more likely to default than those with all paid accounts.

For credit history, those who have always made their payments on time are less likely to have bad credit than those who have a history of critical accounts and paid accounts.Individuals with credit histories of 4 to 6 years or less than a year are less likely to have bad credit compared to those with a credit history of 7 or more years. While unemployment might slightly decrease the odds of bad credit, this effect is not statistically significant.



```
> #Predict Probabilities using Interaction Model
> AData$predicted_prob <- predict(history_employ_interaction, type = "response")
> #Summary of Predicted Probabilities
> interaction_summary <- ADataa %>%
+   group_by(credit_history, employment) %>%
+   summarise(mean_prob = mean(predicted_prob))
`summarise()` has grouped output by 'credit_history'. You can override using the `.groups` argument.
> interaction_summary
# A tibble: 20 × 3
# Groups:   credit_history [4]
   credit_history     employment       mean_prob
   <fct>              <fct>                <dbl>
 1 all paid           1 to 3 years         0.362
 2 all paid           4 to 6 years         0.208
 3 all paid           7 or more years      0.286
 4 all paid           less than 1 year     0.176
 5 all paid           unemployed           0.333
 6 critical accounts  1 to 3 years         0.609
 7 critical accounts  4 to 6 years         0.605
 8 critical accounts  7 or more years      0.788
 9 critical accounts  less than 1 year     0.397
10 critical accounts  unemployed           0.789
11 delays             1 to 3 years         0.380
12 delays             4 to 6 years         0.400
13 delays             7 or more years      0.414
14 delays             less than 1 year     0.121
15 delays             unemployed           0.500
16 existing paid      1 to 3 years         0.466
17 existing paid      4 to 6 years         0.465
18 existing paid      7 or more years      0.656
19 existing paid      less than 1 year     0.469
20 existing paid      unemployed           0.613
```

Figure 104: Interaction Model for Predicted Probabilities

predicted_prob
0.1764706
0.2857143
0.3617021
0.3617021
0.2857143
0.3617021
0.3617021
0.2083333
0.1764706
0.2857143
0.2083333
0.2857143
0.2857143
0.3333333
0.3617021
0.2857143
0.3617021

Figure 105: New Column for Predicted Probabilities

Customers with a history of critical accounts are at the highest risk of bad credit, regardless of their employment status. Those with a history of delays also face a significant risk, albeit lower than the critical accounts.

Customers with existing paid accounts have a lower risk, and those with all paid accounts have the lowest risk of bad credit. Unemployed individuals and those with less than a year of employment are at a higher risk of bad credit. On the other hand, those with 4 to 6 years or 7 or more years of employment are at a lower risk. Therefore, credit risk is greatly influenced by the mix of work position and credit history.

```
> #Bar Plot between Credit History and Employment Status on Predicted Probabilities
> ggplot(interaction_summary, aes(x = credit_history, y = mean_prob, fill = employment)) +
+   geom_bar(stat = "identity", position = "dodge") +
+   geom_text(aes(label = round(mean_prob, 2)),
+             position = position_dodge(width = 0.9), vjust = -0.5) +
+   labs(title = "Interaction Effect between Credit History and Employment on Credit Risk",
+        x = "Credit History",
+        y = "Mean Predicted Probability",
+        fill = "Employment") +
+   theme_minimal()
```

Figure 106: Create Bar Plot

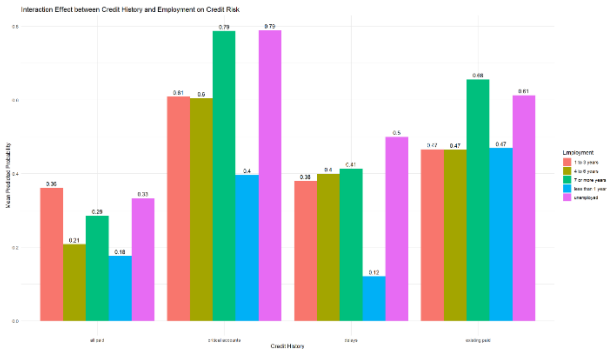


Figure 107: Bar Plot

From the observation on the graph above, critical accounts category has the highest mean predicted mean probabilities across all employment statuses in credit history effect. A less risky profile is suggested by the typically lower expected probabilities for people having a history of existing paid accounts. Within each credit history category, the bars' different heights show the interaction impact. For instance, the impact of having critical accounts is greater among those without jobs than for those with higher employment opportunities.

```
> #4.4.2
> #Interaction between Age Group and Credit Class
> ADATA <- ADATA %>%
+   mutate(age_group = case_when(
+     age <= 30 ~ "< 30",
+     age > 30 & age <= 50 ~ "30-50",
+     age > 50 ~ "50 more"
+   ))
> #Create Tables for Age Group and Credit Class
> age_class_table <- table(ADATA$age_group, ADATA$class)
> age_class_table

          0   1
  < 30    356 263
  30-50   292 355
  50 more  52  82
> #Chi-Square Test
> age_chisq <- chisq.test(age_class_table)
> age_chisq

        Pearson's Chi-squared test

data:  age_class_table
X-squared = 26.823, df = 2, p-value = 1.498e-06
```

Figure 108: Analysis 4.4.2

age_group
50 more
< 30
30-50
30-50
50 more
30-50
50 more
30-50
50 more
< 30
< 30
< 30
< 30
50 more
< 30
30-50
50 more
< 30
30-50
30-50
30-50

Figure 109: New Column for Age Group

Since the p-value of 1.498e-06 is smaller than the significance level of 0.05, we can reject the null hypothesis and assumed that there is a statistically significant association between age group and credit class.

```
> #Bar Plot between Age Group and Credit Class
> ggplot(AData, aes(x = age_group, fill = class)) +
+   geom_bar(position = "dodge") +
+   labs(title = "Age Group and Credit Risk Class",
+        x = "Age Group",
+        y = "Count") +
+   theme_minimal()
```

Figure 110: Create Bar Plot

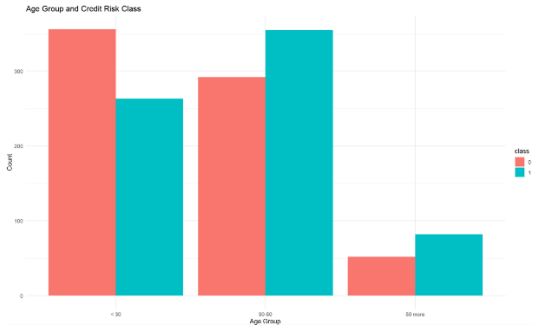


Figure 111: Bar Chart

Most of the individuals in the data set are between 30 and 50 years old, with a smaller number being younger than 30 and an even smaller number being older than 50. While older people, especially those who are over 50 often have a stronger credit history, young people, especially those under 30, are more likely to have bad credit.

```
> #4.4.3
> #Relationship between Saving Status and Employment
> # Grouping savings_status
> AData$savings_status <- recode(AData$savings_status,
+                                "<100" = "<500",
+                                "100<=X<500" = "<500",
+                                "500<=X<1000" = ">=500",
+                                ">=1000" = ">=500",
+                                "no known savings" = "no known savings")
> # Grouping employment
> AData$employment <- recode(AData$employment,
+                            "<1" = "short_term",
+                            "1<=X<4" = "mid_term",
+                            "4<=X<7" = "long_term",
+                            ">=7" = "long_term",
+                            "unemployed" = "unemployed")
```

Figure 112: Grouping

Since the p-value in the Chi-Square test is smaller, which is 0.001013, we can reject the null hypothesis. Therefore, the distribution of saving statuses is not related to employment status.

```
> #Bar Plot between Savings Status and Employment
> ggplot(AData, aes(x = savings_status, fill = employment)) +
+   geom_bar(position = "dodge") +
+   labs(title = "Savings Status and Employment",
+        X = "Savings Status",
+        y = "Count") +
+   theme_minimal()
```

Figure 113: Create Bar Chart

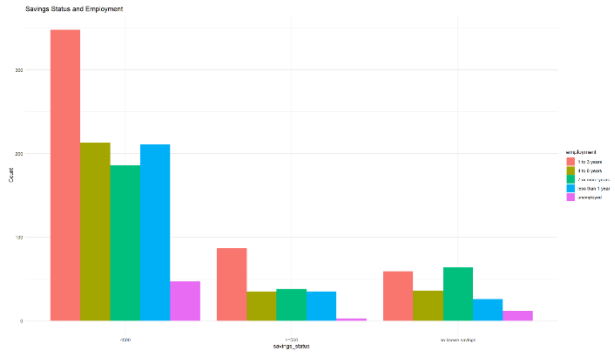


Figure 114: Bar Chart

Each bar represents a specific saving status category, such as less than 500, 500 or more, and no known savings. Those who save less, which are less than 500, often have shorter job histories, which might mean they are less financially stable or earn less money. In contrast, those who save more might have longer job histories, which are 500 or more than 500, showing that they are more financially secure and earn more money.

```
#4.4.4
#Boxplot between Credit Amount and Credit Class
ggplot(AData, aes(x = class, y = credit_amount, fill = class)) +
  geom_boxplot() +
  labs(title = "Credit Amount by Credit Risk Class",
       x = "Credit Class",
       y = "Credit Amount") +
  theme_minimal()
```

Figure 115: Create Boxplot

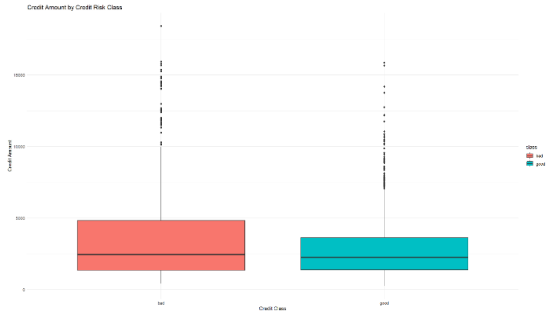


Figure 116: Boxplot

Class 0 has a larger median credit amount, which probably indicates good credit, compared to class 1, which probably indicates bad credit. This suggests that people who represent lesser credit risk are typically granted more credit. The distribution of credit amounts within each class is shown in the boxplots. The box displays the middle 50% of the data, or the interquartile range (IQR). The IQR is shown by the box and includes the middle 50% of the data.

Additional Features – Density Plot

```
> #Density plot for Credit Amount
> ggplot(AData, aes(x = credit_amount)) +
+   geom_density(alpha = 0.3) +
+   labs(title = "Density Plot for Credit Amount by Credit Class",
+        x = "Credit Amount",
+        y = "Density") +
+   theme_minimal()
```

Figure 117: Creating Density Plot

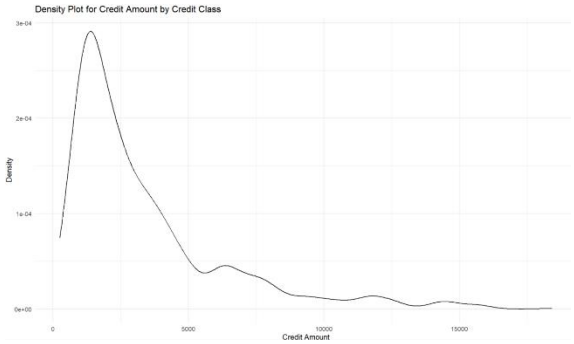


Figure 118: Density Plot

From the graph above, the density plot shows that several credit amounts fall within this range with a peak at a credit amount of around 2000 dollars. With a few big credit amounts setting the tail to the right, the distribution is skewed to the right. The mode, or the most typical credit amount is represented by the peak of the curve.

3.4.5 Is there any relationship between the purpose of having the loan and the credit risk classification?

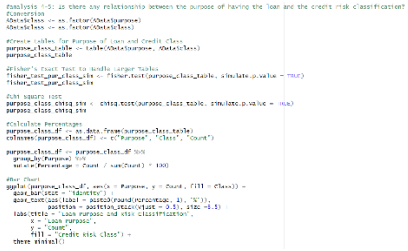


Figure 119: Analysis 4.5

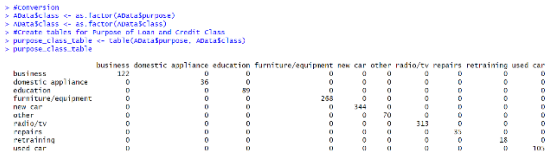


Figure 120: Conversion & Creating Table

To find developments, patterns, and possible risk factors, the analysis uses contingency tables and the conversion of categorical data to factors. The contingency tables show the frequency of each combination of purpose and credit class.

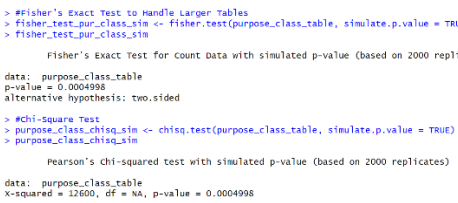


Figure 121: Fisher's Exact Test & Chi-Squared Test

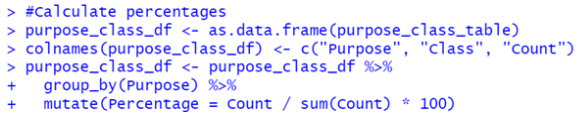


Figure 122: Calculation Percentages

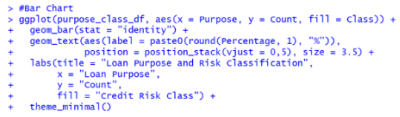


Figure 123: Creating Bar Chart

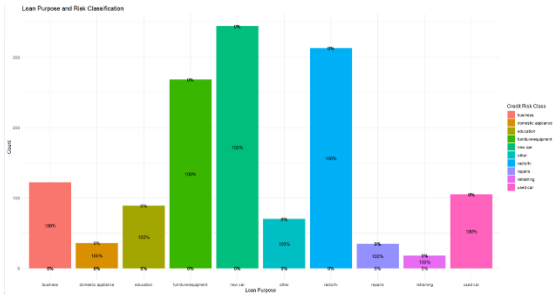


Figure 124: Bar Chart with Percentages

Based on the graph above, some loan purposes, such as new car and furniture equipment have a significantly higher number of loans compared to others loan purposes like retraining and repairs. For most loan purposes, most loans are categorized as good, and it means that a lot of loan purposes generally have a low default rate. Although the general pattern being positive, some loan purposes like repairs, retraining and used car have a comparatively greater proportion of bad loans, which linked to higher credit risk.

3.4.6 Is there any relationship between the categorical variables (e.g., credit history, purpose, and employment status)?

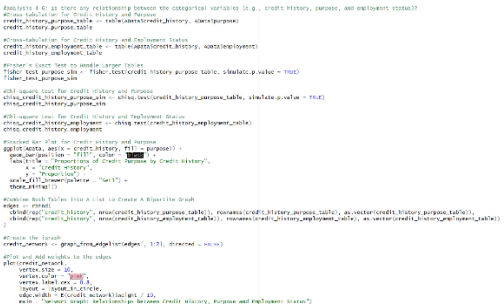


Figure 125: Analysis 4.6

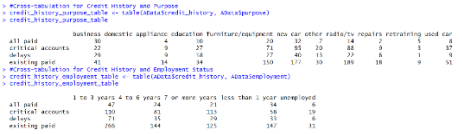


Figure 126: Cross-Tabulation

People who have good credit histories, like they always pay their bills on time, often apply for loans to buy cars or furniture. In contrast, people with bad credit histories, like always having overdue payments or not making payments, might apply for loans for riskier things like starting a business or retraining. For employment status, people with good credit histories often have stable jobs, while those with bad credit histories may be unemployed or have short job histories.

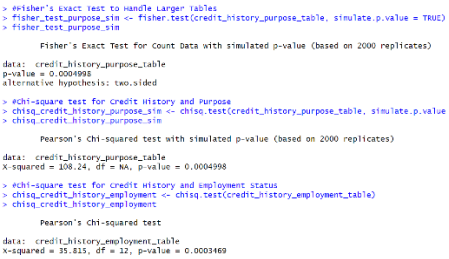


Figure 127: Fisher's Exact Test, Chi-Squared Test

The statistical test that was conducted included the fisher test and chi-squared test. In the fisher test, the p-value is 0.0004998, meaning that there is a statistically significant association between credit history and purpose. In the chi-squared test that is used to analyze the relationship between credit history and purpose, the p-value is 0.0004998, which is the same as the Fisher's Exact Test and reinforces the conclusion that they do have strong relationships with each other.

On the other hand, the p-value in the Chi-squared Test for analyzing the relationship between credit history and employment status, the p-value is 0.0003469.

Additional Features – Stacked Bar Plot

```
> #Stacked Bar Plot for Credit History and Purpose
> ggplot(AData, aes(x = credit_history, fill = purpose)) +
+   geom_bar(position = "fill", color = "black") +
+   labs(title = "Proportions of Credit Purpose by Credit History",
+        x = "Credit History",
+        y = "Proportion") +
+   scale_fill_brewer(palette = "Set3") +
+   theme_minimal()
```

Figure 128: Creating Stacked Bar Plot

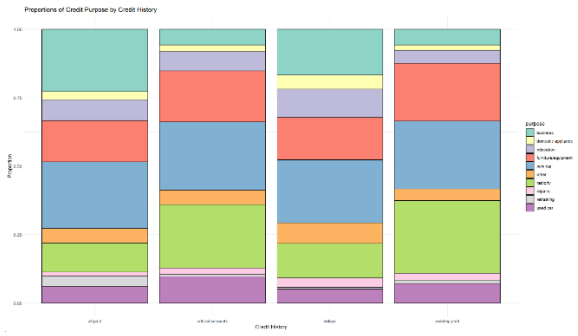


Figure 129: Stacked Bar Plot

In this diagram, it represents the distribution of different credit purposes across various credit history categories. For example, those with all paid credit history seems to have a higher proportion of loans for new car and radio/TV purposes compared to other groups. In contrast, individuals with delays in their credit history have a larger proportion of loans for repairs and retraining. Good credit holders, such as those who have consistently made their bill payments on time, are able to apply for loans for a variety of uses, including the purchase of furniture or cars. However, individuals with poor credit histories, such as those who have late payments, may only be eligible for loans for uses, and the conditions of these loans may be less advantageous.

Additional Features – Network Graph

```
> #Combine Both Tables into A List to Create A Bipartite Graph
> edges <- rbind(
+   cbind(rep("credit_history", nrow(credit_history_purpose_table)), rownames(credit_history_purpose_table), as.vector(credit_history_purpose_table)),
+   cbind(rep("credit_history", nrow(credit_history_employment_table)), rownames(credit_history_employment_table), as.vector(credit_history_employment_table))
+ )
> #Create the igraph
> credit_network <- graph_from_edgelist(edges[, 1:2], directed = FALSE)
> #Plot and Add weights to the edges
> plot(credit_network,
+      vertex.size = 10,
+      vertex.color = "pink",
+      vertex.label.cex = 0.8,
+      layout = layout_in_circle,
+      edge.width = E(credit_network)$weight / 10,
+      main = "Network Graph: Relationships between Credit History, Purpose and Employment Status")
```

Figure 130: Creating Network Graph

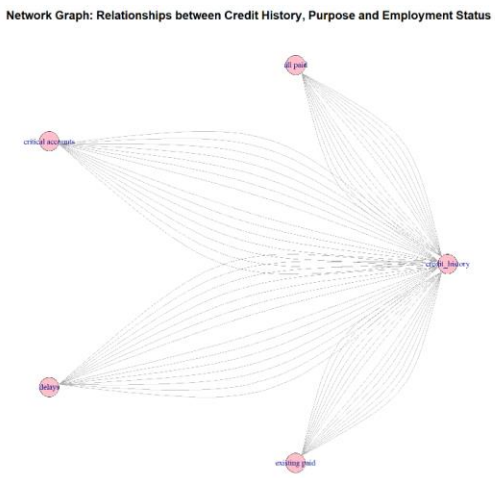


Figure 131: Network Graph

The graph clearly shows pink circles representing nodes, which correspond to categorical variables. The size of the nodes may reflect its relative frequency of connection to other nodes. The gray line, which is the edges representing or associations between the categories.

The graph highlights central nodes like existing paid, all paid which are heavily connected and suggests that they function as key hubs or important points in the network. In contrast, the peripheral nodes like critical accounts and delays are less connected as they may show lower interactions with other categories and highlighting the challenges faced by individuals with bad credit histories.

4.0 Conclusion

4.1 Results

In the analysis, all the main factors from the initial hypothesis were individually evaluated before being analysed collectively.

Demographic Factors: Analysis 3.1.1 shows that an individual's marital status affect their credit risk classification. From the visualization, the female group shares nearly an equal amount in both classes, although it is skewed to the "bad" risks more (54.9%). Meanwhile for males, those who or married or has separated can be seen with more "bad" risks (75.8% for separated males). While those males who are single, majority are still in "good" credit risks (61.5%). In analysis 3.1.2, job skills also do play a role though it's unlikely compared to other demographic factors as it shows little differences in both classes regardless of the individuals' skills. As for analysis 3.1.3 and 3.1.4, indicates the influence of being a foreign work, which does have a significance where foreign worker is mostly classified as "bad" risk, meanwhile gender has little potential with males being more classified as "good" risks.

Loan-Related Factors: Analysis 3.2.1 shows that the loan purpose does affect the credit risk classification with half of the individuals under the "bad" credit risks. The highest "bad" credit risk ratio on loan purpose is spent on vehicles (199 "bad", 145 "good"), furniture (145 "bad", 123 "good), other purposes (63 "bad", 7 "good"). Another loan-related factor that significantly impacts credit risks is the duration of the loan. From analysis 3.2.2, it shows that individuals with a loan duration around 12 to 36 months has a fair share of both credit classes, above 36 months individuals with "bad" credit risks starts to increase, and above 60 months gather a majority of "bad" than "good" credit risks individuals.

Instalment Commitments: Analysis 3.3.1 shows from the Welch's t-test the "bad" credit risk class has a much higher mean value (3.13), than the "good" class (2.92) which indicates that the "bad" credit risks have a higher average in instalment commitments, the low p–value also supports this conclusion. The visualizations also add up to the fact that individuals with a high number of instalment commitments (3 and 4) tends to fall into the "bad" credit risk class, suggesting financial strain within this category.

Credit History: Analysis 3.4.1 shows a chi-squared test done; a low p-value was received indicating the conclusion of the significance relationship. The visualization also shows that "good" credit risk individuals tends to have paid their credit beforehand, while those who have made a delay in their payment are more classified with "bad" risks (>50%).

When we analyse all the four main factors above, we can agree that the hypotheses are falls under acceptance as all these factors plays a strong role in determining credit risk classification. Despite some variables being not as strong as others (e.g., age is not a strong demographic factor to determine, but personal status is a strong factor) the entirety of the main factor still helps to determine the credit risk. For the first hypothesis, demographic factor plays a significant role similar to loan term and credit amount, although job status does not affect directly, it still helps to support the hypothesis if combined with the other factor.

For the second hypothesis, we can agree that this hypothesis falls under acceptance as higher credit amount do indicate "bad" credit risks proven from the analysis 3.2.1. Meanwhile, loan duration also proves to have an influence, where the longer a loan duration is, the more individuals is classified as "bad" credit risk indicated from analysis 3.3.2.

4.2 Recommendations

Based on the results from multiple analysis, our group recommends that banks should adopt a data-driven strategy to improve credit risk management by using customer data to create specific credit classification models. Real-time monitoring systems and advanced analytics techniques like AI can help identify complex patterns in customer data. Customized financial products and services can be provided by dividing the client base based on risk profiles, allowing banks to identify inconspicuous patterns. To reduce risks, banks should set higher standards for lending loans with larger amounts and longer durations and adjust policies for high-risk customers. A risk-based approach to lending is recommended, considering the variables that affect credit risk.

4.3 Limitations and Future Direction

This assignment faced limitations in its credit risk classification variables, including a limited range of relevant factors and time constraints. To improve accuracy and generalization, the analysis could have included a larger range of variables and more processing methods.

Additionally, a more in-depth examination of the associations between factors and their impact on credit risk could have improved the analysis. Despite these limitations, our group is satisfied with the analysis and plans to improve in future studies. The analysis provides valuable insights into data analysis and the importance of considering factors in credit risk classification.