GROUP ASSIGNMENT

CT127-3-2-PFDA

Programming for Data Analysis

INTAKE CODE: APU2F2409SE

LECTURER: DR. KULOTHUNKAN A/L PALASUNDRAM

DATE ASSIGNED: 3 OCTOBER 2024

DATE COMPLETED: 1 DECEMBER 2024

GROUP 18 MEMBERS:

DARYL SIM WEI SHERN (TP068964)

HO SHANE FOONG (TP068496)

CHEONG SHEUE LING (TP069004)

CHOO CHENG DA (TP068973)

JOHN HAR WEY JON (TP068348)

Table of Contents

1.0 Introduction

This assignment requires us to play a role of a data analyst who is working in the banking sector. We are provided with a dataset consisting of the credit risk classification along with the demographic variables and credit behaviors of bank customers who are applying for loans. The dataset contains factors such as age, gender, marital status, loan amount, credit history, loan duration, number of dependents, purpose of loan application, and more. Our task is to conduct data exploration, data cleaning and preprocessing, data analysis, data visualization, and hypothesis formation and testing to present valuable insights to the stakeholders.

1.1 Data Description

duration: This column represents the loan duration in months.

property_magnitude: This column represents the type of asset that is owned by the customer.

credit_amount: This column represents the loan amount that the customer is borrowing.

personal_status: This column represents the gender and marital status of the customer.

age: This column represents the customer's age in years.

1.2 Assumptions

1. Assume that every row of data represents an individual customer.
2. Assume that the first column in the dataset without a column name is the index of the row.
3. Assume that the loan duration is represented in months and expressed as an integer.
4. Assume that "no credits/all paid" in the credit_history column is classified as "all paid".
5. Assume that "radio/tv" in purpose column is classified as "domestic appliance".
6. Assume that the credit amount is expressed as an integer.
7. Assume that the installment commitment is expressed as an integer.
8. Assume that the residence period is measured in years and expressed as an integer.
9. Assume that the age of a customer is measured in years and expressed as an integer.
10. Assume that the blank values in the other payment plans column are missing values.

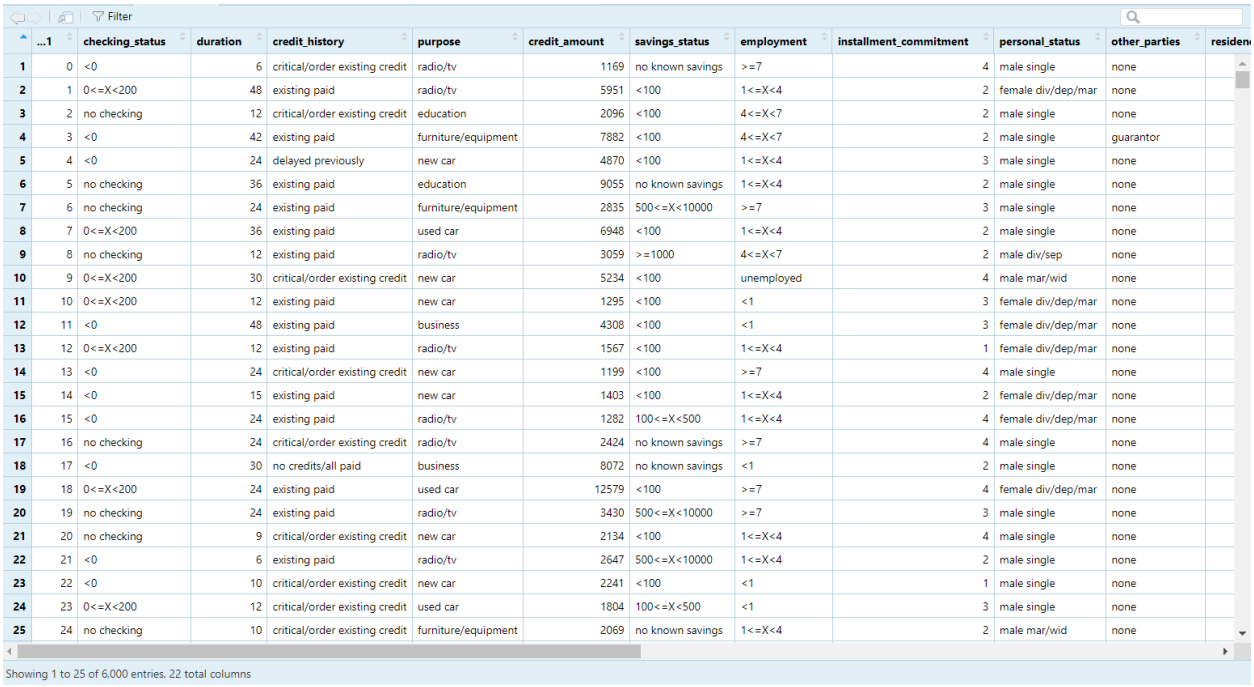
11. Assume that the existing credits are expressed as an integer.
12. Assume that the number of dependents is measured in persons and expressed as an integer.

2.0 Data Preparation

2.1 Data Import

```
# Load the libraries and packages
library(dplyr)
library(readr)
library(ggplot2)
library(DataExplorer)
library(missForest)
library(tidyverse)
library(VIM)
library(caret)
library(caTools)
library(randomForest)
library(vcd)
library(plotly)
library(readxl)
library(dplyr)
library(httr)
library(broom)

# Set the working directory
## Enter your own path
setwd("C:/APU/Degree/Semester 1/Programming for Data Analysis/Assignment/PFDA Assignment")
getwd()


# Import the dataset
## Enter your own file path
filePath = "C:/APU/Degree/Semester 1/Programming for Data Analysis/Assignment/PFDA Assignment/5. credit_risk_classification.csv"

## Read the CSV file into a dataframe
pfda_df = read_csv(filePath)
View(pfda_df)
```

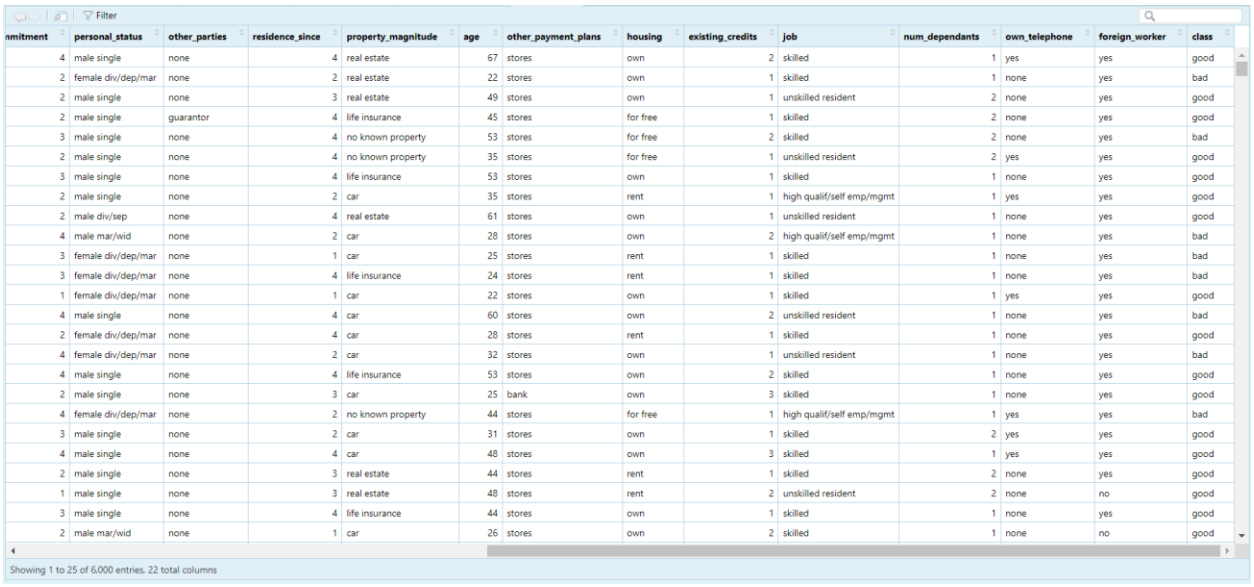
```
> pfda_df = read_csv(filePath)
New names:
• `` -> `...1`
Rows: 6000 Columns: 22
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
chr (14): checking_status, credit_history, purpose, savings_status, employment, personal_status, other_parties, property_magnit...
dbl  (8): ...1, duration, credit_amount, installment_commitment, residence_since, age, existing_credits, num_dependants

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
>
```



Filter

	...1	checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment	installment_commitment	personal_status	other_parties	residen
1	0	<0	6	critical/order existing credit	radio/tv	1169	no known savings	>=7	4	male single	none	
2	1	0<=X<200	48	existing paid	radio/tv	5951	<100	1<=X<4	2	female div/dep/mar	none	
3	2	no checking	12	critical/order existing credit	education	2096	<100	4<=X<7	2	male single	none	
4	3	<0	42	existing paid	furniture/equipment	7882	<100	4<=X<7	2	male single	guarantor	
5	4	<0	24	delayed previously	new car	4870	<100	1<=X<4	3	male single	none	
6	5	no checking	36	existing paid	education	9055	no known savings	1<=X<4	2	male single	none	
7	6	no checking	24	existing paid	furniture/equipment	2835	500<=X<10000	>=7	3	male single	none	
8	7	0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	2	male single	none	
9	8	no checking	12	existing paid	radio/tv	3059	>=1000	4<=X<7	2	male div/sep	none	
10	9	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	4	male mar/wid	none	
11	10	0<=X<200	12	existing paid	new car	1295	<100	<1	3	female div/dep/mar	none	
12	11	<0	48	existing paid	business	4308	<100	<1	3	female div/dep/mar	none	
13	12	0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	1	female div/dep/mar	none	
14	13	<0	24	critical/order existing credit	new car	1199	<100	>=7	4	male single	none	
15	14	<0	15	existing paid	new car	1403	<100	1<=X<4	2	female div/dep/mar	none	
16	15	<0	24	existing paid	radio/tv	1282	100<=X<500	1<=X<4	4	female div/dep/mar	none	
17	16	no checking	24	critical/order existing credit	radio/tv	2424	no known savings	>=7	4	male single	none	
18	17	<0	30	no credits/all paid	business	8072	no known savings	<1	2	male single	none	
19	18	0<=X<200	24	existing paid	used car	12579	<100	>=7	4	female div/dep/mar	none	
20	19	no checking	24	existing paid	radio/tv	3430	500<=X<10000	>=7	3	male single	none	
21	20	no checking	9	critical/order existing credit	new car	2134	<100	1<=X<4	4	male single	none	
22	21	<0	6	existing paid	radio/tv	2647	500<=X<10000	1<=X<4	2	male single	none	
23	22	<0	10	critical/order existing credit	new car	2241	<100	<1	1	male single	none	
24	23	0<=X<200	12	critical/order existing credit	used car	1804	100<=X<500	<1	3	male single	none	
25	24	no checking	10	critical/order existing credit	furniture/equipment	2069	no known savings	1<=X<4	2	male mar/wid	none	

Showing 1 to 25 of 6,000 entries, 22 total columns



Filter

nmitment	personal_status	other_parties	residence_since	property_magnitude	age	other_payment_plans	housing	existing_credits	job	num_dependants	own_telephone	foreign_worker	class
4	male single	none	4	real estate	67	stores	own	2	skilled	1	yes	yes	good
2	female div/dep/mar	none	2	real estate	22	stores	own	1	skilled	1	none	yes	bad
2	male single	none	3	real estate	49	stores	own	1	unskilled resident	2	none	yes	good
2	male single	guarantor	4	life insurance	45	stores	for free	1	skilled	2	none	yes	good
3	male single	none	4	no known property	53	stores	for free	2	skilled	2	none	yes	bad
2	male single	none	4	no known property	35	stores	for free	1	unskilled resident	2	yes	yes	good
3	male single	none	4	life insurance	53	stores	own	1	skilled	1	none	yes	good
2	male single	none	2	car	35	stores	rent	1	high qualif/self emp/mgmt	1	yes	yes	good
2	male div/sep	none	4	real estate	61	stores	own	1	unskilled resident	1	none	yes	good
4	male mar/wid	none	2	car	28	stores	own	2	high qualif/self emp/mgmt	1	none	yes	bad
3	female div/dep/mar	none	1	car	25	stores	rent	1	skilled	1	none	yes	bad
3	female div/dep/mar	none	4	life insurance	24	stores	rent	1	skilled	1	none	yes	bad
1	female div/dep/mar	none	1	car	22	stores	own	1	skilled	1	yes	yes	good
4	male single	none	4	car	60	stores	own	2	unskilled resident	1	none	yes	bad
2	female div/dep/mar	none	4	car	28	stores	rent	1	skilled	1	none	yes	good
4	female div/dep/mar	none	2	car	32	stores	own	1	unskilled resident	1	none	yes	bad
4	male single	none	4	life insurance	53	stores	own	2	skilled	1	none	yes	good
2	male single	none	3	car	25	bank	own	3	skilled	1	none	yes	good
4	female div/dep/mar	none	2	no known property	44	stores	for free	1	high qualif/self emp/mgmt	1	yes	yes	bad
3	male single	none	2	car	31	stores	own	1	skilled	2	yes	yes	good
4	male single	none	4	car	48	stores	own	3	skilled	1	yes	yes	good
2	male single	none	3	real estate	44	stores	rent	1	skilled	2	none	yes	good
1	male single	none	3	real estate	48	stores	rent	2	unskilled resident	2	none	no	good
3	male single	none	4	life insurance	44	stores	own	1	skilled	1	none	yes	good
2	male mar/wid	none	1	car	26	stores	own	2	skilled	1	none	no	good

Showing 1 to 25 of 6,000 entries, 22 total columns

Libraries and packages were loaded for the use of extra features that will assist in data import, data cleaning, data preprocessing and data analysis. The working directory is set, so that R can easily locate files to read, or a default location to save files. Next, **read_csv()** from the **readr** package is used to read the dataset which is a CSV file into a tibble. This **read_csv()** is faster and more efficient than the base R function **read.csv()** as it can automatically detect the column types and handle missing values more intuitively. After importing the dataset into R, it specified that the dataset consists of 6000 rows and 22 columns.

2.2 Data Cleaning / Preprocessing

```
# Data Exploration
dim(pfda_df)
nrow(pfda_df)
ncol(pfda_df)
```

```
> dim(pfda_df)
[1] 6000   22
> nrow(pfda_df)
[1] 6000
> ncol(pfda_df)
[1] 22
```

These three lines of code were used to check the dimension of the dataset. There are 6000 rows and 22 columns.

```
colnames(pfda_df)
```

```
> colnames(pfda_df)
 [1] "...1"              "checking_status"   "duration"              "credit_history"         "purpose"
 [6] "credit_amount"     "savings_status"    "employment"            "installment_commitment" "personal_status"
[11] "other_parties"     "residence_since"   "property_magnitude"    "age"                    "other_payment_plans"
[16] "housing"           "existing_credits"  "job"                   "num_dependants"         "own_telephone"
[21] "foreign_worker"    "class"
```

Figure 1: Checking Column Names Result

This code colnames(pfda_df) is used to check the names of the columns inside the dataset.

```
spec(pfda_df)
```

```
> spec(pfda_df)
cols(
  ...1 = col_double(),
  checking_status = col_character(),
  duration = col_double(),
  credit_history = col_character(),
  purpose = col_character(),
  credit_amount = col_double(),
  savings_status = col_character(),
  employment = col_character(),
  installment_commitment = col_double(),
  personal_status = col_character(),
  other_parties = col_character(),
  residence_since = col_double(),
  property_magnitude = col_character(),
  age = col_double(),
  other_payment_plans = col_character(),
  housing = col_character(),
  existing_credits = col_double(),
  job = col_character(),
  num_dependants = col_double(),
  own_telephone = col_character(),
  foreign_worker = col_character(),
  class = col_character()
)
```

The spec() function is used to display the column specification and how each column's data type was interpreted when the dataset was imported using the read_csv(). There are columns that are **double** data type which indicates that they contain numeric values. For columns that are character data type, they contain text or string data.

```
sapply(pfda_df, class)
str(pfda_df)
glimpse(pfda_df)
summary(pfda_df)
```

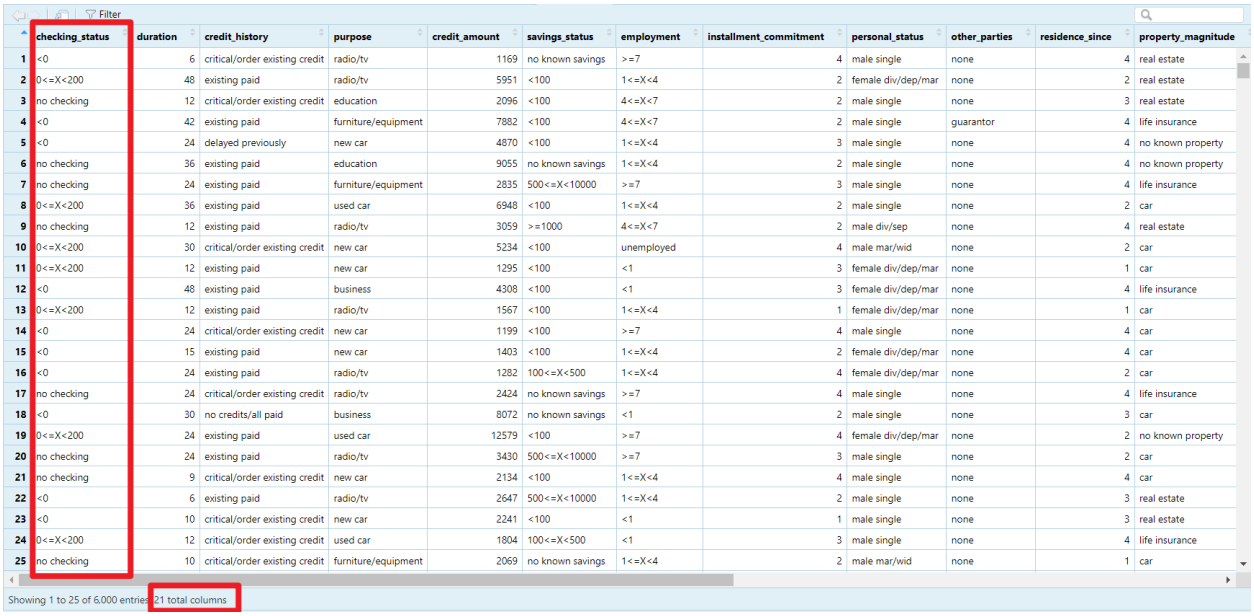
```
> sapply(pfda_df, class)
                ...1       checking_status              duration        credit_history               purpose 
           "numeric"           "character"             "numeric"           "character"           "character" 
       credit_amount       savings_status            employment installment_commitment       personal_status 
           "numeric"           "character"           "character"             "numeric"           "character" 
       other_parties     residence_since  property_magnitude                   age   other_payment_plans 
           "character"             "numeric"           "character"           "numeric"           "character" 
             housing   existing_credits                   job       num_dependants        own_telephone 
           "character"             "numeric"           "character"             "numeric"           "character" 
      foreign_worker               class 
           "character"           "character"
```

```
> str(pfda_df)
spc_tbl_ [6,000 × 22] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
 $ ...1                : num [1:6000] 0 1 2 3 4 5 6 7 8 9 ...
 $ checking_status     : chr [1:6000] "<0" "0<=X<200" "no checking" "<0" ...
 $ duration            : num [1:6000] 6 48 12 42 24 36 24 36 12 30 ...
 $ credit_history      : chr [1:6000] "critical/order existing credit" "existing paid" "critical/order existing credit" "existing paid" ...
 $ purpose             : chr [1:6000] "radio/tv" "radio/tv" "education" "furniture/equipment" ...
 $ credit_amount       : num [1:6000] 1169 5951 2096 7882 4870 ...
 $ savings_status      : chr [1:6000] "no known savings" "<100" "<100" "<100" ...
 $ employment          : chr [1:6000] ">=7" "1<=X<4" "4<=X<7" "4<=X<7" ...
 $ installment_commitment: num [1:6000] 4 2 2 2 3 2 3 2 2 4 ...
 $ personal_status     : chr [1:6000] "male single" "female div/dep/mar" "male single" "male single" ...
 $ other_parties       : chr [1:6000] "none" "none" "none" "guarantor" ...
 $ residence_since     : num [1:6000] 4 2 3 4 4 4 4 2 4 2 ...
 $ property_magnitude  : chr [1:6000] "real estate" "real estate" "real estate" "life insurance" ...
 $ age                 : num [1:6000] 67 22 49 45 53 35 53 35 61 28 ...
 $ other_payment_plans : chr [1:6000] "stores" "stores" "stores" "stores" ...
 $ housing             : chr [1:6000] "own" "own" "own" "for free" ...
 $ existing_credits    : num [1:6000] 2 1 1 1 2 1 1 1 1 2 ...
 $ job                 : chr [1:6000] "skilled" "skilled" "unskilled resident" "skilled" ...
 $ num_dependants      : num [1:6000] 1 1 2 2 2 2 1 1 1 1 ...
 $ own_telephone       : chr [1:6000] "yes" "none" "none" "none" ...
 $ foreign_worker      : chr [1:6000] "yes" "yes" "yes" "yes" ...
 $ class               : chr [1:6000] "good" "bad" "good" "good" ...
```

```
> summary(pfda_df)
      ...1      checking_status       duration     credit_history       purpose          credit_amount   savings_status
 Min.   :   0   Length:6000        Min.   : 4.00   Length:6000        Length:6000        Min.   :  250   Length:6000
 1st Qu.:1500   Class :character   1st Qu.:12.00   Class :character   Class :character   1st Qu.: 1332   Class :character
 Median :3000   Mode  :character   Median :19.65   Mode  :character   Mode  :character   Median : 2290   Mode  :character
 Mean   :3000                      Mean   :22.03                                         Mean   : 3344
 3rd Qu.:4499                      3rd Qu.:27.02                                         3rd Qu.: 4164
 Max.   :5999                      Max.   :72.00                                         Max.   :18424
  employment        installment_commitment personal_status    other_parties      residence_since property_magnitude
 Length:6000        Min.   :1.000          Length:6000        Length:6000        Min.   :1.00    Length:6000
 Class :character   1st Qu.:2.000          Class :character   Class :character   1st Qu.:2.00    Class :character
 Mode  :character   Median :3.353          Mode  :character   Mode  :character   Median :3.00    Mode  :character
                    Mean   :3.058                                                Mean   :2.85
                    3rd Qu.:4.000                                                3rd Qu.:4.00
                    Max.   :4.000                                                Max.   :4.00
      age        other_payment_plans   housing          existing_credits     job            num_dependants  own_telephone
 Min.   :19.00   Length:6000        Length:6000        Min.   :1.000    Length:6000        Min.   :1.000   Length:6000
 1st Qu.:27.00   Class :character   Class :character   1st Qu.:1.000    Class :character   1st Qu.:1.000   Class :character
 Median :32.00   Mode  :character   Mode  :character   Median :1.000    Mode  :character   Median :1.000   Mode  :character
 Mean   :34.95                                         Mean   :1.392                       Mean   :1.158
 3rd Qu.:40.00                                         3rd Qu.:2.000                       3rd Qu.:1.000
 Max.   :75.00                                         Max.   :4.000                       Max.   :2.000
 foreign_worker        class
 Length:6000        Length:6000
 Class :character   Class :character
 Mode  :character   Mode  :character
```

To view the data type of each column and view the structure and details of each column.

```
# Data Cleaning and Preprocessing

## Delete the first column
pfda_df <- pfda_df[, -1]
```

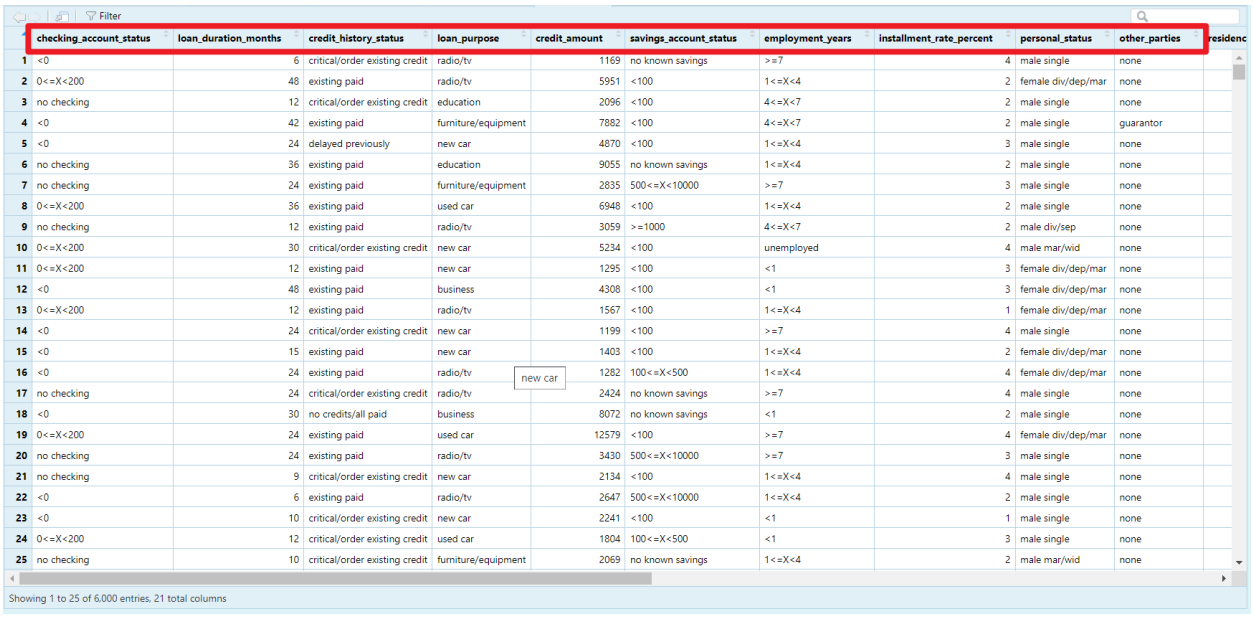


	checking_status	duration	credit_history	purpose	credit_amount	savings_status	employment	installment_commitment	personal_status	other_parties	residence_since	property_magnitude
1	<0	6	critical/order existing credit	radio/tv	1169	no known savings	>=7	4	male single	none	4	real estate
2	0<=X<200	48	existing paid	radio/tv	5951	<100	1<=X<4	2	female div/dep/mar	none	2	real estate
3	no checking	12	critical/order existing credit	education	2096	<100	4<=X<7	2	male single	none	3	real estate
4	<0	42	existing paid	furniture/equipment	7882	<100	4<=X<7	2	male single	guarantor	4	life insurance
5	<0	24	delayed previously	new car	4870	<100	1<=X<4	3	male single	none	4	no known property
6	no checking	36	existing paid	education	9055	no known savings	1<=X<4	2	male single	none	4	no known property
7	no checking	24	existing paid	furniture/equipment	2835	500<=X<10000	>=7	3	male single	none	4	life insurance
8	0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	2	male single	none	2	car
9	no checking	12	existing paid	radio/tv	3059	>=1000	4<=X<7	2	male div/sep	none	4	real estate
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	4	male mar/wid	none	2	car
11	0<=X<200	12	existing paid	new car	1295	<100	<1	3	female div/dep/mar	none	1	car
12	<0	48	existing paid	business	4308	<100	<1	3	female div/dep/mar	none	4	life insurance
13	0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	1	female div/dep/mar	none	1	car
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	4	male single	none	4	car
15	<0	15	existing paid	new car	1403	<100	1<=X<4	2	female div/dep/mar	none	4	car
16	<0	24	existing paid	radio/tv	1282	100<=X<500	1<=X<4	4	female div/dep/mar	none	2	car
17	no checking	24	critical/order existing credit	radio/tv	2424	no known savings	>=7	4	male single	none	4	life insurance
18	<0	30	no credits/all paid	business	8072	no known savings	<1	2	male single	none	3	car
19	0<=X<200	24	existing paid	used car	12579	<100	>=7	4	female div/dep/mar	none	2	no known property
20	no checking	24	existing paid	radio/tv	3430	500<=X<10000	>=7	3	male single	none	2	car
21	no checking	9	critical/order existing credit	new car	2134	<100	1<=X<4	4	male single	none	4	car
22	<0	6	existing paid	radio/tv	2647	500<=X<10000	1<=X<4	2	male single	none	3	real estate
23	<0	10	critical/order existing credit	new car	2241	<100	<1	1	male single	none	3	real estate
24	0<=X<200	12	critical/order existing credit	used car	1804	100<=X<500	<1	3	male single	none	4	life insurance
25	no checking	10	critical/order existing credit	furniture/equipment	2069	no known savings	1<=X<4	2	male mar/wid	none	1	car

Showing 1 to 25 of 6,000 entries 21 total columns

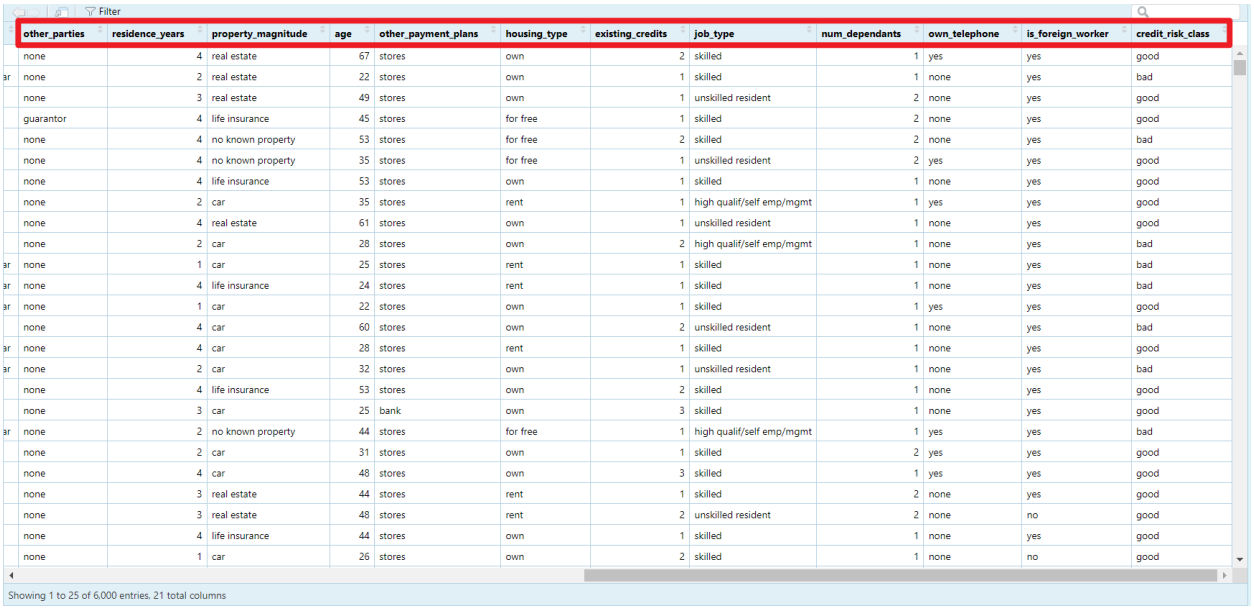
Remove the first column which is the index because it is redundant and RStudio also provides the index at the side which is better for visualization. Not only that, it may affect the data analysis.

```
## Rename the necessary columns
## Exclude credit_amount, personal_status, other_parties, property_magnitude, age, other_payment_plans, existing_credits, num_dependents, own_telephone
pfda_df <- pfda_df %>%
  rename(
    checking_account_status = checking_status,
    loan_duration_months = duration,
    credit_history_status = credit_history,
    loan_purpose = purpose,
    savings_account_status = savings_status,
    employment_years = employment,
    installment_rate_percent = installment_commitment,
    residence_years = residence_since,
    housing_type = housing,
    job_type = job,
    is_foreign_worker = foreign_worker,
    credit_risk_class = class
  )
## Check the column names after renaming
colnames(pfda_df)
```



	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_rate_percent	personal_status	other_parties	residenc
1	<0	6	critical/order existing credit	radio/tv	1169	no known savings	>=7	4	male single	none	
2	0<=X<200	48	existing paid	radio/tv	5951	<100	1<=X<4	2	female div/dep/mar	none	
3	no checking	12	critical/order existing credit	education	2096	<100	4<=X<7	2	male single	none	
4	<0	42	existing paid	furniture/equipment	7882	<100	4<=X<7	2	male single	guarantor	
5	<0	24	delayed previously	new car	4870	<100	1<=X<4	3	male single	none	
6	no checking	36	existing paid	education	9055	no known savings	1<=X<4	2	male single	none	
7	no checking	24	existing paid	furniture/equipment	2835	500<=X<10000	>=7	3	male single	none	
8	0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	2	male single	none	
9	no checking	12	existing paid	radio/tv	3059	>=1000	4<=X<7	2	male div/sep	none	
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	4	male mar/wid	none	
11	0<=X<200	12	existing paid	new car	1295	<100	<1	3	female div/dep/mar	none	
12	<0	48	existing paid	business	4308	<100	<1	3	female div/dep/mar	none	
13	0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	1	female div/dep/mar	none	
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	4	male single	none	
15	<0	15	existing paid	new car	1403	<100	1<=X<4	2	female div/dep/mar	none	
16	<0	24	existing paid	radio/tv	1282	100<=X<500	1<=X<4	4	female div/dep/mar	none	
17	no checking	24	critical/order existing credit	radio/tv	2424	no known savings	>=7	4	male single	none	
18	<0	30	no credits/all paid	business	8072	no known savings	<1	2	male single	none	
19	0<=X<200	24	existing paid	used car	12579	<100	>=7	4	female div/dep/mar	none	
20	no checking	24	existing paid	radio/tv	3430	500<=X<10000	>=7	3	male single	none	
21	no checking	9	critical/order existing credit	new car	2134	<100	1<=X<4	4	male single	none	
22	<0	6	existing paid	radio/tv	2647	500<=X<10000	1<=X<4	2	male single	none	
23	<0	10	critical/order existing credit	new car	2241	<100	<1	1	male single	none	
24	0<=X<200	12	critical/order existing credit	used car	1804	100<=X<500	<1	3	male single	none	
25	no checking	10	critical/order existing credit	furniture/equipment	2069	no known savings	1<=X<4	2	male mar/wid	none	

Showing 1 to 25 of 6,000 entries, 21 total columns



	other_parties	residence_years	property_magnitude	age	other_payment_plans	housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class
	none	4	real estate	67	stores	own	2	skilled	1	yes	yes	good
ar	none	2	real estate	22	stores	own	1	skilled	1	none	yes	bad
	none	3	real estate	49	stores	own	1	unskilled resident	2	none	yes	good
	guarantor	4	life insurance	45	stores	for free	1	skilled	2	none	yes	good
	none	4	no known property	53	stores	for free	2	skilled	2	none	yes	bad
	none	4	no known property	35	stores	for free	1	unskilled resident	2	yes	yes	good
	none	4	life insurance	53	stores	own	1	skilled	1	none	yes	good
	none	2	car	35	stores	rent	1	high qualif/self emp/mgmt	1	yes	yes	good
	none	4	real estate	61	stores	own	1	unskilled resident	1	none	yes	good
	none	2	car	28	stores	own	2	high qualif/self emp/mgmt	1	none	yes	bad
ar	none	1	car	25	stores	rent	1	skilled	1	none	yes	bad
ar	none	4	life insurance	24	stores	rent	1	skilled	1	none	yes	bad
ar	none	1	car	22	stores	own	1	skilled	1	yes	yes	good
	none	4	car	60	stores	own	2	unskilled resident	1	none	yes	bad
ar	none	4	car	28	stores	rent	1	skilled	1	none	yes	good
ar	none	2	car	32	stores	own	1	unskilled resident	1	none	yes	bad
	none	4	life insurance	53	stores	own	2	skilled	1	none	yes	good
	none	3	car	25	bank	own	3	skilled	1	none	yes	good
ar	none	2	no known property	44	stores	for free	1	high qualif/self emp/mgmt	1	yes	yes	bad
	none	2	car	31	stores	own	1	skilled	2	yes	yes	good
	none	4	car	48	stores	own	3	skilled	1	yes	yes	good
	none	3	real estate	44	stores	rent	1	skilled	2	none	yes	good
	none	3	real estate	48	stores	rent	2	unskilled resident	2	none	no	good
	none	4	life insurance	44	stores	own	1	skilled	1	none	yes	good
	none	1	car	26	stores	own	2	skilled	1	none	no	good

Showing 1 to 25 of 6,000 entries, 21 total columns

```
> colnames(pfda_df)
 [1] "checking_account_status"  "loan_duration_months"    "credit_history_status"   "loan_purpose"          "credit_amount"
 [6] "savings_account_status"   "employment_years"        "installment_rate_percent" "personal_status"       "other_parties"
[11] "residence_years"          "property_magnitude"      "age"                     "other_payment_plans"   "housing_type"
[16] "existing_credits"         "job_type"                "num_dependants"          "own_telephone"         "is_foreign_worker"
[21] "credit_risk_class"
```

Rename the column names for better clarity.

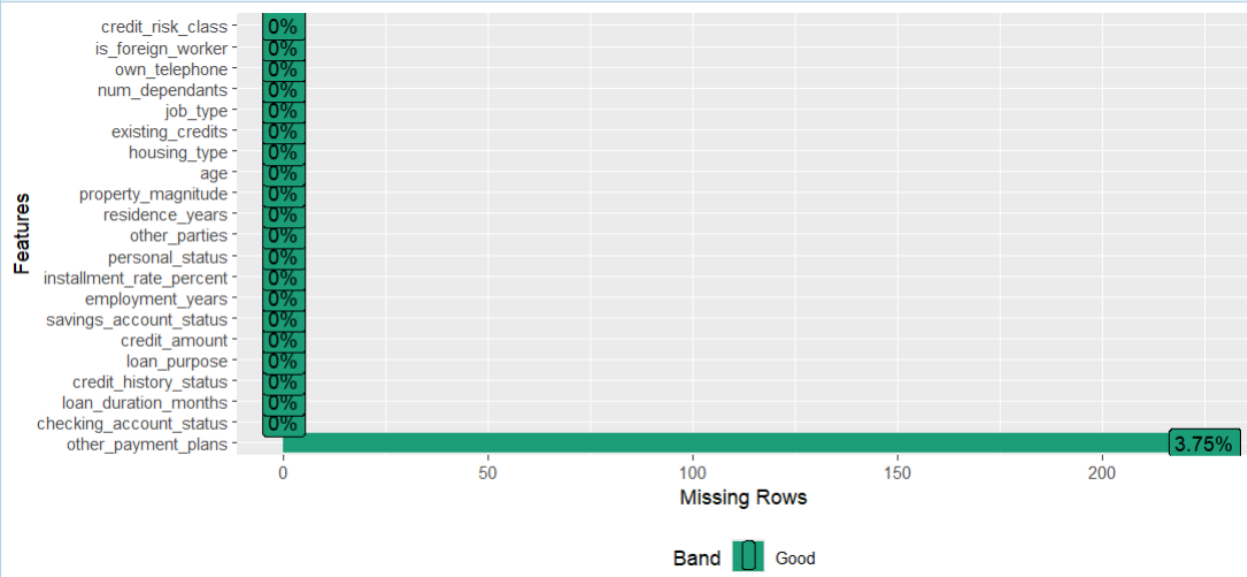
```
## Check the total number of duplicated rows
sum(duplicated(pfda_df))
```

```
> sum(duplicated(pfda_df))
[1] 4600
```

There are 4600 rows of duplicated data. We decided to retain the data since we assume that every row of data represents an individual customer.

```
## Check the total number of missing values for each column
colSums(is.na(pfda_df))
plot_missing(pfda_df)
```

```
> colSums(is.na(pfda_df))
checking_account_status      loan_duration_months      credit_history_status             loan_purpose             credit_amount
                      0                         0                          0                        0                         0
savings_account_status          employment_years installment_rate_percent          personal_status             other_parties
                      0                         0                          0                        0                         0
       residence_years       property_magnitude                        age       other_payment_plans              housing_type
                      0                         0                          0                      225                         0
      existing_credits                 job_type             num_dependants           own_telephone          is_foreign_worker
                      0                         0                          0                        0                         0
     credit_risk_class
                      0
```



There are 225 rows with missing values in the other_payment_plans column.

checking_account_status column (previously known as checking_status)

```
## checking_account_status column (previously known as checking_status)
## Check the details of this column
summary(pfda_df$checking_account_status)
## Check the frequency of each unique category
unique(pfda_df$checking_account_status)
as.data.frame(table(pfda_df$checking_account_status))
```

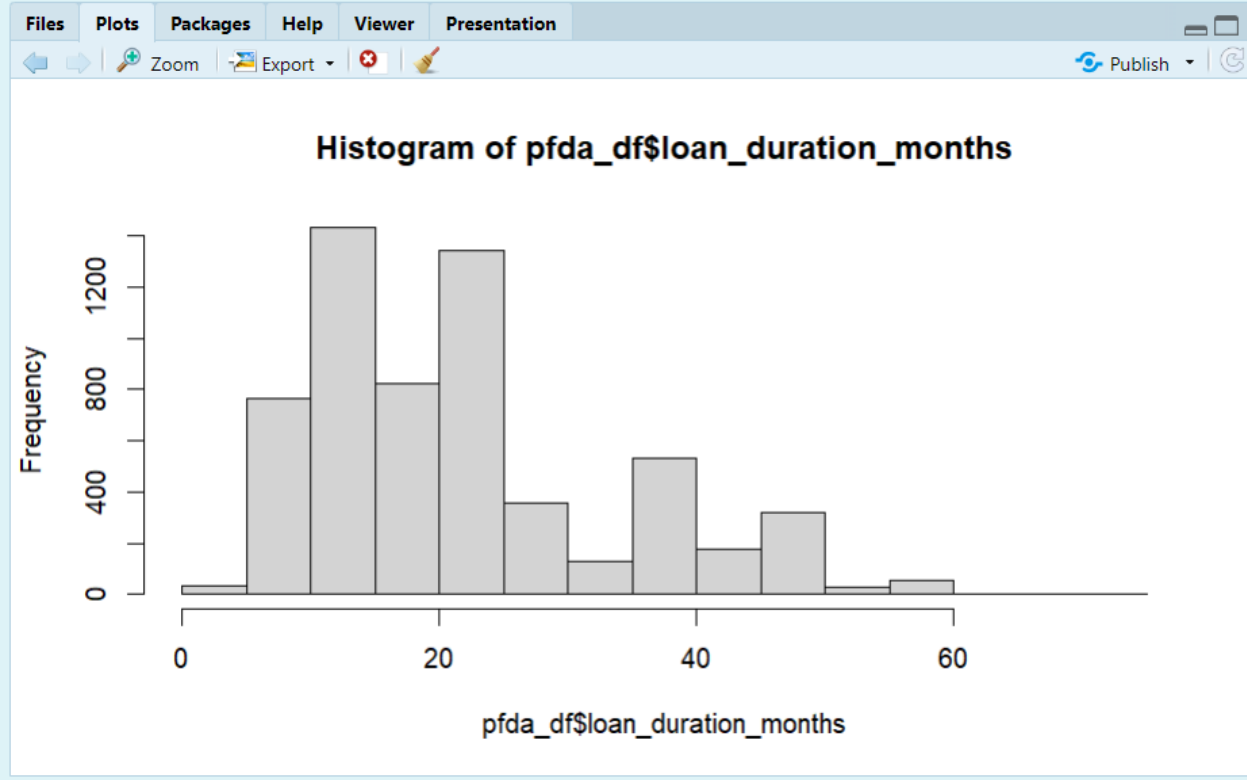
```
> ## checking_account_status column (previously known as checking_status)
> ## Check the details of this column
> summary(pfda_df$checking_account_status)
   Length     Class      Mode
     6000 character character
> ## Check the frequency of each unique category
> unique(pfda_df$checking_account_status)
[1] "<0"          "0<=X<200"    "no checking" ">=200"
> as.data.frame(table(pfda_df$checking_account_status))
         Var1 Freq
1          <0 1861
2       >=200  436
3    0<=X<200 1942
4 no checking 1761
```

Check the details of the checking_account_status column.

loan_duration_months (previously known as duration)

```
## loan_duration_months (previously known as duration)
## Check the details of this column
summary(pfda_df$loan_duration_months)
## Check all unique numbers
unique(pfda_df$loan_duration_months)
```

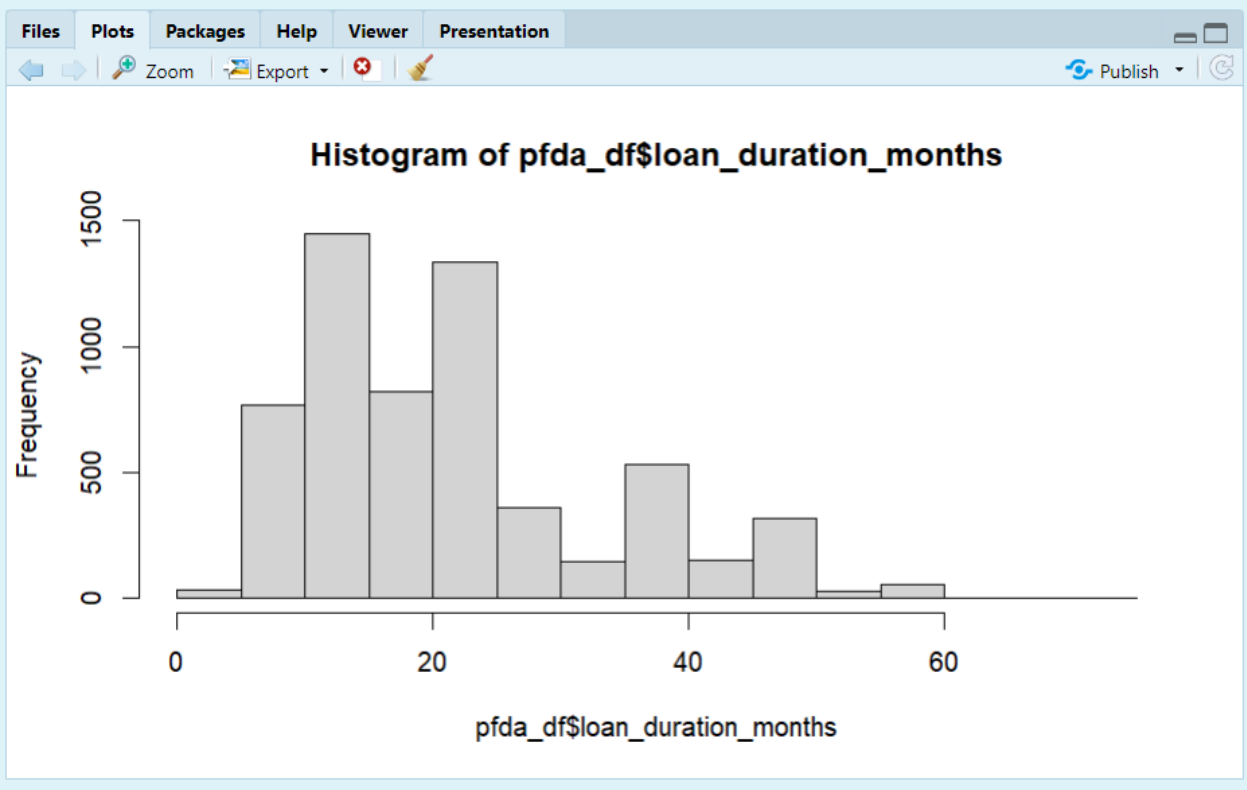
```
> ## loan_duration_months (previously known as duration)
> ## Check the details of this column
> summary(pfda_df$loan_duration_months)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   12.00   19.65   22.03   27.02   72.00
> ## Check all unique numbers
> unique(pfda_df$loan_duration_months)
  [1]  6.000000 48.000000 12.000000 42.000000 24.000000 36.000000 30.000000 15.000000  9.000000 10.000000  7.000000 60.000000 18.000000
 [14] 45.000000 11.000000 27.000000  8.000000 54.000000 20.000000 14.000000 33.000000 21.000000 16.000000  4.000000 47.000000 13.000000
 [27] 22.000000 39.000000 28.000000  5.000000 26.000000 72.000000 40.000000 23.361532 21.048243 12.680297 14.917150 15.317027 21.742746
 [40] 10.922290 18.930250 18.054477 35.033601 12.003016 12.233008 19.451017 14.777505 51.919815 41.482970 18.208489 30.512818 18.947862
 [53] 21.531877  9.901463 41.106825 45.610582 23.817221 35.137288 24.985620 14.392925 29.202249 21.696384 26.247418 34.893998 43.301668
 [66]  9.571686 21.345887 22.873521 38.194392 12.392316 26.170376 10.550089 38.143330 20.614037 29.520944 15.572594 20.603777 34.558024
 [79] 36.153834 11.478575 12.564942 25.868480  9.307342 31.792074  6.514833 21.922174 13.605979 31.797284 18.576707  9.152563 13.359835
 [92] 24.091827 48.798764 16.490641 18.472529 39.481087 17.542660 12.673301 57.491322 46.406387 22.635524  9.597152 21.094515 15.440202
[105] 40.759488 23.663875 13.076163 32.977448 28.309756 37.544543 11.379947 25.068700 20.575069 19.472891 24.350668 15.709523 49.671679
[118] 19.375241 20.179364 27.956253 21.653749 33.213977 45.543123 26.864971 19.140333 24.194517 10.643916 41.188762 24.822243 16.389887
[131] 13.500099 15.483264 45.608099 22.685600 14.863162 35.525588 36.416989 25.147085 13.866813 18.411930 13.530949 11.844271 29.400091
[144] 23.679088 12.497502 30.139996 24.518389 46.060619 28.762186 36.167384  6.499706 21.275112 32.385354 26.677782 19.312224 23.659606
[157] 18.154511 24.670454 34.422213 22.590067 17.811912 22.043566 21.195151 21.577211 33.443451 21.667729 14.379910 44.756907 24.879629
[170] 12.159914 43.801895 23.723448 12.695260 24.717840 14.130167 29.446162 15.551332 11.124544 24.473590 27.152531 27.963030 35.876511
[183] 39.672674 12.285270 11.222922 20.793662 28.936742 19.534730 20.172473 22.837714 41.197736 19.755331 28.145955 11.406328 18.397788
[196] 12.439950 21.776414 22.643580 14.514694 42.723088 21.258700 23.002707 19.766964 11.987494 51.636899 28.901719 29.491526 23.421834
[209] 11.826408 44.843826  7.281073 36.041881 11.660640 11.531516 21.142467 21.247263 34.786883 14.162071 30.541053 29.438890 11.903052
[222] 32.642838 16.598279 11.383309 39.251492 54.799111 36.914405 35.749420  8.513024 34.545178 23.503509 46.547636 42.989006 40.597408
[235] 22.665614 28.205551 20.609349 19.502530 19.284127 42.874927 17.171022  9.952677 35.668603 24.803113 32.794921 37.648621  6.521005
[248] 16.978206 13.108187 13.866448 21.723254 23.703218 43.243366  7.461352 39.896982 11.859311 16.387965 47.822522 30.022018 33.976458
[261] 47.073356 40.406742 12.896735 11.734775 35.877268 20.309603 17.760552 35.383927 19.291990 26.163832 45.660536 23.709453 29.110747
[274] 29.765553 19.297054 39.279464 42.657787 27.089979 15.283236 30.569356 35.530881 19.787757 34.004352 32.233014 32.947077 13.674774
[287] 29.052548 14.772987 20.862416 21.490431 15.950621 31.882098 21.965613 34.790084 44.372011 28.321615 16.447522 19.060065 32.262283
[300] 16.728541 35.193310 14.417419 39.941434  9.534628 28.140992 40.171431 17.415416 31.603271 26.135949 18.634696 16.019403 16.070503
[313] 19.107487 28.330309 13.683229 38.510405 36.681421 18.015570  9.583668 10.027486 19.416699 11.279866 16.968986 17.646193 59.251904
[326] 39.931653 56.805194  7.877007 13.866712 40.497697 47.268668 43.750240 16.765966  9.895131 12.924314 44.177040 13.070452 21.212013
[339] 41.865731 24.172684 32.522503 19.785874 16.316239 14.336798 20.653546  8.480787 14.515668 33.901276 42.591538 33.871988  9.733059
[352] 18.996495 20.035818 35.204302 40.270732 15.073141 22.209810 30.527326 46.998237 15.784463 34.967290 24.218910 19.818231 14.991767
[365] 21.310169
```



Check the details of the loan_duration_months column.

```
## Convert the loan duration in months from double to integer by using standard rounding
pfda_df$loan_duration_months <- round(pfda_df$loan_duration_months)
unique(pfda_df$loan_duration_months)
as.data.frame(table(pfda_df$loan_duration_months))
```

```
> ## Convert the loan duration in months from double to integer by using standard rounding
> pfda_df$loan_duration_months <- round(pfda_df$loan_duration_months)
> unique(pfda_df$loan_duration_months)
 [1]  6 48 12 42 24 36 30 15  9 10  7 60 18 45 11 27  8 54 20 14 33 21 16  4 47 13 22 39 28  5 26 72 40 23 19 35 52
[38] 41 31 46 25 29 43 38 32 49 57 50 34 44 17 55 37 59
> as.data.frame(table(pfda_df$loan_duration_months))
   Var1 Freq
1     4   23
2     5   10
3     6  309
4     7   39
5     8   29
6     9  215
7    10  175
8    11   90
9    12  905
10   13   61
11   14   64
12   15  330
13   16   61
14   17   32
15   18  590
16   19   59
17   20   78
18   21  220
19   22   64
20   23   49
21   24  970
22   25   33
23   26   22
24   27   78
25   28   48
26   29   34
27   30  178
28   31   14
29   32   28
30   33   34
31   34   19
32   35   49
33   36  428
34   37    5
35   38   23
36   39   38
37   40   38
38   41   28
39   42   44
40   43   28
41   44   17
42   45   36
43   46   29
44   47   22
45   48  257
46   49    6
47   50    4
48   52   13
49   54   14
50   55    2
51   57    8
52   59    1
53   60   45
54   72    4
>
```

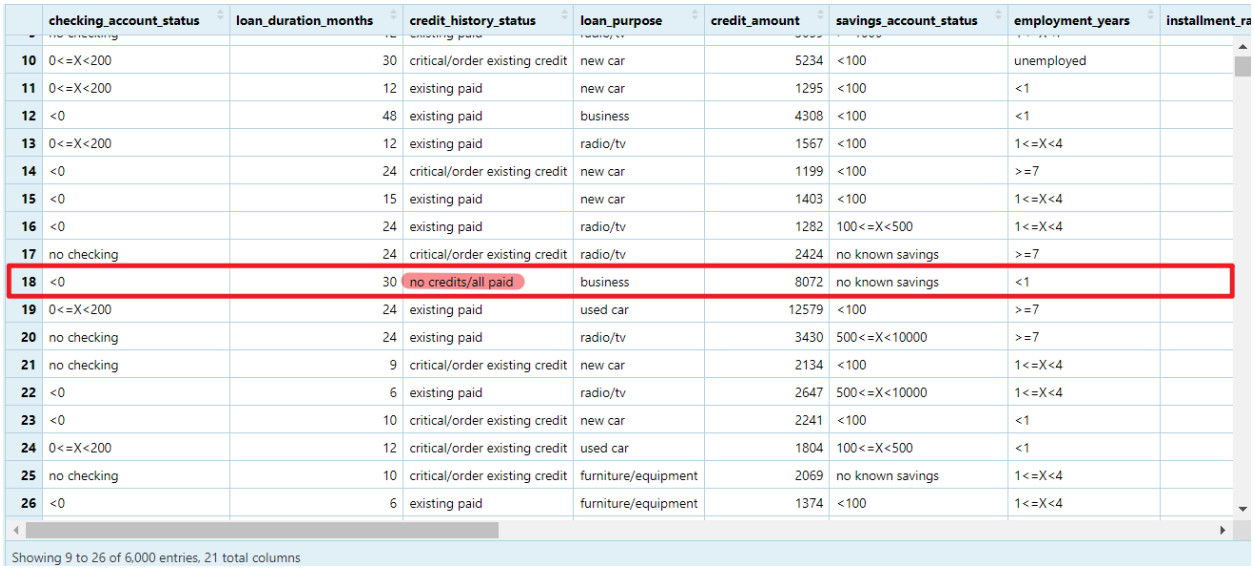


Convert the data type of loan_duration_months column to integer by normal rounding for better distribution of the data.

credit_history_status (previously known as credit_history)

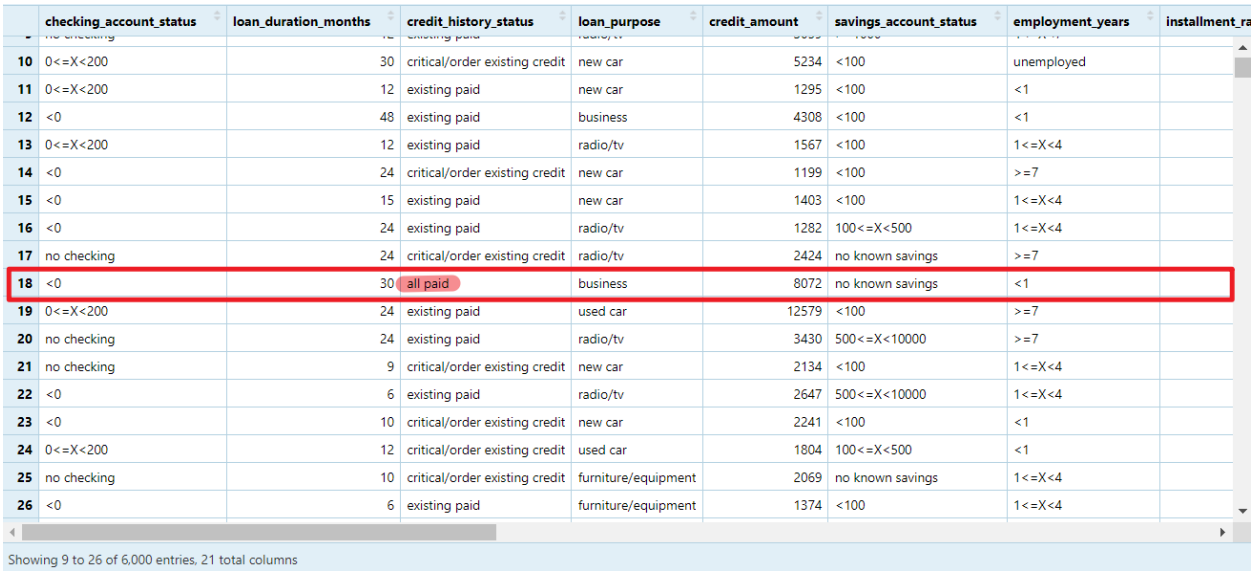
```
> #3.1 Check the details of this column
> summary(pfda_df$credit_history_status)
   Length     Class      Mode
     6000 character character
```

```
> #3.2 Check the frequency of each unique category
> unique(pfda_df$credit_history_status)
[1] "critical/order existing credit" "existing paid"                  "delayed previously"
[4] "no credits/all paid"            "all paid"
> as.data.frame(table(pfda_df$credit_history_status))
                          Var1 Freq
1                     all paid  378
2 critical/order existing credit 1605
3            delayed previously  779
4                 existing paid 3037
5           no credits/all paid  201
```



	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_ra
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	
11	0<=X<200	12	existing paid	new car	1295	<100	<1	
12	<0	48	existing paid	business	4308	<100	<1	
13	0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	
15	<0	15	existing paid	new car	1403	<100	1<=X<4	
16	<0	24	existing paid	radio/tv	1282	100<=X<500	1<=X<4	
17	no checking	24	critical/order existing credit	radio/tv	2424	no known savings	>=7	
18	<0	30	no credits/all paid	business	8072	no known savings	<1	
19	0<=X<200	24	existing paid	used car	12579	<100	>=7	
20	no checking	24	existing paid	radio/tv	3430	500<=X<10000	>=7	
21	no checking	9	critical/order existing credit	new car	2134	<100	1<=X<4	
22	<0	6	existing paid	radio/tv	2647	500<=X<10000	1<=X<4	
23	<0	10	critical/order existing credit	new car	2241	<100	<1	
24	0<=X<200	12	critical/order existing credit	used car	1804	100<=X<500	<1	
25	no checking	10	critical/order existing credit	furniture/equipment	2069	no known savings	1<=X<4	
26	<0	6	existing paid	furniture/equipment	1374	<100	1<=X<4	

Showing 9 to 26 of 6,000 entries, 21 total columns



	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_ra
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	
11	0<=X<200	12	existing paid	new car	1295	<100	<1	
12	<0	48	existing paid	business	4308	<100	<1	
13	0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	
15	<0	15	existing paid	new car	1403	<100	1<=X<4	
16	<0	24	existing paid	radio/tv	1282	100<=X<500	1<=X<4	
17	no checking	24	critical/order existing credit	radio/tv	2424	no known savings	>=7	
18	<0	30	all paid	business	8072	no known savings	<1	
19	0<=X<200	24	existing paid	used car	12579	<100	>=7	
20	no checking	24	existing paid	radio/tv	3430	500<=X<10000	>=7	
21	no checking	9	critical/order existing credit	new car	2134	<100	1<=X<4	
22	<0	6	existing paid	radio/tv	2647	500<=X<10000	1<=X<4	
23	<0	10	critical/order existing credit	new car	2241	<100	<1	
24	0<=X<200	12	critical/order existing credit	used car	1804	100<=X<500	<1	
25	no checking	10	critical/order existing credit	furniture/equipment	2069	no known savings	1<=X<4	
26	<0	6	existing paid	furniture/equipment	1374	<100	1<=X<4	

Showing 9 to 26 of 6,000 entries, 21 total columns

```
> #3.4 Double check the frequency of each unique category
> unique(pfda_df$credit_history_status)
[1] "critical/order existing credit" "existing paid"                  "delayed previously"
[4] "all paid"
> as.data.frame(table(pfda_df$credit_history_status))
                        Var1 Freq
1                   all paid  579
2 critical/order existing credit 1605
3            delayed previously  779
4                 existing paid 3037
```

Replace the “no credits/all paid” with “all paid” as we assume that they are in the same category.

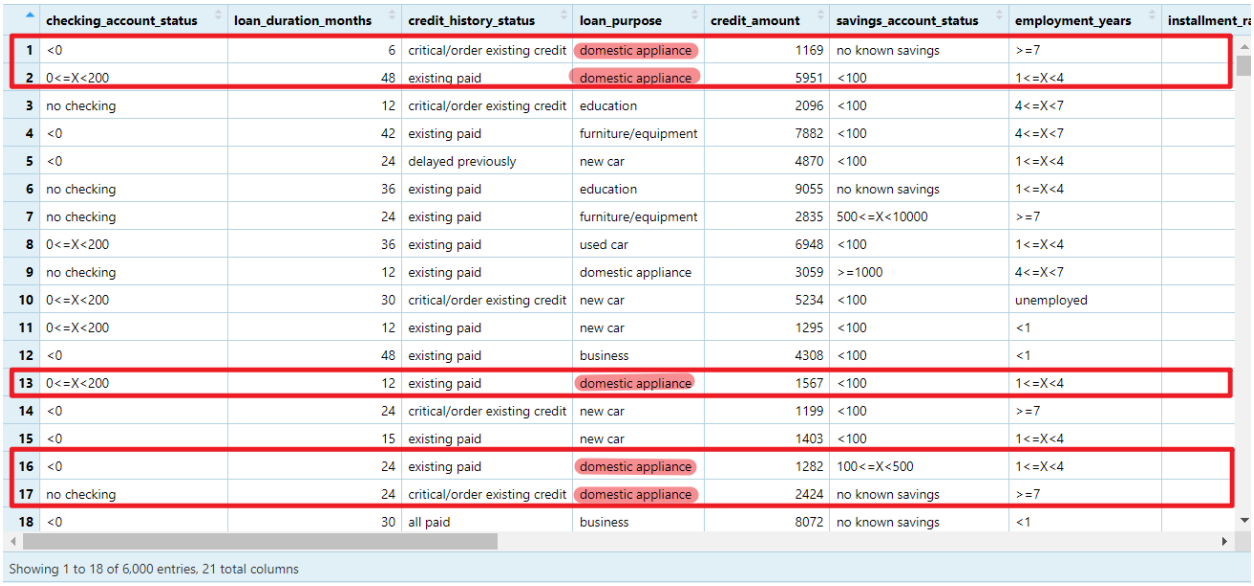
loan_purpose column (previously known as purpose)

```
> ## 4.0 loan_purpose column (previously known as purpose)
> # 4.1 Check the frequency of each unique category
> summary(pfda_df$loan_purpose)
   Length     Class      Mode
     6000 character character
```

```
> # 4.2 Check the frequency of each unique category
> unique(pfda_df$loan_purpose)
 [1] "radio/tv"            "education"           "furniture/equipment" "new car"             "used car"
 [6] "business"            "domestic appliance"  "repairs"             "other"               "retraining"
> as.data.frame(table(pfda_df$loan_purpose))
                 Var1 Freq
1            business  535
2  domestic appliance  154
3           education  358
4  furniture/equipment 1152
5            new car 1474
6               other  290
7            radio/tv 1328
8             repairs  164
9          retraining   91
10           used car  454
```

	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_ra
1	<0	6	critical/order existing credit	radio/tv	1169	no known savings	>=7	
2	0<=X<200	48	existing paid	radio/tv	5951	<100	1<=X<4	
3	no checking	12	critical/order existing credit	education	2096	<100	4<=X<7	
4	<0	42	existing paid	furniture/equipment	7882	<100	4<=X<7	
5	<0	24	delayed previously	new car	4870	<100	1<=X<4	
6	no checking	36	existing paid	education	9055	no known savings	1<=X<4	
7	no checking	24	existing paid	furniture/equipment	2835	500<=X<10000	>=7	
8	0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	
9	no checking	12	existing paid	radio/tv	3059	>=1000	4<=X<7	
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	
11	0<=X<200	12	existing paid	new car	1295	<100	<1	
12	<0	48	existing paid	business	4308	<100	<1	
13	0<=X<200	12	existing paid	radio/tv	1567	<100	1<=X<4	
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	
15	<0	15	existing paid	new car	1403	<100	1<=X<4	
16	<0	24	existing paid	radio/tv	1282	100<=X<500	1<=X<4	
17	no checking	24	critical/order existing credit	radio/tv	2424	no known savings	>=7	
18	<0	30	all paid	business	8072	no known savings	<1	

Showing 1 to 18 of 6,000 entries, 21 total columns



	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_ra
1	<0	6	critical/order existing credit	domestic appliance	1169	no known savings	>=7	
2	0<=X<200	48	existing paid	domestic appliance	5951	<100	1<=X<4	
3	no checking	12	critical/order existing credit	education	2096	<100	4<=X<7	
4	<0	42	existing paid	furniture/equipment	7882	<100	4<=X<7	
5	<0	24	delayed previously	new car	4870	<100	1<=X<4	
6	no checking	36	existing paid	education	9055	no known savings	1<=X<4	
7	no checking	24	existing paid	furniture/equipment	2835	500<=X<10000	>=7	
8	0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	
9	no checking	12	existing paid	domestic appliance	3059	>=1000	4<=X<7	
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	
11	0<=X<200	12	existing paid	new car	1295	<100	<1	
12	<0	48	existing paid	business	4308	<100	<1	
13	0<=X<200	12	existing paid	domestic appliance	1567	<100	1<=X<4	
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	
15	<0	15	existing paid	new car	1403	<100	1<=X<4	
16	<0	24	existing paid	domestic appliance	1282	100<=X<500	1<=X<4	
17	no checking	24	critical/order existing credit	domestic appliance	2424	no known savings	>=7	
18	<0	30	all paid	business	8072	no known savings	<1	

Showing 1 to 18 of 6,000 entries, 21 total columns

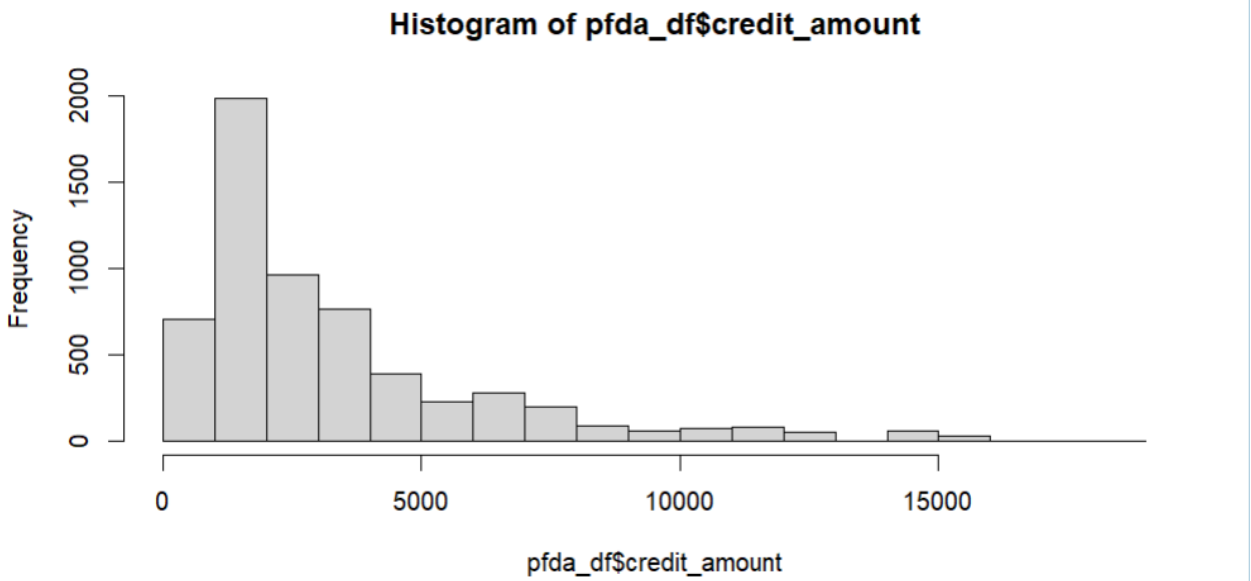
```
> # 4.4 Double check the frequency of each unique category
> unique(pfda_df$loan_purpose)
[1] "domestic appliance"  "education"           "furniture/equipment" "new car"             "used car"
[6] "business"            "repairs"             "other"               "retraining"
> as.data.frame(table(pfda_df$loan_purpose))
                Var1 Freq
1           business  535
2  domestic appliance 1482
3          education  358
4 furniture/equipment 1152
5            new car 1474
6              other  290
7            repairs  164
8         retraining   91
9           used car  454
```

Replace “radio/tv” with “domestic appliance” as we assume that they are in the same category.

credit_amount column

```
> # 5.1 Check the details of this column
> summary(pfda_df$credit_amount)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
    250    1332    2290    3344    4164   18424
```

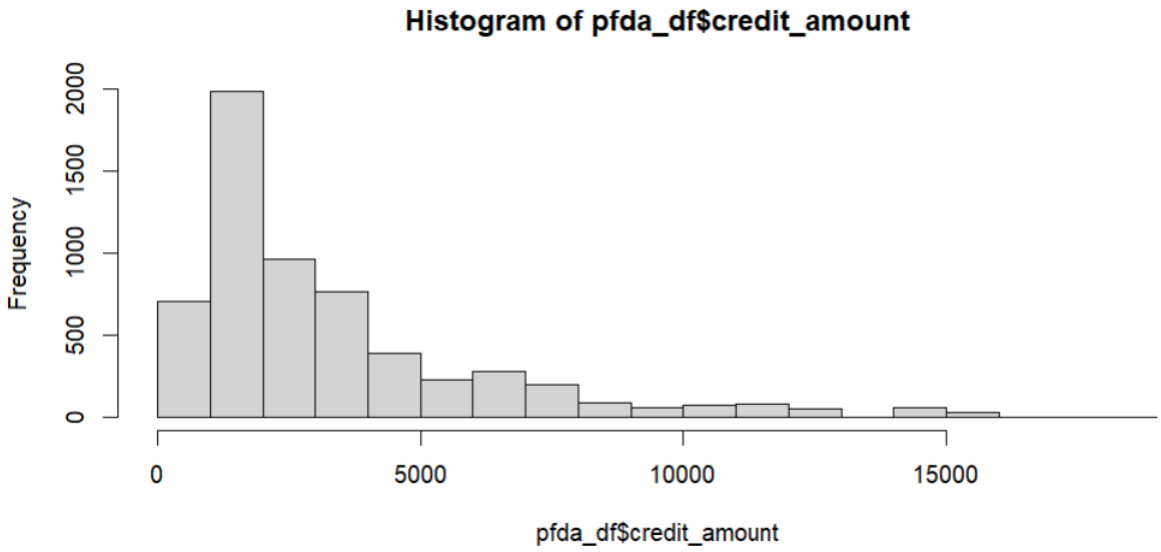
```
> # 5.2 Check all unique numbers
> unique(pfda_df$credit_amount)
   [1]  1169.0000  5951.0000  2096.0000  7882.0000  4870.0000  9055.0000  2835.0000  6948.0000  3059.0000  5234.0000
  [11]  1295.0000  4308.0000  1567.0000  1199.0000  1403.0000  1282.0000  2424.0000  8072.0000 12579.0000  3430.0000
  [21]  2134.0000  2647.0000  2241.0000  1804.0000  2069.0000  1374.0000   426.0000   409.0000  2415.0000  6836.0000
  [31]  1913.0000  4020.0000  5866.0000  1264.0000  1474.0000  4746.0000  6110.0000  2100.0000  1225.0000   458.0000
  [41]  2333.0000  1158.0000  6204.0000  6187.0000  6143.0000  1393.0000  2299.0000  1352.0000  7228.0000  2073.0000
  [51]  5965.0000  1262.0000  3378.0000  2225.0000   783.0000  6468.0000  9566.0000  1961.0000  6229.0000  1391.0000
  [61]  1537.0000  1953.0000 14421.0000  3181.0000  5190.0000  2171.0000  1007.0000  1819.0000  2394.0000  8133.0000
  [71]   730.0000  1164.0000  5954.0000  1977.0000  1526.0000  3965.0000  4771.0000  9436.0000  3832.0000  5943.0000
  [81]  1213.0000  1568.0000  1755.0000  2315.0000  1412.0000 12612.0000  2249.0000  1108.0000   618.0000  1409.0000
  [91]   797.0000  3617.0000  1318.0000 15945.0000  2012.0000  2622.0000  2337.0000  7057.0000  1469.0000  2323.0000
 [101]   932.0000  1919.0000  2445.0000 11938.0000  6458.0000  6078.0000  7721.0000  1410.0000  1449.0000   392.0000
 [111]  6260.0000  7855.0000  1680.0000  3578.0000  7174.0000  2132.0000  4281.0000  2366.0000  1835.0000  3868.0000
 [121]  1768.0000   781.0000  1924.0000  2121.0000   701.0000   639.0000  1860.0000  3499.0000  8487.0000  6887.0000
 [131]  2708.0000  1984.0000 10144.0000  1240.0000  8613.0000   766.0000  2728.0000  1881.0000   709.0000  4795.0000
 [141]  3416.0000  2462.0000  2288.0000  3566.0000   860.0000   682.0000  5371.0000  1582.0000  1346.0000  5848.0000
 [151]  7758.0000  6967.0000  1288.0000   339.0000  3512.0000  1898.0000  2872.0000  1055.0000  7308.0000   909.0000
 [161]  2978.0000  1131.0000  1577.0000  3972.0000  1935.0000   950.0000   763.0000  2064.0000  1414.0000  3414.0000
 [171]  7485.0000  2577.0000   338.0000  1963.0000   571.0000  9572.0000  4455.0000  1647.0000  3777.0000   884.0000
 [181]  1360.0000  5129.0000  1175.0000   674.0000  3244.0000  4591.0000  3844.0000  3915.0000  2108.0000  3031.0000
 [191]  1501.0000  1382.0000   951.0000  2760.0000  4297.0000   936.0000  1168.0000  5117.0000   902.0000  1495.0000
 [201] 10623.0000  1424.0000  6568.0000  1413.0000  3074.0000  3835.0000  5293.0000  1908.0000  3342.0000  3104.0000
 [211]  3913.0000  3021.0000  1364.0000   625.0000  1200.0000   707.0000  4657.0000  2613.0000 10961.0000  7865.0000
 [221]  1478.0000  3149.0000  4210.0000  2507.0000  2141.0000   866.0000  1544.0000  1823.0000 14555.0000  2767.0000
 [231]  1291.0000  2522.0000   915.0000  1595.0000  4605.0000  1185.0000  3447.0000  1258.0000   717.0000  1204.0000
 [241]  1925.0000   433.0000   666.0000  2251.0000  2150.0000  4151.0000  2030.0000  7418.0000  2684.0000  2149.0000
 [251]  3812.0000  1154.0000  1657.0000  1603.0000  5302.0000  2748.0000  1231.0000   802.0000  6304.0000  1533.0000
 [261]  8978.0000   999.0000  2662.0000  1402.0000 12169.0000  3060.0000 11998.0000  2697.0000  2404.0000  4611.0000
 [271]  1901.0000  3368.0000  1574.0000  1445.0000  1520.0000  3878.0000 10722.0000  4788.0000  7582.0000  1092.0000
 [281]  1024.0000  1076.0000  9398.0000  6419.0000  4796.0000  7629.0000  9960.0000  4675.0000  1287.0000  2515.0000
 [291]  2745.0000   672.0000  3804.0000  1344.0000  1038.0000 10127.0000  1543.0000  4811.0000   727.0000  1237.0000
 [301]   276.0000  5381.0000  5511.0000  3749.0000   685.0000  1494.0000  2746.0000   708.0000  4351.0000  3643.0000
 [311]  4249.0000  1938.0000  2910.0000  2659.0000  1028.0000  3398.0000  5801.0000  1525.0000  4473.0000  1068.0000
 [321]  6615.0000  1864.0000  7408.0000 11590.0000  4110.0000  3384.0000  2101.0000  1275.0000  4169.0000  1521.0000
 [331]  5743.0000  3599.0000  3213.0000  4439.0000  3949.0000  1459.0000   882.0000  3758.0000  1743.0000  1136.0000
 [341]  1236.0000   959.0000  3229.0000  6199.0000  1246.0000  2331.0000  4463.0000   776.0000  2406.0000  1239.0000
 [621]  2303.0000  8086.0000  2346.0000  3973.0000   888.0000 10222.0000  4221.0000  6361.0000  1297.0000   900.0000
 [631]  1050.0000  1047.0000  6314.0000  3496.0000  3609.0000  4843.0000  4139.0000  5742.0000 10366.0000  2080.0000
 [641]  2580.0000  4530.0000  5150.0000  5595.0000  1453.0000  1538.0000  2279.0000  5103.0000  9857.0000  6527.0000
 [651]  1347.0000  2862.0000  2753.0000  3651.0000   975.0000  2896.0000  4716.0000  2284.0000  1103.0000   926.0000
 [661]  1800.0000  1905.0000  1123.0000  6331.0000  1377.0000  2503.0000  2528.0000  5324.0000  6560.0000  2969.0000
 [671]  1206.0000  2118.0000   629.0000  1198.0000  2476.0000  1138.0000 14027.0000  7596.0000  1505.0000  3148.0000
 [681]  6148.0000  1337.0000  1228.0000   790.0000  2570.0000   250.0000  1316.0000  1882.0000  6416.0000  6403.0000
 [691]  1987.0000   760.0000  2603.0000  3380.0000  3990.0000 11560.0000  4380.0000  6761.0000  4280.0000  2325.0000
 [701]  1048.0000  3160.0000  2483.0000 14179.0000  1797.0000  2511.0000  1274.0000  5248.0000  3029.0000   428.0000
 [711]   976.0000   841.0000  5771.0000  1555.0000  1285.0000  1299.0000  1271.0000   691.0000  5045.0000  2124.0000
 [721]  2214.0000 12680.0000  2463.0000  1155.0000  3108.0000  2901.0000  1655.0000  2812.0000  8065.0000  3275.0000
 [731]  2223.0000  1480.0000  1371.0000  3535.0000  3509.0000  5711.0000  3872.0000  4933.0000  1940.0000   836.0000
 [741]  1941.0000  2675.0000  2751.0000  6224.0000  5998.0000  1188.0000  6313.0000  1221.0000  2892.0000  3062.0000
 [751]  2301.0000  7511.0000  1549.0000  1795.0000  7472.0000  9271.0000   590.0000   930.0000  9283.0000  1778.0000
 [761]   907.0000   484.0000  9629.0000  3051.0000  3931.0000  7432.0000  1338.0000  1554.0000 15857.0000  1345.0000
 [771]  1101.0000  3016.0000  2712.0000   731.0000  3780.0000  1602.0000  3966.0000  4165.0000  8335.0000  6681.0000
 [781]  2375.0000 11816.0000  5084.0000  2327.0000   886.0000   601.0000  2957.0000  2611.0000  5179.0000  2993.0000
 [791]  1943.0000  1559.0000  3422.0000  3976.0000  1249.0000  2235.0000  1471.0000 10875.0000   894.0000  3343.0000
 [801]  3959.0000  3577.0000  5804.0000  2169.0000  2439.0000  2210.0000  2221.0000  2389.0000  3331.0000  7409.0000
 [811]   652.0000  7678.0000  1343.0000   874.0000  3590.0000  1322.0000  3595.0000  1422.0000  6742.0000  7814.0000
 [821]  9277.0000  2181.0000  1098.0000  4057.0000   795.0000  2825.0000 15672.0000  6614.0000  7824.0000  2442.0000
 [831]  1829.0000  5800.0000  8947.0000  2606.0000  1592.0000  2186.0000  4153.0000  2625.0000  3485.0000 10477.0000
 [841]  1278.0000  1107.0000  3763.0000  3711.0000  3594.0000  3195.0000  4454.0000  4736.0000  2991.0000  2142.0000
 [851]  3161.0000 18424.0000  2848.0000 14896.0000  2359.0000  3345.0000  1817.0000 12749.0000  1366.0000  2002.0000
 [861]  6872.0000   697.0000  1049.0000 10297.0000  1867.0000  1747.0000  1670.0000  1224.0000   522.0000  1498.0000
 [871]   745.0000  2063.0000  6288.0000  6842.0000  3527.0000   929.0000  1455.0000  1845.0000  8358.0000  2859.0000
 [881]  3621.0000  2145.0000  4113.0000 10974.0000  1893.0000  3656.0000  4006.0000  3069.0000  1740.0000  2353.0000
 [891]  3556.0000  2397.0000   454.0000  1715.0000  2520.0000  3568.0000  7166.0000  3939.0000  1514.0000  7393.0000
 [901]  1193.0000  7297.0000  2831.0000   753.0000  2427.0000  2538.0000  8386.0000  4844.0000  2923.0000  8229.0000
 [911]  1433.0000  6289.0000  6579.0000  3565.0000  1569.0000  1936.0000  2390.0000  1736.0000  3857.0000   804.0000
 [921]  4576.0000  1284.2906  2543.4102  1357.2892  1479.2379  1271.0829  1982.9479  6284.0833 15240.0446  3393.8446
 [931]  4156.4727  2774.8547  8254.2874  1838.7893  1664.2442  1554.0291  1440.4837  6463.3708 15381.2804   674.8443
 [941]  4640.5721  4163.1931 12475.8804  1039.1564  3927.6955  1350.8865   956.5961  6029.2678 11644.1601  4111.3099
 [951]  4287.9913  4591.5749  1533.6012  2730.1386  1282.4404  2561.9725  3445.0783  1843.2208  1320.9928  8038.5225
 [961]  1281.5345  1369.6434  1358.8366  1381.4694  5163.8221  1080.3000  3141.8283   664.3339  4290.2580  3460.3079
 [971]  3956.9005   996.8302   720.2692  4213.7251  5186.8206  5133.5900  1332.3182  2373.6364  1366.3073  3355.4872
 [981] 14886.0809  1958.6793 10159.8248  1351.5370  1272.3641  2388.5529  2469.7801  1336.6949  2057.4777 11566.7034
 [991]  6228.2601  2579.4990  4857.5632  6330.4058  2579.6159  1294.4389  6836.7805  3899.3919  2317.9101   651.0142
 [ reached getOption("max.print") -- omitted 314 entries ]
>
```




```
> # 5.4 Convert the credit amount from double to integer by using standard rounding
> pfda_df$credit_amount <- round(pfda_df$credit_amount)
> unique(pfda_df$credit_amount)
   [1]  1169  5951  2096  7882  4870  9055  2835  6948  3059  5234  1295  4308  1567  1199  1403  1282  2424  8072
  [19] 12579  3430  2134  2647  2241  1804  2069  1374   426   409  2415  6836  1913  4020  5866  1264  1474  4746
  [37]  6110  2100  1225   458  2333  1158  6204  6187  6143  1393  2299  1352  7228  2073  5965  1262  3378  2225
  [55]   783  6468  9566  1961  6229  1391  1537  1953 14421  3181  5190  2171  1007  1819  2394  8133   730  1164
  [73]  5954  1977  1526  3965  4771  9436  3832  5943  1213  1568  1755  2315  1412 12612  2249  1108   618  1409
  [91]   797  3617  1318 15945  2012  2622  2337  7057  1469  2323   932  1919  2445 11938  6458  6078  7721  1410
 [109]  1449   392  6260  7855  1680  3578  7174  2132  4281  2366  1835  3868  1768   781  1924  2121   701   639
 [127]  1860  3499  8487  6887  2708  1984 10144  1240  8613   766  2728  1881   709  4795  3416  2462  2288  3566
 [145]   860   682  5371  1582  1346  5848  7758  6967  1288   339  3512  1898  2872  1055  7308   909  2978  1131
 [163]  1577  3972  1935   950   763  2064  1414  3414  7485  2577   338  1963   571  9572  4455  1647  3777   884
 [181]  1360  5129  1175   674  3244  4591  3844  3915  2108  3031  1501  1382   951  2760  4297   936  1168  5117
 [199]   902  1495 10623  1424  6568  1413  3074  3835  5293  1908  3342  3104  3913  3021  1364   625  1200   707
 [217]  4657  2613 10961  7865  1478  3149  4210  2507  2141   866  1544  1823 14555  2767  1291  2522   915  1595
 [235]  4605  1185  3447  1258   717  1204  1925   433   666  2251  2150  4151  2030  7418  2684  2149  3812  1154
 [253]  1657  1603  5302  2748  1231   802  6304  1533  8978   999  2662  1402 12169  3060 11998  2697  2404  4611
 [271]  1901  3368  1574  1445  1520  3878 10722  4788  7582  1092  1024  1076  9398  6419  4796  7629  9960  4675
 [289]  1287  2515  2745   672  3804  1344  1038 10127  1543  4811   727  1237   276  5381  5511  3749   685  1494
 [307]  2746   708  4351  3643  4249  1938  2910  2659  1028  3398  5801  1525  4473  1068  6615  1864  7408 11590
 [325]  4110  3384  2101  1275  4169  1521  5743  3599  3213  4439  3949  1459   882  3758  1743  1136  1236   959
 [343]  3229  6199  1246  2331  4463   776  2406  1239  3399  2247  1766  2473  1542  3850  3650  3446  3001  3079
 [361]  6070  2146 13756 14782  7685  2320   846 14318   362  2212 12976  1283  1330  4272  2238  1126  7374  2326
 [379]  1820   983  3249  1957 11760  2578  2348  1223  1516  1473  1887  8648  2899  2039  2197  1053  3235   939
 [397]  1967  7253  2292  1597  1381  5842  2579  8471  2782  1042  3186  2028   958  1591  2762  2779  2743  1149
 [415]  1313  1190  3448 11328  1872  2058  2136  1484   660  3394   609  1884  1620  2629   719  5096  1244  1842
 [433]  2576  1512 11054   518  2759  2670  4817  2679  3905  3386   343  4594  3620  1721  3017   754  1950  2924
 [451]  1659  7238  2764  4679  3092   448   654  1238  1245  3114  2569  5152  1037  3573  1201  3622   960  1163
 [469]  1209  3077  3757  1418  3518  1934  8318   368  2122  2996  9034  1585  1301  1323  3123  5493  1216  1207
 [487]  1309  2360  6850  8588   759  4686  2687   585  2255  1361  7127  1203   700  5507  3190  7119  3488  1113
 [505]  7966  1532  1503  2302   662  2273  2631  1311  3105  2319  3612  7763  3049  1534  2032  6350  2864  1255
 [523]  1333  2022  1552   626  8858   996  1750  6999  1995  1331  2278  5003  3552  1928  2964  1546   683 12389
 [541]  4712  1553  1372  3979  6758  3234  5433   806  1082  2788  2930  1927  2820   937  1056  3124  1388  2384
 [559]  2133  2799  1289  1217  2246   385  1965  1572  2718  1358   931  1442  4241  2775  3863  2329   918  1837
 [577]  3349  2828  4526  2671  2051  1300   741  3357  3632  1808 12204  9157  3676  3441   640  3652  1530  3914
```

```
> as.data.frame(table(pfda_df$credit_amount))
    Var1 Freq
1    250    6
2    276    2
3    338    3
4    339    3
5    343    4
6    362    4
7    368    5
8    385    7
9    392    2
10   409    6
11   426    8
12   428    7
13   433   12
14   437    8
15   446    2
16   448    3
17   454    6
18   458    3
19   461    3
20   484    8
21   518    1
22   522    3
23   538    5
24   571    4
25   585    3
26   590    5
27   601    1
28   609    8
29   610    4
30   618    6
31   622    5
32   625    5
33   626    6
34   629    7
35   634    3
36   639    8
```

```
463 1963    2
464 1965    7
465 1967    3
466 1977   10
467 1978    4
468 1979    6
469 1980    4
470 1983    5
471 1984    4
472 1987    7
473 1995    6
474 2002    5
475 2012    2
476 2022    2
477 2028    9
478 2030    3
479 2032    3
480 2039    9
481 2040    4
482 2051    4
483 2057    5
484 2058    1
485 2061    5
486 2063    3
487 2064    6
488 2069    7
489 2073    4
490 2080    3
491 2092    2
492 2096    1
493 2100    5
494 2101    2
495 2108    7
496 2116    2
497 2118    8
498 2121    1
499 2122    4
500 2124    8
```



Convert the data type of credit_amount column to integer by normal rounding for better distribution of the data.

savings_account_status column (previously known as savings_status)

```
> # 6.1 Check the details of this column
> summary(pfda_df$savings_account_status)
   Length     Class      Mode
     6000 character character
```

```
> # 6.2 Check the frequency of each unique category
> unique(pfda_df$savings_account_status)
[1] "no known savings" "<100"             "500<=X<10000"     ">=1000"           "100<=X<500"
> as.data.frame(table(pfda_df$savings_account_status))
              Var1 Freq
1             <100 3586
2          >=1000  379
3       100<=X<500  652
4     500<=X<10000  473
5 no known savings  910
```

Check the details of the savings_account_status column.

employment_years column (previously known as employment)

```
> # 7.1 Check the details of this column
> summary(pfda_df$employment_years)
   Length     Class      Mode
     6000 character character
```

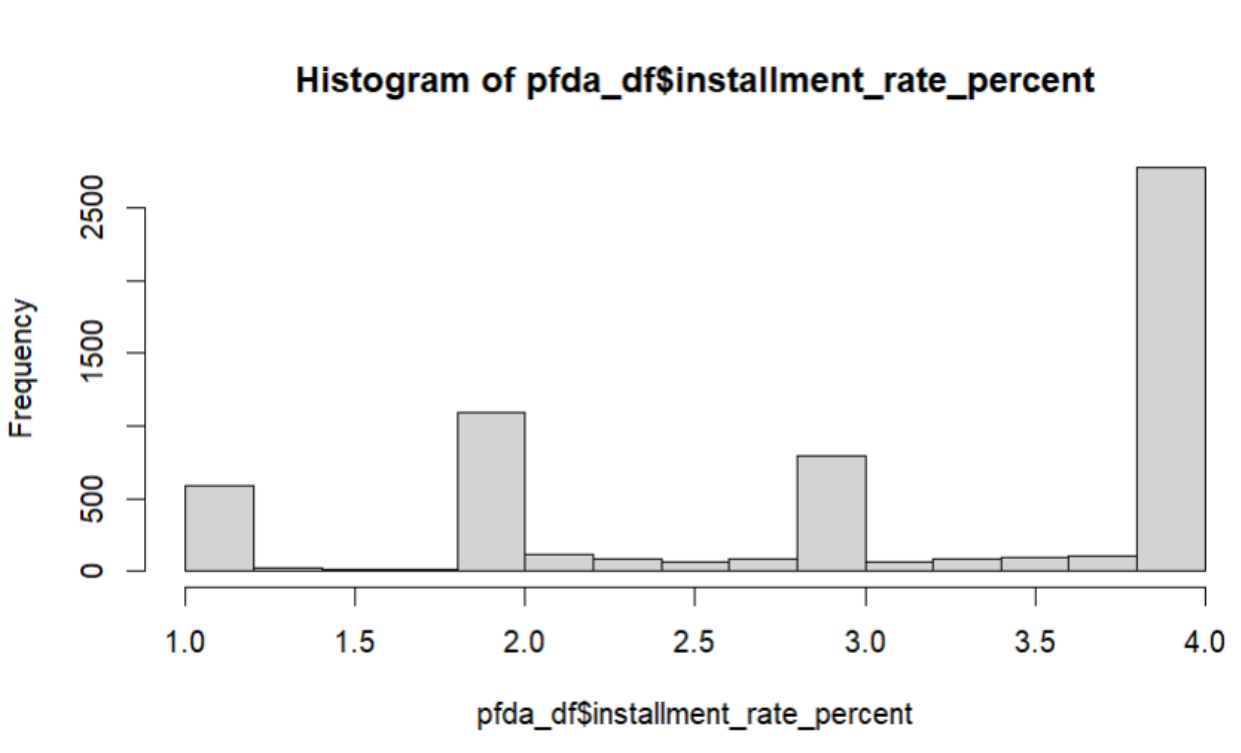
```
> # 7.2 Check the frequency of each unique category
> unique(pfda_df$employment_years)
[1] ">=7"        "1<=X<4"     "4<=X<7"     "unemployed" "<1"
> as.data.frame(table(pfda_df$employment_years))
        Var1 Freq
1         <1 1120
2        >=7 1227
3     1<=X<4 2158
4     4<=X<7 1199
5 unemployed  296
```

Check the details of the employment_years column.

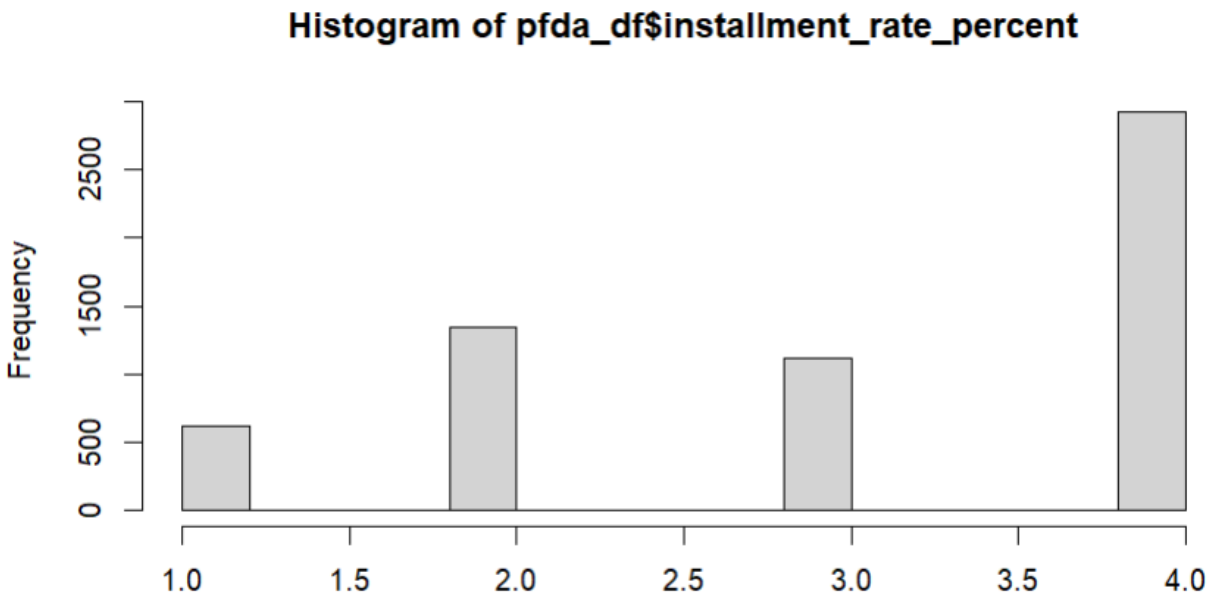
installment_rate_percent column (previously known as installment_commitment)

```
> # 8.1 Check the details of this column
> summary(pfda_df$installment_rate_percent)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   2.000   3.353   3.058   4.000   4.000
```

```
> # 8.2 Check all unique numbers
> unique(pfda_df$installment_rate_percent)
  [1] 4.000000 2.000000 3.000000 1.000000 2.226766 1.655948 1.328153 2.155042 2.036318 3.961165 3.393469 2.326651
 [13] 2.344262 2.728457 2.034748 2.371795 3.411354 1.901463 1.867255 2.969537 3.714595 2.041068 2.196463 3.433521
 [25] 3.250861 3.631333 3.311500 2.058194 3.915214 3.239787 3.812253 3.634268 3.934614 3.276541 1.685777 3.730039
 [37] 3.047716 2.301888 3.639506 3.316241 2.183598 2.188314 1.934240 3.020489 3.824670 1.051483 2.386174 3.346304
 [49] 3.649774 3.898291 2.226639 2.478213 3.056108 2.888503 2.727193 2.853266 2.188796 2.358721 1.374060 2.179573
 [61] 2.651485 2.663446 3.229207 3.363227 3.478611 2.511888 2.614219 2.954990 2.078373 1.835892 2.967581 3.424075
 [73] 3.463296 1.867117 2.320155 2.476472 3.068655 2.503439 3.220764 3.152857 3.517479 2.506974 2.397559 3.838584
 [85] 2.886535 3.792787 1.051504 2.163036 1.399394 3.515442 2.852183 2.189955 3.675691 2.466456 2.273723 1.358920
 [97] 1.710056 2.184613 3.408111 2.963334 3.115253 3.751276 2.405202 3.428527 3.740974 3.681039 2.477715 3.056740
[109] 2.872672 2.395782 2.044199 3.443995 2.645373 2.161769 3.390822 3.849131 3.995831 3.231596 3.474201 2.789044
[121] 3.893244 2.830320 3.353152 2.164805 2.229395 2.531876 3.480100 3.453241 3.720236 2.233620 2.191654 3.187127
[133] 2.256512 3.838353 3.505497 2.822510 2.558890 2.985286 2.376739 2.250422 3.607011 3.487493 2.861837 2.127024
[145] 1.401556 1.041016 3.749911 1.260503 2.547758 3.622149 3.241085 3.725268 3.378317 3.878221 3.513998 1.731328
[157] 2.044369 3.662743 3.076668 2.208182 3.179912 2.579464 3.784668 3.278723 3.903151 2.765473 3.608088 2.890369
[169] 2.817996 3.179191 3.094893 3.302358 3.796472 2.378040 3.526274 3.620806 3.372608 2.738950 2.462737 3.584766
[181] 3.599700 2.606997 2.765016 1.125682 3.030472 3.919331 2.776151 1.821791 2.964752 2.347619 2.902569 2.396729
[193] 2.257497 2.643166 2.465029 2.622714 1.007785 2.736602 3.986257 1.463409 2.075958 3.468734 2.937659 3.852261
[205] 1.938503 3.155559 3.250384 1.030472 3.734430 2.708293 2.525190 2.840710 2.178409 3.535336 2.685854 3.648823
[217] 2.831601 1.826929 3.128917 3.378773 3.337447 3.899647 3.244353 3.498248 3.669652 3.984931 2.450666 2.903930
[229] 1.512190 3.552453 2.175775 2.111307 1.473058 3.896729 3.945273 2.606077 3.997256 3.103390
```




```
> # 8.4 Convert the installment rate from double to integer by using standard rounding
> pfda_df$installment_rate_percent <- round(pfda_df$installment_rate_percent)
> unique(pfda_df$installment_rate_percent)
[1] 4 2 3 1
> as.data.frame(table(pfda_df$installment_rate_percent))
  Var1 Freq
1    1  621
2    2 1340
3    3 1114
4    4 2925
```



Convert the data type of installment_rate_percent column to integer by normal rounding for better distribution of the data.

personal_status column

```
> # 9.1 Check the details of this column
> summary(pfda_df$personal_status)
   Length     Class      Mode
     6000 character character
> # 9.2 Check the frequency of each unique category
> unique(pfda_df$personal_status)
[1] "male single"       "female div/dep/mar" "male div/sep"       "male mar/wid"
> as.data.frame(table(pfda_df$personal_status))
                Var1 Freq
1 female div/dep/mar 1932
2       male div/sep  517
3       male mar/wid  768
4        male single 2783
```

Check the details of the personal_status column.

other_parties column

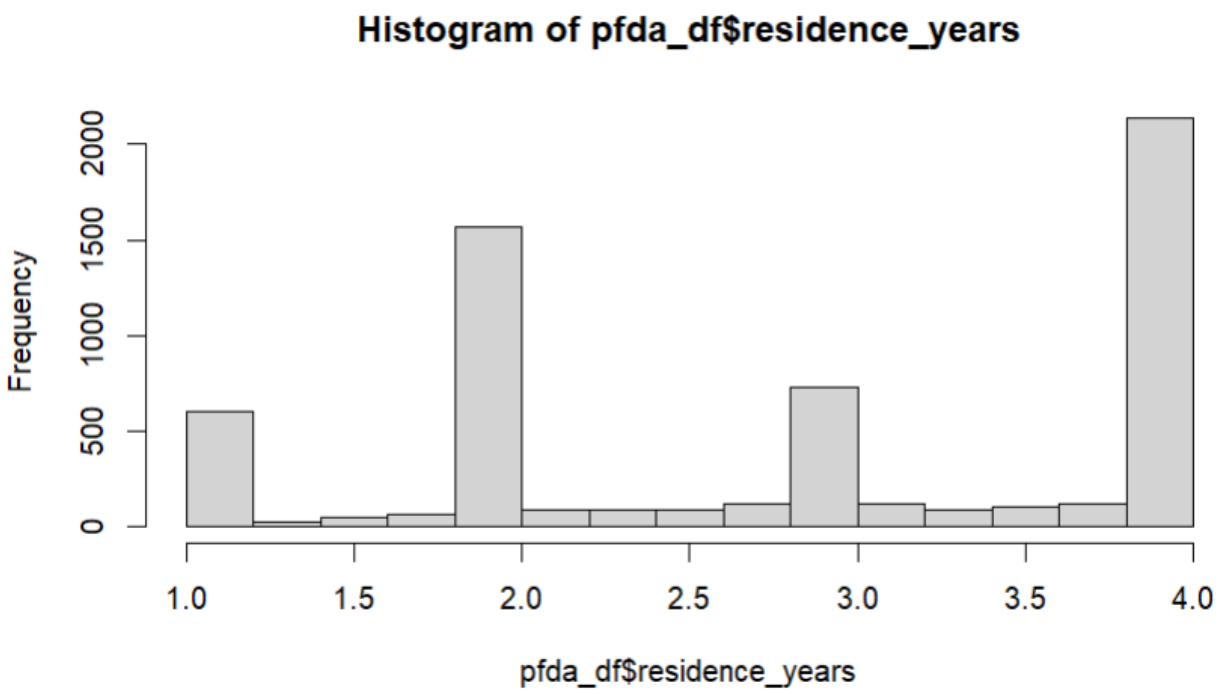
```
> # 10.1 Check the details of this column
> summary(pfda_df$other_parties)
   Length     Class      Mode
     6000 character character
> # 10.2 Check the frequency of each unique category
> unique(pfda_df$other_parties)
[1] "none"         "guarantor"    "co applicant"
> as.data.frame(table(pfda_df$other_parties))
          Var1 Freq
1 co applicant  226
2    guarantor  403
3         none 5371
```

Check the details of other_parties columns.

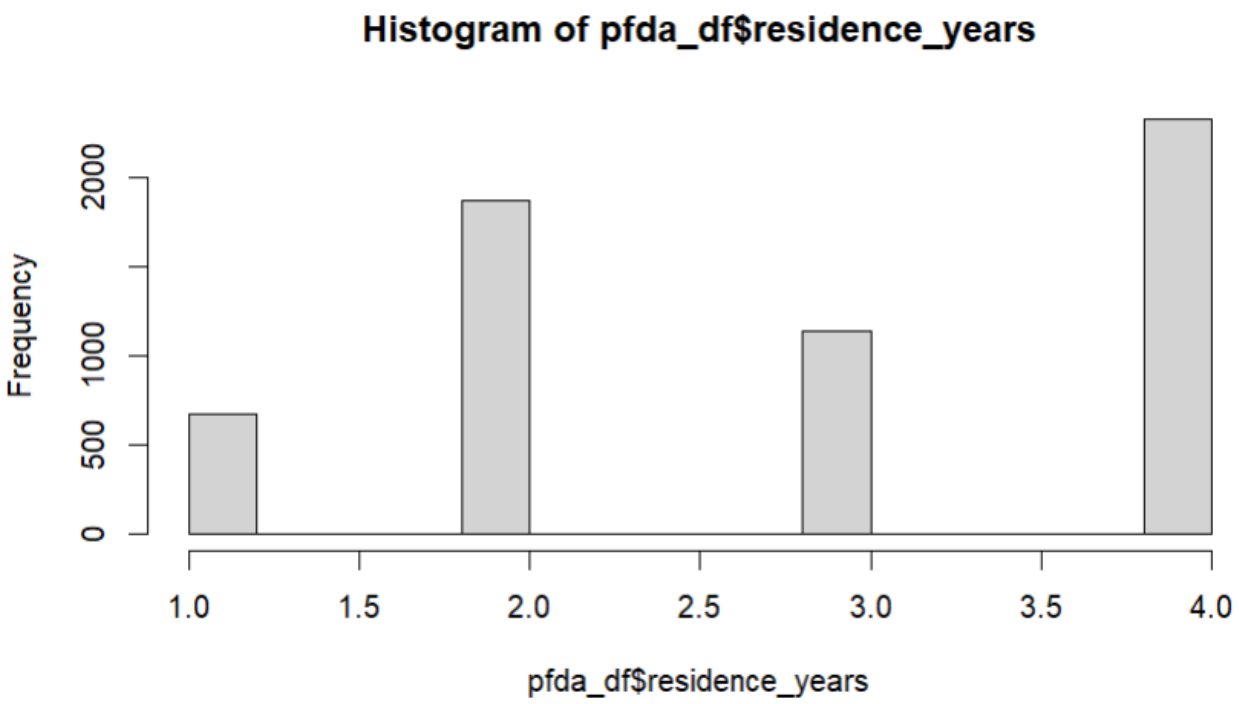
residence_years column (previously known as residence_since)

```
> # 11.1 Check the details of this column
> summary(pfda_df$residence_years)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    2.00    3.00    2.85    4.00    4.00
```

```
> # 11.2 Check all unique numbers
> unique(pfda_df$residence_years)
  [1] 4.000000 2.000000 3.000000 1.000000 3.858118 1.603216 2.949231 1.340149 3.917150 2.842270 3.015542 2.310083
 [13] 3.922331 3.696734 2.388753 2.653302 3.172131 2.069496 3.085470 2.315954 3.611240 3.588646 2.600976 3.138353
 [25] 3.601764 3.969537 3.041068 3.488085 3.250861 2.368667 3.838254 3.143371 2.623000 3.825419 3.557648 2.182866
 [37] 3.934614 2.723459 1.966607 2.692981 2.867926 3.841354 2.240329 2.341880 3.816402 1.188314 1.934240 3.959021
 [49] 2.350661 3.965678 2.386174 3.653696 2.929348 1.350226 1.144177 1.949146 3.773361 1.739106 3.954086 1.498960
 [61] 1.615851 1.168325 3.888503 1.867199 3.363596 2.800949 3.871500 3.622408 1.538082 2.820427 2.697030 1.793316
 [73] 3.452078 2.429178 2.618254 1.655723 1.614603 2.021874 2.992075 2.307110 2.696673 2.126704 1.417946 2.064839
 [85] 3.547972 3.431480 3.194943 3.616931 2.499967 3.601350 1.960466 3.739909 3.377729 1.534327 2.496561 2.595270
 [97] 2.053485 2.152857 1.354905 1.129597 3.353540 3.034958 2.493026 3.750147 2.183408 2.795118 3.361816 3.892389
[109] 3.659606 3.948496 3.474071 3.545916 2.062696 2.688030 3.163036 3.866869 2.838481 1.426091 1.793303 3.486674
[121] 3.300316 2.418695 3.726277 3.723081 2.591889 2.291819 2.481667 3.377988 2.589831 3.746318 2.784394 3.571473
[133] 2.740974 2.744212 2.475013 3.599777 2.433145 2.082891 1.436336 3.556005 3.322686 1.043401 1.080884 1.609178
[145] 3.166215 2.216809 3.991663 3.474201 1.254237 2.394522 2.006980 2.830320 3.646848 3.223463 3.541211 2.360345
[157] 3.740050 3.093518 1.064632 1.279764 2.397218 3.458085 1.228601 1.178405 3.583361 2.256512 3.838353 3.917252
[169] 2.876661 2.670331 2.588745 1.044142 3.246522 2.500843 2.785979 1.430919 2.127024 1.036822 1.401556 3.986328
[181] 2.075359 3.691714 3.068693 2.749911 1.086834 3.273879 2.485838 1.622149 2.482170 3.794955 3.251442 2.585545
[193] 2.243559 2.243001 1.953104 2.022185 3.523789 2.168629 2.615554 2.032026 3.734775 2.416364 1.820088 2.841072
[205] 2.721277 3.415134 1.382736 2.039075 3.636644 3.109631 2.182004 1.726397 2.452554 3.549090 3.910471 1.893878
[217] 2.302358 3.592944 3.540653 3.867528 3.379194 2.694827 3.372608 3.159300 1.658437 3.074527 3.079762 3.169532
[229] 3.199399 3.093003 2.360135 2.765016 1.377047 3.030472 2.053779 2.601935 1.776151 2.643581 2.690165 2.695238
[241] 2.063609 3.068847 3.105783 1.669900 2.907709 3.023314 2.313955 2.754572 3.994810 1.868301 3.830966 1.402722
[253] 3.958771 2.609772 3.236117 3.759955 3.062532 3.875317 2.852261 1.938503 3.688881 2.030472 1.132785 2.645853
[265] 3.790262 3.262595 3.182284 2.563965 3.840710 3.070671 3.233216 3.657073 3.526030 3.355104 2.946469 1.479582
[277] 2.752598 3.263178 3.442258 2.621227 2.337447 3.799295 3.215110 2.488706 2.501752 3.008955 2.132616 1.015069
[289] 1.675998 1.487810 3.552453 3.824225 2.805865 3.833040 2.315372 2.036485 1.696961 1.997256
```




```
> # 11.4 Convert the residence period from double to integer by using standard rounding
> pfda_df$residence_years <- round(pfda_df$residence_years)
> unique(pfda_df$residence_years)
[1] 4 2 3 1
> as.data.frame(table(pfda_df$residence_years))
  Var1 Freq
1    1  670
2    2 1870
3    3 1134
4    4 2326
```



Convert the data type of residence_years column to integer by normal rounding for better distribution of data.

property_magnitude column

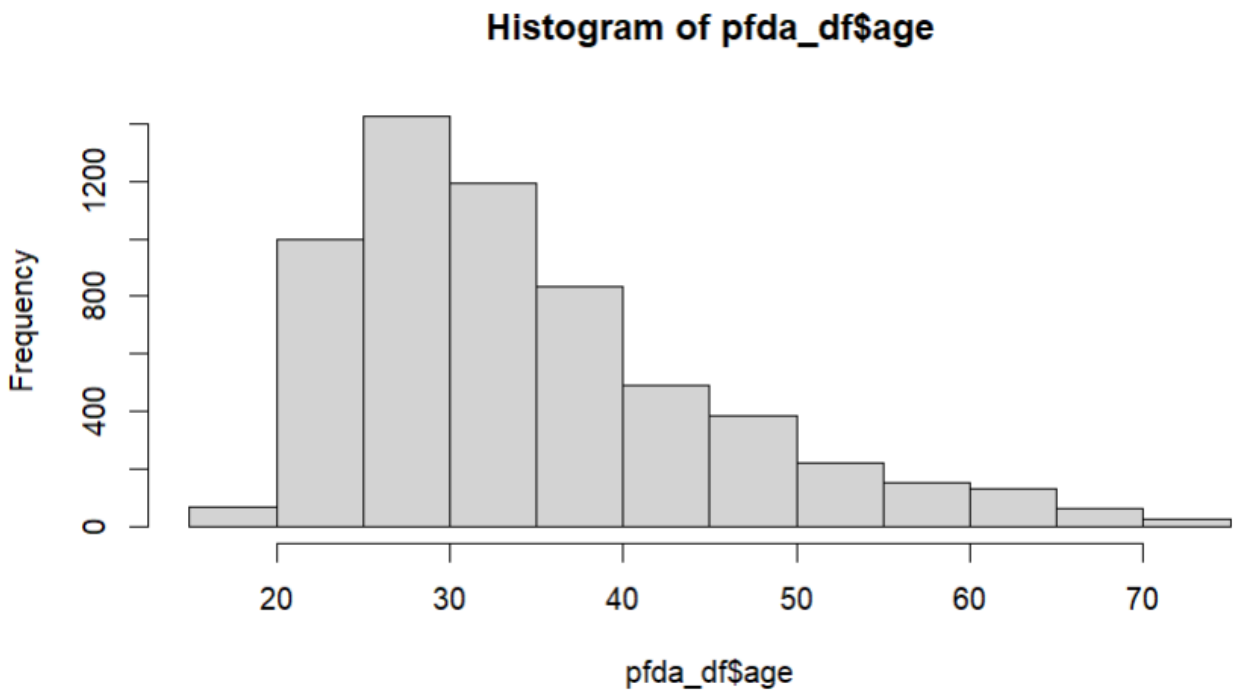
```
> # 12.1 Check the details of this column
> summary(pfda_df$property_magnitude)
   Length     Class      Mode
     6000 character character
> # 12.2 Check the frequency of each unique category
> unique(pfda_df$property_magnitude)
[1] "real estate"       "life insurance"    "no known property" "car"
> as.data.frame(table(pfda_df$property_magnitude))
               Var1 Freq
1               car 2156
2    life insurance 1576
3 no known property 1002
4       real estate 1266
```

Check the details of the property_magnitude column.

age column

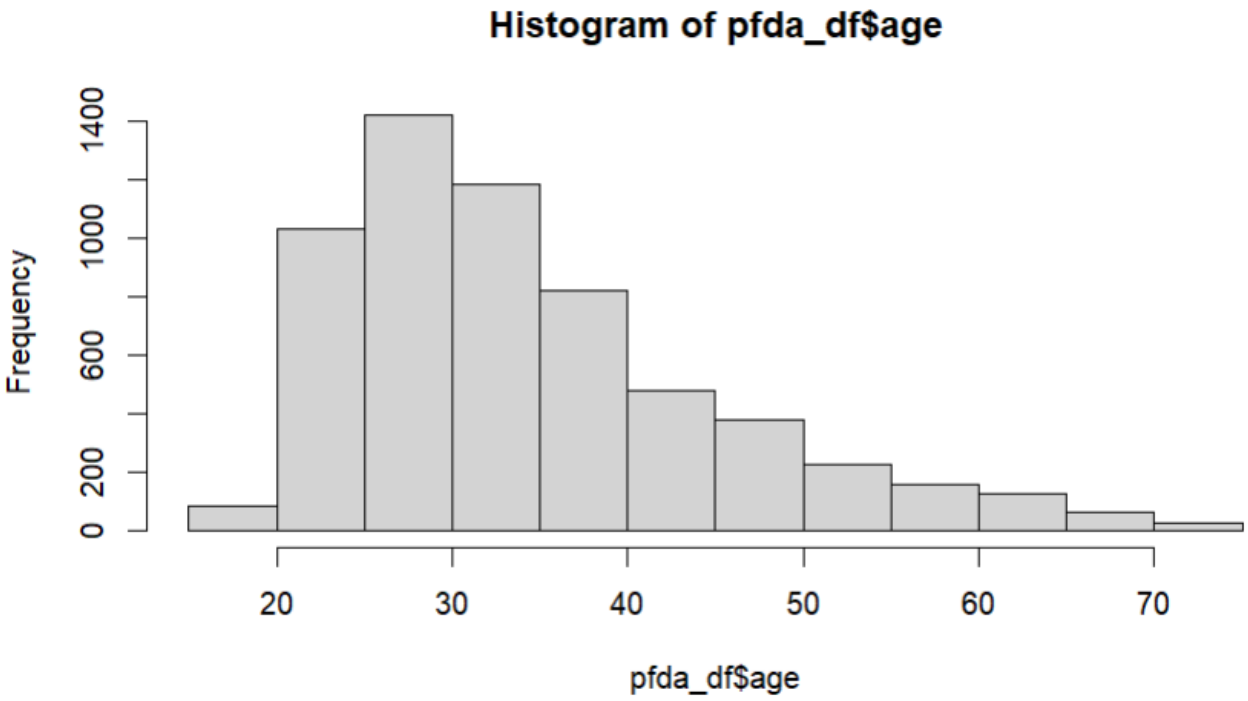
```
> # 13.1 Check the details of this column
> summary(pfda_df$age)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  19.00   27.00   32.00   34.95   40.00   75.00
```

```
> # 13.2 Check all unique numbers
> unique(pfda_df$age)
  [1] 67.00000 22.00000 49.00000 45.00000 53.00000 35.00000 61.00000 28.00000 25.00000 24.00000 60.00000 32.00000 44.00000
 [14] 31.00000 48.00000 26.00000 36.00000 39.00000 42.00000 34.00000 63.00000 27.00000 30.00000 57.00000 33.00000 37.00000
 [27] 58.00000 23.00000 29.00000 52.00000 50.00000 46.00000 51.00000 41.00000 40.00000 66.00000 47.00000 56.00000 54.00000
 [40] 20.00000 21.00000 38.00000 70.00000 65.00000 74.00000 68.00000 43.00000 55.00000 64.00000 75.00000 19.00000 62.00000
 [53] 59.00000 31.43247 26.96784 39.84769 25.79368 38.94477 19.84227 47.75390 64.72026 38.09325 33.98446 31.54477 29.70453
 [66] 34.41879 26.50452 54.10680 24.81959 45.98207 35.08610 24.82787 24.27154 35.72201 40.20086 35.36965 44.44496 25.41135
 [79] 28.39902 30.18212 25.32745 23.57880 25.42919 52.76797 30.97642 24.76873 29.95234 27.52237 38.12899 23.56611 29.99820
 [92] 25.62300 20.98929 31.36472 34.15514 29.18287 46.82305 23.86643 23.34289 20.87135 38.49051 29.33401 33.84907 39.73083
[105] 27.20165 28.68376 69.86904 22.37663 29.88587 22.26636 29.22731 67.81123 56.03065 43.03456 25.34630 22.67663 33.70045
[118] 24.90388 33.38975 21.35984 30.86928 23.16835 49.46749 36.99168 48.39377 34.47640 37.92681 25.39276 61.38330 38.53120
[131] 27.18158 28.39810 36.38584 33.55980 21.51400 61.01455 26.25552 23.37406 31.35915 43.89506 24.44679 58.05513 25.50212
[144] 26.73645 35.49348 23.38175 28.34622 33.98404 34.90667 28.17936 30.25313 23.52139 47.33754 23.53555 28.34834 29.83720
[157] 37.19593 33.83589 22.29178 25.80811 27.65818 27.84260 30.12153 31.49990 29.70816 25.32883 33.81947 24.43158 57.56124
[170] 22.04036 40.66765 30.73363 33.22088 32.99312 36.55847 23.70236 24.19982 36.41166 53.71479 43.31857 28.41962 44.39521
[183] 24.64646 29.65731 32.01395 23.24985 41.55022 48.20488 29.64444 24.42546 38.30053 30.03166 26.88654 24.27628 20.36053
[196] 47.92269 43.84075 31.81126 49.76415 33.47022 38.44015 32.95643 33.13313 36.54633 37.55166 31.32968 42.97293 30.22531
[209] 28.84009 40.73165 27.58131 67.84123 34.80360 33.52067 30.63077 33.71356 33.99564 29.41636 29.07333 23.48805 60.24407
[222] 28.76148 35.36722 27.77897 26.88844 23.57147 31.70487 27.69842 25.12652 40.16317 31.59480 36.19990 48.19510 58.82172
[235] 26.58024 23.94326 29.67237 25.59367 35.77901 23.55601 26.93612 27.17360 38.67821 25.67646 25.78289 38.90822 34.24565
[248] 30.50406 36.08405 29.34955 40.94581 44.78770 31.15798 34.47911 33.06780 31.79351 28.23146 23.97261 22.89324 49.30774
[261] 54.06120 23.64685 29.28212 30.54121 30.20219 23.63965 29.44030 27.09352 20.25853 33.55953 31.46724 30.58887 24.54192
[274] 58.26637 41.39040 25.88106 41.66722 29.90581 27.32329 30.66901 34.76661 37.06747 30.65934 29.17749 36.88643 33.36259
[287] 24.92643 30.65108 32.74452 37.42804 27.65833 31.04929 19.76214 41.77907 31.66148 23.23242 23.75359 29.62428 35.48085
[300] 40.99893 31.65266 25.48910 27.35673 23.75708 28.91206 27.16976 31.07568 37.97115 34.07688 30.82891 23.75644 26.45801
[313] 23.32828 29.38797 22.11092 33.47982 59.19074 30.72160 32.30999 51.49157 27.96797 40.02751 30.82910 36.79912 40.11858
[326] 46.55982 36.35400 45.22018 29.51576 28.97557 31.95956 25.81678 24.89037 43.90797 32.37765 38.67150 33.50910 39.55236
[339] 31.55511 28.75963 41.67406 32.16825 47.54065 27.57909 24.38649 23.52989 39.73378 34.31209 27.91552 23.79087 33.53545
[352] 28.34156 50.87875 34.14107 41.37345 39.19489 36.41866 27.76215 37.78601 41.29329 32.74318 25.18283 28.43023 35.24258
[365] 30.26777 32.03970 41.58442 27.91428 45.84098 30.06156 25.32788 30.72810 48.37518 29.48501 28.33010 32.92950 38.89251
[378] 29.74511 22.93665 37.81474 25.82560 44.95958 26.01298 32.41961 45.32386 25.19456 26.97251 26.14636 42.37979 27.94447
[391] 29.72013 29.21861 34.96929 57.81805 35.62513 30.99351 30.43583 35.62224 27.06094 24.60164 29.64585 42.68792 27.73408
[404] 28.93892 39.01025 32.43603 28.77870 26.80284 27.71732 32.36398 54.62585 27.84382 33.67344 29.54176 36.83666 38.84199
[417] 29.42093 43.01968 40.27882 26.69338 53.18209 30.72426 23.20071 38.93599 31.17659 30.49124 33.65174 28.13262 20.03767
[430] 25.53857 29.60971 39.19346 22.84564 39.41759 37.89046 25.31537 49.63855 44.05473 24.90912 24.00274 25.51695
```



```
> # 13.4 Convert the age from double to integer by using standard rounding
> pfda_df$age <- round(pfda_df$age)
> unique(pfda_df$age)
 [1] 67 22 49 45 53 35 61 28 25 24 60 32 44 31 48 26 36 39 42 34 63 27 30 57 33 37 58 23 29
[30] 52 50 46 51 41 40 66 47 56 54 20 21 38 70 65 74 68 43 55 64 75 19 62 59
```

```
> as.data.frame(table(pfda_df$age))
   Var1 Freq
1    19    6
2    20   82
3    21   76
4    22  145
5    23  233
6    24  314
7    25  261
8    26  284
9    27  298
10   28  304
11   29  240
12   30  294
13   31  286
14   32  248
15   33  198
16   34  240
17   35  211
18   36  204
19   37  166
20   38  152
21   39  145
22   40  155
23   41   88
24   42  114
25   43   96
26   44   94
27   45   89
28   46   74
29   47   92
30   48   83
31   49   67
32   50   61
33   51   45
34   52   39
35   53   37
36   54   57
37   55   47
38   56   20
39   57   39
40   58   47
41   59   25
42   60   26
43   61   43
44   62   12
45   63   29
46   64   18
47   65   25
48   66   20
49   67   16
50   68   22
51   70    7
52   74   19
53   75    7
```



Convert the data type of the age column to integer by normal rounding for better distribution of the data.

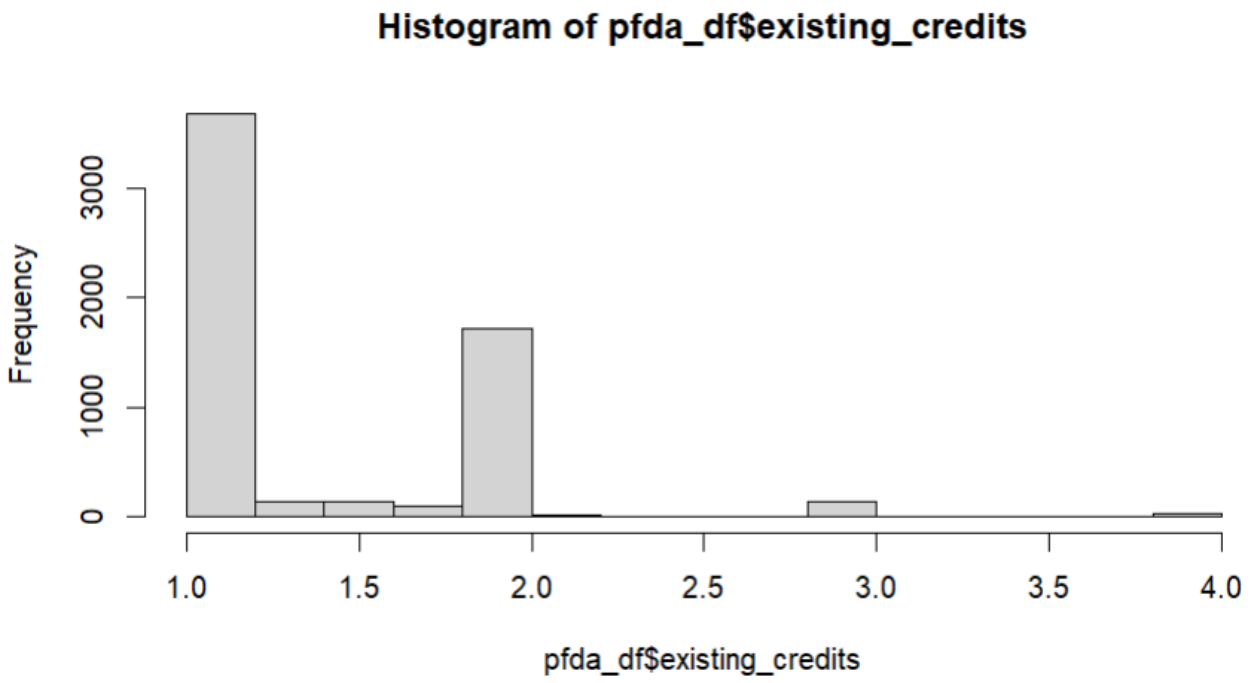
housing_type column (previously known as housing)

```
> # 15.1 Check the details of this column
> summary(pfda_df$housing_type)
   Length     Class      Mode
     6000 character character
> # 15.2 Check the frequency of each unique category
> unique(pfda_df$housing_type)
[1] "own"      "for free" "rent"
> as.data.frame(table(pfda_df$housing_type))
      Var1 Freq
1 for free  850
2      own 4310
3     rent  840
```

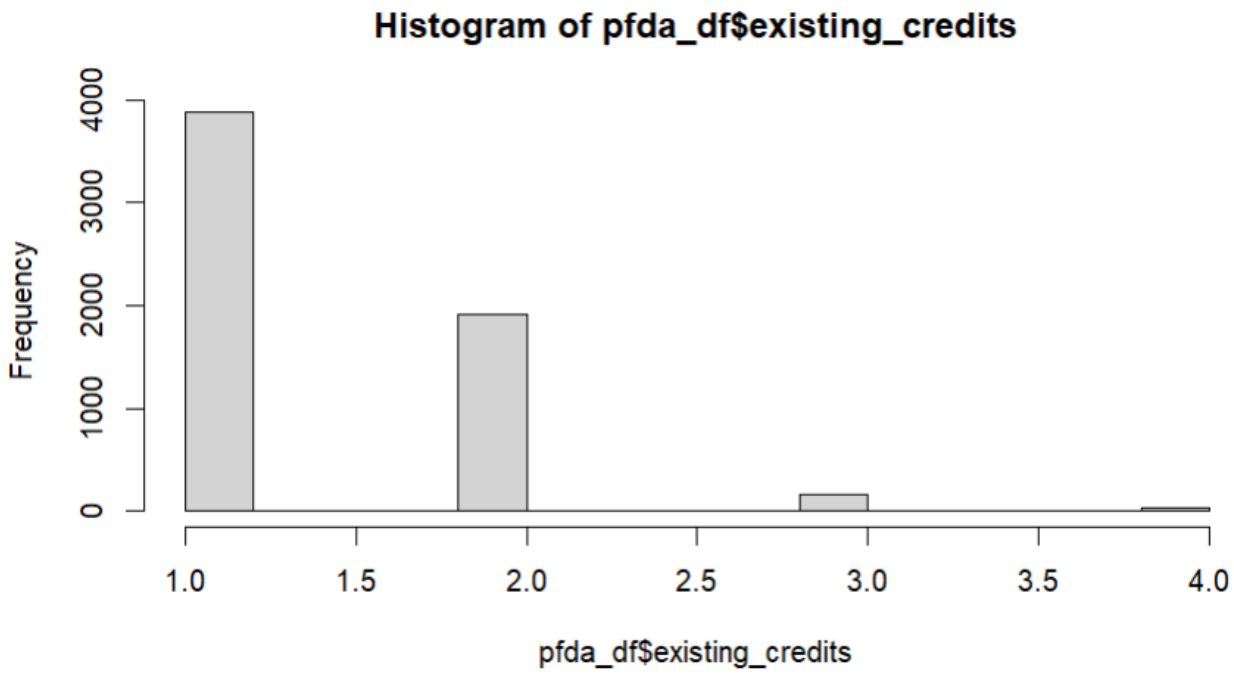
Check the details of the housing _type column.

existing_credits column

```
> # 16.1 Check the details of this column
> summary(pfda_df$existing_credits)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.392   2.000   4.000

> # 16.2 Check all unique numbers
> unique(pfda_df$existing_credits)
  [1] 2.000000 1.000000 3.000000 4.000000 1.603216 1.050769 1.113383 1.972383 1.578865
 [10] 1.561526 1.328153 1.250251 1.696734 1.462918 1.326651 1.728457 1.965252 1.901463
 [19] 1.867255 1.969537 1.714595 2.958932 1.250861 1.428314 1.557648 1.657111 1.269961
 [28] 1.240329 1.658120 1.386174 1.350226 1.949146 1.984695 1.944252 1.322832 1.573367
 [37] 1.128500 1.811204 1.179361 1.348515 2.961037 1.336554 1.614603 1.340625 1.063352
 [46] 1.417946 1.731648 1.127690 1.867117 1.438133 1.320155 1.260091 2.755458 1.534327
 [55] 1.946515 1.958541 1.323230 1.482521 1.091704 2.892389 1.113465 1.930929 1.995931
 [64] 3.211107 1.031348 1.163036 2.733737 1.573909 1.777424 1.853395 1.581305 2.880360
 [73] 1.236685 1.518333 1.622012 1.232805 1.237506 1.400223 1.477715 1.563664 1.977901
 [82] 1.924643 1.919116 1.924565 1.696925 1.615798 1.262899 1.084746 1.830320 1.388268
 [91] 1.265938 1.279764 1.588793 1.397218 1.729042 1.133185 1.583361 2.256512 1.917252
[100] 1.164834 1.411255 1.829164 1.430919 1.986328 1.465654 1.250089 1.173668 1.273879
[109] 1.758915 1.251442 1.207228 1.756999 1.731328 2.168629 1.967974 1.519954 1.420536
[118] 1.096849 1.617264 1.019537 1.182004 2.452554 1.089529 1.297959 1.203528 2.157909
[127] 1.091723 2.159300 1.537263 1.415234 1.125682 1.821791 1.968195 1.105783 1.678417
[136] 1.092291 2.518855 1.377286 3.772125 1.236117 1.240045 1.468734 1.720516 1.937659
[145] 1.938503 1.602836 1.104869 1.817716 1.563965 2.159290 1.464664 1.657073 1.442258
[154] 1.662553 1.100353 1.225333 1.194135 1.157686 1.896610
```



```
> # 16.4 Convert the existing credits from double to integer by using standard rounding
> pfda_df$existing_credits <- round(pfda_df$existing_credits)
> unique(pfda_df$existing_credits)
[1] 2 1 3 4
> as.data.frame(table(pfda_df$existing_credits))
  Var1 Freq
1    1 3886
2    2 1921
3    3  158
4    4   35
```



Convert the data type of the existing_credits column to integer by normal rounding for better distribution of the data.

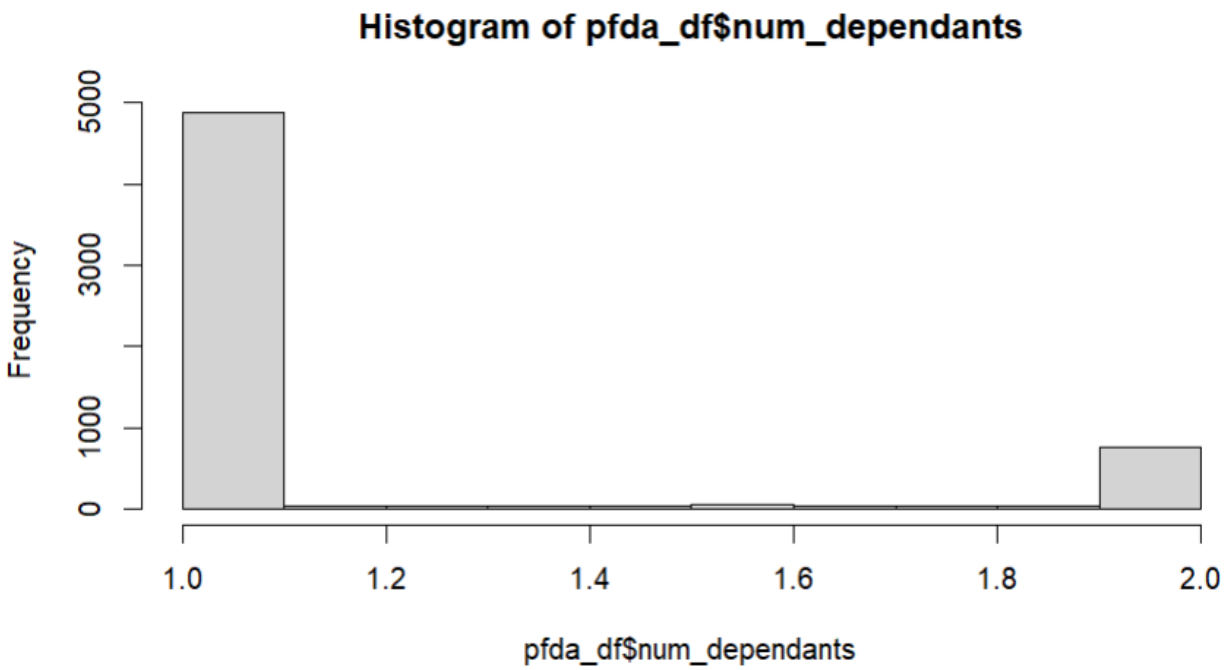
job_type column (previously known as job)

```
> # 17.1 Check the details of this column
> summary(pfda_df$job_type)
   Length     Class      Mode
     6000 character character
> # 17.2 Check the frequency of each unique category
> unique(pfda_df$job_type)
[1] "skilled"                   "unskilled resident"        "high qualif/self emp/mgmt"
[4] "unemp/unskilled non res"
> as.data.frame(table(pfda_df$job_type))
                        Var1 Freq
1 high qualif/self emp/mgmt 1019
2                   skilled 3645
3   unemp/unskilled non res  311
4         unskilled resident 1025
```

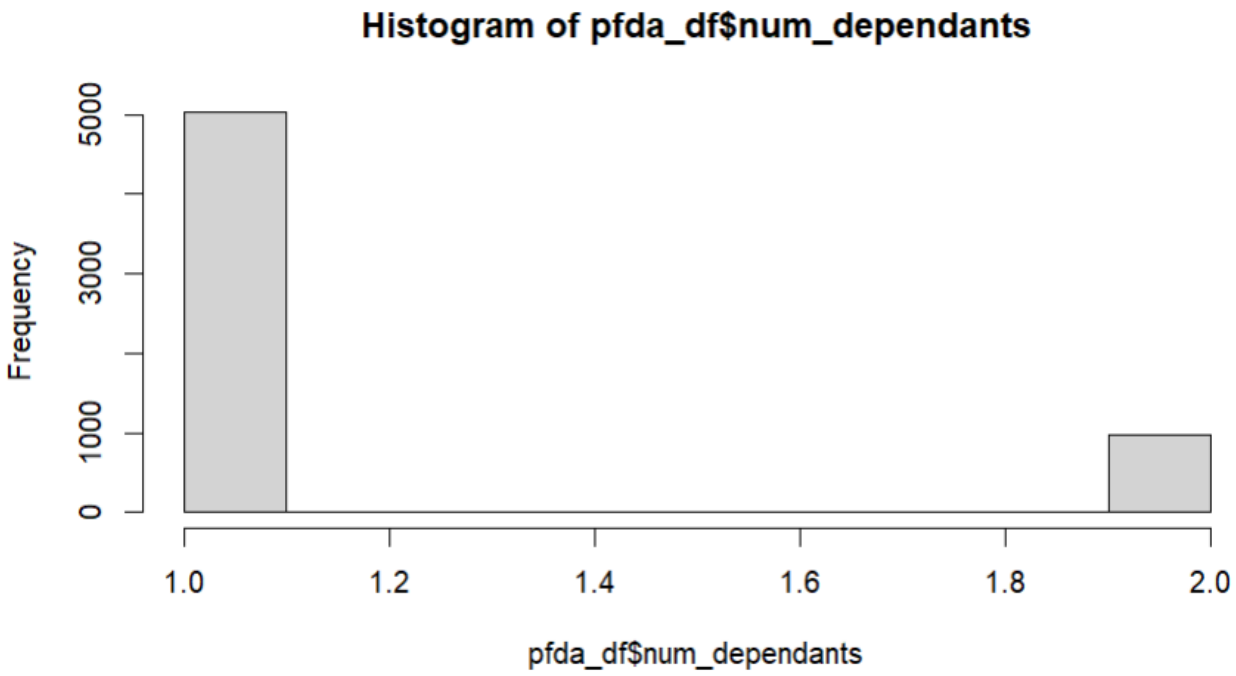
Check the details of the job_type column.

num_dependants column

```
> # 18.1 Check the details of this column
> summary(pfda_df$num_dependants)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   1.000   1.000   1.158   1.000   2.000

> # 18.2 Check all unique numbers
> unique(pfda_df$num_dependants)
  [1] 1.000000 2.000000 1.396784 1.050769 1.113383 1.421135 1.838934 1.250251 1.961165
 [10] 1.462918 1.728457 1.965252 1.805620 1.969537 1.041068 1.398821 1.250861 1.815666
 [19] 1.080873 1.557648 1.730039 1.047716 1.841354 1.908201 1.193087 1.539377 1.055748
 [28] 1.179361 1.614603 1.731648 1.580544 1.438133 1.761764 1.622271 1.534327 1.389618
 [37] 1.702365 1.946515 1.958541 1.323230 1.946195 1.781296 1.838481 1.777424 1.108103
 [46] 1.863139 1.649782 1.681039 1.237506 1.082891 1.977901 1.924643 1.705506 1.995831
 [55] 1.057864 1.397218 1.076200 1.430919 1.036822 1.075359 1.757081 1.251442 1.725268
 [64] 1.878221 1.307777 1.911592 1.208182 1.519954 1.805045 1.903151 1.608088 1.450910
 [73] 1.910471 1.297959 1.540653 1.579650 1.539881 1.599700 1.697668 1.696501 1.588339
 [82] 1.347619 1.031805 1.511657 1.156978 1.915483 1.402722 1.240045 1.720516 1.749616
 [91] 1.602836 1.262595 1.563965 1.840710 1.644896 1.415801 1.442258 1.662553 1.244353
[100] 1.066308 1.512190 1.850818 1.087888 1.194135 1.018242 1.103390
```




```
> # 18.4 Convert the number of dependants from double to integer by using standard rounding
> pfda_df$num_dependants <- round(pfda_df$num_dependants)
> unique(pfda_df$num_dependants)
[1] 1 2
> as.data.frame(table(pfda_df$num_dependants))
  Var1 Freq
1    1 5035
2    2  965
```



Convert the data type of the num_dependants column to integer by normal rounding for better distribution of the data.

own_telephone column

```
> # 19.1 Check the details of this column
> summary(pfda_df$own_telephone)
   Length     Class      Mode
     6000 character character
> # 19.2 Check the frequency of each unique category
> unique(pfda_df$own_telephone)
[1] "yes"  "none"
> as.data.frame(table(pfda_df$own_telephone))
  Var1 Freq
1 none 3981
2  yes 2019
```

Check the details of the own_telephone column.

is_foreign_worker column (previously known as foreign_worker)

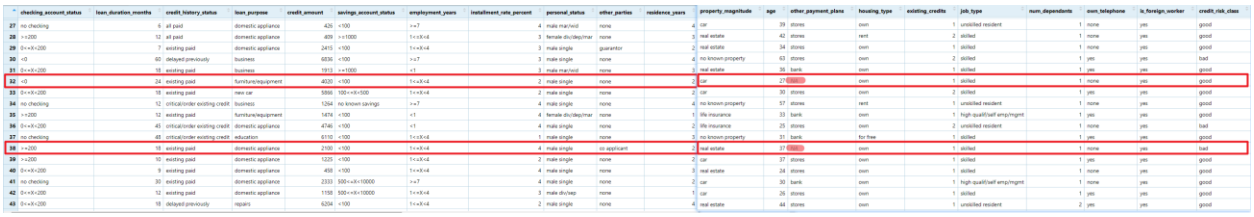
```
> # 20.1 Check the details of this column
> summary(pfda_df$is_foreign_worker)
   Length     Class      Mode
     6000 character character
>
> # 20.2 Check the frequency of each unique category
> unique(pfda_df$is_foreign_worker)
[1] "yes" "no"
> as.data.frame(table(pfda_df$is_foreign_worker))
  Var1 Freq
1   no  158
2  yes 5842
```

Check the details of the is_foreign_worker column.

other_payment_plans column

```
> # 14.1 Check the details of this column
> summary(pfda_df$other_payment_plans)
   Length     Class      Mode
     6000 character character
>
> # 14.2 Check the frequency of each unique category
> unique(pfda_df$other_payment_plans)
[1] "stores" "bank"   NA
> as.data.frame(table(pfda_df$other_payment_plans))
    Var1 Freq
1   bank 1124
2 stores 4651
>
> # 14.3 Check the number of NA values
> sum(is.na(pfda_df$other_payment_plans))
[1] 225
```

There are 225 rows of data with missing values in other_payment_plans column. We have to perform data imputation.



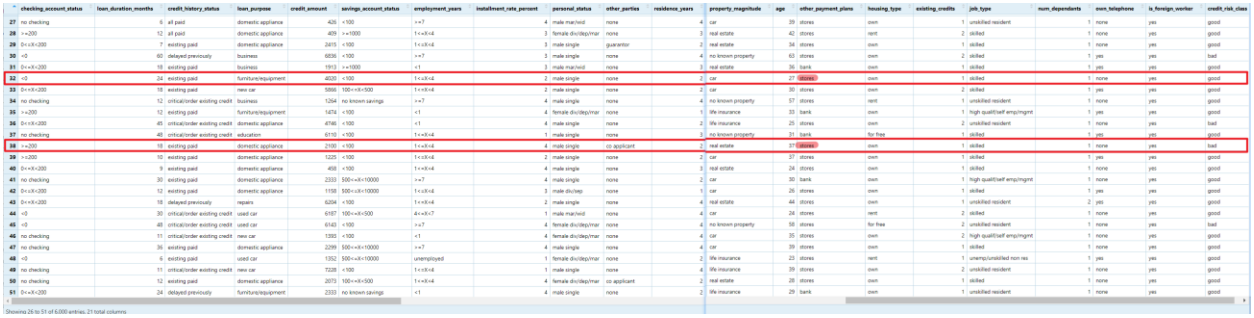
	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_rate_percent	personal_status	other_parties	residence_years
27	no checking	6	all paid	domestic appliance	426	<100	>=7	4	male mar/wid	none	4
28	>=200	12	all paid	domestic appliance	409	>=1000	1<=X<4	3	female div/dep/mar	none	3
29	0<=X<200	7	existing paid	domestic appliance	2415	<100	1<=X<4	3	male single	guarantor	2
30	<0	60	delayed previously	business	6836	<100	>=7	3	male single	none	4
31	0<=X<200	18	existing paid	business	1913	>=1000	<1	3	male mar/wid	none	3
32	<0	24	existing paid	furniture/equipment	4020	<100	1<=X<4	2	male single	none	2
33	0<=X<200	18	existing paid	new car	5866	100<=X<500	1<=X<4	2	male single	none	2
34	no checking	12	critical/order existing credit	business	1264	no known savings	>=7	4	male single	none	4
35	>=200	12	existing paid	furniture/equipment	1474	<100	<1	4	female div/dep/mar	none	1
36	0<=X<200	45	critical/order existing credit	domestic appliance	4746	<100	<1	4	male single	none	2
37	no checking	48	critical/order existing credit	education	6110	<100	1<=X<4	1	male single	none	3
38	>=200	18	existing paid	domestic appliance	2100	<100	1<=X<4	4	male single	co applicant	2
39	>=200	10	existing paid	domestic appliance	1225	<100	1<=X<4	2	male single	none	2
40	0<=X<200	9	existing paid	domestic appliance	458	<100	1<=X<4	4	male single	none	3
41	no checking	30	existing paid	domestic appliance	2333	500<=X<10000	>=7	4	male single	none	2
42	0<=X<200	12	existing paid	domestic appliance	1158	500<=X<10000	1<=X<4	3	male div/sep	none	1
43	0<=X<200	18	delayed previously	repairs	6204	<100	1<=X<4	2	male single	none	4

property_magnitude	age	other_payment_plans	housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class
car	39	stores	own	1	unskilled resident	1	none	yes	good
real estate	42	stores	rent	2	skilled	1	none	yes	good
real estate	34	stores	own	1	skilled	1	none	yes	good
no known property	63	stores	own	2	skilled	1	yes	yes	bad
real estate	36	bank	own	1	skilled	1	yes	yes	good
car	27	NA	own	1	skilled	1	none	yes	good
car	30	stores	own	2	skilled	1	yes	yes	good
no known property	57	stores	rent	1	unskilled resident	1	none	yes	good
life insurance	33	bank	own	1	high qualif/self emp/mgmt	1	yes	yes	good
life insurance	25	stores	own	2	unskilled resident	1	none	yes	bad
no known property	31	bank	for free	1	skilled	1	yes	yes	good
real estate	37	NA	own	1	skilled	1	none	yes	bad
car	37	stores	own	1	skilled	1	yes	yes	good
real estate	24	stores	own	1	skilled	1	none	yes	good
car	30	bank	own	1	high qualif/self emp/mgmt	1	none	yes	good
car	26	stores	own	1	skilled	1	yes	yes	good
real estate	44	stores	own	1	unskilled resident	2	yes	yes	good

```
> # 14.4 Convert the data type of all categorical columns from character to factor
> # To prepare the data for Random Forest imputation
> pfda_df <- pfda_df %>%
+   mutate(across(where(is.character), as.factor))
> str(pfda_df)
tibble [6,000 × 21] (S3: tbl_df/tbl/data.frame)
 $ checking_account_status : Factor w/ 4 levels "<0",">=200","0<=X<200",..: 1 3 4 1 1 4 4 3 4 3 ...
 $ loan_duration_months    : int [1:6000] 6 48 12 42 24 36 24 36 12 30 ...
 $ credit_history_status   : Factor w/ 4 levels "all paid","critical/order existing credit",..: 2 4 2 4 3 4 4 4 4 2 ...
 $ loan_purpose            : Factor w/ 9 levels "business","domestic appliance",..: 2 2 3 4 5 3 4 9 2 5 ...
 $ credit_amount           : int [1:6000] 1169 5951 2096 7882 4870 9055 2835 6948 3059 5234 ...
 $ savings_account_status  : Factor w/ 5 levels "<100",">=1000",..: 5 1 1 1 1 5 4 1 2 1 ...
 $ employment_years        : Factor w/ 5 levels "<1",">=7","1<=X<4",..: 2 3 4 4 3 3 2 3 4 5 ...
 $ installment_rate_percent: int [1:6000] 4 2 2 2 3 2 3 2 2 4 ...
 $ personal_status         : Factor w/ 4 levels "female div/dep/mar",..: 4 1 4 4 4 4 4 4 2 3 ...
 $ other_parties           : Factor w/ 3 levels "co applicant",..: 3 3 3 2 3 3 3 3 3 3 ...
 $ residence_years         : int [1:6000] 4 2 3 4 4 4 4 2 4 2 ...
 $ property_magnitude      : Factor w/ 4 levels "car","life insurance",..: 4 4 4 2 3 3 2 1 4 1 ...
 $ age                     : int [1:6000] 67 22 49 45 53 35 53 35 61 28 ...
 $ other_payment_plans     : Factor w/ 2 levels "bank","stores": 2 2 2 2 2 2 2 2 2 2 ...
 $ housing_type            : Factor w/ 3 levels "for free","own",..: 2 2 2 1 1 1 2 3 2 2 ...
 $ existing_credits        : int [1:6000] 2 1 1 1 2 1 1 1 1 2 ...
 $ job_type                : Factor w/ 4 levels "high qualif/self emp/mgmt",..: 2 2 4 2 2 4 2 1 4 1 ...
 $ num_dependants          : int [1:6000] 1 1 2 2 2 2 1 1 1 1 ...
 $ own_telephone           : Factor w/ 2 levels "none","yes": 2 1 1 1 1 2 1 2 1 1 ...
 $ is_foreign_worker       : Factor w/ 2 levels "no","yes": 2 2 2 2 2 2 2 2 2 2 ...
 $ credit_risk_class       : Factor w/ 2 levels "bad","good": 2 1 2 2 1 2 2 2 2 1 ...
> sapply(pfda_df, class)
 checking_account_status    loan_duration_months    credit_history_status            loan_purpose           credit_amount
                "factor"               "integer"                 "factor"                 "factor"               "integer"
  savings_account_status       employment_years installment_rate_percent         personal_status           other_parties
                "factor"                 "factor"                "integer"                 "factor"                 "factor"
          residence_years     property_magnitude                     age     other_payment_plans            housing_type
                "integer"                 "factor"                "integer"                 "factor"                 "factor"
        existing_credits                job_type          num_dependants           own_telephone       is_foreign_worker
                "integer"                 "factor"                "integer"                 "factor"                 "factor"
      credit_risk_class
                "factor"
```

Convert all the data type of all categorical columns from character to factor to prepare for data imputation.

```
# 14.5 Replace missing or empty strings with NA
pfda_df[pfda_df == ""] <- NA

# 14.6 Convert pfda_df from tibble to data frame
pfda_df = as.data.frame(pfda_df)

# 14.7 Perform Random Forest imputation
imputed_data <- missForest(pfda_df)
pfda_df <- imputed_data$ximp

# 14.8 Check whether there are still any missing values in other_payment_plans column
sum(is.na(pfda_df$other_payment_plans))
unique(pfda_df$other_payment_plans)
as.data.frame(table(pfda_df$other_payment_plans))
```



	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_rate_percent	personal_status	other_parties	residence_years
27	no checking	6	all paid	domestic appliance	426	<100	>=7	4	male mar/wid	none	4
28	>=200	12	all paid	domestic appliance	409	>=1000	1<=X<4	3	female div/dep/mar	none	3
29	0<=X<200	7	existing paid	domestic appliance	2415	<100	1<=X<4	3	male single	guarantor	2
30	<0	60	delayed previously	business	6836	<100	>=7	3	male single	none	4
31	0<=X<200	18	existing paid	business	1913	>=1000	<1	3	male mar/wid	none	3
32	<0	24	existing paid	furniture/equipment	4020	<100	1<=X<4	2	male single	none	2
33	0<=X<200	18	existing paid	new car	5866	100<=X<500	1<=X<4	2	male single	none	2
34	no checking	12	critical/order existing credit	business	1264	no known savings	>=7	4	male single	none	4
35	>=200	12	existing paid	furniture/equipment	1474	<100	<1	4	female div/dep/mar	none	1
36	0<=X<200	45	critical/order existing credit	domestic appliance	4746	<100	<1	4	male single	none	2
37	no checking	48	critical/order existing credit	education	6110	<100	1<=X<4	1	male single	none	3
38	>=200	18	existing paid	domestic appliance	2100	<100	1<=X<4	4	male single	co applicant	2
39	>=200	10	existing paid	domestic appliance	1225	<100	1<=X<4	2	male single	none	2
40	0<=X<200	9	existing paid	domestic appliance	458	<100	1<=X<4	4	male single	none	3
41	no checking	30	existing paid	domestic appliance	2333	500<=X<10000	>=7	4	male single	none	2
42	0<=X<200	12	existing paid	domestic appliance	1158	500<=X<10000	1<=X<4	3	male div/sep	none	1
43	0<=X<200	18	delayed previously	repairs	6204	<100	1<=X<4	2	male single	none	4
44	<0	30	critical/order existing credit	used car	6187	100<=X<500	4<=X<7	1	male mar/wid	none	4
45	<0	48	critical/order existing credit	used car	6143	<100	>=7	4	female div/dep/mar	none	4
46	no checking	11	critical/order existing credit	new car	1393	<100	<1	4	female div/dep/mar	none	4
47	no checking	36	existing paid	domestic appliance	2299	500<=X<10000	>=7	4	male single	none	4
48	<0	6	existing paid	used car	1352	500<=X<10000	unemployed	1	female div/dep/mar	none	2
49	no checking	11	critical/order existing credit	new car	7228	<100	1<=X<4	1	male single	none	4
50	no checking	12	existing paid	domestic appliance	2073	100<=X<500	1<=X<4	4	female div/dep/mar	co applicant	2
51	0<=X<200	24	delayed previously	furniture/equipment	2333	no known savings	<1	4	male single	none	2

Showing 26 to 51 of 6,000 entries, 21 total columns

property_magnitude	age	other_payment_plans	housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class
car	39	stores	own	1	unskilled resident	1	none	yes	good
real estate	42	stores	rent	2	skilled	1	none	yes	good
real estate	34	stores	own	1	skilled	1	none	yes	good
no known property	63	stores	own	2	skilled	1	yes	yes	bad
real estate	36	bank	own	1	skilled	1	yes	yes	good
car	27	stores	own	1	skilled	1	none	yes	good
car	30	stores	own	2	skilled	1	yes	yes	good
no known property	57	stores	rent	1	unskilled resident	1	none	yes	good
life insurance	33	bank	own	1	high qualif/self emp/mgmt	1	yes	yes	good
life insurance	25	stores	own	2	unskilled resident	1	none	yes	bad
no known property	31	bank	for free	1	skilled	1	yes	yes	good
real estate	37	stores	own	1	skilled	1	none	yes	bad
car	37	stores	own	1	skilled	1	yes	yes	good
real estate	24	stores	own	1	skilled	1	none	yes	good
car	30	bank	own	1	high qualif/self emp/mgmt	1	none	yes	good
car	26	stores	own	1	skilled	1	yes	yes	good
real estate	44	stores	own	1	unskilled resident	2	yes	yes	good
car	24	stores	rent	2	skilled	1	none	yes	good
no known property	58	stores	for free	2	unskilled resident	1	none	yes	bad
car	35	stores	own	2	high qualif/self emp/mgmt	1	none	yes	good
car	39	stores	own	1	skilled	1	none	yes	good
life insurance	23	stores	rent	1	unemp/unskilled non res	1	yes	yes	good
life insurance	39	stores	own	2	unskilled resident	1	none	yes	good
real estate	28	stores	own	1	skilled	1	none	yes	good
life insurance	29	bank	own	1	unskilled resident	1	none	yes	good

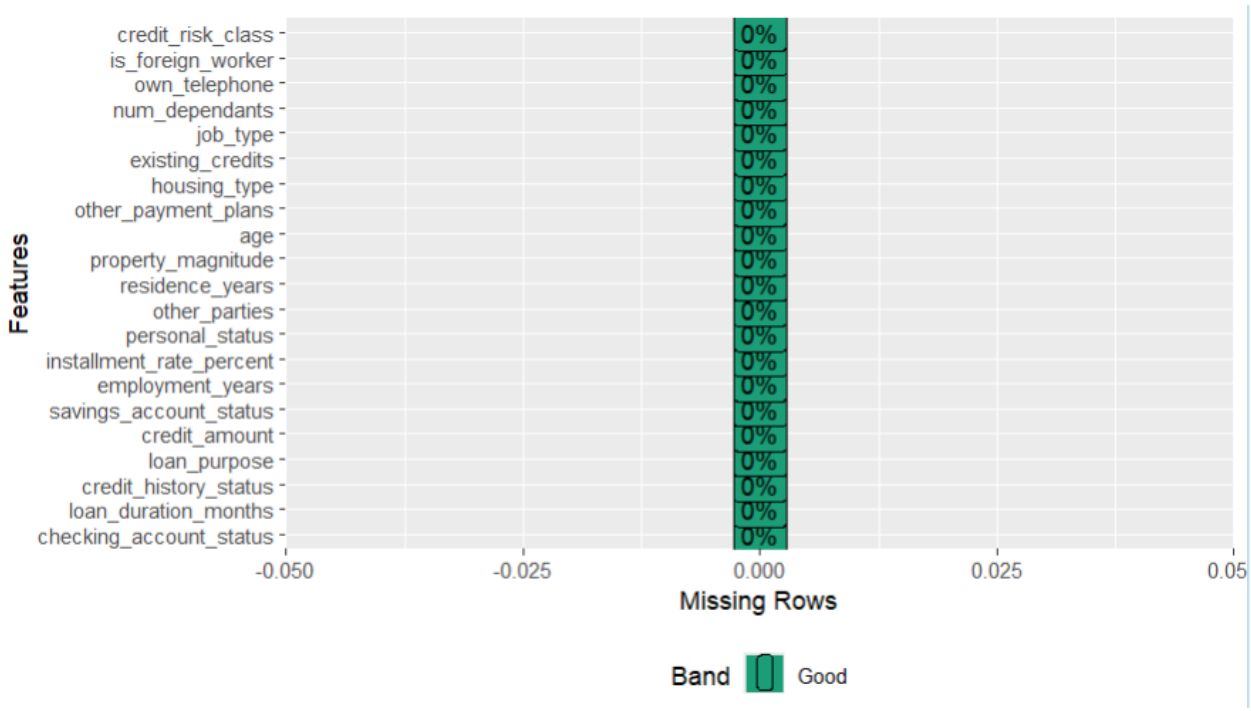
```
> # 14.8 Check whether there are still any missing values in other_payment_plans column
> sum(is.na(pfda_df$other_payment_plans))
[1] 0
> unique(pfda_df$other_payment_plans)
[1] stores bank
Levels: bank stores
> as.data.frame(table(pfda_df$other_payment_plans))
    Var1 Freq
1   bank 1124
2 stores 4876
```

Replace missing or empty strings with NA and convert the pdfa_df from tibble to data frame to prepare for Random Forest data imputation. Then, perform Random Forest imputation. After conducting the data imputation, there are no more missing values in other_payment_plans column and left with two categories which are bank and stores.

2.3 Data Validation

```
> # Data Validation
>
> summary(pfda_df)
 checking_account_status loan_duration_months credit_history_status loan_purpose       credit_amount   savings_account_status employment_years
 Length:6000             Min.   : 4.00        Length:6000           Length:6000        Min.   :  250   Length:6000            Length:6000
 Class :character        1st Qu.:12.00        Class :character      Class :character   1st Qu.: 1332   Class :character       Class :character
 Mode  :character        Median :20.00        Mode  :character      Mode  :character   Median : 2290   Mode  :character       Mode  :character
                         Mean   :22.03                                                 Mean   : 3344
                         3rd Qu.:27.00                                                 3rd Qu.: 4164
                         Max.   :72.00                                                 Max.   :18424
 installment_rate_percent personal_status    other_parties      residence_years property_magnitude      age        other_payment_plans housing_type
 Min.   :1.000            Length:6000        Length:6000        Min.   :1.000   Length:6000        Min.   :19.00   Length:6000         Length:6000
 1st Qu.:2.000            Class :character   Class :character   1st Qu.:2.000   Class :character   1st Qu.:27.00   Class :character    Class :character
 Median :3.000            Mode  :character   Mode  :character   Median :3.000   Mode  :character   Median :32.00   Mode  :character    Mode  :character
 Mean   :3.057                                                  Mean   :2.853                      Mean   :34.95
 3rd Qu.:4.000                                                  3rd Qu.:4.000                      3rd Qu.:40.00
 Max.   :4.000                                                  Max.   :4.000                      Max.   :75.00
 existing_credits   job_type         num_dependants  own_telephone      is_foreign_worker  credit_risk_class
 Min.   :1.00     Length:6000        Min.   :1.000   Length:6000        Length:6000        Length:6000
 1st Qu.:1.00     Class :character   1st Qu.:1.000   Class :character   Class :character   Class :character
 Median :1.00     Mode  :character   Median :1.000   Mode  :character   Mode  :character   Mode  :character
 Mean   :1.39                        Mean   :1.161
 3rd Qu.:2.00                        3rd Qu.:1.000
 Max.   :4.00                        Max.   :2.000
>
```

```
>
> ## Check the total number of missing values for each column
> as.data.frame(colSums(is.na(pfda_df)))
                          colSums(is.na(pfda_df))
checking_account_status                         0
loan_duration_months                            0
credit_history_status                           0
loan_purpose                                    0
credit_amount                                   0
savings_account_status                          0
employment_years                                0
installment_rate_percent                        0
personal_status                                 0
other_parties                                   0
residence_years                                 0
property_magnitude                              0
age                                             0
other_payment_plans                             0
housing_type                                    0
existing_credits                                0
job_type                                        0
num_dependants                                  0
own_telephone                                   0
is_foreign_worker                               0
credit_risk_class                               0
> plot_missing(pfda_df)
```



After performing data cleaning and preprocessing, data validation is conducted to check all the columns again and ensure that no errors were missed out. According to the graph, there are no columns with missing values and the dataset is ready for analysis.

```
# Export the data frame containing the cleaned data to a new CSV file
current_directory = getwd()
export_path = file.path(current_directory, "(cleaned) credit_risk_classification.csv")
write.csv(pfda_df, export_path, row.names = FALSE)
```

Export the cleaned dataset that is stored in the data frame into a CSV file for documentation and further analysis to be done individually.

3.0 Data Analysis

3.1 Daryl Sim Wei Shern TP068964

Objective: To determine whether the loan duration impacts the credit risk classification of a customer.

```
# Load the libraries and packages
library(dplyr)
library(readr)
library(ggplot2)
library(DataExplorer)
library(missForest)
library(VIM)
library(caret)
library(caTools)
library(randomForest)
library(vcd)

# Set the working directory
## Enter your own path
setwd("C:/APU/Degree/Semester 1/Programming for Data Analysis/Assignment/PFDA Assignment")
getwd()


# Import the dataset
## Enter your own file path
filePath = "C:/APU/Degree/Semester 1/Programming for Data Analysis/Assignment/PFDA Assignment/(cleaned) credit_risk_classification.csv"

## Read the CSV file into a dataframe
pfda_df = read.csv(filePath)

# Convert the data type of all categorical columns from character to factor
pfda_df <- pfda_df %>%
  mutate(across(where(is.character), as.factor))
str(pfda_df)
sapply(pfda_df, class)

# Duplicate the dataset for individual analysis
pfda_df_Daryl <- pfda_df
View(pfda_df_Daryl)
```

Loading the necessary packages and libraries for data exploration and data analysis. Setting the working directory and importing the dataset. Then, convert the data type of all categorical columns from character to factor for data analysis convenience. Duplicate the main dataset for individual analysis purposes.

Exploratory Data Analysis

Analysis 1.1: Analyze the general information of the loan_duration_months and credit_risk_class columns

```
# Data Analysis
#----------------------------------------------------------------------Daryl Sim Wei Shern TP068964----------

# Objective 1: To determine whether the loan duration impacts the credit risk classification of a customer.
# Independent Variable: loan_duration_months (previously known as duration)


# Analysis 1.1: Analyze the general information of the loan_duration_months and credit_risk_class columns

# Summarize the dataset
summary(pfda_df_Daryl$loan_duration_months)
table(pfda_df_Daryl$credit_risk_class)

# Summary statistics grouped by credit_risk_class
pfda_df_Daryl %>%
  group_by(credit_risk_class) %>%
  summarise(mean_duration = mean(loan_duration_months),
            median_duration = median(loan_duration_months),
            sd_duration = sd(loan_duration_months),
            count = n())
```

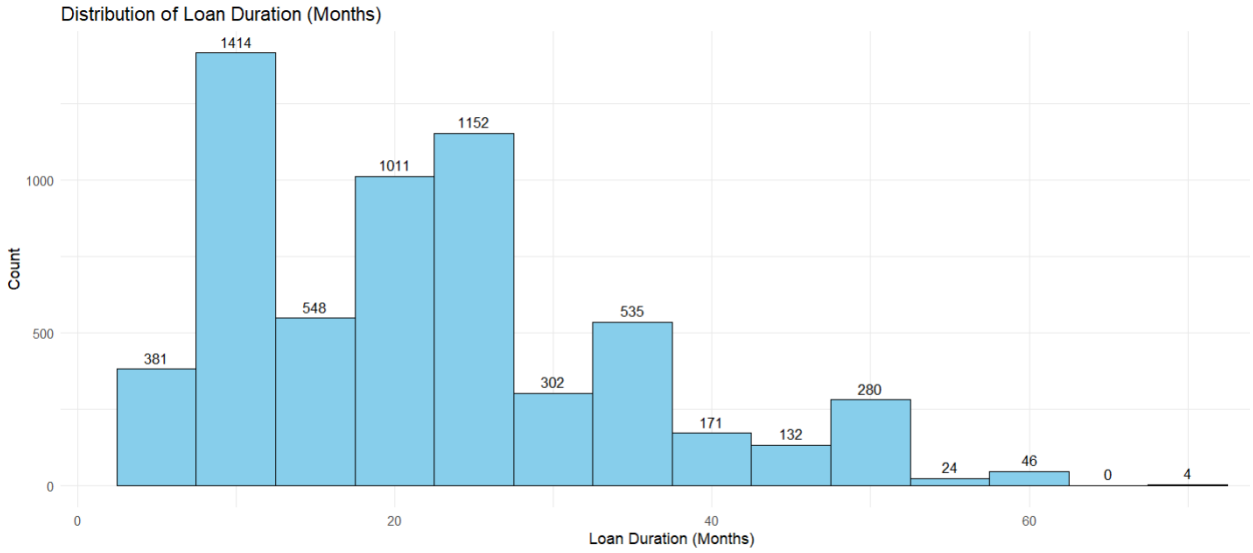
```
> # Analysis 1.1: Analyze the general information of the loan_duration_months and credit_risk_class columns
>
> # Summarize the dataset
> summary(pfda_df_Daryl$loan_duration_months)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   4.00   12.00   20.00   22.03   27.00   72.00
> table(pfda_df_Daryl$credit_risk_class)

 bad good
3000 3000
>
> # Summary statistics grouped by credit_risk_class
> pfda_df_Daryl %>%
+   group_by(credit_risk_class) %>%
+   summarise(mean_duration = mean(loan_duration_months),
+             median_duration = median(loan_duration_months),
+             sd_duration = sd(loan_duration_months),
+             count = n())
# A tibble: 2 × 5
  credit_risk_class mean_duration median_duration sd_duration count
  <fct>                    <dbl>           <dbl>       <dbl> <int>
1 bad                       24.8            23.5        12.2  3000
2 good                      19.3            18          11.0  3000
```

Based on the results, the data is evenly split between two classes as there are 3000 rows with good credit risk classification and 3000 rows with bad credit risk classification, and the total number of rows is 6000. The data in loan_duration_months column ranges between 4 months, which is the minimum value, and 72 months, which is the maximum value. Next, the median value is 20 months, and the mean value is 22.03 months, which is slightly higher and indicates a slight positive skew. The first quartile is 12 months, while the third quartile is 27 months. This indicates that the interquartile range is between 12 months and 27 months, with most loan durations falling within this range. For the bad credit risk classification, the mean loan duration is 24.8 months, with a median of 23.5 months and a standard deviation of 12.2 months. For the good credit risk classification, the mean loan duration is 19.3 months, with a median of 18 months and a standard deviation of 11 months. This shows that the loan durations that are classified as bad tend to have

longer durations compared to those with shorter durations. In addition, the loan durations in the bad credit risk classification have a higher variability as its standard deviation is higher than the loan durations in the good credit risk classification.

Analysis 1.2: What is the distribution of loan_duration_months?

```
# Analysis 1.2: How is the distribution of loan_duration_months?

# Visualize distribution of loan_duration_months
loan_duration_distribution <- ggplot(pfda_df_Daryl, aes(x = loan_duration_months)) +
  geom_histogram(binwidth = 5, fill = "skyblue", color = "black") +
  stat_bin(binwidth = 5,
           geom = "text",
           aes(label = ..count..),
           vjust = -0.5,
           size = 3.5) +
  labs(title = "Distribution of Loan Duration (Months)",
       x = "Loan Duration (Months)",
       y = "Count") +
  theme_minimal()
print(loan_duration_distribution)

# Save the plot
ggsave("Distribution of Loan Duration (Months).png", loan_duration_distribution, width = 12, height = 8, dpi = 300, bg="white")
```



A histogram is used to display distribution of the loan durations in months. According to the histogram, there is more data with loan durations below 30 months.

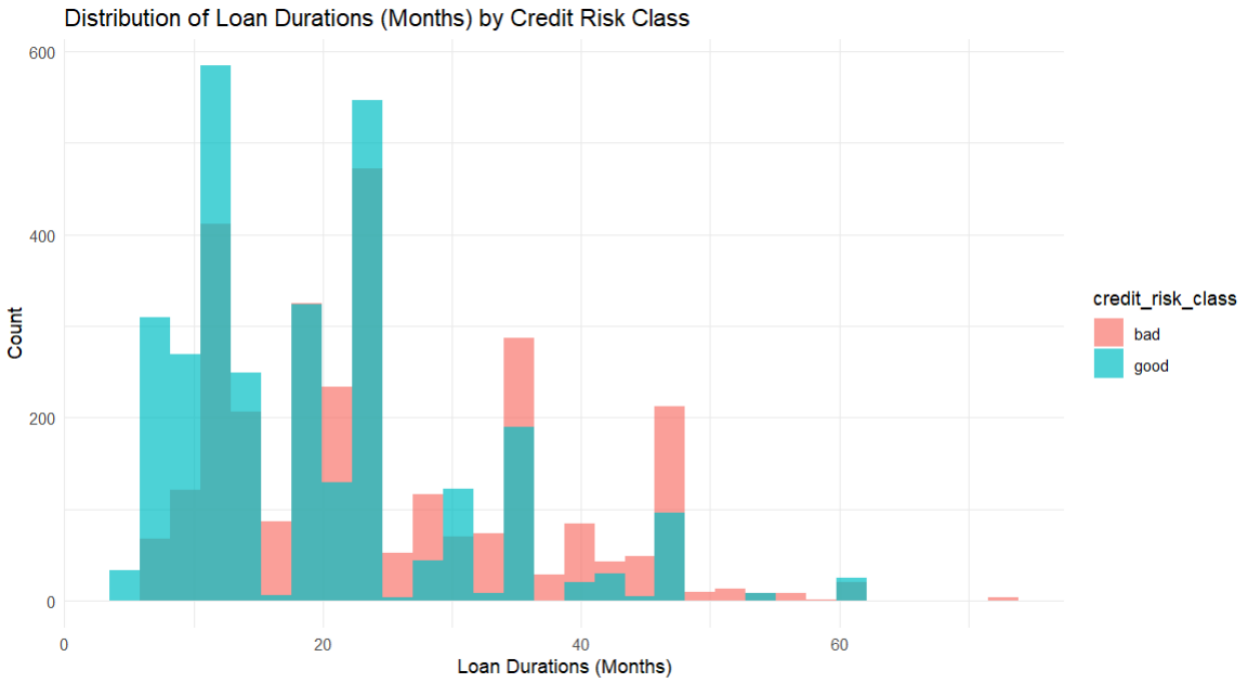
Analysis 1.3: Are there outliers in loan_duration_months by credit_risk_class?

```r
# Analysis 1.3: Are there outliers in loan_duration_months by credit_risk_class?

# Visualize the loan_duration_months by credit_risk_class
loan_duration_box <- ggplot(pfda_df_Daryl, aes(x = credit_risk_class, y = loan_duration_months, fill = credit_risk_class)) +
  geom_boxplot() +
  labs(title = "Loan Duration by Credit Risk Class", x = "Credit Risk Class", y = "Loan Duration (Months)") +
  theme_minimal()
print(loan_duration_box)

# Save the plot
ggsave("Boxplot of Loan Duration by Credit Risk Class.png", loan_duration_box, width = 12, height = 8, dpi = 300, bg="White")
```

A boxplot is used to view the distribution and central tendency like mean and median of the loan_duration_months column by the credit_risk_class. It is also used to display the interquartile range and help identify outliers. Based on the boxplot, the loan durations in good credit risk classification have more outliers than the loan durations in bad credit risk classification. However, these outliers are accepted and will not be changed or removed. This is because there are different kinds of loans with different purposes, which may result in different loan durations, thus they will be retained in the dataset for analysis.

Analysis 1.4: What is the distribution of loan_duration_months by credit_risk_class?

```
# Analysis 1.4: What is the distribution of loan_duration_months by credit_risk_class?

# Visualize the distribution of loan_duration_months by credit_risk_class
loan_duration_hist <- ggplot(pfda_df_Daryl, aes(x = loan_duration_months, fill = credit_risk_class)) +
  geom_histogram(position = "identity", alpha = 0.7, bins = 30) +
  labs(title = "Distribution of Loan Durations (Months) by Credit Risk Class",
       x = "Loan Durations (Months)",
       y = "Count") +
  theme_minimal()
print(loan_duration_hist)

# Save the plot
ggsave("Distribution of Loan Durations (Months) by Credit Risk Class.png", loan_duration_hist, width = 12, height = 8, dpi = 300, bg="white")
```



A histogram is used to display the distribution of loan_duration_months by credit_risk_class. Based on the histogram, loan durations that are below 25 months are mainly classified as good credit risk classification. As the loan duration increases, cases of good credit risk classification decrease.

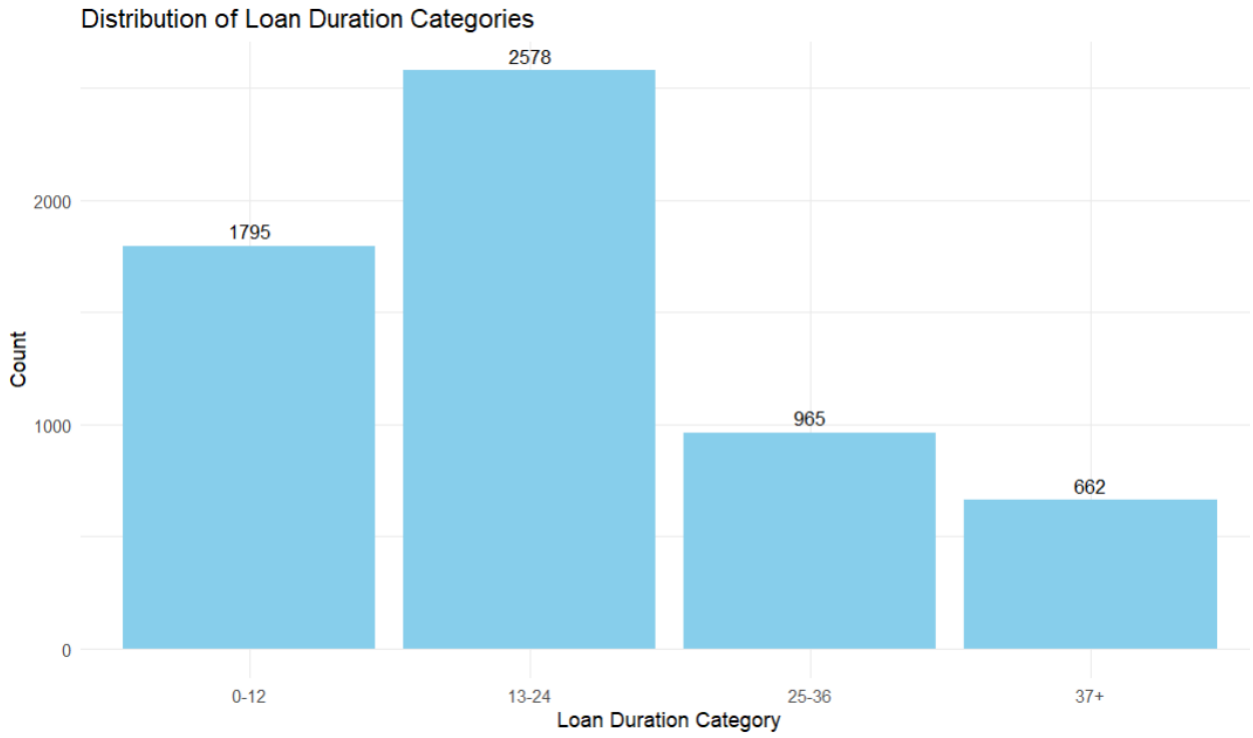
Analysis 1.5: What is the distribution of loan_duration_months when it is sorted by ranges?

```
# Analysis 1.5: What is the distribution of loan_duration_months when it is sorted by ranges?

# Categorize the loan_duration_months for easier visualization
pfda_df_Daryl$loan_duration_category <- cut(pfda_df_Daryl$loan_duration_months, breaks = c(0, 12, 24, 36, Inf),
                                            labels = c("0-12", "13-24", "25-36", "37+"))
summary(pfda_df_Daryl$loan_duration_category)

# Calculate the number of rows for each loan_duration_category
loan_category_count <- pfda_df_Daryl %>%
  group_by(loan_duration_category) %>%
  count()
View(loan_category_count)

# Visualize the distribution of loan_duration_category
loan_duration_category_distribution <-
  ggplot(loan_category_count, aes(x = loan_duration_category, y = n, fill = loan_duration_category)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9), fill = "skyblue") +
  geom_text(aes(label = n),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5) +
  labs(title = "Distribution of Loan Duration Categories", x = "Loan Duration Category", y = "Count") +
  theme_minimal()
print(loan_duration_category_distribution)

# Save the plot
ggsave("Distribution of Loan Duration Categories.png", loan_duration_category_distribution, width = 12, height = 8, dpi = 300, bg="white")
```

ns	housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class	loan_duration_category	loan_duration_category_less_than_24
	own	2	skilled	1	yes	yes	good	0-12	0-24
	own	1	skilled	1	none	yes	bad	37+	24+
	own	1	unskilled resident	2	none	yes	good	0-12	0-24
	for free	1	skilled	2	none	yes	good	37+	24+
	for free	2	skilled	2	none	yes	bad	13-24	0-24
	for free	1	unskilled resident	2	yes	yes	good	25-36	24+
	own	1	skilled	1	none	yes	good	13-24	0-24
	rent	1	high qualif/self emp/mgmt	1	yes	yes	good	25-36	24+
	own	1	unskilled resident	1	none	yes	good	0-12	0-24
	own	2	high qualif/self emp/mgmt	1	none	yes	bad	25-36	24+
	rent	1	skilled	1	none	yes	bad	0-12	0-24
	rent	1	skilled	1	none	yes	bad	37+	24+
	own	1	skilled	1	yes	yes	good	0-12	0-24
	own	2	unskilled resident	1	none	yes	bad	13-24	0-24
	rent	1	skilled	1	none	yes	good	13-24	0-24
	own	1	unskilled resident	1	none	yes	bad	13-24	0-24
	own	2	skilled	1	none	yes	good	13-24	0-24
	own	3	skilled	1	none	yes	good	25-36	24+
	for free	1	high qualif/self emp/mgmt	1	yes	yes	bad	13-24	0-24
	own	1	skilled	2	yes	yes	good	13-24	0-24
	own	3	skilled	1	yes	yes	good	0-12	0-24
	rent	1	skilled	2	none	yes	good	0-12	0-24
	rent	2	unskilled resident	2	none	no	good	0-12	0-24
	own	1	skilled	1	none	yes	good	0-12	0-24
	own	2	skilled	1	none	no	good	0-12	0-24

```
> summary(pfda_df_Daryl$loan_duration_category)
 0-12 13-24 25-36   37+
 1795  2578   965   662
```

	loan_duration_category	n
1	0-12	1795
2	13-24	2578
3	25-36	965
4	37+	662



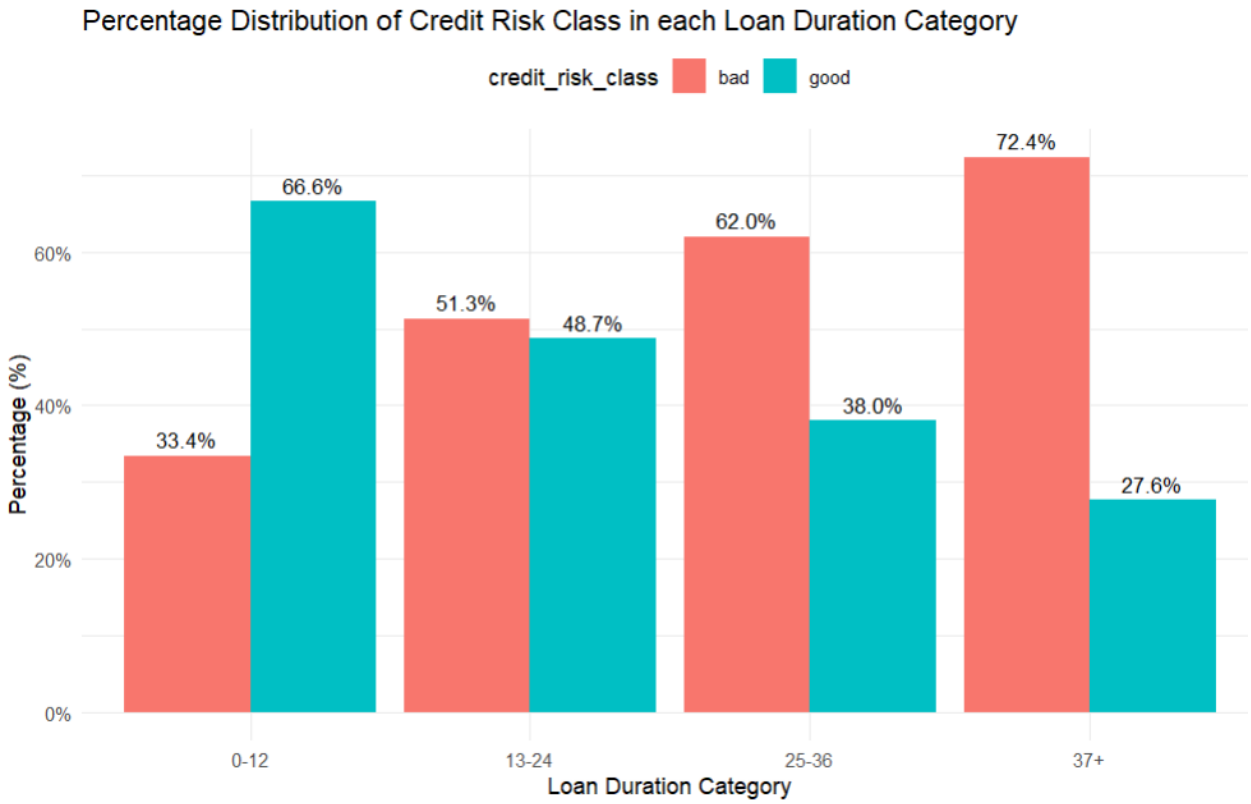
The loan durations are categorized into 4 categories, which are “0-12”, “13-24”, “25-36”, and “37+” for further analysis. A new column, loan_duration_category is created to store the value for each row. A bar chart is normally used to analyze categorical variables. Based on the bar chart that shows the distribution of the loan duration categories, the “13-24” loan duration range has the most data, followed by the “0-12”, “25-36”, and “37+” loan duration ranges.

Analysis 1.6: What is the distribution of good and bad credit_risk_class in each loan_duration_category?

```
# Analysis 1.6: What is the distribution of good and bad credit_risk_class in each loan_duration_category?

# Calculate the number and percentage of good and bad credit_risk_class in each loan_duration_category
proportion_loan_duration_category <- pfda_df_Daryl %>%
  group_by(loan_duration_category, credit_risk_class) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(loan_duration_category) %>%
  mutate(percent = count / sum(count) * 100)
View(proportion_loan_duration_category)

# Visualize the distribution of good and bad credit_risk_class in each loan_duration_category
loan_duration_category_bar <-
  ggplot(proportion_loan_duration_category, aes(x = loan_duration_category, y = count, fill = credit_risk_class)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  geom_text(aes(label = count),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5) +
  labs(title = "Distribution of Good and Bad Credit Risk Class in each Loan Duration Category",
       x = "Loan Duration Category",
       y = "Count") +
  theme_minimal()
print(loan_duration_category_bar)

# Save the plot
ggsave("Distribution of Good and Bad Credit Risk Class in each Loan Duration Category.png", loan_duration_category_bar, width = 12, height = 8, dpi = 300, bg="White")
```

	loan_duration_category	credit_risk_class	count	percent
1	0-12	bad	600	33.42618
2	0-12	good	1195	66.57382
3	13-24	bad	1323	51.31885
4	13-24	good	1255	48.68115
5	25-36	bad	598	61.96891
6	25-36	good	367	38.03109
7	37+	bad	479	72.35650
8	37+	good	183	27.64350


```
# Visualize the percentage of good and bad credit_risk_class in each loan_duration_category
proportion_loan_duration_category_bar <-
  ggplot(proportion_loan_duration_category, aes(x = loan_duration_category, y = percent, fill = credit_risk_class)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  geom_text(aes(label = sprintf("%.1f%%", percent)),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5) +
  labs(title = "Percentage Distribution of Credit Risk Class in each Loan Duration Category",
       x = "Loan Duration Category",
       y = "Percentage (%)") +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal() +
  theme(legend.position = "top")
print(proportion_loan_duration_category_bar)

# Save the plot
ggsave("Percentage Distribution of Credit Risk Class in each Loan Duration Category.png", proportion_loan_duration_category_bar, width = 12, height = 8, dpi = 300, bg="white")
```



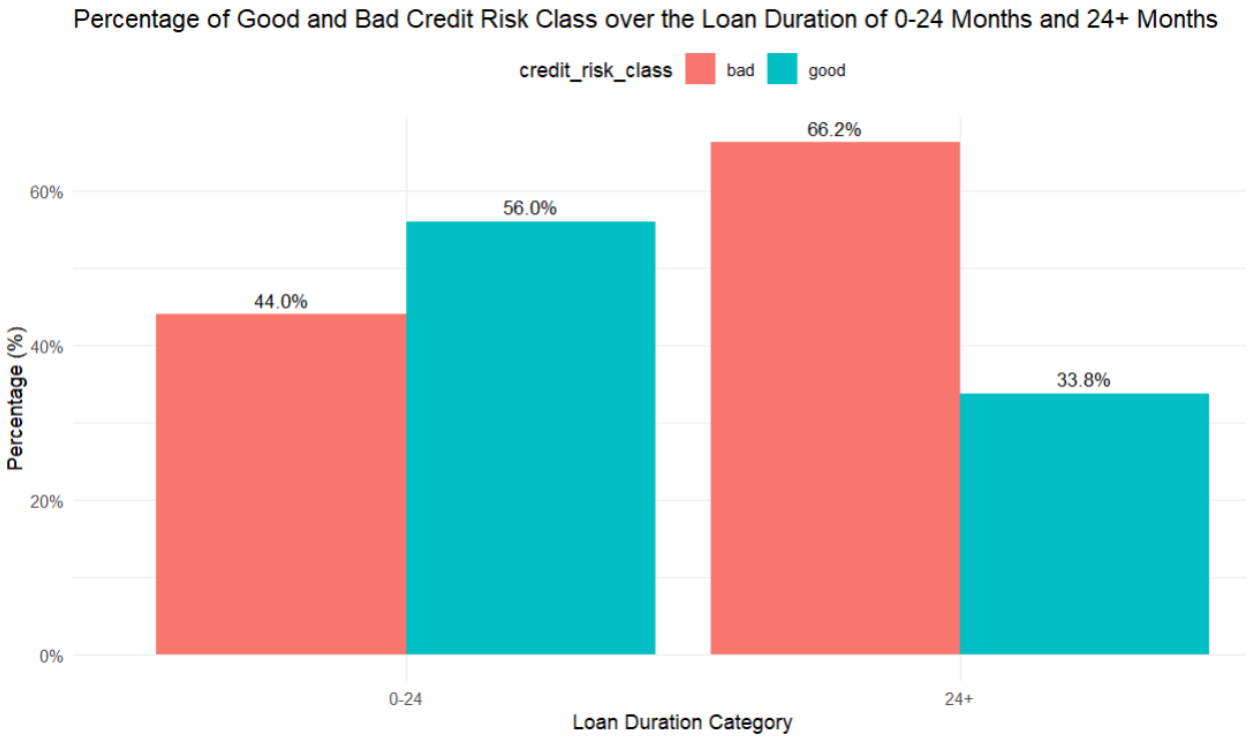
A stacked bar chart but displayed side-by-side for better visualization is used to show the distribution of good and bad credit risk classification in each loan duration category. Based on the stacked bar chart, the “0-12” loan duration range has the highest number of good credit risk classification with 66.6%, followed by the “13-24" loan duration range with 48.7%. The distribution in “13-24” loan duration range is fairly balanced. On the other hand, “0-12” loan duration range has more good credit risk classification cases, while “25-36” and “37+” loan duration ranges have more bad credit risk classification cases.

Analysis 1.7: What is the distribution of good and bad credit_risk_class for loan duration less than or equal to 24 months and for loan terms greater than 24 months?

```r
# Analysis 1.7: What is the distribution of good and bad credit_risk_class for loan duration less than or equal to 24 months and for loan terms greater than 24 months?

# Categorize the loan_duration_months for easier visualization
pfda_df_Daryl$loan_duration_category_less_than_24 <- cut(pfda_df_Daryl$loan_duration_months, breaks = c(0, 24, Inf),
                                                         labels = c("0-24", "24+"))
summary(pfda_df_Daryl$loan_duration_category_less_than_24)
levels(pfda_df_Daryl$loan_duration_category_less_than_24)

# Calculate the percentage of good and bad credit_risk_class in 0-24 and 24+ months loan duration
proportion_loan_duration_category_less_than_24 <- pfda_df_Daryl %>%
  group_by(loan_duration_category_less_than_24, credit_risk_class) %>%
  summarise(count = n(), .groups = 'drop') %>%
  group_by(loan_duration_category_less_than_24) %>%
  mutate(percent = count / sum(count) * 100)
View(proportion_loan_duration_category_less_than_24)

# Visualize the percentage of good and bad credit_risk_class over the loan duration of 0-24 months and 24+ months
proportion_loan_duration_category_less_than_24_bar <-
  ggplot(proportion_loan_duration_category_less_than_24, aes(x = loan_duration_category_less_than_24, y = percent, fill = credit_risk_class)) +
  geom_bar(stat = "identity", position = position_dodge(width = 0.9)) +
  geom_text(aes(label = sprintf("%.1f%%", percent)),
            position = position_dodge(width = 0.9),
            vjust = -0.5,
            size = 3.5) +
  labs(title = "Percentage of Good and Bad Credit Risk Class over the Loan Duration of 0-24 Months and 24+ Months",
       x = "Loan Duration Category",
       y = "Percentage (%)") +
  scale_y_continuous(labels = scales::percent_format(scale = 1)) +
  theme_minimal() +
  theme(legend.position = "top")
print(proportion_loan_duration_category_less_than_24_bar)

# Save the plot
ggsave("Percentage of Good and Bad Credit Risk Class over the Loan Duration of 0-24 Months and 24+ Months.png",
       proportion_loan_duration_category_less_than_24_bar, width = 12, height = 8, dpi = 300, bg="white")
```

```
> summary(pfda_df_Daryl$loan_duration_category_less_than_24)
0-24  24+
4373 1627
> levels(pfda_df_Daryl$loan_duration_category_less_than_24)
[1] "0-24" "24+"
```

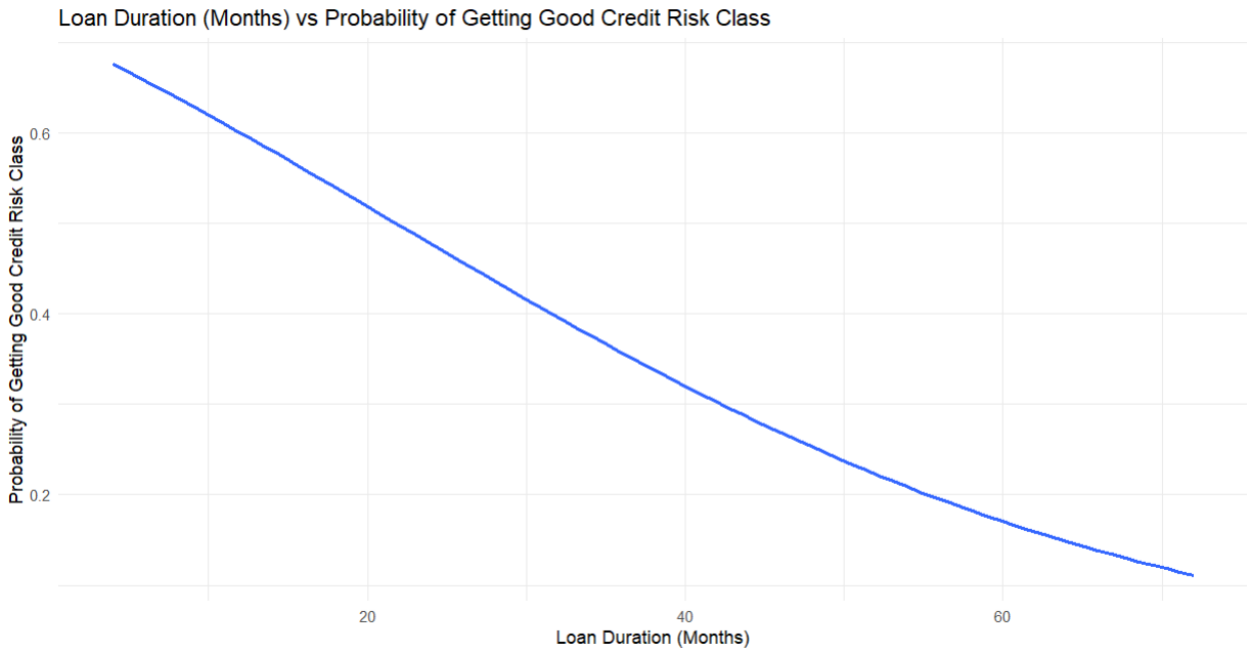
	loan_duration_category_less_than_24	credit_risk_class	count	percent
1	0-24	bad	1923	43.97439
2	0-24	good	2450	56.02561
3	24+	bad	1077	66.19545
4	24+	good	550	33.80455



The loan durations are categorized into “0-24” and “24+” loan duration ranges to specifically analyze the distribution of good and bad credit risk classification cases for loan durations that are less than or equal to 24 months and greater than 24 months. A stacked bar chart that is displayed side-by-side is also used to show the difference between the two loan duration categories. Based on the stacked bar chart, the “0-24” loan duration range has 2450 good credit risk classification cases, which is 56% of its total cases. On the other hand, the “24+” loan duration range has 550 good credit risk classification, which is 33.8% of its total cases.

Analysis 1.8: What is the relationship between loan_duration_months and the probability of getting good credit_risk_class?

```
# Analysis 1.8: What is the relationship between loan_duration_months and the probability of getting good credit_risk_class?

# Visualize the relationship between loan_duration_months and the probability of getting good credit_risk_class
loan_duration_vs_good_probability <- ggplot(pfda_df_Daryl, aes(x = loan_duration_months, y = as.numeric(credit_risk_class) - 1)) +
  geom_smooth(method = "glm", method.args = list(family = "binomial"), se = FALSE) +
  labs(title = "Loan Duration (Months) vs Probability of Getting Good Credit Risk Class",
       x = "Loan Duration (Months)",
       y = "Probability of Getting Good Credit Risk Class") +
  theme_minimal()
print(loan_duration_vs_good_probability)

# Save the plot
ggsave("Loan Duration (Months) vs Probability of Getting Good Credit Risk Class.png",
       loan_duration_vs_good_probability, width = 12, height = 8, dpi = 300, bg="white")
```



A line graph is used to examine the relationship between loan duration and the probability of getting good credit risk classification. This line graph also uses the generalized linear model (GLM) and binomial logistic regression to predict binary outcomes, which is the probability of getting good credit risk classification here. Based on the line graph, the curve slopes downwards, which indicates that as the loan duration increases, the probability of being classified as good credit risk decreases. This also shows that there is a significant relationship between loan duration and credit risk classification.

Literature Review

In 2002, Jiménez and Saurina conducted research to analyze the relationship between loan characteristics and credit risk. Loan maturity is one of the loan characteristics that has been examined. Jiménez and Saurina (2002) found that longer loan maturity will increase the probability of default (PD) of a customer, which leads to higher credit risk and a greater likelihood of the customer being classified as a bad credit risk.

In 2020, Calem et al. conducted research to investigate how long-term auto borrowing and auto-loan default are associated with the observable characteristics of a borrower and economic indicators. Calem et al. (2020) found that borrowers who go for long-term auto loans often exhibit characteristics that are connected to higher chances of credit risk. For instance, higher utilization of credit card, earning lower incomes, and residing in areas with higher unemployment rates. However, the research emphasized that longer loan duration is not the sole factor that contributes to the increase in credit risk, unobserved factors such as changes in the profile or financial condition of a borrower are the main contributors to an increase in default rates (Calem et al., 2020).

Hypothesis Formation

Based on the exploratory data analysis and literature review, as the loan duration increases, the credit risk also increases. Thus, it contributes to the hypothesis formation.

Null Hypothesis, H_0: Customers who apply for a loan with a duration of **less than or equal to 24 months** have a **less than or equal to 50% probability** of receiving a good credit risk classification compared to customers who apply for a loan with a duration of more than 24 months.

Alternative Hypothesis, H_1: Customers who apply for a loan with a duration of **less than or equal to 24 months** have a **greater than 50% probability** of receiving a good credit risk classification compared to customers who apply for a loan with a duration of more than 24 months.

After forming the hypotheses, accepting and rejecting criteria should be determined. 0.05 will be used as the significance level (α).

If p-value is less than or equal (≤) to 0.05, reject null hypothesis, H_0.

If p-value is greater than (>) 0.05, fail to reject null hypothesis, H_0.

Hypothesis Testing

Two Sample t-test

```
# Two sample t-test
loan_duration_t_test <- t.test(loan_duration_months ~ credit_risk_class, data = pfda_df_Daryl)
print(loan_duration_t_test)
```

```
> # Two sample t-test
> loan_duration_t_test <- t.test(loan_duration_months ~ credit_risk_class, data = pfda_df_Daryl)
> print(loan_duration_t_test)

        Welch Two Sample t-test

data:  loan_duration_months by credit_risk_class
t = 18.484, df = 5940.2, p-value < 2.2e-16
alternative hypothesis: true difference in means between group bad and group good is not equal to 0
95 percent confidence interval:
 4.950371 6.124962
sample estimates:
mean in group bad mean in group good
         24.80167          19.26400
```

The large t-value of **18.484** and an extremely small p-value of **less than 2.2e-16** indicate a statistically significant difference in the mean loan durations between bad and good credit risk classification. The positive t-value also shows that the loan durations in the bad credit risk classification are greater than those in the good credit risk classification. Next, the mean loan duration of bad credit risk classification is **24.80** months, while the mean loan duration of good credit risk classification is **19.26** months. Based on the confidence interval, we can be **95% confident** that the true difference in average loan duration for bad credit risk classification is between **4.95 months** and **6.12 months** greater than the average loan duration for good credit risk classification. From the mean loan duration for both categories, the loan durations in bad credit risk classification are higher compared to the loan durations in good credit risk classification.

Chi-square Test of Independence

```
# Chi-square test of independence
loan_duration_chisq_test <- chisq.test(table(pfda_df_Daryl$loan_duration_category, pfda_df_Daryl$credit_risk_class))
print(loan_duration_chisq_test)
```

```
> # Chi-square test of independence
> loan_duration_chisq_test <- chisq.test(table(pfda_df_Daryl$loan_duration_category, pfda_df_Daryl$credit_risk_class))
> print(loan_duration_chisq_test)

        Pearson's Chi-squared test

data:  table(pfda_df_Daryl$loan_duration_category, pfda_df_Daryl$credit_risk_class)
X-squared = 386.67, df = 3, p-value < 2.2e-16
```

The Chi-square (X^2) value which is **386.67** indicates a substantial deviation from what would be expected if loan duration category and credit risk classification were independent. This large test statistic leads to a very small p-value, which is **lesser than 2.2e-16**. This provides strong evidence to confirm that there is a **statistically significant association** between loan duration and credit risk classification. Besides that, the p-value is much smaller than the significance level of 0.05, indicating **strong evidence against the null hypothesis**.

Cramér's V

```
# Cramér's V
loan_duration_cramers_v <- assocstats(table(pfda_df_Daryl$loan_duration_category, pfda_df_Daryl$credit_risk_class))$cramer
print(loan_duration_cramers_v)
```

```
> # Cramér's V
> loan_duration_cramers_v <- assocstats(table(pfda_df_Daryl$loan_duration_category, pfda_df_Daryl$credit_risk_class))$cramer
> print(loan_duration_cramers_v)
[1] 0.2538598
```

Cramér's V is used to measure the strength of the association between loan duration category and credit risk classification. Based on the result, the strength of the relationship is **0.2538598**, approximately **0.254**. This number falls in the category of **weak to moderate association**.

Logistic Regression for all Loan Duration Categories

```
# To check the levels of credit_risk_class and to ensure that the order is "bad" to good"
levels(pfda_df_Daryl$credit_risk_class)


# Logistic Regression with all categories
logistic_loan_duration_category_log_model <- glm(credit_risk_class ~ loan_duration_category, data = pfda_df_Daryl, family = binomial)
logistic_loan_duration_category_log_model
summary(logistic_loan_duration_category_log_model)
```

```
> # To check the levels of credit_risk_class and to ensure that the order is "bad" to good"
> levels(pfda_df_Daryl$credit_risk_class)
[1] "bad"  "good"
```

This is to ensure that the logistic regression model will be using the bad credit risk classification as the reference category and the good credit risk classification as the target category.

```
> summary(logistic_loan_duration_category_log_model)

Call:
glm(formula = credit_risk_class ~ loan_duration_category, family = binomial,
    data = pfda_df_Daryl)

Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)                0.68897    0.05003   13.77   <2e-16 ***
loan_duration_category13-24 -0.74174    0.06369  -11.65   <2e-16 ***
loan_duration_category25-36 -1.17720    0.08307  -14.17   <2e-16 ***
loan_duration_category37+  -1.65119    0.10028  -16.47   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 7922.0  on 5996  degrees of freedom
AIC: 7930

Number of Fisher Scoring iterations: 4
```

```
## 0-12 months loan duration
# Calculate the odds ratio for 0-12 months loan duration
logistic_loan_duration_category_12_or <- exp(coef(logistic_loan_duration_category_log_model)["(Intercept)"])
print(paste("Odds Ratio for 0-12 months loan duration:", logistic_loan_duration_category_12_or))

# Good credit risk class probability for 0-12 months loan duration
logistic_duration_probability_0_12 <- logistic_loan_duration_category_12_or / (1 + logistic_loan_duration_category_12_or)
logistic_duration_probability_0_12 <- round(logistic_duration_probability_0_12 * 100, 1)
logistic_duration_probability_0_12_text <- paste0(logistic_duration_probability_0_12, "%")
print(paste("Good credit risk class probability for 0-12 months loan duration:", logistic_duration_probability_0_12_text))
```

```
> ## 0-12 months loan duration
> # Calculate the odds ratio for 0-12 months loan duration
> logistic_loan_duration_category_12_or <- exp(coef(logistic_loan_duration_category_log_model)["(Intercept)"])
> print(paste("Odds Ratio for 0-12 months loan duration:", logistic_loan_duration_category_12_or))
[1] "Odds Ratio for 0-12 months loan duration: 1.99166666666636"
> # Good credit risk class probability for 0-12 months loan duration
> logistic_duration_probability_0_12 <- logistic_loan_duration_category_12_or / (1 + logistic_loan_duration_category_12_or)
> logistic_duration_probability_0_12 <- round(logistic_duration_probability_0_12 * 100, 1)
> logistic_duration_probability_0_12_text <- paste0(logistic_duration_probability_0_12, "%")
> print(paste("Good credit risk class probability for 0-12 months loan duration:", logistic_duration_probability_0_12_text))
[1] "Good credit risk class probability for 0-12 months loan duration: 66.6%"
```

This “Intercept” is representing the “0-12” months loan duration category. The p-value **(< 2e-16)** is much smaller than the standard significance level of 0.05. This provides very strong evidence to

reject the null hypothesis. Then, the calculated probability of getting good credit risk classification for "0-12" months loan duration category is **66.6%**.

```
## 13-24 months loan duration
## Calculate the log-odds for 13-24 months loan duration
logistic_loan_duration_category_24_or <- coef(logistic_loan_duration_category_log_model)["(Intercept)"] +
                                         coef(logistic_loan_duration_category_log_model)["loan_duration_category13-24"]
# Convert the log-odds to odds ratio
logistic_loan_duration_category_24_or <- exp(logistic_loan_duration_category_24_or)
print(paste("Odds Ratio for 13-24 months loan duration:", logistic_loan_duration_category_24_or))

# Convert odds ratio to probability
logistic_duration_probability_13_24 <- logistic_loan_duration_category_24_or / (1 + logistic_loan_duration_category_24_or)
logistic_duration_probability_13_24 <- round(logistic_duration_probability_13_24 * 100, 1)

# Good credit risk class probability for 13-24 months loan duration
logistic_duration_probability_13_24_text <- paste0(logistic_duration_probability_13_24, "%")
print(paste("Good credit risk class probability for 13-24 months loan duration:", logistic_duration_probability_13_24_text))
```

```
> ## 13-24 months loan duration
> ## Calculate the log-odds for 13-24 months loan duration
> logistic_loan_duration_category_24_or <- coef(logistic_loan_duration_category_log_model)["(Intercept)"] +
+                                          coef(logistic_loan_duration_category_log_model)["loan_duration_category13-24"]
> # Convert the log-odds to odds ratio
> logistic_loan_duration_category_24_or <- exp(logistic_loan_duration_category_24_or)
> print(paste("Odds Ratio for 13-24 months loan duration:", logistic_loan_duration_category_24_or))
[1] "Odds Ratio for 13-24 months loan duration: 0.94860166288738"
>
> # Convert odds ratio to probability
> logistic_duration_probability_13_24 <- logistic_loan_duration_category_24_or / (1 + logistic_loan_duration_category_24_or)
> logistic_duration_probability_13_24 <- round(logistic_duration_probability_13_24 * 100, 1)
>
> # Good credit risk class probability for 13-24 months loan duration
> logistic_duration_probability_13_24_text <- paste0(logistic_duration_probability_13_24, "%")
> print(paste("Good credit risk class probability for 13-24 months loan duration:", logistic_duration_probability_13_24_text))
[1] "Good credit risk class probability for 13-24 months loan duration: 48.7%"
```

```
> exp(-0.74174)
[1] 0.4762845
```

This "loan_duration_category13-24" is representing the "13-24" months loan duration category. The p-value **(< 2e-16)** is much smaller than the standard significance level of 0.05. This provides very strong evidence to reject the null hypothesis. Then, the calculated probability of getting good credit risk classification for "13-24" months loan duration category is **48.7%**. The negative coefficient, **-0.74174**, means that longer loan durations are associated with a lower probability of getting a good credit risk classification. The odds of "13-24" months loan duration category is:

$e^{-0.74174} = 0.4762845 \approx 0.4760$

So, the odds are about **47.6%** of the odds of "0-12" months loan duration category. This also indicates that the odds of getting a good credit risk classification are about **52.4%** [(1-0.4760) * 100] lower than the odds of "0-12" months loan duration category.

```
## 25-36 months loan duration
## Calculate the log-odds for 25-36 months loan duration
logistic_loan_duration_category_36_or <- coef(logistic_loan_duration_category_log_model)["(Intercept)"] +
  coef(logistic_loan_duration_category_log_model)["loan_duration_category25-36"]

# Convert the log-odds to odds ratio
logistic_loan_duration_category_36_or <- exp(logistic_loan_duration_category_36_or)
print(paste("Odds Ratio for 25-36 months loan duration:", logistic_loan_duration_category_36_or))

# Convert odds ratio to probability
logistic_duration_probability_25_36 <- logistic_loan_duration_category_36_or / (1 + logistic_loan_duration_category_36_or)
logistic_duration_probability_25_36 <- round(logistic_duration_probability_25_36 * 100, 1)

# Good credit risk class probability for 25-36 months loan duration
logistic_duration_probability_25_36_text <- paste0(logistic_duration_probability_25_36, "%")
print(paste("Good credit risk class probability for 25-36 months loan duration:", logistic_duration_probability_25_36_text))
```

```
> ## 25-36 months loan duration
> ## Calculate the log-odds for 25-36 months loan duration
> logistic_loan_duration_category_36_or <- coef(logistic_loan_duration_category_log_model)["(Intercept)"] +
+   coef(logistic_loan_duration_category_log_model)["loan_duration_category25-36"]
>
> # Convert the log-odds to odds ratio
> logistic_loan_duration_category_36_or <- exp(logistic_loan_duration_category_36_or)
> print(paste("Odds Ratio for 25-36 months loan duration:", logistic_loan_duration_category_36_or))
[1] "Odds Ratio for 25-36 months loan duration: 0.613712374581993"
>
> # Convert odds ratio to probability
> logistic_duration_probability_25_36 <- logistic_loan_duration_category_36_or / (1 + logistic_loan_duration_category_36_or)
> logistic_duration_probability_25_36 <- round(logistic_duration_probability_25_36 * 100, 1)
>
> # Good credit risk class probability for 25-36 months loan duration
> logistic_duration_probability_25_36_text <- paste0(logistic_duration_probability_25_36, "%")
> print(paste("Good credit risk class probability for 25-36 months loan duration:", logistic_duration_probability_25_36_text))
[1] "Good credit risk class probability for 25-36 months loan duration: 38%"
```

```
> exp(-1.17720)
[1] 0.3081403
```

This “loan_duration_category25-36” is representing the “25-36” months loan duration category. The p-value **(< 2e-16)** is much smaller than the standard significance level of 0.05. This provides very strong evidence to reject the null hypothesis. Then, the calculated probability of getting good credit risk classification for “25-36” months loan duration category is **38%**. The negative coefficient, **-1.17720**, means that longer loan durations are associated with a lower probability of getting a good credit risk classification. The odds of “25-36” months loan duration category is:

$e^{-1.17720} = 0.3081403 \approx 0.3081$

So, the odds are about **30.81%** of the odds of “0-12” months loan duration category. This also indicates that the odds of getting a good credit risk classification are about **69.19%** [(1-0.3081) * 100] lower than the odds of “0-12” months loan duration category.

```
## 37+ months loan duration
## Calculate the log-odds for 37+ months loan duration
logistic_loan_duration_category_37plus_or <- coef(logistic_loan_duration_category_log_model)["(Intercept)"] +
  coef(logistic_loan_duration_category_log_model)["loan_duration_category37+"]

# Convert the log-odds to odds ratio
logistic_loan_duration_category_37plus_or <- exp(logistic_loan_duration_category_37plus_or)
print(paste("Odds Ratio for 37+ months loan duration:", logistic_loan_duration_category_37plus_or))

# Convert odds ratio to probability
logistic_duration_probability_37plus <- logistic_loan_duration_category_37plus_or / (1 + logistic_loan_duration_category_37plus_or)
logistic_duration_probability_37plus <- round(logistic_duration_probability_37plus * 100, 1)

# Good credit risk class probability for 37+ months loan duration
logistic_duration_probability_37plus_text <- paste0(logistic_duration_probability_37plus, "%")
print(paste("Good credit risk class probability for 37+ months loan duration:", logistic_duration_probability_37plus_text))
```

```
> ## 37+ months loan duration
> ## Calculate the log-odds for 37+ months loan duration
> logistic_loan_duration_category_37plus_or <- coef(logistic_loan_duration_category_log_model)["(Intercept)"] +
+   coef(logistic_loan_duration_category_log_model)["loan_duration_category37+"]
>
> # Convert the log-odds to odds ratio
> logistic_loan_duration_category_37plus_or <- exp(logistic_loan_duration_category_37plus_or)
> print(paste("Odds Ratio for 37+ months loan duration:", logistic_loan_duration_category_37plus_or))
[1] "Odds Ratio for 37+ months loan duration: 0.382045929019736"
>
> # Convert odds ratio to probability
> logistic_duration_probability_37plus <- logistic_loan_duration_category_37plus_or / (1 + logistic_loan_duration_category_37plus_or)
> logistic_duration_probability_37plus <- round(logistic_duration_probability_37plus * 100, 1)
>
> # Good credit risk class probability for 37+ months loan duration
> logistic_duration_probability_37plus_text <- paste0(logistic_duration_probability_37plus, "%")
> print(paste("Good credit risk class probability for 37+ months loan duration:", logistic_duration_probability_37plus_text))
[1] "Good credit risk class probability for 37+ months loan duration: 27.6%"
```

```
> exp(-1.65119)
[1] 0.1918215
```

This “loan_duration_category37+” is representing the “37+” months loan duration category. The p-value **(< 2e-16)** is much smaller than the standard significance level of 0.05. This provides very strong evidence to reject the null hypothesis. Then, the calculated probability of getting good credit risk classification for “37+” months loan duration category is **27.6%**. The negative coefficient, **-1.65119**, means that longer loan durations are associated with a lower probability of getting a good credit risk classification. The odds of “37+” months loan duration category is:

$e^{-1.65119} = 0.1918215 \approx 0.1918$

So, the odds are about **19.18%** of the odds of “0-12” months loan duration category. This also indicates that the odds of getting a good credit risk classification are about **80.82%** [(1-0.3081) * 100] lower than the odds of “0-12” months loan duration category.

```
# Calculate the weighted probability for loan duration below 25 months
weighted_probability_0_12_13_24 <- (logistic_duration_probability_0_12 + logistic_duration_probability_13_24) / 2
weighted_probability_0_12_13_24

# Calculate the weighted probability for loan duration above 24 months
weighted_probability_25_36_37plus <- (logistic_duration_probability_25_36 + logistic_duration_probability_37plus) / 2
weighted_probability_25_36_37plus
```

```
# Calculate the percentage difference between the probability for loan duration below 25 months and above 24 months
percentage_difference_probability_loan_duration_all <- (weighted_probability_0_12_13_24 - weighted_probability_25_36_37plus) / weighted_probability_25_36_37plus
percentage_difference_probability_loan_duration_all <- round(percentage_difference_probability_loan_duration_all * 100, 1)
percentage_difference_probability_loan_duration_all_text <- paste0(percentage_difference_probability_loan_duration_all, "%")
print(paste("The percentage difference between loan duration below 25 months and above 24 months:", percentage_difference_probability_loan_duration_all_text))
```

```
> # Calculate the weighted probability for loan duration below 25 months
> weighted_probability_0_12_13_24 <- (logistic_duration_probability_0_12 + logistic_duration_probability_13_24) / 2
> weighted_probability_0_12_13_24
(Intercept)
      57.65
>
> # Calculate the weighted probability for loan duration above 24 months
> weighted_probability_25_36_37plus <- (logistic_duration_probability_25_36 + logistic_duration_probability_37plus) / 2
> weighted_probability_25_36_37plus
(Intercept)
      32.8
>
> # Calculate the percentage difference between the probability for loan duration below 25 months and above 24 months
> percentage_difference_probability_loan_duration_all <- (weighted_probability_0_12_13_24 - weighted_probability_25_36_37plus) / weighted_probability_25_36_37plus
> percentage_difference_probability_loan_duration_all <- round(percentage_difference_probability_loan_duration_all * 100, 1)
> percentage_difference_probability_loan_duration_all_text <- paste0(percentage_difference_probability_loan_duration_all, "%")
> print(paste("The percentage difference between loan duration below 25 months and above 24 months:", percentage_difference_probability_loan_duration_all_text))
[1] "The percentage difference between loan duration below 25 months and above 24 months: 75.8%"
>
```

These calculations were performed to find the percentage difference between the probability of getting good credit risk classification for the loan durations that are less than or equal to 24 months and above 24 months. The weighted probability for the loan durations that are less than or equal to 24 months is **57.65%**. On the other hand, the weighted probability for the loan durations that are above 24 months is **32.8%**. Then, their percentage difference is **75.8%**, indicating that the loan durations that are less than or equal to 24 months have a 75.8% chance to get a good credit risk classification than the loan durations that are above 24 months. This proves the hypothesis that indicates a more than 50% chance of getting good credit risk classification.

Logistic Regression for “0-24” and “24+” Loan Duration Categories

```
# Logistic Regression with 0-24 and 24+ months categories
logistic_loan_duration_log_model <- glm(credit_risk_class ~ loan_duration_category_less_than_24, data = pfda_df_Daryl, family = binomial)
logistic_loan_duration_log_model
summary(logistic_loan_duration_log_model)
```

```
> summary(logistic_loan_duration_log_model)

Call:
glm(formula = credit_risk_class ~ loan_duration_category_less_than_24,
    family = binomial, data = pfda_df_Daryl)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        0.24220    0.03047    7.95 1.87e-15 ***
loan_duration_category_less_than_24+ -0.91422    0.06062  -15.08  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 8080.3  on 5998  degrees of freedom
AIC: 8084.3

Number of Fisher Scoring iterations: 4
```

```
## 0-24 months loan duration
# Calculate the odds ratio for 0-24 months loan duration
logistic_loan_duration_0_24_odds_ratio <- exp(coef(logistic_loan_duration_log_model)["(Intercept)"])
print(paste("Odds Ratio for 0-24 months loan duration:", logistic_loan_duration_0_24_odds_ratio))

# Good credit risk class probability for 0-24 months loan duration
logistic_duration_probability_0_24 <- logistic_loan_duration_0_24_odds_ratio / (1 + logistic_loan_duration_0_24_odds_ratio)
logistic_duration_probability_0_24 <- round(logistic_duration_probability_0_24 * 100, 1)
logistic_duration_probability_0_24_text <- paste0(logistic_duration_probability_0_24, "%")
print(paste("Good credit risk class probability for 0-24 months loan duration:", logistic_duration_probability_0_24_text))
```

```
> ## 0-24 months loan duration
> # Calculate the odds ratio for 0-24 months loan duration
> logistic_loan_duration_0_24_odds_ratio <- exp(coef(logistic_loan_duration_log_model)["(Intercept)"])
> print(paste("Odds Ratio for 0-24 months loan duration:", logistic_loan_duration_0_24_odds_ratio))
[1] "Odds Ratio for 0-24 months loan duration: 1.27405096203851"
>
> # Good credit risk class probability for 0-24 months loan duration
> logistic_duration_probability_0_24 <- logistic_loan_duration_0_24_odds_ratio / (1 + logistic_loan_duration_0_24_odds_ratio)
> logistic_duration_probability_0_24 <- round(logistic_duration_probability_0_24 * 100, 1)
> logistic_duration_probability_0_24_text <- paste0(logistic_duration_probability_0_24, "%")
> print(paste("Good credit risk class probability for 0-24 months loan duration:", logistic_duration_probability_0_24_text))
[1] "Good credit risk class probability for 0-24 months loan duration: 56%"
```

This “Intercept” is representing the “0-24” months loan duration category. The p-value (1.87e-15) is much smaller than the standard significance level of 0.05. This provides very strong evidence to reject the null hypothesis. Then, the calculated probability of getting good credit risk classification for “0-24” months loan duration category is **56%**.

```
## 24+ months loan duration
## Calculate the log-odds for 24+ months loan duration
logistic_loan_duration_24plus_odds_ratio <- coef(logistic_loan_duration_log_model)["(Intercept)"] +
                                            coef(logistic_loan_duration_log_model)["loan_duration_category_less_than_2424+"]

# Convert the log-odds to odds ratio
logistic_loan_duration_24plus_odds_ratio <- exp(logistic_loan_duration_24plus_odds_ratio)
print(paste("Odds Ratio for 24+ months loan duration:", logistic_loan_duration_24plus_odds_ratio))

# Convert odds ratio to probability
logistic_duration_probability_24plus <- logistic_loan_duration_24plus_odds_ratio / (1 + logistic_loan_duration_24plus_odds_ratio)
logistic_duration_probability_24plus <- round(logistic_duration_probability_24plus * 100, 1)

# Good credit risk class probability for 24+ months loan duration
logistic_duration_probability_24plus_text <- paste0(logistic_duration_probability_24plus, "%")
print(paste("Good credit risk class probability for 24+ months loan duration:", logistic_duration_probability_24plus_text))
```

```
> ## 24+ months loan duration
> ## Calculate the log-odds for 24+ months loan duration
> logistic_loan_duration_24plus_odds_ratio <- coef(logistic_loan_duration_log_model)["(Intercept)"] +
+                                             coef(logistic_loan_duration_log_model)["loan_duration_category_less_than_2424+"]
>
> # Convert the log-odds to odds ratio
> logistic_loan_duration_24plus_odds_ratio <- exp(logistic_loan_duration_24plus_odds_ratio)
> print(paste("Odds Ratio for 24+ months loan duration:", logistic_loan_duration_24plus_odds_ratio))
[1] "Odds Ratio for 24+ months loan duration: 0.510677808728014"
>
> # Convert odds ratio to probability
> logistic_duration_probability_24plus <- logistic_loan_duration_24plus_odds_ratio / (1 + logistic_loan_duration_24plus_odds_ratio)
> logistic_duration_probability_24plus <- round(logistic_duration_probability_24plus * 100, 1)
>
> # Good credit risk class probability for 24+ months loan duration
> logistic_duration_probability_24plus_text <- paste0(logistic_duration_probability_24plus, "%")
> print(paste("Good credit risk class probability for 24+ months loan duration:", logistic_duration_probability_24plus_text))
[1] "Good credit risk class probability for 24+ months loan duration: 33.8%"
```

```
> exp(-0.91422)
[1] 0.4008292
```

This “loan_duration_category_less_than_2424+” is representing the “24+” months loan duration category. The p-value **(< 2e-16)** is much smaller than the standard significance level of 0.05. This provides very strong evidence to reject the null hypothesis. Then, the calculated probability of getting good credit risk classification for “24+” months loan duration category is **33.8%**. The negative coefficient, **-0.91422**, means that longer loan durations are associated with a lower probability of getting a good credit risk classification. The odds of “24+” months loan duration category is:

$e^{-0.91422} = 0.4008292 \approx 0.4008$

So, the odds are about **40.08%** of the odds of “0-24” months loan duration category. This also indicates that the odds of getting a good credit risk classification are about **59.92%** [(1-0.4008) * 100] lower than the odds of “0-24” months loan duration category.

```
# Calculate the percentage difference between the probability for 0-24 and 24+ months loan duration
percentage_difference_probability_loan_duration <- (logistic_duration_probability_0_24 - logistic_duration_probability_24plus) / logistic_duration_probability_24plus
percentage_difference_probability_loan_duration <- round(percentage_difference_probability_loan_duration * 100, 1)
percentage_difference_probability_loan_duration_text <- paste0(percentage_difference_probability_loan_duration, "%")
print(paste("The percentage difference between 0-24 and 24+ months loan duration:", percentage_difference_probability_loan_duration_text))
```

```
> # Calculate the percentage difference between the probability for 0-24 and 24+ months loan duration
> percentage_difference_probability_loan_duration <- (logistic_duration_probability_0_24 - logistic_duration_probability_24plus) / logistic_duration_probability_24plus
> percentage_difference_probability_loan_duration <- round(percentage_difference_probability_loan_duration * 100, 1)
> percentage_difference_probability_loan_duration_text <- paste0(percentage_difference_probability_loan_duration, "%")
> print(paste("The percentage difference between 0-24 and 24+ months loan duration:", percentage_difference_probability_loan_duration_text))
[1] "The percentage difference between 0-24 and 24+ months loan duration: 65.7%"
```

These calculations were performed to find the percentage difference between the probability of getting good credit risk classification for the loan durations that are less than or equal to 24 months and above 24 months. The percentage difference is **65.7%**, indicating that the loan durations that are less than or equal to 24 months have a 65.7% chance to get a good credit risk classification than the loan durations that are above 24 months. This proves the hypothesis that indicates a more than 50% chance of getting good credit risk classification.

Extra Feature 1: One-Sample Proportion Test

```r
# Extra Feature 1 - One-Sample Proportion Test

# Filter the dataset for loan_duration_months <= 24
duration_less_than_24 <- pfda_df_Daryl %>% filter(loan_duration_months <= 24)

# Count the number of good and total cases for 0-24 months loan duration
good_count <- sum(duration_less_than_24$credit_risk_class == "good")
total_count <- nrow(duration_less_than_24)
good_count
total_count

# Perform one-sample proportion test
prop_test <- prop.test(good_count, total_count, p = 0.5, alternative = "greater")
print(prop_test)
```

```
> # Extra Feature 1 - One-Sample Proportion Test
>
> # Filter the dataset for loan_duration_months <= 24
> duration_less_than_24 <- pfda_df_Daryl %>% filter(loan_duration_months <= 24)
>
> # Count the number of good and total cases for 0-24 months loan duration
> good_count <- sum(duration_less_than_24$credit_risk_class == "good")
> total_count <- nrow(duration_less_than_24)
> good_count
[1] 2450
> total_count
[1] 4373
>
> # Perform one-sample proportion test
> prop_test <- prop.test(good_count, total_count, p = 0.5, alternative = "greater")
> print(prop_test)

        1-sample proportions test with continuity correction

data:  good_count out of total_count, null probability 0.5
X-squared = 63.269, df = 1, p-value = 9.015e-16
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5477619 1.0000000
sample estimates:
        p
0.5602561
```

The One-Sample Proportion Test is used to determine whether the loan durations that are less than or equal to 24 months have more than 50% of good credit risk classification cases. The Chi-squared (X^2) test statistic of **63.269** indicate a substantial deviation from what would be expected if loan duration category and credit risk classification were independent. This large test statistic results in an extremely small p-value, which is **9.015e-16**, contributing to strong evidence that there is statistically significant connection between loan duration category and credit risk classification in the dataset. The small p-value also contributes significantly to rejecting the null hypothesis as it is

much lower than the standard significance level which is 0.05. Based on the confidence interval, we are 95% confident that the true proportion of good credit risk classification cases lies between **54.8%** and **100%**. Additionally, the observed proportion of good credit risk cases is **56.03%** and is greater than 50%. For the loan durations that are less than or equal to 24 months, the chance of getting a good credit risk classification is higher than 50%, showing that shorter loan durations are associated with better credit risk classification outcomes.

Extra Feature 2: Two-Sample Proportion Test

```
# Extra Feature 2 - Two-Sample Proportion Test

# Count the good and total cases for 0-24 and 24+ months loan duration
two_sample_loan_counts <- pfda_df_Daryl %>%
  group_by(loan_duration_category_less_than_24) %>%
  summarise(
    two_sample_good_count = sum(credit_risk_class == "good"),
    two_sample_total_count = n()
    )
View(two_sample_loan_counts)
```

	loan_duration_category_less_than_24	two_sample_good_count	two_sample_total_count
1	0-24	2450	4373
2	24+	550	1627

```
# Extract the numbers (counts) for 0-24 and 24+ months loan duration
two_sample_0_24_good <- two_sample_loan_counts$two_sample_good_count[two_sample_loan_counts$loan_duration_category_less_than_24 == "0-24"]
two_sample_0_24_total <- two_sample_loan_counts$two_sample_total_count[two_sample_loan_counts$loan_duration_category_less_than_24 == "0-24"]
two_sample_24plus_good <- two_sample_loan_counts$two_sample_good_count[two_sample_loan_counts$loan_duration_category_less_than_24 == "24+"]
two_sample_24plus_total <- two_sample_loan_counts$two_sample_total_count[two_sample_loan_counts$loan_duration_category_less_than_24 == "24+"]

# Perform two-sample proportion test
two_sample_prop_test <- prop.test(
  x = c(two_sample_0_24_good, two_sample_24plus_good),
  n = c(two_sample_0_24_total, two_sample_24plus_total),
  alternative = "greater"
)
print(two_sample_prop_test)
```

```
> print(two_sample_prop_test)

        2-sample test for equality of proportions with continuity correction

data:  c(two_sample_0_24_good, two_sample_24plus_good) out of c(two_sample_0_24_total, two_sample_24plus_total)
X-squared = 233.32, df = 1, p-value < 2.2e-16
alternative hypothesis: greater
95 percent confidence interval:
 0.1988862 1.0000000
sample estimates:
   prop 1    prop 2
0.5602561 0.3380455
```

The Two-Sample Proportion Test is used to compare the proportion of cases with good credit risk classification in the category of loan duration less than or equal to 24 months and the category of loan duration greater than 24 months. The Chi-squared (X^2) test statistic of **233.32** indicate a substantial deviation from what would be expected if loan duration category and credit risk classification were independent. This large test statistic results in an extremely small p-value, which is **less than 2.2e-16**, contributing to strong evidence that there is statistically significant connection between loan duration category and credit risk classification in the dataset. The small p-value also contributes significantly to rejecting the null hypothesis as it is much lower than the standard significance level which is 0.05. According to the test results, there are **56.03%** cases with good credit risk classification in the category of loan duration less than or equal to 24 months. For loan duration category of greater than 24 months, there are **33.8%** cases with good credit risk classification. Comparing the statistics, loan duration category of less than or equal to 24 months has more good credit risk classification cases than loan duration category of greater than 24 months. This is also supported by the 95% confidence interval which indicates that there is at least **19.89%** true difference in the good credit risk classification cases between the two categories. With this Two-Sample Proportion Test, it enhances the evidence from One-Sample Proportion Test that shorter loan durations are associated with better credit risk classification outcomes.

Conclusion

Based on the data analysis and hypothesis testing, it is shown that there is a statistically significant relationship between loan duration and credit risk classification. Besides that, all the tests return a p-value that is much smaller than the standard significance level of 0.05. Therefore, there is sufficient proof to **reject the null hypothesis (H_0)** and **accept the alternative hypothesis (H_1)**.

3.2 Choo Cheng Da TP068973

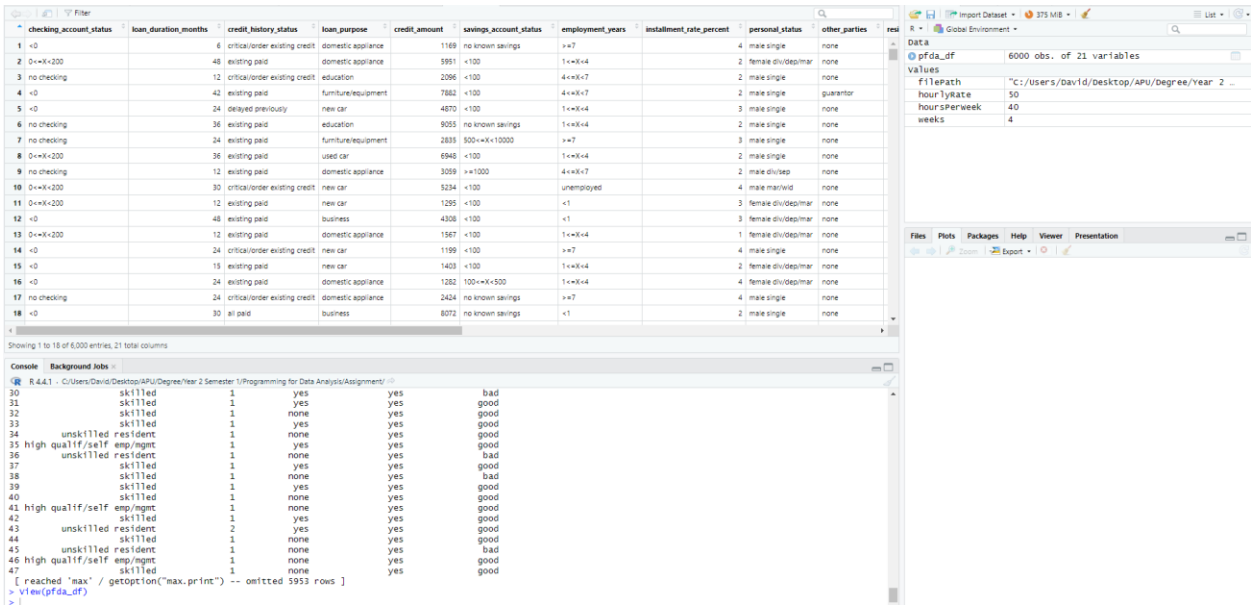
Objective: To investigate whether personal status of an individual will impact the credit risk classification of a customer.

```
# Load the libraries and packages
library(dplyr)
library(readr)
library(ggplot2)
library(DataExplorer)
library(missForest)
library(VIM)
library(caret)
library(caTools)
library(randomForest)
library(plotly)
```

First, load the libraries required to run the function of the R code.

```
# Set the working directory
## Enter your own path
setwd("C:/Users/David/Desktop/APU/Degree/Year 2 Semester 1/Programming for Data Analysis/Assignment")
getwd()


# Import the dataset from working directory
## Enter your own file path
filePath = "C:/Users/David/Desktop/APU/Degree/Year 2 Semester 1/Programming for Data Analysis/Assignment/(cleaned) credit_risk_classification.csv"

## Read the CSV file into a dataframe
pfda_df = read.csv(filePath)

# Convert the data type of all categorical columns from character to factor
pfda_df <- pfda_df %>%
  mutate(across(where(is.character), as.factor))
str(pfda_df)
sapply(pfda_df, class)

pfda_df
view(pfda_df)
```

Set the working directory from the folder location path on local PC using setwd command which sets the working directory where R will look for files to read/write. while getwd() command Retrieves and displays the current working directory, and it is useful to verify the directory has been set correctly. Then import the cleaned dataset from the folder location path. Next, reads the CSV file into a dataframe. Next, convert categorical columns from character to factor using mutate() that used to modify or create new column in a dataframe. While across() is to apply a transformation to multiple column in a dataframe and as.factor converts the selected columns to factor type to categorize the data for easier analysis. Str() is a to display column name, data type and entries of column. Sapply() is to applies a function to each column in the dataframe and retrieves the data types (class) of each column. Last, view() command is used to show the entire dataframe in a new Rstudio windows as shown in the below figure.



After all the data was imported into Rstudio. The next step is proceed with data analysis. The objective of the data analysis is to determine if the personal status will impacts good or bad credit risk classification of a customer.

Analysis 1: Show the number of customers in each personal status and the total number of customers in good or bad credit risk class.

```
# Summarize the dataset
summary(pfda_df$personal_status)
table(pfda_df$credit_risk_class)
```

First, summarize the dataset of personal status and credit risk class.

```
> summary(pfda_df$personal_status)
female div/dep/mar       male div/sep       male mar/wid         male single
              1932                517                768                2783
> table(pfda_df$credit_risk_class)

 bad good
3000 3000
```

As the result shown, there are 4 categories of personal status which are Male single, Male Married/Widowed, Male Divorced/Separated and Female Divorced/Dependent/Married. Where each bad and good credit class includes 3000 customer respectively.

```
# Analyse the credit risk classification
pfda_df <- pfda_df[order(pfda_df$personal_status, pfda_df$credit_risk_class), ]

# categorize the personal status.
pfda_df$personal_status_group <- ifelse(grepl("single", pfda_df$personal_status), "Male single",
                                ifelse(grepl("mar|vid", pfda_df$personal_status), "Male Married/Widowed",
                                       ifelse(grepl("div|sep", pfda_df$personal_status), "Male Divorced/Separated",
                                              ifelse(grepl("female div/dep/mar", pfda_df$personal_status), "Female Divorced/Dependent/Married",
                                                     "Unknown"))))
```

Analyze the credit risk classification and personal status and categories in the 4 different grouping of personal status and each grouping have good and bad credit class. Next, categorize the personal status with ifelse() command.

Analysis 2, using visualization techniques for further analysis with bar chart.

```
# Summarize data to calculate the total people of each category
bar_chart <- pfda_df %>%
  group_by(personal_status, credit_risk_class) %>%
  summarise(count = n(), .groups = "drop") %>%
  mutate(total = sum(count),
         percentage = (count / total) * 100)

# Bar Chart
plot <- ggplot(pfda_df, aes(x = personal_status, fill = credit_risk_class)) +
  geom_bar(position = "dodge") +
  geom_text(stat = "count",
            aes(label = after_stat(count)),
            position = position_dodge(width = 0.9),  # Aligns text with bars
            vjust = -0.5) +  # Places text above bars
  labs(title = "Credit Risk Classification by Personal Status",
       x = "Personal Status",
       y = "Number of Customers",
       fill = "Credit Risk") +
  theme_minimal()

print(plot)
```

A bar chart is a chart that is used to present numerical differences between categories. Summarize the data is the first step with input pfda_df, groups the data by personal status and credit risk class. Summarize the number of rows (customers) for each group and Ensures the resulting dataframe is ungrouped after summarizing using drop. Then mutate total number of customers.

This bar chart visualizes the distribution of credit risk classifications (good vs. bad) across different personal statuses of individuals. The x-axis represents the personal status categories, while the y-axis shows the count of customers in each group. The bars are color-coded to distinguish between good and bad credit risks.

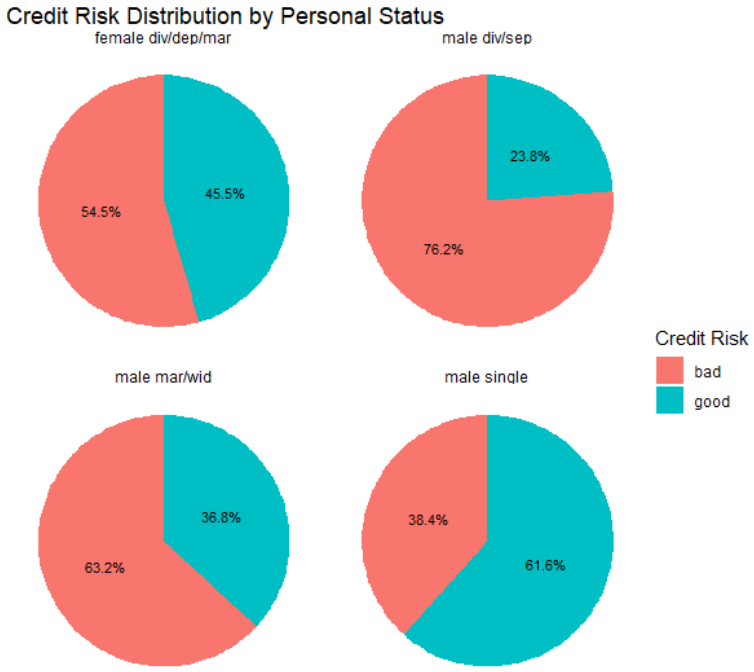
Key observation

The male single category has the highest count of individuals with a good credit risk classification (1715), followed by female div/dep/mar (879). The lowest count of good credit risks is seen in the male div/sep group (123). While on the bad credit risk is highest for the male single category (1068), and lowest for male div/sep (394). In the female div/dep/mar category, the number of individuals with a bad credit risk (1053) exceeds those with a good credit risk (879), indicating that this group tends to have more credit issues. Conversely, in the male single group, individuals classified as having a good credit risk significantly outnumber those with a bad credit risk.

Analysis 3, using visualization techniques for further analysis with pie chart.

```
# Normalize data to ensure each pie chart sums to 100%
pie_chart_data <- pfda_df %>%
  group_by(personal_status, credit_risk_class) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(personal_status) %>%
  mutate(total = sum(count),
         percentage = (count / total) * 100)

# Pie Chart
pie_chart <- ggplot(pie_chart_data, aes(x = "", y = percentage, fill = credit_risk_class)) +
  geom_bar(stat = "identity", width = 1) +
  coord_polar(theta = "y") +  # Transform to pie chart
  facet_wrap(~ personal_status) +  # Create one pie chart per personal status
  labs(title = "Credit Risk Distribution by Personal Status",
       fill = "Credit Risk",
       y = "",
       x = "") +
  theme_void() +  # Removes axis and gridlines for clean pie chart
  geom_text(aes(label = paste0(round(percentage, 1), "%")),
            position = position_stack(vjust = 0.5),  # Center text inside slices
            size = 3)

print(pie_chart)
```

Pie chart offers a good visual to show the percentage of the credit risk distribution by personal status. With the code above, summarizing the data is the first step with input pfda_df, groups the data by personal status and credit risk class. Summarize the number of rows (customers) for each group and ensure the resulting dataframe is ungrouped after summarizing using drop. Then mutate total number of customers and use the percentage formula to get the percentage number of customers accurately.



The pie charts display the proportion of credit risk classifications (good vs. bad) for each personal status category. The proportions are represented as percentages, with bad credit risks shown in red and good credit risks shown in light blue.

Key observations with 4 different personal status

a) Female div/dep/mar

Distribution:

- Bad Credit Risk: 54.5%
- Good Credit Risk: 45.5%

Explanations:

The female div/dep/mar group has a slightly higher proportion of individuals classified as bad credit risks compared to good risks. This indicates that this group might face moderate financial challenges, with a majority leaning toward riskier credit behavior.

b) Male div/sep

Distribution:

- Bad Credit Risk: 76.2%
- Good Credit Risk: 23.8%

Explanations:

The male div/sep group shows the highest percentage of bad credit risks, with over three-quarters of individuals classified as bad. This suggests that this group is the most financially vulnerable or prone to credit issues.

c) Male mar/wid

Distribution:

- Bad Credit Risk: 63.2%
- Good Credit Risk: 36.8%

Explanations:

The male mar/wid group also has a majority of bad credit risks (over 60%), indicating significant financial challenges, though not as severe as the male div/sep group.

d) Male single

Distribution:

- Bad Credit Risk: 38.4%
- Good Credit Risk: 61.6%

Explanations:

The male single group is the least financially risky, with a clear majority (over 60%) classified as good credit risks. This suggests better financial stability or repayment behavior among this demographic.

Conclusion

After two charts is shown here, the bar chart provides a detailed view of credit risk classifications across personal status while the pie chart highlights the varying proportions of credit risks among personal status categories. Both analysis underscores the influence of personal status on financial behavior and offers actionable insights for targeted financial strategies.

Insights

The female div/dep/mar category appears to be a riskier group overall, as it has a higher proportion of bad credit risks compared to good ones. Besides that, male div/sep group also shows a higher count of bad credit risks compared to good risks, although this group has the lowest overall counts. On the other hand, the male single category stands out as having the most favorable credit risk profile, with the highest number of good credit risks.

Implications

These findings show that personal status may play a role in determining credit risk of a bank. For example, single males might have more financial stability or better repayment behaviors compared to divorced or widowed individuals. The higher proportion of bad credit risks in the female div/dep/mar category might point to socioeconomic challenges faced by individuals in this group.

Analysis 4: Analyse the data using Logistic regression model.

```
#Data analysis
#Test the data about credit risk classification (Good = 0, Bad = 1) for each group using a logistic regression model
# Check the structure of the data
str(data)

# Fit logistic regression model
log_model <- glm(credit_risk_class ~ personal_status, data = pfda_df, family = binomial)
summary(log_model)


#Interpretation of the Results and convert to probability
odds <- exp(0.2897)            # Convert log-odds to odds
probability <- odds / (1 + odds)  # Convert odds to probability
print(probability)
```

Logistic Regression is a model to predict the probability of "Bad" credit risk based on personal status. First, Create a dataset for all the personal status. A binary variable where 1 represents bad credit risk, and 0 represents good credit risk. Then fit the logistic regression model to the data. This code fits a logistic regression model to analyze the relationship between credit risk class (the dependent variable) and personal status (the independent variable) using the dataset pfda_df. Using glm() is to fit regression models and use the formula of credit_risk_class ~ personal_status

to estimate the effect of different level of personal_status of classified as bad credit risk class. Lastly, display the result including coefficients and dispersion.

```
> # Check the structure of the data
> str(data)
'data.frame':   2783 obs. of  2 variables:
 $ credit_risk    : num  1 1 1 1 1 1 1 1 1 1 ...
 $ personal_status: chr  "male single" "male single" "male single" "male single" ...
>
> # Fit logistic regression model
> log_model <- glm(credit_risk_class ~ personal_status, data = pfda_df, family = binomial)
> summary(log_model)

Call:
glm(formula = credit_risk_class ~ personal_status, family = binomial,
    data = pfda_df)

Coefficients:
                             Estimate Std. Error z value Pr(>|z|)
(Intercept)                 -0.18061    0.04569  -3.953 7.71e-05 ***
personal_statusmale div/sep -0.98355    0.11294  -8.709  < 2e-16 ***
personal_statusmale mar/wid -0.35809    0.08765  -4.085 4.40e-05 ***
personal_statusmale single   0.65424    0.06006  10.894  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 7947.1  on 5996  degrees of freedom
AIC: 7955.1

Number of Fisher Scoring iterations: 4

>
>
> #Interpretation of the Results and convert to probability
> odds <- exp(0.2897)            # Convert log-odds to odds
> probability <- odds / (1 + odds)  # Convert odds to probability
> print(probability)
[1] 0.5719227
>
```

From the output, the dataset contains 2783 observations and 2 variables. Call function is to Confirms the formula used in the model where the dependent variable is credit risk class while the independent variable is personal status. The logistic regression model is fitted to predict credit risk class based on personal status while the family = binomial indicates that this is a logistic regression model.

Coefficients:

Intercept (-0.18061)

The intercept represents the log-odds of credit_risk_class = 1 (bad) when personal_status is the reference category. The **log-odds** of a "bad" credit risk for the baseline group are -0.18061. Odds

$e^{0.18061} \approx 0.834$e^{-0.18061} meaning the baseline group has a 16.6% lower odds of having bad credit risk compared to having good credit risk.

personal_status of male div/sep (-0.98355)

For male div/sep, the coefficient is negative (-0.98355), meaning that this group is less likely to have bad credit risk compared to the reference group. Odds $e^{-0.98355} \approx 0.374$e^{-0.98355}, meaning male div/sep has approximately 63% lower odds of having bad credit risk compared to the baseline group.

personal_status of male mar/wid (-0.35809)

For male mar/wid, the coefficient is also negative (-0.35809), meaning that this group is less likely to have bad credit risk than the reference group. Odds: $e{-0.35809} \approx 0.699$e^{-0.35809}, meaning male mar/wid has 30.1% lower odds of having bad credit risk compared to the baseline.

personal_status for male single (0.65424)

For male single, the coefficient is positive (0.65424), meaning that this group is more likely to have bad credit risk compared to the reference group. **Odds**: $e^{0.65424} \approx 1.924$e^{0.65424, meaning male single has 92.4% higher odds of having bad credit risk compared to the baseline.

For statistical significance (p-values), (Intercept) 7.71e-05, personal_statusmale div/sep < 2e-16, personal_statusmale mar/wid 4.40e-05 and personal_statusmale single < 2e-1 □ All coefficients are highly statistically significant, with **p-values** less than 0.05 (some even much smaller). This suggests that all personal_status categories (compared to the baseline) are significantly related to the likelihood of having bad credit risk.

Null deviance: The deviance (a measure of model fit) when no predictors are included (intercept-only model). Residual deviance: The deviance after including personal_status as a predictor. Lower residual deviance indicates a better model fit. Here, the residual deviance decreased from 8317.8 to 7947.1, indicating that the model with personal_status as a predictor is a better fit. AIC (Akaike Information Criterion) indicates a better model. Here, the AIC is 7955.1, which reflects a good fit.

Literature review

This paper explores the relationship between credit constraints and marriage in France, focusing on how marital status impacts access to loans. Using data from the 2001 housing survey, it examines whether marriage serves as a signal of stability to lenders. Results reveal no significant difference in credit constraints between married and unmarried couples but show that unmarried couples are more

likely to feel discouraged from borrowing. Married couples appear disadvantaged in terms of loan terms, suggesting that unobservable characteristics may influence credit outcomes. The study connects theories of marriage with credit accessibility. (Leturcq, n.d.)

Source: Leturcq, M. (n.d.). *Do bankers prefer married couples?* [online] Available at: https://shs.hal.science/halshs-00655584/document [Accessed 1 Dec. 2024].

Hypothesis

On hypothesis, the target on the personal status will be male single. There are two hypothesis which are Null hypothesis and Alternative Hypothesis.

Null Hypothesis (H_0):

There is no significant difference in the distribution of credit risk classification (Good vs. Bad) for the "male single" group.

Alternative Hypothesis (H_1):

There is a significant difference in the distribution of credit risk classification (Good vs. Bad) for the "male single" group.

To test and prove which hypothesis can be use, chi-square test of independence. The Chi-Square test is a statistical procedure for determining the difference between observed and expected data (Avijeet Biswal, 2021).

```
# Create a contingency table
contingency_table <- table(pfda_df$personal_status, pfda_df$credit_risk_class)

# Display the table
print(contingency_table)
```

To specifically show the table for better investigate of each personal status, create a contingency table to show good and bad credit class as shown as the output below.

```
# Create a contingency table
contingency_table <- table(pfda_df$personal_status, pfda_df$credit_risk_class)

# Display the table
print(contingency_table)
```

```
                   bad good
female div/dep/mar 1053  879
male div/sep        394  123
male mar/wid        485  283
male single        1068 1715
```

Next, observe the frequencies for male single personal status to match the target hypothesis.

```
# Observed frequencies for male single
observed <- c(1068, 1715)

# Assign labels for clarity
names(observed) <- c("Bad", "Good")

# Display the observed data
print(observed)
```

```
> # Observed frequencies for male single
> observed <- c(1068, 1715)
>
> # Assign labels for clarity
> names(observed) <- c("Bad", "Good")
> # Display the observed data
> print(observed)
 Bad Good
1068 1715
```

Then, set the expected data of total count of male single individuals and calculate the expected frequencies under null hypothesis (50% 50%).

```
#set the expected data
# Total count of male single individuals
total <- sum(observed)

# Expected frequencies under null hypothesis (50%-50%)
expected <- c(total / 2, total / 2)

# Display the expected data
print(expected)
```

```
> # Expected frequencies under null hypothesis (50%-50%)
> expected <- c(total / 2, total / 2)
>
> # Display the expected data
> print(expected)
[1] 1391.5 1391.5
```

The final step comes to perform a chi-square test and the output is shown.

```
# Perform Chi-Square Test
chi_result <- chisq.test(observed, p = c(0.5, 0.5)) # 50%-50% expected proportions

# Display the test results
print(chi_result)
```

```
> # Perform Chi-Square Test
> chi_result <- chisq.test(observed, p = c(0.5, 0.5)) # 50%-50% expected proportions
>
> # Display the test results
> print(chi_result)

        Chi-squared test for given probabilities

data:  observed
X-squared = 150.42, df = 1, p-value < 2.2e-16
```

Analysis the test result of the chi-square test:

Test Statistic (X-squared): 150.42

1. This is the value of the chi-squared test statistic, which measures how much the observed data deviates from the expected data under the null hypothesis.

Degrees of Freedom (df): 1

- This is the number of independent categories minus one. Here, df = 1 indicates one degree of freedom (likely comparing two groups: "Good" vs. "Bad" credit risk).

P-value (< 2.2e-16):

- This is the probability of observing a chi-squared statistic as extreme as 150.42. The p-value is **< 2.2e-16**, which is smaller than the common significance level (e.g., $\alpha = 0.05$ or 0.01). This means that the probability of observing a chi-squared statistic as large as 150.42, assuming the null hypothesis is true, is essentially zero.

In conclusion, null hypothesis is rejected since the p-value is extremely small (< 0.05). There is strong evidence to support the alternative hypothesis (H_1). The distribution of credit risk (Good vs. Bad) is not independent of the group being analyzed which is male single. In practical terms, this means that personal status ("male single") significantly affects the likelihood of credit risk classification.

3.3 Cheong Sheue Ling TP069004

Objective: To investigate whether property magnitude impacts the credit risk classification of a customer.

Analysis

Probability of each category of property magnitude

Exploratory Data Analysis

```
# 1.1 Check the details of this column
summary(pfda_df$property_magnitude)
# 1.2 Check the frequency of each unique category
unique(pfda_df$property_magnitude)
as.data.frame(table(pfda_df$property_magnitude))
```

summary(pfda_df$property_magnitude): Summarizes the property_magnitude column.

```
> summary(pfda_df$property_magnitude)
   Length     Class      Mode
     6000 character character
```

unique(pfda_df$property_magnitude): Lists the unique values in the property_magnitude column.

```
> unique(pfda_df$property_magnitude)
[1] "real estate"       "life insurance"    "no known property"
[4] "car"
```

as.data.frame(table(pfda_df$property_magnitude)): Displays the frequency of each unique value in property_magnitude.

```
> as.data.frame(table(pfda_df$property_magnitude))
               Var1 Freq
1               car 2156
2    life insurance 1576
3 no known property 1002
4       real estate 1266
```

```
joint_counts <- table(pfda_df$property_magnitude,pfda_df$class )

View(joint_counts)
```

joint_counts <- table(pfda_df$property_magnitude, pfda_df$class): Creates a contingency table showing counts of different property_magnitude categories for each class (good or bad).

view(joint_counts): Views the joint frequency table.

	Var1	Var2	Freq
1	car	bad	1150
2	life insurance	bad	871
3	no known property	bad	640
4	real estate	bad	339
5	car	good	1006
6	life insurance	good	705
7	no known property	good	362
8	real estate	good	927

```
good_class_counts <- joint_counts[, "good"]
View(good_class_counts)
```

good_class_counts <- joint_counts[, "good"]: Extracts the counts for the good class.

view(good_class_counts): Views the counts of good class.

good_class_counts	integer [4]	1006 705 362 927
car	integer [1]	1006
life insurance	integer [1]	705
no known property	integer [1]	362
real estate	integer [1]	927

```
category_frequencies <- rowSums(joint_counts)
View(category_frequencies)


#Extract the probability of each category
category_probabilities <- good_class_counts / category_frequencies
View(category_probabilities)
```

category_frequencies <- rowSums(joint_counts): Calculates the total frequency for each property magnitude category.

view(category_frequencies): Views the category frequency.

category_frequencies	double [4]	2156 1576 1002 1266
car	double [1]	2156
life insurance	double [1]	1576
no known property	double [1]	1002
real estate	double [1]	1266

category_probabilities <- good_class_counts / category_frequencies: Calculates the probability of each category belonging to the good class.

view(category_probabilities): Views the probabilities.

category_probabilities	double [4]	0.467 0.447 0.361 0.732
car	double [1]	0.4666048
life insurance	double [1]	0.447335
no known property	double [1]	0.3612774
real estate	double [1]	0.7322275

```
total_good_class <- sum(good_class_counts)
View(total_good_class)
```

total_good_class <- sum(good_class_counts): Calculates the total number of good class entries.

view(total_good_class): Views the total number of good class.

total_good_class	integer [1]	3000

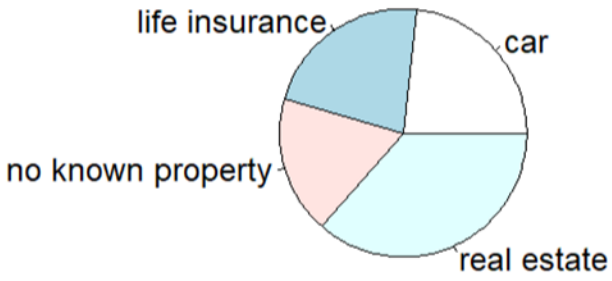
```
property_info <- data.frame(
  PropertyMagnitude = names(category_frequencies),
  CategoryFrequency = category_frequencies,
  GoodClassFrequency = good_class_counts,
  ProbabilityGoodClass = category_probabilities
)
View(property_info)
```

The code combines the frequency and probability information into a new data frame property_info, which is then viewed.

	PropertyMagnitude	CategoryFrequency	GoodClassFrequency	ProbabilityGoodClass
car	car	2156	1006	0.4666048
life insurance	life insurance	1576	705	0.4473350
no known property	no known property	1002	362	0.3612774
real estate	real estate	1266	927	0.7322275

```
probability<-c(0.4666048,0.4473350,0.3612774,0.7322275)
labels <- c("car","life insurance", "no known property", "real estate")
pie(probability, labels)
```

A pie chart is created to visually represent the probabilities of different categories (car, life insurance, no known property, and real estate).



Literature Review

Real estate ownership is a proxy for wealth and financial stability. Borrowers with real estate are often seen as having better financial management skills and access to stable income sources, which are critical for timely debt repayments. Financial stability correlates positively with higher credit scores, as highlighted by research in the field of consumer finance (Campbell & Cocco, 2015).

Hypothesis

Alternative Hypothesis: Users whose property magnitude is real estate have higher probability to get good class.

Null hypothesis: Users whose property magnitude is real estate do not have higher probability to get good class.

```
pfda_df$class_binary <- ifelse(pfda_df$class == "good", 1, 0)
propertymagnitude_category_log_model <- glm(class_binary ~ property_magnitude, data = pfda_df, family = binomial)
summary(propertymagnitude_category_log_model)
```

pfda_df$class_binary <- ifelse(pfda_df$class == "good", 1, 0): Creates a binary class variable, where good is 1 and other classes are 0.

propertymagnitude_category_log_model <- glm(class_binary ~ property_magnitude, data = pfda_df, family = binomial): Fits a logistic regression model predicting the binary class (good or bad) based on property_magnitude.

summary(propertymagnitude_category_log_model): Summarizes the logistic regression model results, showing the significance of the predictor.

```
> summary(propertymagnitude_category_log_model)

Call:
glm(formula = class_binary ~ property_magnitude, family = binomial,
    data = pfda_df)

Coefficients:
                                Estimate Std. Error z value
(Intercept)                     -0.13378    0.04317  -3.099
property_magnitudelife insurance    -0.07766    0.06656  -1.167
property_magnitudeno known property -0.43604    0.07867  -5.543
property_magnitudereal estate        1.13973    0.07676  14.848
                                Pr(>|z|)
(Intercept)                      0.00194 **
property_magnitudelife insurance     0.24327
property_magnitudeno known property 2.98e-08 ***
property_magnitudereal estate        < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 7928.6  on 5996  degrees of freedom
AIC: 7936.6

Number of Fisher Scoring iterations: 4
```

Conclusion

Based on the model output, the null hypothesis is rejected, and the alternative hypothesis is accepted. This means users with real estate property magnitude have a higher probability of being classified as good.

3.4 Ho Shane Foong TP068496

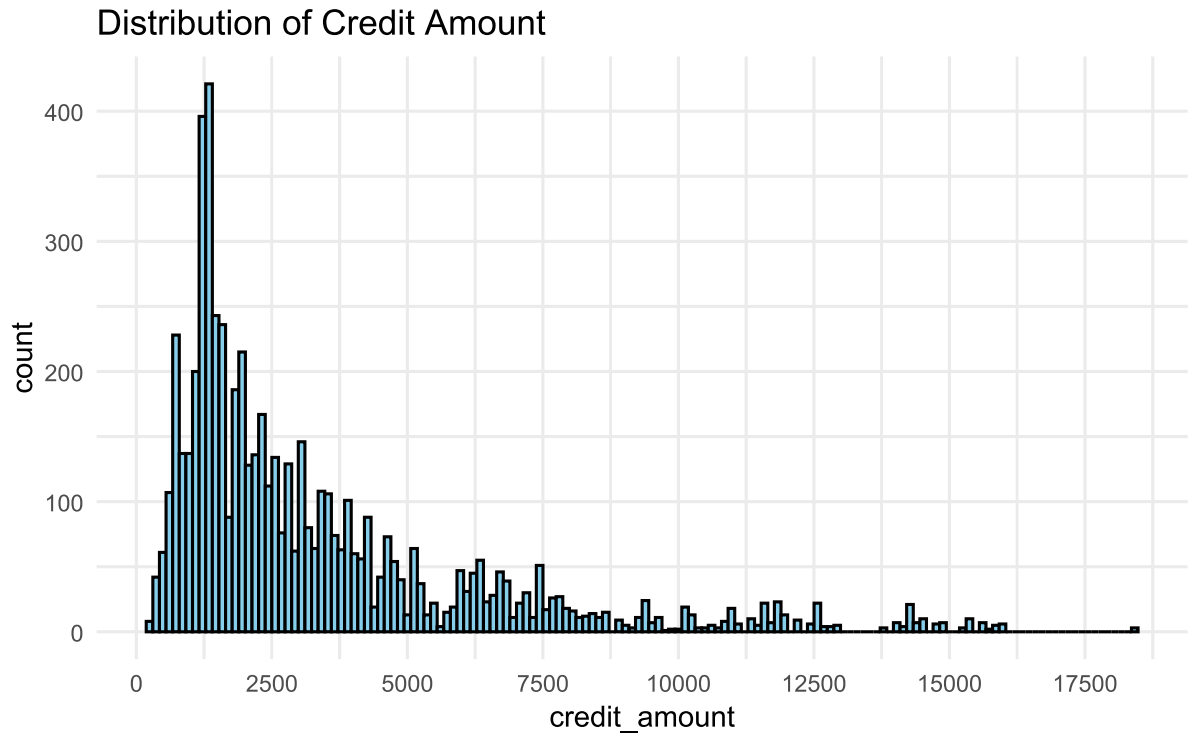
Objective: To determine whether the credit amount has effect on the individual's credit risk class.

Analysis

Probability of Credit amount below 10,000 (10k) to have effect on the individual's credit risk class.

Exploratory Data Analysis

```
## duplicate raw_data
raw_data_credit_amount <- raw_data

#### Histogram for credit amount ####
plot_credit_amount <- ggplot(raw_data_credit_amount, aes(x=credit_amount)) +
  scale_x_continuous(breaks = seq(0,max(raw_data$credit_amount),2500)) +
  geom_histogram(bins = 150 ,fill = "skyblue", color = "black") +
  labs(title = "Distribution of Credit Amount") +
  theme_minimal()
plot_credit_amount
```



On the code snipped in figure 10, the *raw_data_credit_amount* will receive a copy of *raw_data*. This is to ensure that any change made to the *raw_data_credit_amount* will not affect the original data.

Now using “ggplot” to create a histogram shown in figure 11, where it is based on the *raw_data_credit_amount* which it show us the distribution of the credit amount inside the data.

By using *scale_x_continuous*, it allow the x-axis to be shown with the scale from 0 to the maximum credit amount with break every 2500.

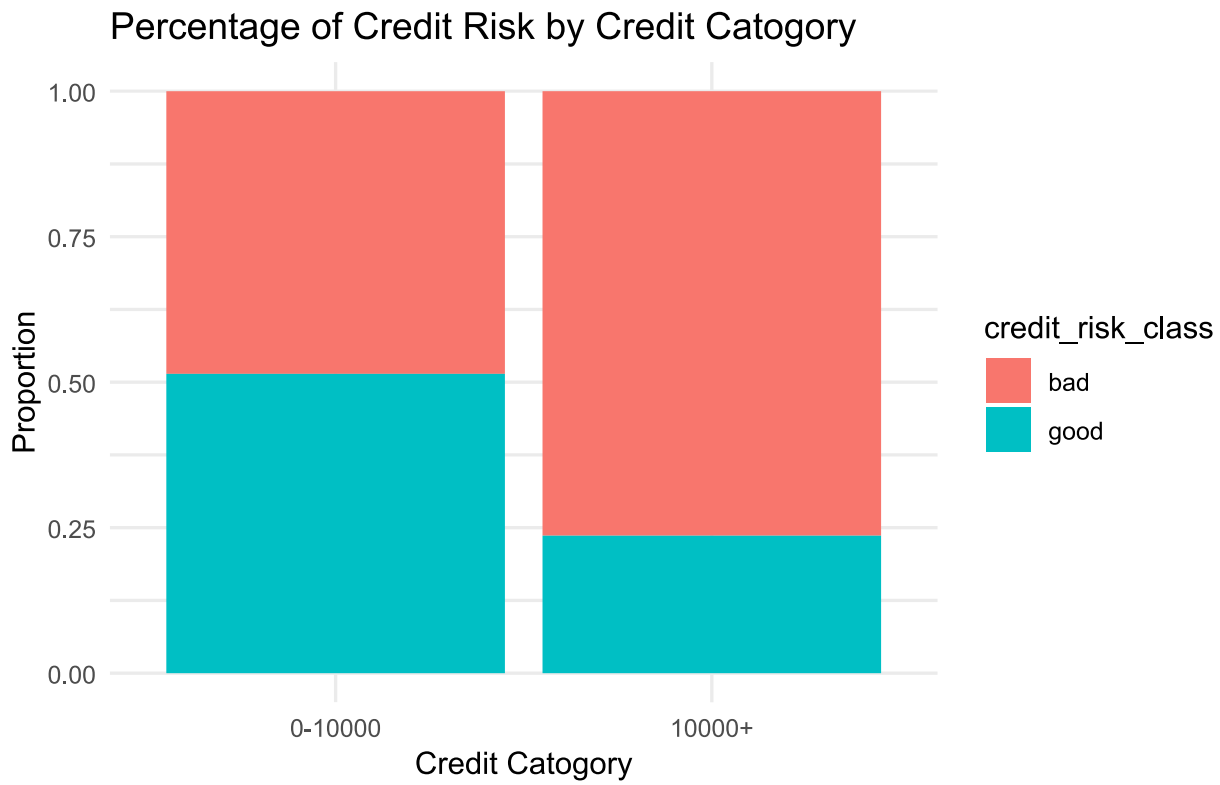
```
## add category to differnece with 0-10000 and 10000+
raw_data_credit_amount$cato_by_credit_less_than_10k <- cut(raw_data$credit_amount,
                                                           breaks = c(0, 10000, Inf),
                                                           labels = c("0-10000", "10000+"))

## filter those row where credit amount is less than 10000
raw_data_credit_amount_below_10k <- raw_data_credit_amount %>%
  filter(credit_amount <= 10000)
```

Running the “cut” function is to categorize the credit amount into two categories, “0-10000” and “10000+” where corresponding to any credit amount below 10000 will be set to “0-10000” and “10000+” when it is above 10000. This will put the category into another column inside the data frame.

Then created another variable called *raw_data_credit_amount_below_10k* where it clones the data from the *raw_data_credit_amount* but it is filter when the credit is below 10000.

```
## calcuate the catogory of 0-10000 and 10000+
summary(raw_data_credit_amount$cato_by_credit_less_than_10k)
> summary(raw_data_credit_amount$cato_by_credit_less_than_10k)
0-10000  10000+
   5691     309
```

Now can use summary to the count of “0-10000” and “10000+” inside our data. Reason why we categorize the credit amount is because our analysis will be based on the credit amount that is less than 10k.




```
#### Proportion on credit risk by loan purpose ####
raw_data_credit_amount_filted_pct <- raw_data_credit_amount %>%
  group_by(cato_by_credit_less_than_10k, credit_risk_class) %>%
  summarise(count = n(), .groups = "drop") %>%
  group_by(cato_by_credit_less_than_10k) %>%
  mutate(percentage = count / sum(count) * 100)

# Plot the percentages
plot_credit_amount_bar_pct <- ggplot(raw_data_credit_amount_filted_pct,
                                     aes(x = cato_by_credit_less_than_10k, y = percentage,
                                         fill = credit_risk_class)) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Percentage of Credit Risk by Credit Catogory",
       x = "Credit Catogory",
       y = "Proportion") +
  theme_minimal()
plot_credit_amount_bar_pct
```

Now we can determine the percentage of each credit amount category with its credit risk class. By grouping then from *cato_by_credit_less_than_10k* and *credit_risk_class* from the *raw_data_credit_amount* then summarize them and regroup them based on the *cato_by_credit_less_then_10k* and mutate the percentage into it.

Then using “ggplot” to represent the data in the *raw_data_credit_amount_filted_pct* to proportion in each category with each credit risk.

Literature Review

In the article from Wells Fargo, the amount that you borrow will affect the interest rate based on the borrowed amount. With the interest rate being determined by how much you have borrowed, the loan approver will need to consider how long does the borrower is able to pay back during its lifetime. (*Understand the Total Cost of Borrowing – Wells Fargo*, 2024)

From the article of Bluebrick explaining on how personal loan process in Malaysia, the bank will rely on the ones credit score. Having a high credit score will lead to a successful borrowing of loan. But in Malaysia, they will do one step extra by cross-checking established credit rating agencies like, Central Credit Reference Information System (CCRIS) and CTOS. These agencies check with ones loan usage and repayment to determine approval of loan being approved. (Wilson, 2023)

Hypothesis

- Null: credit amount is less than 10000 do not have more probability of receiving a good credit risk
- Alternative: credit amount is less than 10000 have 50% more probability of receiving a good credit risk

One-Sample Proportions Testing

```
#### one sample prop testing ####
prop_test_below_10k <- prop.test(sum(raw_data_credit_amount_below_10k$credit_risk_class == "good"),
                                 nrow(raw_data_credit_amount_below_10k),
                                 p = 0.5,
                                 alternative = "greater")
prop_test_below_10k
```

```
        1-sample proportions test with continuity correction

data:  sum(raw_data_credit_amount_below_10k$credit_risk_class == "good") out of nrow(raw_data_credit_amount_bel
ow_10k), null probability 0.5
X-squared = 4.6115, df = 1, p-value = 0.01588
alternative hypothesis: true p is greater than 0.5
95 percent confidence interval:
 0.5033313 1.0000000
sample estimates:
        p
0.5143209
```

Using the One sample proportions testing, we can provide it with the total row of credit risk class of good and the total row from *raw_data_credit_amount_below_10k,* data that we created and filter. Providing the P as 0.5 and alternative as “greater” to make sure that the proportions testing have more than 50%.

By running the test returning us with the ‘**P**’ gave us **0.5143209** where converted to percentage is more than 50%. With the return of the data, the **p-value** is 0.01588, compare to our significant value as 0.05, we will reject the null hypothesis and accept the alternative hypothesis.

Two-Sample t-test

```
#### two sample prop testing (t-test) ####
test_result_credit_amount <- t.test(credit_amount ~ credit_risk_class, data = raw_data_credit_amount)
test_result_credit_amount
```

```
        Welch Two Sample t-test

data:  credit_amount by credit_risk_class
t = 9.8411, df = 5423.4, p-value < 2.2e-16
alternative hypothesis: true difference in means between group bad and group good is not equal to 0
95 percent confidence interval:
 600.5134 899.2819
sample estimates:
 mean in group bad mean in group good
          3719.243           2969.345
```

Using t-test by providing the credit amount and the credit risk class where it compares the average credit amount with the credit risk class. The **p-value** is 2.2e-16 whereby the significant value of 0.05 is smaller than the **p-value**. But this testing does not provide any information for the hypothesis. Hence, we will not be accepting or rejecting any hypothesis.

Logistic Regression and Odd ratio

```
#### Log regression with odd ratio and precentage ####
# Factorise credit risk class
raw_data_credit_amount$credit_risk_class <- factor(raw_data_credit_amount$credit_risk_class,
                                                   levels = c("bad", "good"))

# Relevel the catogory by credit less than 10k
raw_data_credit_amount$cato_by_credit_less_than_10k <- relevel(raw_data_credit_amount$cato_by_credit_less_than_10k,
                                                               ref = "10000+")

# Perform Log regression
glm_model_credit <- glm( credit_risk_class ~ cato_by_credit_less_than_10k,
                         family = binomial,
                         data = raw_data_credit_amount)
summary(glm_model_credit)
```

```
Call:
glm(formula = credit_risk_class ~ cato_by_credit_less_than_10k,
    family = binomial, data = raw_data_credit_amount)

Coefficients:
                                  Estimate Std. Error z value Pr(>|z|)
(Intercept)                        -1.1734     0.1339  -8.761   <2e-16 ***
cato_by_credit_less_than_10k0-10000   1.2307     0.1365   9.014   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 8222.6  on 5998  degrees of freedom
AIC: 8226.6

Number of Fisher Scoring iterations: 4
```

Using Logistic Regression allow us to find the odd ratio for credit risk class and category of credit amount. This testing model the relationship between credit risk class and category of credit amount. But first we have to convert the column *credit_risk_class* into a factor by using the factor function

with the label of "bad" and "good". We also provide a relevel function to change the reference value to '10000+' as we going to use '10000+' as the baseline of our Logistic Regression against '0-10000'. Once we run the Logistic Regression, we are going to take the coefficient of '*cato_by_credit_less_than:10k0-10000'* as this measure the probability of "good" credit risk with the category level of "0-10000". We could take the estimate of **1.2307** and calculate the probability. We could see the **p-value** is **2e-16** where is lower than our significance value of 0.05 to be one step forward to reject the null hypothesis.

```
# Calculate odd ratio and probability
ratio_credit_amount <- exp(coef(glm_model_credit)["cato_by_credit_less_than_10k0-10000"])
ratio_credit_amount

probability_credit_less_than_10k <- ratio_credit_amount * 100 / (1 + ratio_credit_amount)
probability_credit_less_than_10k
```

```
> # Calculate odd ratio and probability
> ratio_credit_amount <- exp(coef(glm_model_credit)["cato_by_credit_less_than_10k0-10000"])
> ratio_credit_amount
cato_by_credit_less_than_10k0-10000
                           3.423528
>
> probability_credit_less_than_10k <- ratio_credit_amount * 100 / (1 + ratio_credit_amount)
> probability_credit_less_than_10k
cato_by_credit_less_than_10k0-10000
                           77.39361
```

To calculate the probability, we can hand in the coeffience estimate value of **1.2307** from *cato_by_credit_less_than_10k0-10000* and put it into exponent funtion where it return us with **3.423528**. Putting the value into the formula *"ratio_credit_amount * 100 / (1+ratio_credit_amount)*" return us with the probability of 77%.

With the **p-value** of **2e-16** and the probability of **77%**, where it is **p-value** is lower than the significant value of 0.05 and the probability is more than 50% whereby we can reject the null hypothesis and accept the alternative hypothesis.

3.5 John Har Wey Jon TP068348

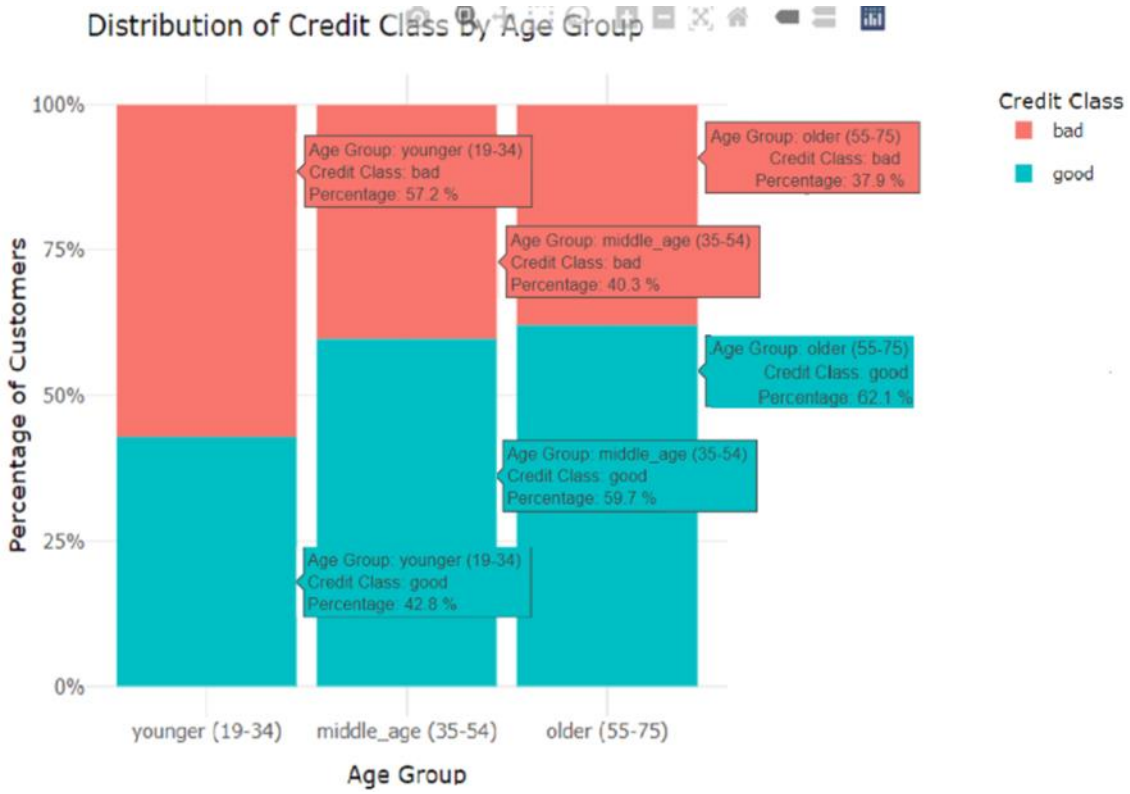
Objective: To determine how age impacts the credit risk classification of a customer.

```
# Analyse the credit risk classification differ between younger, middle_age, and older customers.
sort_age_pfda_df <- pfda_df[order(pfda_df$age, pfda_df$credit_risk_class), ]

# This code is to categorize the age to 3 intervals
sort_age_pfda_df$age_group <- cut(sort_age_pfda_df$age,
                                  breaks = c(18,34,54,75),
                                  right = TRUE,
                                  labels = c("younger (19-34)", "middle_age (35-54)", "older (55-75)"),
                                   include.lowest = TRUE)
```

```
# Summarize data to get counts and percentages
bar_chart_summary <- sort_age_pfda_df %>%
  group_by(age_group, credit_risk_class) %>%
  summarise(count = n()) %>%
  mutate(total = sum(count),
         percentage = (count / total) * 100)
```

```
# Bar Chart
bar_plot <- ggplot(bar_chart_summary, aes(x = age_group, y = count, fill = credit_risk_class,
                                          text = paste("Age Group:", age_group,
                                                       "<br>Credit Class:", credit_risk_class,
                                                       "<br>Percentage:", round(percentage, 1), "%"))) +
  geom_bar(stat = "identity", position = "fill") +
  labs(title = "Distribution of Credit Class by Age Group",
       x = "Age Group",
       y = "Percentage of Customers",
       fill = "Credit Class") +
  scale_y_continuous(labels = scales::percent) +
  theme_minimal()

bar_plot

# This extra feature is to help  analyse the bar chart accurately by hovering over the bar chart and it will list the percentage, counts, credit class, and the age group.
interactive_plot <- ggplotly(bar_plot, tooltip = "text")

interactive_plot
```

During the process of my analysis, I had chosen visualization techniques to analyse the dataset. The main objective of this code is to identify whether there is relationship between the age as the independent variable and the credit risk classification as the dependent variable by generating a interactive stacked bar chart that displays the percentage of the “good” and “bad” credit risk within the categorized age group. To arrange the data into age intervals and credit risk categories, the code implemented separates the cleaned dataset into age_group and classes. After that it summarises the total number of customers of each age group and credit risk class. Following a new column of (%) is implemented for the code to display the percentage of customers for each age group that are categorized between “good” or “bad” credit risk. Lastly, the use of the code “library ggplot2” is to allow the generation of the stacked bar chart with the categorized age group and credit risk class (good or bad) indicated in each bar with this two information it allows the bar chart to show the percentage of the categorized customers in each category. The bar chart uses the position = “fill” to indicate a 100% stack bar chart in the bar as percentages. The y-axis is labelled as “Percentage of Customers”, and numbers are labelled as percentages for easy analysis.



The generated bar chart shown in the figure above shows the distribution of credit class in age group. The result shown above as the conclusion, it allows us to analyze that the bar representing the last age group (Older) which the customers ages between 55-75 have more reliability and higher rate of getting a good credit risk compared to the age group younger. Giving the last result, it states that the age group of younger customers have a higher rate of getting a “bad” credit risk, where most of the older customers have “good” credit risk. This chart shows that when the customers age increases, the rate of having a “good” credit risk increase, this information proves that there is a positive relationship between the age and the credit risk class.

Hypothesis

In this bar chart, it shows that older customers have a higher probability of having a “good” credit risk classification comparing to the customers with the younger age group, which indicates there is a positive relationship between the independent variable age and dependent variable credit risk classification.

Literature Review

Supporting this idea, Romana Korez Vide, Vesna Sesar, and Ivica Zdrilic conducted the study where age affects main financial attitudes and behaviors.

The sentence written in the literature review like “Older, more established reports generally indicate lower risk.” This shows that due to customers being older more reports of the customer are available therefore allowing us to have a clearer analyzation and directly supporting my hypothesis which older customers have a lower risk of having a “bad” credit risk. (Fishelson-Holstine, 2003)

Hypothesis

H_0 = The age group of 55 - 75 does not have at least a probability of 60% of getting a good credit risk classification

H_A = The age group of 55 - 75 does have at least a probability of 60% of getting a good credit risk classification

Hypothesis Testing

```
Coefficients:
                          Estimate Std. Error z value Pr(>|z|)
(Intercept)               -0.28984    0.03412  -8.495  < 2e-16 ***
age_groupmiddle_age (35-54)  0.68445    0.05633  12.150  < 2e-16 ***
age_groupolder (55-75)       0.78301    0.10598   7.389 1.48e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8317.8  on 5999  degrees of freedom
Residual deviance: 8140.9  on 5997  degrees of freedom
AIC: 8146.9

Number of Fisher Scoring iterations: 4
```

The positive coefficient is 0.78301. The p value is (1.48e – 13) which is smaller than the significant level of 0.05. This shows very convincing evidence to reject the null hypothesis and accept the alternative hypothesis. The positive coefficient of 0.78301 states that the older age group of 55 - 75 does not have at least a probability of 60% of getting a good credit risk classification increase

by a factor of $e^{0.78301} \approx 2.19$ and in percentage of 119%. In conclusion, I have sufficient evidence to reject the null hypothesis and accept the alternative hypothesis.

4.0 Group Hypothesis

These are the individual hypotheses from the data analysis section:

1. Customers who apply for a loan with a duration of less than or equal to 24 months have a greater than 50% probability of receiving a good credit risk classification compared to customers who apply for a loan with a duration of more than 24 months.
2. There is a significant difference in the distribution of credit risk classification (Good vs. Bad) for the "male single" group.
3. Users whose property magnitude is real estate have higher probability to get good class.
4. Credit amount that is less than 10000 has 50% more probability of receiving a good credit risk.
5. The age group of 55 - 75 does have at least a probability of 60% of getting a good credit risk classification

The combined (complex) hypothesis and null hypothesis are:

Null Hypothesis, H_0: Customers who are single males, aged between 55 and 75, own real estate, and apply for a loan of less than 10000 with a duration of less than or equal to 24 months do not have a significantly higher probability of being classified as a good credit risk compared to customers who do not meet these criteria.

Alternative Hypothesis, H_1: Customers who are single males, aged between 55 and 75, own real estate, and apply for a loan of less than 10000 with a duration of less than or equal to 24 months have higher probability of being classified as good credit risk compared to customers who do not meet these criteria.

After forming the hypotheses, accepting and rejecting criteria should be determined. 0.05 will be used as the significance level (α).

If p-value is less than or equal (≤) to 0.05, reject null hypothesis, H_0.

If p-value is greater than (>) 0.05, fail to reject null hypothesis, H_0.

Data Preparation

```
# Data Preparation

# Duplicate the dataset
pfda_df_Group <- pfda_df
View(pfda_df_Group)
```

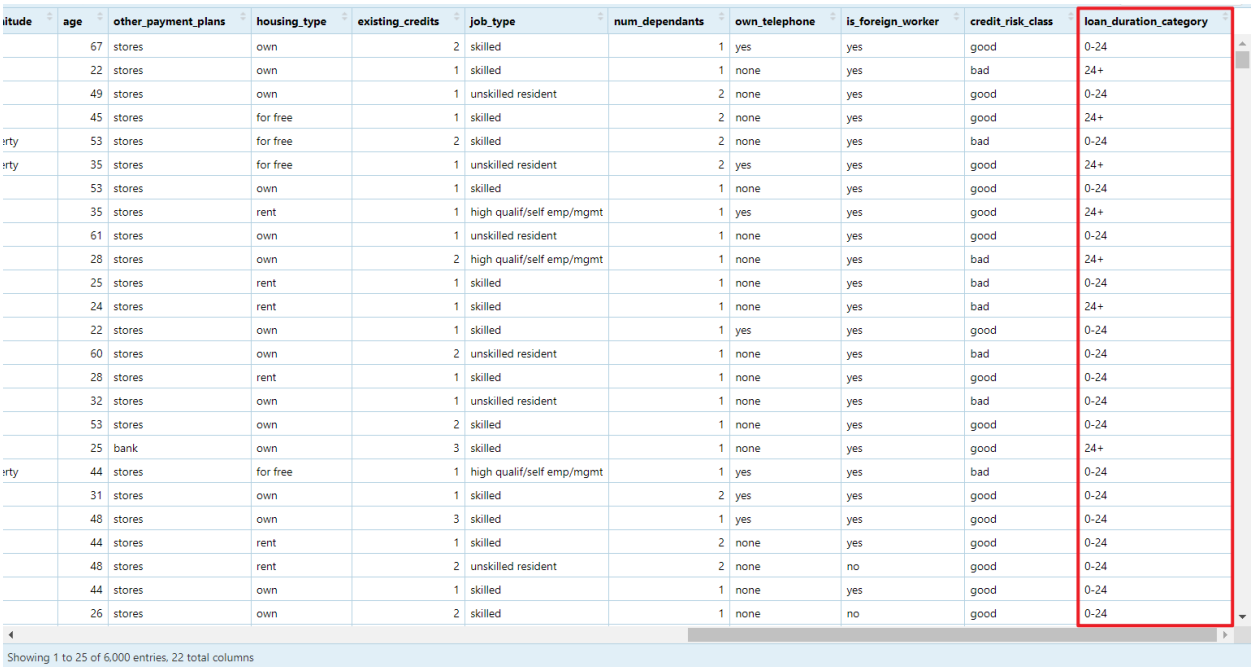
	checking_account_status	loan_duration_months	credit_history_status	loan_purpose	credit_amount	savings_account_status	employment_years	installment_rate_percent	personal_status
1	<0	6	critical/order existing credit	domestic appliance	1169	no known savings	>=7	4	male single
2	0<=X<200	48	existing paid	domestic appliance	5951	<100	1<=X<4	2	female div/dep/
3	no checking	12	critical/order existing credit	education	2096	<100	4<=X<7	2	male single
4	<0	42	existing paid	furniture/equipment	7882	<100	4<=X<7	2	male single
5	<0	24	delayed previously	new car	4870	<100	1<=X<4	3	male single
6	no checking	36	existing paid	education	9055	no known savings	1<=X<4	2	male single
7	no checking	24	existing paid	furniture/equipment	2835	500<=X<10000	>=7	3	male single
8	0<=X<200	36	existing paid	used car	6948	<100	1<=X<4	2	male single
9	no checking	12	existing paid	domestic appliance	3059	>=1000	4<=X<7	2	male div/sep
10	0<=X<200	30	critical/order existing credit	new car	5234	<100	unemployed	4	male mar/wid
11	0<=X<200	12	existing paid	new car	1295	<100	<1	3	female div/dep/
12	<0	48	existing paid	business	4308	<100	<1	3	female div/dep/
13	0<=X<200	12	existing paid	domestic appliance	1567	<100	1<=X<4	1	female div/dep/
14	<0	24	critical/order existing credit	new car	1199	<100	>=7	4	male single
15	<0	15	existing paid	new car	1403	<100	1<=X<4	2	female div/dep/
16	<0	24	existing paid	domestic appliance	1282	100<=X<500	1<=X<4	4	female div/dep/
17	no checking	24	critical/order existing credit	domestic appliance	2424	no known savings	>=7	4	male single
18	<0	30	all paid	business	8072	no known savings	<1	2	male single
19	0<=X<200	24	existing paid	used car	12579	<100	>=7	4	female div/dep/
20	no checking	24	existing paid	domestic appliance	3430	500<=X<10000	>=7	3	male single
21	no checking	9	critical/order existing credit	new car	2134	<100	1<=X<4	4	male single
22	<0	6	existing paid	domestic appliance	2647	500<=X<10000	1<=X<4	2	male single
23	<0	10	critical/order existing credit	new car	2241	<100	<1	1	male single
24	0<=X<200	12	critical/order existing credit	used car	1804	100<=X<500	<1	3	male single
25	no checking	10	critical/order existing credit	furniture/equipment	2069	no known savings	1<=X<4	2	male mar/wid

Showing 1 to 25 of 6,000 entries, 21 total columns

Duplicate the dataset for analysis and not affect the original dataset.

```
# Categorize the loan_duration_months into 0-24 and 24+ months loan duration categories
pfda_df_Group$loan_duration_category <- cut(pfda_df_Group$loan_duration_months, breaks = c(0, 24, Inf), labels = c("0-24", "24+"))
summary(pfda_df_Group$loan_duration_category)
levels(pfda_df_Group$loan_duration_category)
```

```
> # Categorize the loan_duration_months into 0-24 and 24+ months loan duration categories
> pfda_df_Group$loan_duration_category <- cut(pfda_df_Group$loan_duration_months, breaks = c(0, 24, Inf), labels = c("0-24", "24+"))
> summary(pfda_df_Group$loan_duration_category)
0-24  24+
4373 1627
> levels(pfda_df_Group$loan_duration_category)
[1] "0-24" "24+"
```



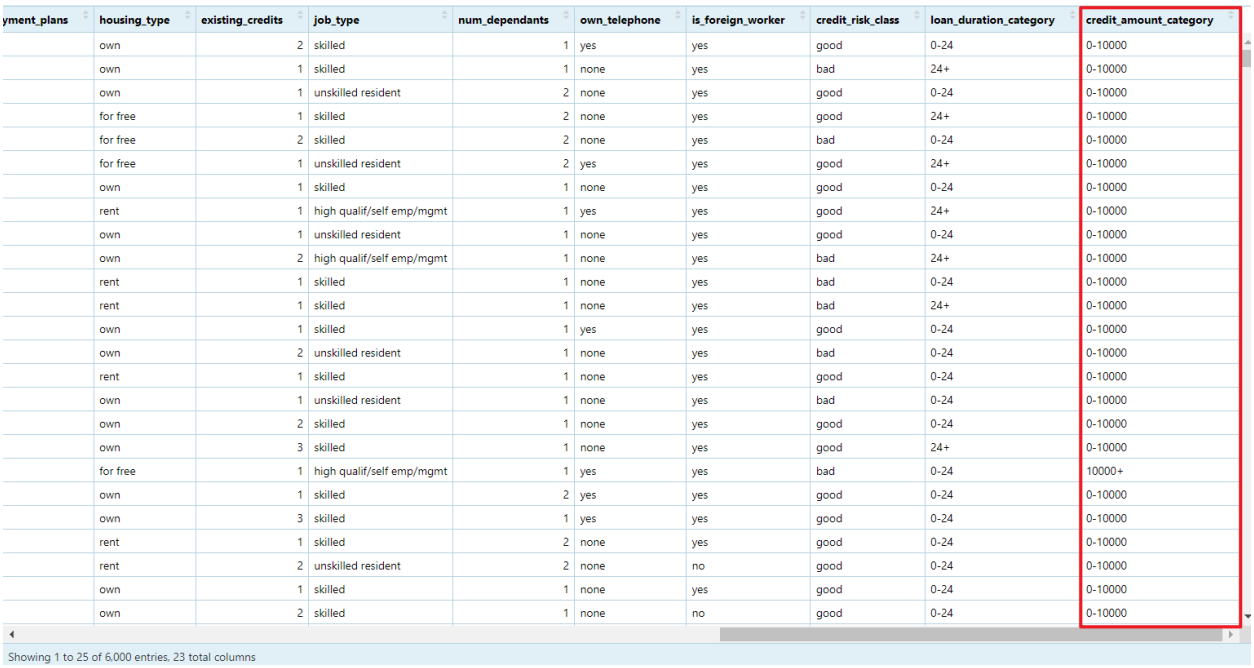
itude	age	other_payment_plans	housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class	loan_duration_category
	67	stores	own	2	skilled	1	yes	yes	good	0-24
	22	stores	own	1	skilled	1	none	yes	bad	24+
	49	stores	own	1	unskilled resident	2	none	yes	good	0-24
	45	stores	for free	1	skilled	2	none	yes	good	24+
erty	53	stores	for free	2	skilled	2	none	yes	bad	0-24
erty	35	stores	for free	1	unskilled resident	2	yes	yes	good	24+
	53	stores	own	1	skilled	1	none	yes	good	0-24
	35	stores	rent	1	high qualif/self emp/mgmt	1	yes	yes	good	24+
	61	stores	own	1	unskilled resident	1	none	yes	good	0-24
	28	stores	own	2	high qualif/self emp/mgmt	1	none	yes	bad	24+
	25	stores	rent	1	skilled	1	none	yes	bad	0-24
	24	stores	rent	1	skilled	1	none	yes	bad	24+
	22	stores	own	1	skilled	1	yes	yes	good	0-24
	60	stores	own	2	unskilled resident	1	none	yes	bad	0-24
	28	stores	rent	1	skilled	1	none	yes	good	0-24
	32	stores	own	1	unskilled resident	1	none	yes	bad	0-24
	53	stores	own	2	skilled	1	none	yes	good	0-24
	25	bank	own	3	skilled	1	none	yes	good	24+
erty	44	stores	for free	1	high qualif/self emp/mgmt	1	yes	yes	bad	0-24
	31	stores	own	1	skilled	2	yes	yes	good	0-24
	48	stores	own	3	skilled	1	yes	yes	good	0-24
	44	stores	rent	1	skilled	2	none	yes	good	0-24
	48	stores	rent	2	unskilled resident	2	none	no	good	0-24
	44	stores	own	1	skilled	1	none	yes	good	0-24
	26	stores	own	2	skilled	1	none	no	good	0-24

Showing 1 to 25 of 6,000 entries, 22 total columns

The loan durations are categorized into “0-24” and “24+” loan duration ranges to specifically analyze the distribution of good and bad credit risk classification cases for loan durations that are less than or equal to 24 months and greater than 24 months.

```
# Categorize the credit_amount into 0-10000 and 10000+ credit amount categories
pfda_df_Group$credit_amount_category <- cut(pfda_df_Group$credit_amount, breaks = c(0, 10000, Inf), labels = c("0-10000", "10000+"))
summary(pfda_df_Group$credit_amount_category)
levels(pfda_df_Group$credit_amount_category)
```

```
> # Categorize the credit_amount into 0-10000 and 10000+ credit amount categories
> pfda_df_Group$credit_amount_category <- cut(pfda_df_Group$credit_amount, breaks = c(0, 10000, Inf), labels = c("0-10000", "10000+"))
> summary(pfda_df_Group$credit_amount_category)
0-10000  10000+
   5691     309
> levels(pfda_df_Group$credit_amount_category)
[1] "0-10000" "10000+"
```



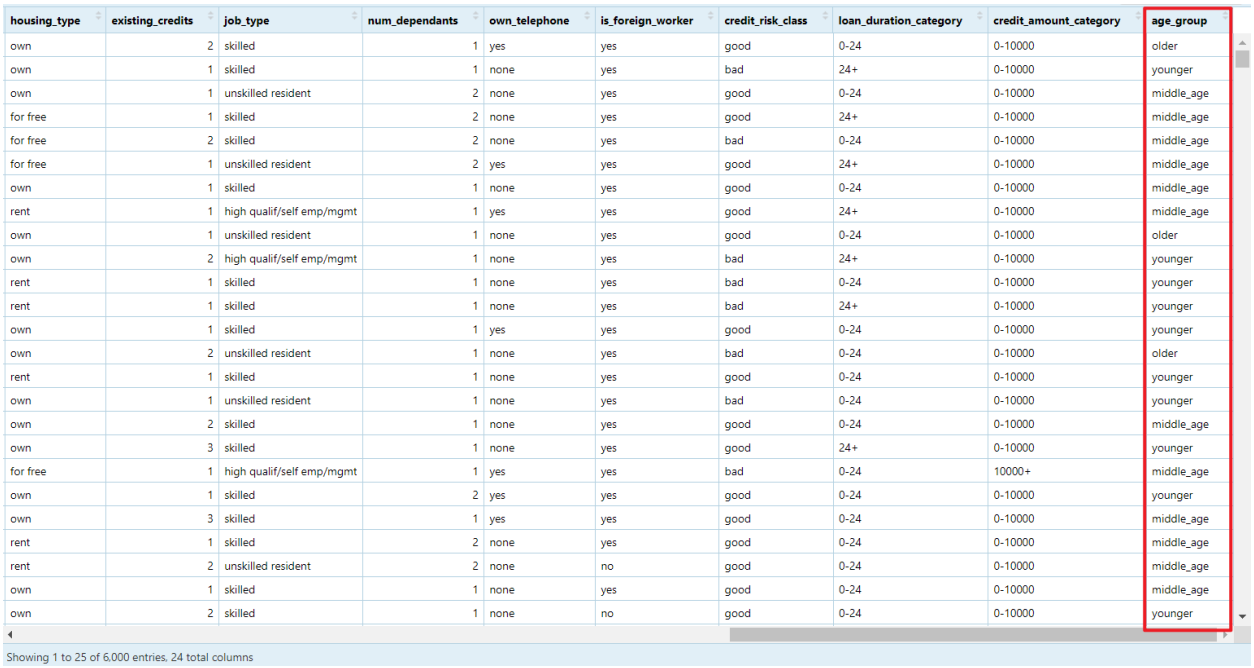
yment_plans	housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class	loan_duration_category	credit_amount_category
	own	2	skilled	1	yes	yes	good	0-24	0-10000
	own	1	skilled	1	none	yes	bad	24+	0-10000
	own	1	unskilled resident	2	none	yes	good	0-24	0-10000
	for free	1	skilled	2	none	yes	good	24+	0-10000
	for free	2	skilled	2	none	yes	bad	0-24	0-10000
	for free	1	unskilled resident	2	yes	yes	good	24+	0-10000
	own	1	skilled	1	none	yes	good	0-24	0-10000
	rent	1	high qualif/self emp/mgmt	1	yes	yes	good	24+	0-10000
	own	1	unskilled resident	1	none	yes	good	0-24	0-10000
	own	2	high qualif/self emp/mgmt	1	none	yes	bad	24+	0-10000
	rent	1	skilled	1	none	yes	bad	0-24	0-10000
	rent	1	skilled	1	none	yes	bad	24+	0-10000
	own	1	skilled	1	yes	yes	good	0-24	0-10000
	own	2	unskilled resident	1	none	yes	bad	0-24	0-10000
	rent	1	skilled	1	none	yes	good	0-24	0-10000
	own	1	unskilled resident	1	none	yes	bad	0-24	0-10000
	own	2	skilled	1	none	yes	good	0-24	0-10000
	own	3	skilled	1	none	yes	good	24+	0-10000
	for free	1	high qualif/self emp/mgmt	1	yes	yes	bad	0-24	10000+
	own	1	skilled	2	yes	yes	good	0-24	0-10000
	own	3	skilled	1	yes	yes	good	0-24	0-10000
	rent	1	skilled	2	none	yes	good	0-24	0-10000
	rent	2	unskilled resident	2	none	no	good	0-24	0-10000
	own	1	skilled	1	none	yes	good	0-24	0-10000
	own	2	skilled	1	none	no	good	0-24	0-10000

Showing 1 to 25 of 6,000 entries, 23 total columns

The credit amounts are categorized into “0-10000” and “10000+” credit amount ranges to specifically analyze the distribution of good and bad credit risk classification cases for credit amounts that are less than or equal to 10000 and greater than 10000.

```
# Categorize age into 3 categories: younger (18-34), middle_age (35-54), and older (55-75)
pfda_df_Group$age_group <- cut(pfda_df_Group$age,
                               breaks = c(18,34,54,75),
                               right = TRUE,
                               labels = c("younger", "middle_age", "older"),
                               include.lowest = TRUE)
summary(pfda_df_Group$age_group)
levels(pfda_df_Group$age_group)
```

```
> # Categorize age into 3 categories: younger (18-34), middle_age (35-54), and older (55-75)
> pfda_df_Group$age_group <- cut(pfda_df_Group$age,
+                                breaks = c(18,34,54,75),
+                                right = TRUE,
+                                labels = c("younger", "middle_age", "older"),
+                                include.lowest = TRUE)
> summary(pfda_df_Group$age_group)
   younger middle_age      older
      3509       2069        422
> levels(pfda_df_Group$age_group)
[1] "younger"    "middle_age" "older"
```



housing_type	existing_credits	job_type	num_dependants	own_telephone	is_foreign_worker	credit_risk_class	loan_duration_category	credit_amount_category	age_group
own	2	skilled	1	yes	yes	good	0-24	0-10000	older
own	1	skilled	1	none	yes	bad	24+	0-10000	younger
own	1	unskilled resident	2	none	yes	good	0-24	0-10000	middle_age
for free	1	skilled	2	none	yes	good	24+	0-10000	middle_age
for free	2	skilled	2	none	yes	bad	0-24	0-10000	middle_age
for free	1	unskilled resident	2	yes	yes	good	24+	0-10000	middle_age
own	1	skilled	1	none	yes	good	0-24	0-10000	middle_age
rent	1	high qualif/self emp/mgmt	1	yes	yes	good	24+	0-10000	middle_age
own	1	unskilled resident	1	none	yes	good	0-24	0-10000	older
own	2	high qualif/self emp/mgmt	1	none	yes	bad	24+	0-10000	younger
rent	1	skilled	1	none	yes	bad	0-24	0-10000	younger
rent	1	skilled	1	none	yes	bad	24+	0-10000	younger
own	1	skilled	1	yes	yes	good	0-24	0-10000	younger
own	2	unskilled resident	1	none	yes	bad	0-24	0-10000	older
rent	1	skilled	1	none	yes	good	0-24	0-10000	younger
own	1	unskilled resident	1	none	yes	bad	0-24	0-10000	younger
own	2	skilled	1	none	yes	good	0-24	0-10000	middle_age
own	3	skilled	1	none	yes	good	24+	0-10000	younger
for free	1	high qualif/self emp/mgmt	1	yes	yes	bad	0-24	10000+	middle_age
own	1	skilled	2	yes	yes	good	0-24	0-10000	younger
own	3	skilled	1	yes	yes	good	0-24	0-10000	middle_age
rent	1	skilled	2	none	yes	good	0-24	0-10000	middle_age
rent	2	unskilled resident	2	none	no	good	0-24	0-10000	middle_age
own	1	skilled	1	none	yes	good	0-24	0-10000	middle_age
own	2	skilled	1	none	no	good	0-24	0-10000	younger

Showing 1 to 25 of 6,000 entries, 24 total columns

The age values are categorized into “younger”, “middle_age”, and “older” age groups to specifically analyze the distribution of good and bad credit risk classification cases for different age ranges. The age range between 18 and 34 is classified as “younger”, the age range between 35 and 54 is classified as “middle_age”, and the age range between 55 and 75 is classified as “older”.

Hypothesis Testing

```
# Hypothesis Testing

# Checking the levels of each variable
levels(pfda_df_Group$loan_duration_category)
levels(pfda_df_Group$personal_status)
levels(pfda_df_Group$property_magnitude)
levels(pfda_df_Group$credit_amount_category)
levels(pfda_df_Group$age_group)
levels(pfda_df_Group$credit_risk_class)
```

```
> # Checking the levels of each variable
> levels(pfda_df_Group$loan_duration_category)
[1] "0-24" "24+"
> levels(pfda_df_Group$personal_status)
[1] "female div/dep/mar" "male div/sep"       "male mar/wid"       "male single"
> levels(pfda_df_Group$property_magnitude)
[1] "car"               "life insurance"    "no known property" "real estate"
> levels(pfda_df_Group$credit_amount_category)
[1] "0-10000" "10000+"
> levels(pfda_df_Group$age_group)
[1] "younger"    "middle_age" "older"
> levels(pfda_df_Group$credit_risk_class)
[1] "bad"  "good"
```

Based on the level of each variable, the reference group of loan_duration_category is “0-24” months loan duration, personal_status is “female div/dep/mar”, property_magnitude is “car”, credit_amount_category is “0-10000”, age_group is “younger”, and credit_risk_class is “bad”.

Logistic Regression

```
# Logistic Regression
logistic_model_Group <- glm(
  credit_risk_class ~ loan_duration_category * personal_status * property_magnitude * credit_amount_category * age_group,
  family = binomial(link = "logit"),
  data = pfda_df_Group
)
summary(logistic_model_Group)

# Save the logistic regression results into a .txt file
sink("logistic_model_Group.txt")
print(summary(logistic_model_Group))
sink()
```

(Notepad++ is used to view the full results of the logistic regression model which are saved in a .txt file for better visualization. Specific results will be highlighted at the bottom.)

```
Call:
glm(formula = credit_risk_class ~ loan_duration_category * personal_status *
    property_magnitude * credit_amount_category * age_group,
    family = binomial(link = "logit"), data = pfda_df_Group)

Coefficients: (81 not defined because of singularities)
```

	Estimate	Std. Error	z value
(Intercept)	-1.585e+15	3.465e+06	-457444119
loan_duration_category24+	-1.194e+15	6.629e+06	-180036244
personal_statusmale div/sep	-2.162e+14	7.410e+06	-29175051
personal_statusmale mar/wid	2.422e+15	6.051e+06	400237930
personal_statusmale single	2.189e+15	5.103e+06	428884940
property_magnitudelife insurance	1.448e+15	4.948e+06	292637626
property_magnitudeno known property	-1.469e+15	7.153e+06	-205405218
property_magnitudereal estate	3.670e+15	5.533e+06	663157898
credit_amount_category10000+	-2.918e+15	3.373e+07	-86512505
age_groupmiddle_age	5.255e+15	8.223e+06	639083280
age_groupolder	6.089e+15	1.171e+07	519999025
loan_duration_category24+:personal_statusmale div/sep	-1.509e+15	1.352e+07	-111622623
loan_duration_category24+:personal_statusmale mar/wid	-1.445e+15	1.346e+07	-107333691
loan_duration_category24+:personal_statusmale single	-3.682e+13	9.300e+06	-3959227
loan_duration_category24+:property_magnitudelife insurance	1.740e+15	1.117e+07	155858983
loan_duration_category24+:property_magnitudeno known property	1.453e+15	1.544e+07	94091985
loan_duration_category24+:property_magnitudereal estate	-1.963e+15	1.303e+07	-150652526
personal_statusmale div/sep:property_magnitudelife insurance	-2.185e+15	1.222e+07	-178845496
personal_statusmale mar/wid:property_magnitudelife insurance	-3.164e+15	1.020e+07	-310184143
personal_statusmale single:property_magnitudelife insurance	-1.437e+15	8.006e+06	-179504626
personal_statusmale div/sep:property_magnitudeno known property	3.271e+15	1.855e+07	176281705
personal_statusmale mar/wid:property_magnitudeno known property	6.326e+14	1.362e+07	46432255
personal_statusmale single:property_magnitudeno known property	-6.351e+14	1.034e+07	-61429451
personal_statusmale div/sep:property_magnitudereal estate	-1.085e+15	1.641e+07	-66108442
personal_statusmale mar/wid:property_magnitudereal estate	-4.427e+15	9.733e+06	-454867355
personal_statusmale single:property_magnitudereal estate	-3.483e+15	8.025e+06	-433983808
loan_duration_category24+:credit_amount_category10000+	1.194e+15	3.633e+07	32851724
personal_statusmale div/sep:credit_amount_category10000+	1.725e+15	3.207e+07	53782413
personal_statusmale mar/wid:credit_amount_category10000+	2.696e+15	1.423e+08	18940107
personal_statusmale single:credit_amount_category10000+	-2.189e+15	4.237e+07	-51651605
property_magnitudelife insurance:credit_amount_category10000+	-2.142e+15	6.275e+07	-34135657
property_magnitudeno known property:credit_amount_category10000+	-7.449e+14	1.043e+08	-7145661
property_magnitudereal estate:credit_amount_category10000+	2.797e+15	3.450e+07	81082423
loan_duration_category24+:age_groupmiddle_age	-3.704e+15	1.744e+07	-212365948
loan_duration_category24+:age_groupolder	8.787e+15	4.165e+07	210957469
personal_statusmale div/sep:age_groupmiddle_age	-4.955e+15	1.727e+07	-286941140
personal_statusmale mar/wid:age_groupmiddle_age	-6.357e+15	1.259e+07	-504980683
personal_statusmale single:age_groupmiddle_age	-5.637e+15	1.000e+07	-563393593
personal_statusmale div/sep:age_groupolder	-8.791e+15	3.614e+07	-243267710
personal_statusmale mar/wid:age_groupolder	-6.926e+15	2.692e+07	-257269089
personal_statusmale single:age_groupolder	-6.692e+15	2.080e+07	-321747604
property_magnitudelife insurance:age_groupmiddle_age	-3.392e+15	1.105e+07	-306873874
property_magnitudeno known property:age_groupmiddle_age	-9.135e+14	1.636e+07	-55841557
property_magnitudereal estate:age_groupmiddle_age	-7.852e+15	1.197e+07	-656116970
property_magnitudelife insurance:age_groupolder	-3.950e+15	1.999e+07	-197555008
property_magnitudeno known property:age_groupolder	-4.432e+15	1.821e+07	-243391662
property_magnitudereal estate:age_groupolder	-3.670e+15	1.656e+07	-221592165
credit_amount_category10000+:age_groupmiddle_age	-1.271e+15	1.196e+08	-10624052
credit_amount_category10000+:age_groupolder	-5.108e+15	9.730e+07	-52501719
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance	3.500e+15	2.308e+07	151630312
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance	-1.542e+15	2.086e+07	-73894069
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance	-2.283e+15	1.541e+07	-148199138
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property	1.249e+15	2.920e+07	42788847
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property	1.185e+15	3.238e+07	36608418
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property	1.202e+15	1.997e+07	60197716
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate	5.520e+15	3.023e+07	182600771
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate	1.861e+13	5.097e+07	365104
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate	4.732e+14	2.113e+07	22393158
loan_duration_category24+:personal_statusmale div/sep:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+	1.626e+15	4.624e+07	35169199
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+	-1.046e+15	5.850e+07	-17887778
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+	-2.443e+15	4.783e+07	-51085259
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+	-1.315e+15	4.796e+07	-27427388
personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+	2.005e+15	7.127e+07	28139153
personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+	6.635e+15	3.166e+07	209569290
personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+	-1.315e+15	1.038e+08	-12669876
personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+	1.434e+15	1.632e+08	8789658
personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+	2.849e+15	9.510e+07	29963544
personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:age_groupmiddle_age	7.264e+15	2.813e+07	258212704
loan_duration_category24+:personal_statusmale mar/wid:age_groupmiddle_age	2.104e+15	2.753e+07	76418969
loan_duration_category24+:personal_statusmale single:age_groupmiddle_age	6.318e+15	2.024e+07	312105759
loan_duration_category24+:personal_statusmale div/sep:age_groupolder	-4.494e+15	6.307e+07	-71247294
loan_duration_category24+:personal_statusmale mar/wid:age_groupolder	NA	NA	NA

loan_duration_category24+:personal_statusmale mar/wid:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:age_groupolder	-7.557e+15	3.333e+07	-226737131
loan_duration_category24+:property_magnitudelife insurance:age_groupmiddle_age	-4.981e+14	3.272e+07	-15225360
loan_duration_category24+:property_magnitudeno known property:age_groupmiddle_age	-2.345e+15	3.270e+07	-71732642
loan_duration_category24+:property_magnitudereal estate:age_groupmiddle_age	6.731e+15	2.844e+07	236665803
loan_duration_category24+:property_magnitudelife insurance:age_groupolder	-1.134e+16	6.571e+07	-172507890
loan_duration_category24+:property_magnitudeno known property:age_groupolder	-7.649e+15	3.124e+07	-244848908
loan_duration_category24+:property_magnitudereal estate:age_groupolder	-2.519e+15	4.928e+07	-51111662
personal_statusmale div/sep:property_magnitudelife insurance:age_groupmiddle_age	4.729e+15	2.569e+07	184122077
personal_statusmale mar/wid:property_magnitudelife insurance:age_groupmiddle_age	4.086e+15	2.428e+07	168256289
personal_statusmale single:property_magnitudelife insurance:age_groupmiddle_age	3.916e+15	1.402e+07	279204543
personal_statusmale div/sep:property_magnitudeno known property:age_groupmiddle_age	-8.879e+14	3.103e+07	-28615962
personal_statusmale mar/wid:property_magnitudeno known property:age_groupmiddle_age	1.905e+15	2.404e+07	79241050
personal_statusmale single:property_magnitudeno known property:age_groupmiddle_age	5.258e+15	1.928e+07	272703143
personal_statusmale div/sep:property_magnitudereal estate:age_groupmiddle_age	6.936e+15	2.714e+07	255583056
personal_statusmale mar/wid:property_magnitudereal estate:age_groupmiddle_age	1.169e+16	1.855e+07	630317637
personal_statusmale single:property_magnitudereal estate:age_groupmiddle_age	9.626e+15	1.467e+07	656040448
personal_statusmale div/sep:property_magnitudelife insurance:age_groupolder	9.191e+15	4.501e+07	204197034
personal_statusmale mar/wid:property_magnitudelife insurance:age_groupolder	5.666e+15	4.656e+07	121672864
personal_statusmale single:property_magnitudelife insurance:age_groupolder	5.741e+15	3.195e+07	179647844
personal_statusmale div/sep:property_magnitudeno known property:age_groupolder	NA	NA	NA
personal_statusmale mar/wid:property_magnitudeno known property:age_groupolder	NA	NA	NA
personal_statusmale single:property_magnitudeno known property:age_groupolder	8.479e+15	2.752e+07	308163408
personal_statusmale div/sep:property_magnitudereal estate:age_groupolder	1.009e+16	4.363e+07	231300300
personal_statusmale mar/wid:property_magnitudereal estate:age_groupolder	NA	NA	NA
personal_statusmale single:property_magnitudereal estate:age_groupolder	2.988e+15	2.567e+07	116415959
loan_duration_category24+:credit_amount_category10000+:age_groupmiddle_age	3.000e+14	4.714e+07	6364666
loan_duration_category24+:credit_amount_category10000+:age_groupolder	5.518e+15	4.877e+07	113143760
personal_statusmale div/sep:credit_amount_category10000+:age_groupmiddle_age	1.614e+15	1.446e+08	11161669
personal_statusmale mar/wid:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale single:credit_amount_category10000+:age_groupmiddle_age	1.653e+15	1.320e+08	12519211
personal_statusmale div/sep:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale mar/wid:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale single:credit_amount_category10000+:age_groupolder	1.664e+15	8.809e+07	18893004
property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	-1.519e+15	4.783e+07	-31768702
property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	3.648e+15	2.959e+07	123259926
property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	5.422e+14	1.478e+08	3668484
property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:age_groupmiddle_age	-7.402e+15	4.772e+07	-155095247
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:age_groupmiddle_age	4.024e+15	4.969e+07	80967418
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:age_groupmiddle_age	-1.420e+15	3.609e+07	-39348665
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:age_groupmiddle_age	2.865e+14	6.121e+07	4680705
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:age_groupmiddle_age	4.056e+15	5.070e+07	79998526
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:age_groupmiddle_age	-5.957e+15	3.713e+07	-160422635
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:age_groupmiddle_age	-1.582e+16	5.651e+07	-279845672
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:age_groupmiddle_age	-8.807e+15	3.592e+07	-245190024
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:age_groupolder	5.573e+15	7.460e+07	74711951
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	-4.291e+15	7.924e+07	-54149834
personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA

personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupolder	NA	NA	NA

```
                                                                                          Pr(>|z|)
(Intercept)                                                                                 <2e-16 ***
loan_duration_category24+                                                                   <2e-16 ***
personal_statusmale div/sep                                                                 <2e-16 ***
personal_statusmale mar/wid                                                                 <2e-16 ***
personal_statusmale single                                                                  <2e-16 ***
property_magnitudelife insurance                                                            <2e-16 ***
property_magnitudeno known property                                                         <2e-16 ***
property_magnitudereal estate                                                               <2e-16 ***
credit_amount_category10000+                                                                <2e-16 ***
age_groupmiddle_age                                                                         <2e-16 ***
age_groupolder                                                                              <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep                                       <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid                                       <2e-16 ***
loan_duration_category24+:personal_statusmale single                                        <2e-16 ***
loan_duration_category24+:property_magnitudelife insurance                                  <2e-16 ***
loan_duration_category24+:property_magnitudeno known property                               <2e-16 ***
loan_duration_category24+:property_magnitudereal estate                                     <2e-16 ***
personal_statusmale div/sep:property_magnitudelife insurance                                <2e-16 ***
personal_statusmale mar/wid:property_magnitudelife insurance                                <2e-16 ***
personal_statusmale single:property_magnitudelife insurance                                 <2e-16 ***
personal_statusmale div/sep:property_magnitudeno known property                             <2e-16 ***
personal_statusmale mar/wid:property_magnitudeno known property                             <2e-16 ***
personal_statusmale single:property_magnitudeno known property                              <2e-16 ***
personal_statusmale div/sep:property_magnitudereal estate                                   <2e-16 ***
personal_statusmale mar/wid:property_magnitudereal estate                                   <2e-16 ***
personal_statusmale single:property_magnitudereal estate                                    <2e-16 ***
loan_duration_category24+:credit_amount_category10000+                                      <2e-16 ***
personal_statusmale div/sep:credit_amount_category10000+                                    <2e-16 ***
personal_statusmale mar/wid:credit_amount_category10000+                                    <2e-16 ***
personal_statusmale single:credit_amount_category10000+                                     <2e-16 ***
property_magnitudelife insurance:credit_amount_category10000+                               <2e-16 ***
property_magnitudeno known property:credit_amount_category10000+                            <2e-16 ***
property_magnitudereal estate:credit_amount_category10000+                                  <2e-16 ***
loan_duration_category24+:age_groupmiddle_age                                               <2e-16 ***
loan_duration_category24+:age_groupolder                                                    <2e-16 ***
personal_statusmale div/sep:age_groupmiddle_age                                             <2e-16 ***
personal_statusmale mar/wid:age_groupmiddle_age                                             <2e-16 ***
personal_statusmale single:age_groupmiddle_age                                              <2e-16 ***
personal_statusmale div/sep:age_groupolder                                                  <2e-16 ***
personal_statusmale mar/wid:age_groupolder                                                  <2e-16 ***
```

```
personal_statusmale single:age_groupolder                                                   <2e-16 ***
property_magnitudelife insurance:age_groupmiddle_age                                        <2e-16 ***
property_magnitudeno known property:age_groupmiddle_age                                     <2e-16 ***
property_magnitudereal estate:age_groupmiddle_age                                           <2e-16 ***
property_magnitudelife insurance:age_groupolder                                             <2e-16 ***
property_magnitudeno known property:age_groupolder                                          <2e-16 ***
property_magnitudereal estate:age_groupolder                                                <2e-16 ***
credit_amount_category10000+:age_groupmiddle_age                                            <2e-16 ***
credit_amount_category10000+:age_groupolder                                                 <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance      <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance      <2e-16 ***
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance       <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property   <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property   <2e-16 ***
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property    <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate         <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate         <2e-16 ***
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate          <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:credit_amount_category10000+              NA
loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+              NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+           <2e-16 ***
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+     <2e-16 ***
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+  <2e-16 ***
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+            NA
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+   <2e-16 ***
personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+   <2e-16 ***
personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+    <2e-16 ***
personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+ <2e-16 ***
personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+ <2e-16 ***
personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+ <2e-16 ***
personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+          NA
personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+          NA
personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+           NA
loan_duration_category24+:personal_statusmale div/sep:age_groupmiddle_age                  <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:age_groupmiddle_age                  <2e-16 ***
loan_duration_category24+:personal_statusmale single:age_groupmiddle_age                   <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:age_groupolder                       <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:age_groupolder                           NA
loan_duration_category24+:personal_statusmale single:age_groupolder                        <2e-16 ***
loan_duration_category24+:property_magnitudelife insurance:age_groupmiddle_age             <2e-16 ***
loan_duration_category24+:property_magnitudeno known property:age_groupmiddle_age          <2e-16 ***
loan_duration_category24+:property_magnitudereal estate:age_groupmiddle_age                <2e-16 ***
```

```
loan_duration_category24+:property_magnitudelife insurance:age_groupolder                                                                  <2e-16 ***
loan_duration_category24+:property_magnitudeno known property:age_groupolder                                                               <2e-16 ***
loan_duration_category24+:property_magnitudereal estate:age_groupolder                                                                      <2e-16 ***
personal_statusmale div/sep:property_magnitudelife insurance:age_groupmiddle_age                                                            <2e-16 ***
personal_statusmale mar/wid:property_magnitudelife insurance:age_groupmiddle_age                                                            <2e-16 ***
personal_statusmale single:property_magnitudelife insurance:age_groupmiddle_age                                                             <2e-16 ***
personal_statusmale div/sep:property_magnitudeno known property:age_groupmiddle_age                                                         <2e-16 ***
personal_statusmale mar/wid:property_magnitudeno known property:age_groupmiddle_age                                                         <2e-16 ***
personal_statusmale single:property_magnitudeno known property:age_groupmiddle_age                                                          <2e-16 ***
personal_statusmale div/sep:property_magnitudereal estate:age_groupmiddle_age                                                               <2e-16 ***
personal_statusmale mar/wid:property_magnitudereal estate:age_groupmiddle_age                                                               <2e-16 ***
personal_statusmale single:property_magnitudereal estate:age_groupmiddle_age                                                                <2e-16 ***
personal_statusmale div/sep:property_magnitudelife insurance:age_groupolder                                                                 <2e-16 ***
personal_statusmale mar/wid:property_magnitudelife insurance:age_groupolder                                                                 <2e-16 ***
personal_statusmale single:property_magnitudelife insurance:age_groupolder                                                                  <2e-16 ***
personal_statusmale div/sep:property_magnitudeno known property:age_groupolder                                                                 NA
personal_statusmale mar/wid:property_magnitudeno known property:age_groupolder                                                                 NA
personal_statusmale single:property_magnitudeno known property:age_groupolder                                                              <2e-16 ***
personal_statusmale div/sep:property_magnitudereal estate:age_groupolder                                                                    <2e-16 ***
personal_statusmale mar/wid:property_magnitudereal estate:age_groupolder                                                                        NA
personal_statusmale single:property_magnitudereal estate:age_groupolder                                                                     <2e-16 ***
loan_duration_category24+:credit_amount_category10000+:age_groupmiddle_age                                                                  <2e-16 ***
loan_duration_category24+:credit_amount_category10000+:age_groupolder                                                                       <2e-16 ***
personal_statusmale div/sep:credit_amount_category10000+:age_groupmiddle_age                                                                <2e-16 ***
personal_statusmale mar/wid:credit_amount_category10000+:age_groupmiddle_age                                                                    NA
personal_statusmale single:credit_amount_category10000+:age_groupmiddle_age                                                                 <2e-16 ***
personal_statusmale div/sep:credit_amount_category10000+:age_groupolder                                                                         NA
personal_statusmale mar/wid:credit_amount_category10000+:age_groupolder                                                                         NA
personal_statusmale single:credit_amount_category10000+:age_groupolder                                                                      <2e-16 ***
property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age                                                           <2e-16 ***
property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age                                                        <2e-16 ***
property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age                                                              <2e-16 ***
property_magnitudelife insurance:credit_amount_category10000+:age_groupolder                                                                    NA
property_magnitudeno known property:credit_amount_category10000+:age_groupolder                                                                 NA
property_magnitudereal estate:credit_amount_category10000+:age_groupolder                                                                       NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+                            NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+                            NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+                             NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+                         NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+                         NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+                          NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+                            NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+                            NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+                             NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:age_groupmiddle_age                                 <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:age_groupmiddle_age                                 <2e-16 ***
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:age_groupmiddle_age                                  <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:age_groupmiddle_age                              <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:age_groupmiddle_age                              <2e-16 ***
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:age_groupmiddle_age                               <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:age_groupmiddle_age                                    <2e-16 ***
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:age_groupmiddle_age                                         NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:age_groupmiddle_age                                      <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:age_groupolder                                          NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:age_groupolder                                          NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:age_groupolder                                        <2e-16 ***
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:age_groupolder                                        NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:age_groupolder                                        NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:age_groupolder                                         NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:age_groupolder                                              NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:age_groupolder                                              NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:age_groupolder                                              NA
loan_duration_category24+:personal_statusmale div/sep:credit_amount_category10000+:age_groupmiddle_age                                         NA
loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+:age_groupmiddle_age                                         NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+:age_groupmiddle_age                                          NA
loan_duration_category24+:personal_statusmale div/sep:credit_amount_category10000+:age_groupolder                                              NA
loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+:age_groupolder                                              NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+:age_groupolder                                               NA
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age                                    NA
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age                                 NA
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age                                       NA
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder                                         NA
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+:age_groupolder                                      NA
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+:age_groupolder                                            NA
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age                                  NA
personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age                                  NA
personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age                                   NA
personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age                            <2e-16 ***
personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age                               NA
personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age                                NA
personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age                                     NA
personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age                                     NA
personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age                                      NA
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder                                       NA
```

```
personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder                                  NA
personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder                                  NA
personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder                                   NA
personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupolder                               NA
personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupolder                               NA
personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupolder                                NA
personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupolder                                     NA
personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupolder                                     NA
personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupolder                                      NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age  NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age  NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupmiddle_age   NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupmiddle_age  NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age       NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age       NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupmiddle_age        NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder        NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder        NA
loan_duration_category24+:personal_statusmale single:property_magnitudelife insurance:credit_amount_category10000+:age_groupolder         NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudeno known property:credit_amount_category10000+:age_groupolder     NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudeno known property:credit_amount_category10000+:age_groupolder     NA
loan_duration_category24+:personal_statusmale single:property_magnitudeno known property:credit_amount_category10000+:age_groupolder      NA
loan_duration_category24+:personal_statusmale div/sep:property_magnitudereal estate:credit_amount_category10000+:age_groupolder           NA
loan_duration_category24+:personal_statusmale mar/wid:property_magnitudereal estate:credit_amount_category10000+:age_groupolder           NA
loan_duration_category24+:personal_statusmale single:property_magnitudereal estate:credit_amount_category10000+:age_groupolder            NA
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance:   8317.8  on 5999  degrees of freedom
Residual deviance: 148571.9  on 5889  degrees of freedom
AIC: 148794

Number of Fisher Scoring iterations: 25
```

A logistic regression test is used to test the complex hypothesis. Based on the level of each variable, we should focus on the results that **specify** “personal_statusmale single”, “property_magnitudereal_estate”, and “age_groupolder”. In addition, we should also focus on results that **do not specify** “loan_duration_category24+” and “credit_amount_category10000+”. This is because both of the categories are not aligned with the criteria of the complex hypothesis. Besides that, “loan_duration_category0-24” and “credit_amount_category0-10000” are the **reference group** of their variable, therefore they would not be appearing in the results line.

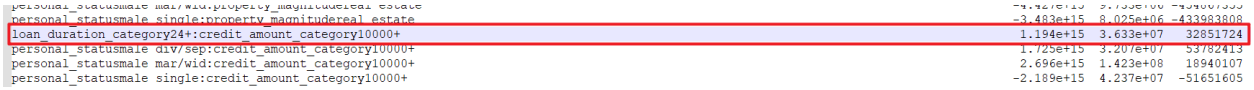
Specified Result

```
personal_statusmale mar/wid:property_magnitudeno known property:age_groupolder                 NA       NA       NA
personal_statusmale single:property_magnitudeno known property:age_groupolder          8.479e+15 2.752e+07 308163408
personal_statusmale div/sep:property_magnitudereal estate:age_groupolder                1.009e+16 4.363e+07 231300300
personal_statusmale mar/wid:property_magnitudereal estate:age_groupolder                    NA       NA       NA
personal_statusmale single:property_magnitudereal estate:age_groupolder                 2.988e+15 2.567e+07 116415959
loan_duration_category24+:credit_amount_category10000+:age_groupmiddle_age              3.000e+14 4.714e+07   6364666
loan_duration_category24+:credit_amount_category10000+:age_groupolder                   5.518e+15 4.877e+07 113143760
personal_statusmale div/sep:credit_amount_category10000+:age_groupmiddle_age            1.614e+15 1.446e+08  11161669
personal_statusmale mar/wid:credit_amount_category10000+:age_groupmiddle_age                NA       NA       NA
```

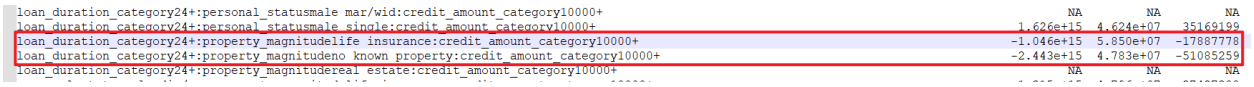
```
personal_statusmale mar/wid:property_magnitudelife insurance:age_groupolder                <2e-16 ***
personal_statusmale single:property_magnitudelife insurance:age_groupolder                 <2e-16 ***
personal_statusmale div/sep:property_magnitudeno known property:age_groupolder                NA
personal_statusmale mar/wid:property_magnitudeno known property:age_groupolder                NA
personal_statusmale single:property_magnitudeno known property:age_groupolder             <2e-16 ***
personal_statusmale div/sep:property_magnitudereal estate:age_groupolder                   <2e-16 ***
personal_statusmale mar/wid:property_magnitudereal estate:age_groupolder                      NA
personal_statusmale single:property_magnitudereal estate:age_groupolder                    <2e-16 ***
loan_duration_category24+:credit_amount_category10000+:age_groupmiddle_age                 <2e-16 ***
loan_duration_category24+:credit_amount_category10000+:age_groupolder                      <2e-16 ***
personal_statusmale div/sep:credit_amount_category10000+:age_groupmiddle_age               <2e-16 ***
personal_statusmale mar/wid:credit_amount_category10000+:age_groupmiddle_age                  NA
personal_statusmale single:credit_amount_category10000+:age_groupmiddle_age                <2e-16 ***
personal_statusmale div/sep:credit_amount_category10000+:age_groupolder                       NA
personal_statusmale mar/wid:credit_amount_category10000+:age_groupolder                       NA
personal_statusmale single:credit_amount_category10000+:age_groupolder                     <2e-16 ***
```

The large positive coefficient value, **2.988e+15**, indicates that the **probability** of getting a **good credit risk increases significantly** when the loan duration is less than or equal to 24 months, the customer's personal status is a single male, the customer owns real estate, the credit amount is less than or equal to 10000, the age of the customer is between 55 and 75. The p-value **(< 2e-16)** is much smaller than the standard significance level of 0.05. This provides very strong evidence to reject the null hypothesis and proves that the relationship between the variables is statistically significant.

Comparing Other Results



personal_statusmale single:property_magnitudereal estate	-3.483e+15	8.025e+06	-433983808
loan_duration_category24+:credit_amount_category10000+	1.194e+15	3.633e+07	32851724
personal_statusmale div/sep:credit_amount_category10000+	1.725e+15	3.207e+07	53782413
personal_statusmale mar/wid:credit_amount_category10000+	2.696e+15	1.423e+08	18940107
personal_statusmale single:credit_amount_category10000+	-2.189e+15	4.237e+07	-51651605



loan_duration_category24+:personal_statusmale mar/wid:credit_amount_category10000+	NA	NA	NA
loan_duration_category24+:personal_statusmale single:credit_amount_category10000+	1.626e+15	4.624e+07	35169199
loan_duration_category24+:property_magnitudelife insurance:credit_amount_category10000+	-1.046e+15	5.850e+07	-17887778
loan_duration_category24+:property_magnitudeno known property:credit_amount_category10000+	-2.443e+15	4.783e+07	-51085259
loan_duration_category24+:property_magnitudereal estate:credit_amount_category10000+	NA	NA	NA

Here are some identified results that did not meet the criteria of the complex hypothesis:

1st Result:

loan_duration_category: 24+

personal_status: female div/dep/mar

property_magnitude: car

credit_amount_category: 10000+

age_group: younger

Coefficient: **1.194e+15**

2nd Result:

loan_duration_category: 24+

personal_status: female div/dep/mar

property_magnitude: life insurance

credit_amount_category: 10000+

age_group: younger

Coefficient: **-1.046e+15**

3rd Result:

loan_duration_category: 24+

personal_status: female div/dep/mar

property_magnitude: no known property

credit_amount_category: 10000+

age_group: younger

Coefficient: **-2.443e+15**

Based on the other results, the identified coefficients are **1.194e+15, -1.046e+15,** and **-2.443e+15,** they are smaller than the coefficient value of the result that satisfy all the criteria of the complex hypothesis, which is **2.988e+15**. The two negative coefficient values indicate that the probability of getting good credit risk classification decreases when the loan duration is above 24 months, the personal status of the customer is not a single male, the customer does not own a real estate, the credit amount is more than 10000, and the customer is younger than 55 years old.

Conclusion

Based on the logistic regression test findings and extremely small p-value, we have **strong evidence** to support our complex hypothesis. Therefore, we **reject** the **null hypothesis (H_0)** and **accept** the **alternative hypothesis (H_1)**.

5.0 Overall Conclusion

To recap the objectives of our group:

1. To determine whether the loan duration impacts the credit risk classification of a customer.
2. To investigate whether personal status of an individual will impact the credit risk classification of a customer.
3. To investigate whether property magnitude impacts the credit risk classification of a customer.
4. To determine whether the credit amount has effect on the individual's credit risk class.
5. To determine how age impacts the credit risk classification of a customer.

Based on the data analysis of each member who was working on each objective, the findings and results show that all the independent variables play a significant part in impacting the credit risk classification of a customer. For loan duration variable, loan durations that are less than or equal to 24 months have higher probability to get a good credit risk than the loan durations that are greater than 24 months. For personal status variable, a customer who is a single male has a higher probability of being classified as good credit risk classification. For property magnitude variable, a customer who owns real estate has a higher likelihood of being classified as good credit risk classification. For credit amount variable, a customer who applies for a loan with the credit amount of less than or equal to 10000 has higher chances of getting a good credit risk. For age variable, a customer who aged between 55 and 75 has higher chances of getting a good credit risk.

Recommendation

The data analysis can be performed on a larger dataset for more sample size, leading to more accurate and reliable findings and insights. Besides that, the exploratory data analysis (EDA) and literature review can be done in a more detailed manner to form a firm hypothesis. Furthermore, more advanced methods like machine learning techniques can be used in data cleaning and preprocessing, data analysis, and hypothesis testing to produce stronger evidence against the null hypothesis.

Limitation and Future Direction

The scope of the dataset and the sample size are limited, and it may not thoroughly represent or reflect the diversity of factors impacting the credit risk classification. Next, there are a lot of independent variables, also known as predictors, incorporating complex interactions and associations, causing the interpretation and visualization to be more challenging. Not only that, these limitations may lead to bias in the dataset. Therefore, expanding the dataset with a wide variety of data and ensuring that the data is evenly distributed to prevent potential bias. Moreover, test advanced models like deep learning techniques should be utilized to examine the complex relationship between variables and produce better interpretations of the dataset. In addition, prediction analysis and prescriptive analysis can be conducted to generate more valuable insights and contribute to decision making.

Word Count

9756 words.

6.0 Workload Matrix

Workload Matrix					
Components / Members	Daryl Sim Wei Shern TP068964	Ho Shane Foong TP068496	Cheong Sheue Ling TP069004	Choo Cheng Da TP068973	John Har Wey Jon TP068348
Introduction (Group)	20%	20%	20%	20%	20%
Data Preparation (Group)	20%	20%	20%	20%	20%
Data Analysis (Individual)	100%	100%	100%	100%	100%
Group Hypothesis (Group)	40%	15%	15%	15%	15%
Overall Conclusion (Group)	20%	20%	20%	20%	20%

7.0 References

Jiménez, G., & Saurina, J. (2002). *Loan Characteristics and Credit Risk*. https://www.bis.org/bcbs/events/wkshop0303/p03jimesaur.pdf

Calem, P., Ramasamy, C., & Wang, J. (2020). What Explains the Post–2011 Trends of Longer Maturities and Rising Default Rates on Auto Loans? *Discussion Papers (Federal Reserve Bank of Philadelphia)*. https://doi.org/10.21799/frbp.dp.2020.02

Campbell, J. Y., & Cocco, J. F. (2003). Household Risk Management and Optimal Mortgage Choice. *The Quarterly Journal of Economics*, *118*(4), 1449–1494. https://doi.org/10.1162/003355303322552847

Understand the Total Cost of Borrowing – Wells Fargo. (2024). Wells Fargo. https://www.wellsfargo.com/goals-credit/smarter-credit/manage-your-debt/total-cost-of-borrowing/#:~:text=The%20amount%20of%20money%20you,keep%20your%20monthly%20payments%20manageable.

Wilson. (2023, November 10). *What is the Personal Loan Process in Malaysia*. BlueBricks Holding. https://www.bluebricks.com.my/personal-loan-process-malaysia/

Fishelson-Holstine, H. (2003). *The Role of Credit Scoring in Increasing Homeownership for Underserved Populations*. https://www.jchs.harvard.edu/sites/default/files/media/imp/babc_04-12.pdf