
# Data Visualisation Report

**CMP5352 Semester 4**
**June 14, 2024**

**Word Count: 2760**

**Abhash Rai 23140736**
Abhash.Rai@mail.bcu.ac.uk
School of Computing and Digital Technology
Birmingham City University

*This work is original and independent.*

# Contents

## List of Figures

## Abstract

A crucial aspect in financial industry is assessing the likelihood of a customer repaying their debt. In an era where data-driven decision-making is highly important, understanding what factors have impact on creditworthiness is vital for both lenders and borrowers.

This report analyzes a real-world dataset using data visualization and predictive modeling technique to examine relationships between factors like demographic characteristics, financial factors and customer characteristics and their effect on creditworthiness to answer five non-trivial research questions. The report also depicts the steps of a data science project lifecycle from data collection, data preprocessing, exploratory data analysis, feature engineering, modeling to reporting the insights.

The insights outlined in this report offer valuable information for financial institutions seeking to improve their risk management practices and for individuals aiming to make informed financial decisions.

# 1 Introduction

Data visualization play a crucial role in the information society to make informed and data driven decision for both professionals and non-professionals across diverse fields (Inastrilla 2023). It is a useful tool for understanding complex data to identify trends, extract insights and communicate them effectively. Humans, being visually oriented, easily understand compelling stories of abstract data when transformed into visualizations (Qin et al. 2020). Shakeel et al. (2022) states people have natural affinity for effective visual representation.

This report focuses on analyzing a real-world 'Credit Card Eligibility Dataset' obtained from Kaggle to provide answer and justification to several non-trivial research questions. The aim of the report is to identify and analyze key factors influencing credit card eligibility and financial well-being of individuals, and present the findings in a clear and compelling manner.

The report is structured in four sections. An explanation of the selected dataset and reasoning for its selection is provided in the 'Motivation and Objectives' section, along with the non-trivial research questions, each justified for its complexity and significance. Then, 'Dataset Wrangling' section outlines the wrangling steps taken including code to clean the dataset which is then used for exploratory data analysis. Following that, the 'Experimental Results' section summarizes the results using high-quality graphs, each of which has a brief caption addressing the non-trivial research questions. Finally, the 'Conclusion' section concludes by going over the questions, outlining the key takeaways from the research, and drawing conclusions.

# 2 Motivation and Objectives

## 2.1 Dataset Selection

Accurate assessment of credit risk is important, especially in case of financial services industry. In order to minimize the risk of default, financial institutions have to establish a credit risk management framework for taking informed decisions when issuing credit cards. This report provides the research and analysis work on 'Credit Card Eligibility Dataset' which can be utilized to address such challenges.

The dataset is publicly available and contains a rich set of variables which was among the reason for its selection. The dataset was primarily selected to potentially provide financial providers and customers understanding of factors which impact financial well-being of individuals and aid in making informed decision like credit card issuance.

## 2.2 Dataset Overview

The 'Credit Card Eligibility Dataset', sourced from Kaggle, contains information about customer demographics, financial status, employment details, lifestyle and credit risk status. The dataset contains discrete, continuous, and categorical data. It has 36457 rows and 18 columns. The columns are:

1. Gender
2. Has.a.car
3. Has.a.property
4. Children.count
5. Income
6. Employment.status
7. Education.level
8. Marital.status
9. Dwelling
10. Customer.relative.age
11. Employment.relative.length
12. Has.a.work.phone
13. Has.a.phone
14. Has.an.email
15. Job.title
16. Family.member.count
17. Account.relative.age
18. Is.high.risk

In the dataset, columns like age and employment length contain negative values to serve as large reference points in the dataset rather than actual values. With age, a negative value, for example, -30, means that the person is 30 units younger since a specific reference date, so essentially, the number of units the account has been around, but counted backwards from now. Also, employment length is measured from the present day and positive numbers mean unemployment periods.

### 2.3 Non-Trivial Research Questions

The non-trivial research questions explored in this report are outlined below with brief description and justification as to why they are non-trivial for every question.

1. **What factors greatly contribute to risk status of customers for credit card?**
   This question focuses on factors which directly affect the classificaiton of customers as either high risk or low risk. Finding the most influential features contributing to credit risk is highly informative. The goal is to rank these features based on their impact on credit risk status which can help financial institutions to improve risk assessment and develop strategies to enhancing financial stability.

2. **Which professions have a higher chance of being classified as high risk?**
   This question examines proportions of high-risk individuals associated with job titles, finding any jobs that may disproportionately contribute to the high-risk pool. Identifying these trends helps lenders refine risk models and offer tailored financial advice to at-risk professional groups.

3. **How does employment duration and status influence income levels among customers?**
   This question is aimed to investigate the association among the work experience length, work status, and income levels. Duration of employment indicates job security and career advancement and status in turn may very well impact earning capabilities. This is an important question due to personal career trajectories in a fast-moving labour market. Insights can then help inform policymakers, financial planners in addressing income disparity and job market policies.

4. **Does asset ownership significantly impact income levels, and does this impact differ between high-risk and low-risk customers?**
   This question studies whether the ownership of assets determines levels of income and if this impact varies between high-risk and low-risk customers. Results can be used to inform strategies from financial providers to design products that support asset growth and income stability, especially for high credit risk customers, by tailoring offerings like secured loans or asset-backed credit lines.

5. **Does higher education level mean high income?**
   This research explores the relationship between education level and income. While higher education is assumed to lead to higher earnings, the study quantifies this relationship and examines exceptions. These insights can help individuals make informed educational investments.

# 3 Dataset Wrangling

## 3.1 Data Preprocessing

```
#install.packages("tidyverse")
library(tidyverse)
df <- read.csv("credit_eligibility.csv")
names(df)
```

```
##  [1] "ID"                  "Gender"              "Has.a.car"
##  [4] "Has.a.property"      "Children.count"      "Income"
##  [7] "Employment.status"   "Education.level"     "Marital.status"
## [10] "Dwelling"            "Age"                 "Employment.length"
## [13] "Has.a.mobile.phone"  "Has.a.work.phone"    "Has.a.phone"
## [16] "Has.an.email"        "Job.title"           "Family.member.count"
## [19] "Account.age"         "Is.high.risk"
```

```
dim(df)
```

```
## [1] 36457    20
```

The dataset was first loaded into variable 'df' for initiating the data wrangling process.

```
problems(df)
str(df)
summary(df)
```

Initial exploration of the dataset was performed by identifying potential issues, summarizing the structure and statistical properties of the data. Some columns represent relative units which should be set to have accurate names to prevent any misinterpretation. On a closer look, the dataset contains boolean features such as 'Gender', 'Has.a.car' and 'Has.a.property' which have values either 'Y' or 'N'. Similarly, other boolean feature columns 'Has.a.mobile.phone', 'Has.a.work.phone', 'Has.a.phone', 'Has.an.email' and 'Is.high.risk' have values '0' or '1'. So these columns must be handled appropriately to set the correct values.

```
df <- df %>% select(-ID)
```

The 'ID' column was removed as it does not provide any significant impact during analysis.

```
# Some columns represent durations measured relative to a reference point
# Giving those columns an appropriate name
names(df)[names(df) == "Age"] <- "Customer.relative.age"
names(df)[names(df) == "Account.age"] <- "Account.relative.age"
names(df)[names(df) == "Employment.length"] <- "Employment.relative.length"
```

The relative unit columns were assigned proper names.

```
df <- df %>%
  mutate(Gender = ifelse(Gender == "M", "Male", "Female"))
```

The 'Gender' column was assigned proper values: 'Male' and 'Female'.

```
df <- df %>%
  mutate(
    Has.a.car = ifelse(Has.a.car == "Y", TRUE, FALSE),
```

```
    Has.a.property = ifelse(Has.a.property == "Y", TRUE, FALSE),
    Has.a.work.phone = ifelse(Has.a.work.phone == 1, TRUE, FALSE),
    Has.a.phone = ifelse(Has.a.phone == 1, TRUE, FALSE),
    Has.an.email = ifelse(Has.an.email == 1, TRUE, FALSE),
    Is.high.risk = ifelse(Is.high.risk == 1, TRUE, FALSE)
)
```

The binary features in the dataset were modified to have boolean values 'True' or 'False' and data type set to logical.

```
df$Family.member.count <- as.integer(
  df$Family.member.count)
df$Account.relative.age <- as.numeric(
  df$Account.relative.age)
df$Employment.relative.length <- as.numeric(
  df$Employment.relative.length)
df$Customer.relative.age <- as.numeric(
  df$Customer.relative.age)
```

The data types of above columns were set to 'integer' and 'numeric' based on their nature.

```
unique(df$Has.a.mobile.phone)

## [1] 1

df <- df %>% select(-Has.a.mobile.phone)
```

'Has.a.mobile.phone' column has only a single value. So, it was removed.

```
df <- df %>%
  mutate(across(where(is.character), ~ str_to_title(as.character(.))))
```

The values of categorical columns were transformed to title case in the dataframe.

```
dim(df)

## [1] 36457    18
```

## 3.2 Exploratory Data Analysis

### 3.2.1 Null Values

```
df[df == ""] <- NA # Converting empty strings to NA values
result <- colSums(is.na(df))
result_df <- data.frame(Column_Names = names(result), NA_Count = result)
rownames(result_df) <- NULL
result_df

##                  Column_Names NA_Count
## 1                      Gender        0
## 2                   Has.a.car        0
## 3              Has.a.property        0
## 4              Children.count        0
## 5                      Income        0
## 6           Employment.status        0
```

```
## 7             Education.level        0
## 8              Marital.status        0
## 9                    Dwelling        0
## 10     Customer.relative.age        0
## 11 Employment.relative.length        0
## 12           Has.a.work.phone        0
## 13                Has.a.phone        0
## 14               Has.an.email        0
## 15                  Job.title    11323
## 16       Family.member.count        0
## 17      Account.relative.age        0
## 18               Is.high.risk        0
```

Only 'Job.title' column contains null values with a total null value count of 11323.

### 3.2.2 Categorical Features

```
character_columns <- df %>% select_if(is.character)
names(character_columns)

## [1] "Gender"            "Employment.status" "Education.level"
## [4] "Marital.status"    "Dwelling"          "Job.title"
```

To explore above listed categorical columns two custom functions were defined: one for visualizing data distribution of the feature, and another for analyzing the feature.

```
# install.packages("ggplot2")
# install.packages("ggrepel")
library(ggplot2)
library(ggrepel)

create_pie_chart_for_category <- function(df, column) {

  df_clean <- df[!is.na(df[[column]]), ]

  value_count_df <- data.frame(table(df_clean[[column]]))
  colnames(value_count_df) <- c("group", "value")

  value_count_df <- value_count_df[value_count_df$value != 0, ]

  label_position <- value_count_df %>%
    mutate(csum = rev(cumsum(rev(value))),
           pos = value/2 + lead(csum, 1),
           pos = if_else(is.na(pos), value/2, pos))

  ggplot(
    value_count_df,
    aes(x = "" , y = value, fill = fct_inorder(group))) +
    geom_col(width = 1, color = 1) +
    coord_polar(theta = "y") +
    scale_fill_brewer(palette = "Accent") +
    geom_label_repel(
      data = label_position,
      aes(
        y = pos,
        label = paste0(
          round((value/sum(value_count_df$value))*100, 2), "%"
```

```
        )
      ),
      size = 4.5, nudge_x = 1, show.legend = FALSE
    ) +
    guides(fill = guide_legend(title = "")) +
    theme_void() +
    ggtitle(paste0("Distribution of ", column)) +
    theme(
      plot.title = element_text(
        hjust = 0.5, margin = margin(t = 0, b = 10), face = "bold"
      )
    )
}
```

Above function used 'ggplot2' and 'ggrepel' to create pie charts, given a dataframe and the column name as parameter respectively.

```
analyze_character_column <- function(df, col_name) {

  if (!col_name %in% names(df)) {
    stop("Column '", col_name, "' not found in the data frame.")
  }

  df_clean <- df[!is.na(df[[col_name]]), ]
  column_data <- df_clean[[col_name]]

  if (!is.character(column_data)) {
    stop("Column '", col_name, "' is not of character type.")
  }

  cat("Column:", col_name, "\n")
  unique_vals <- unique(column_data)
  num_unique_vals <- length(unique_vals)
  cat("Number of unique values:", num_unique_vals, "\n")
  cat("Value counts:\n")

  value_count_df <- data.frame(table(column_data))
  colnames(value_count_df) <- c("Category", "Count")
  value_count_df$Percentage <- round(
    (value_count_df$Count / sum(value_count_df$Count)) * 100, 2
  )
  value_count_df <- value_count_df[
    order(value_count_df$Percentage, decreasing = TRUE),
  ]
  rownames(value_count_df) <- NULL
  print(value_count_df)
  cat("\n")
}
```

Above second function analyzes a specified character column in a data frame, printing unique values, their counts, and their percentage distributions.

```
create_pie_chart_for_category(df, "Gender")
```

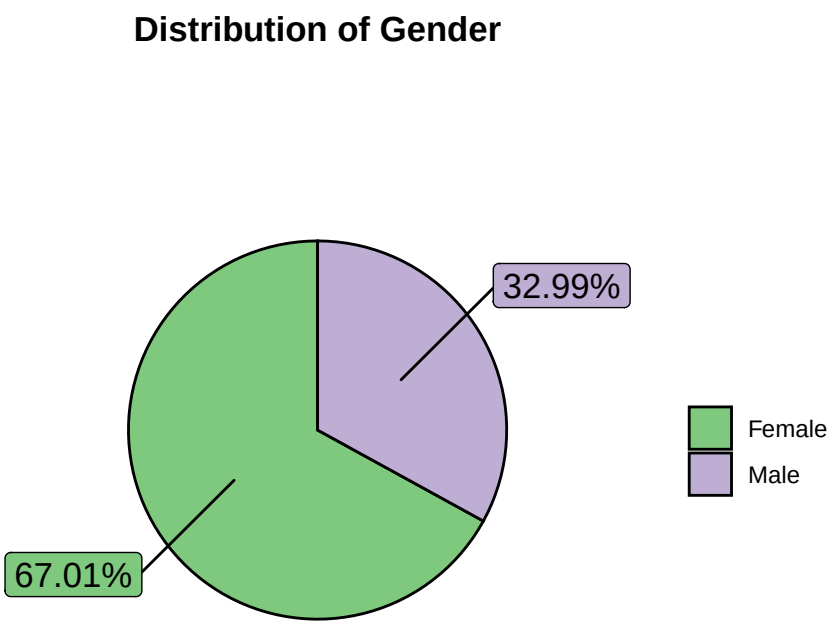


Figure 1: Distribution of Gender

```
analyze_character_column(df, "Gender")

## Column: Gender
## Number of unique values: 2
## Value counts:
##   Category Count Percentage
## 1   Female 24430      67.01
## 2     Male 12027      32.99
```

The dataset containes data of significantly more male as compared with female customers.

```r
ggplot(df, aes(x = Employment.status, fill = Employment.status)) +
  geom_bar() +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5) +
  theme_classic() +
  xlab("Employment Status") +
  ylab("Count") +
  ggtitle("Distribution of Employment Status") +
  scale_fill_brewer(palette = "Set2") +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", margin = margin(b = 10)),
    legend.position = "bottom",
    legend.title = element_blank()
  ) +
  coord_cartesian(ylim = c(0, 20000))
```

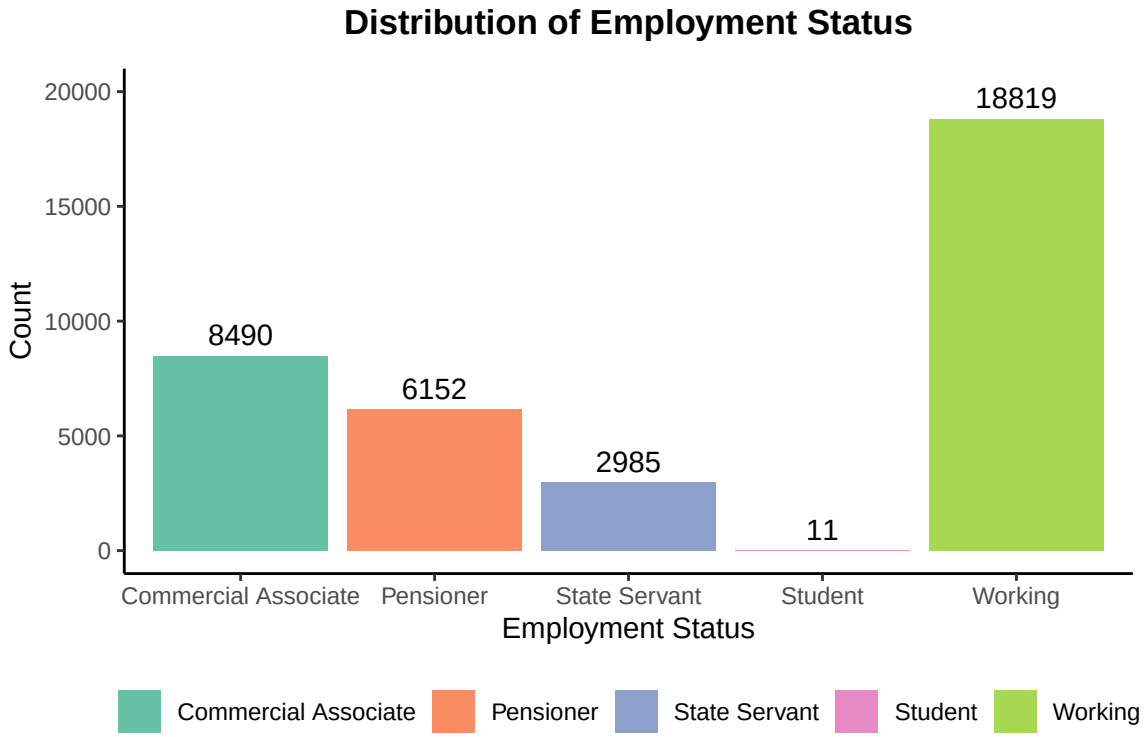


Figure 2: Distribution of Customer Employment Status

```r
analyze_character_column(df, "Employment.status")
```

```
## Column: Employment.status
## Number of unique values: 5
## Value counts:
##                Category Count Percentage
## 1               Working 18819      51.62
## 2 Commercial Associate  8490      23.29
## 3             Pensioner  6152      16.87
## 4         State Servant  2985       8.19
## 5               Student    11       0.03
```

There are total of five categories within the employment status of the customers. Half of the customers are employed and working. The customers with lowest count in the dataset are students.

```
create_pie_chart_for_category(df, "Education.level")
```

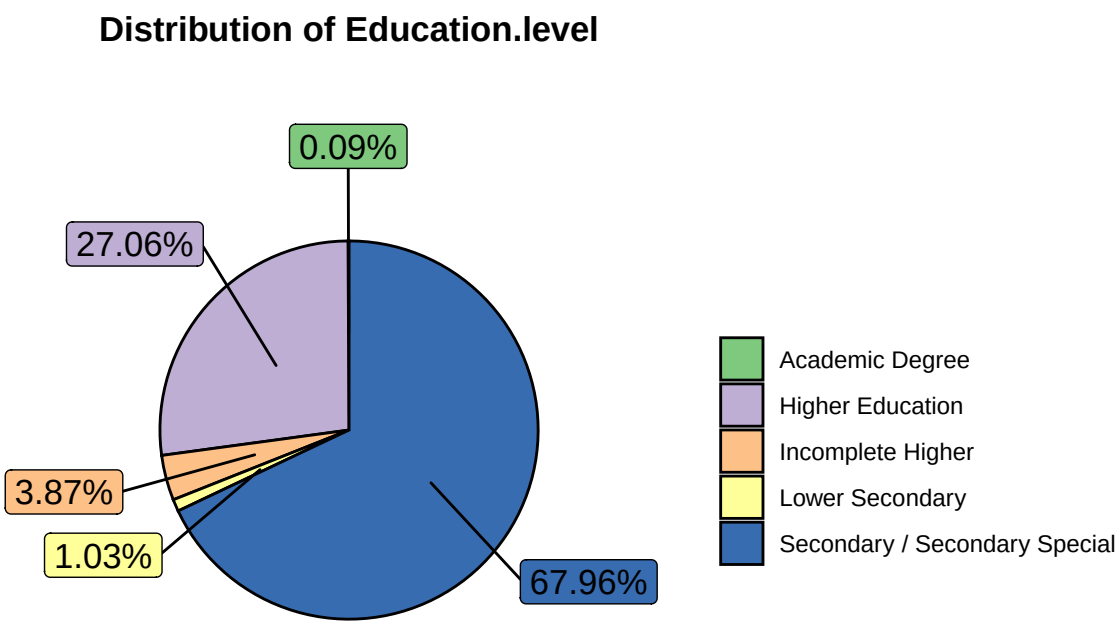


Figure 3: Distribution of Education Level of Customers

```
analyze_character_column(df, "Education.level")

## Column: Education.level
## Number of unique values: 5
## Value counts:
##                          Category Count Percentage
## 1 Secondary / Secondary Special 24777      67.96
## 2              Higher Education  9864      27.06
## 3             Incomplete Higher  1410       3.87
## 4               Lower Secondary   374       1.03
## 5               Academic Degree    32       0.09
```

The vast majority of customers have pursued education up to at least the secondary level or higher. There are some customer who pursed high education but dropped out before completion.

```r
marital_counts <- table(df$Marital.status)
marital_df <- as.data.frame(marital_counts)
names(marital_df) <- c("Marital.status", "Count")

ggplot(marital_df, aes(x = reorder(Marital.status, Count), y = Count)) +
  geom_point(aes(size = Count), color = "steelblue") +
  geom_text(aes(label = Count), vjust = -1, size = 3) +
  theme_classic() +
  xlab("Marital Status") +
  ylab("Count") +
  ggtitle("Distribution of Marital Status") +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  coord_cartesian(ylim = c(0, 30000))
```

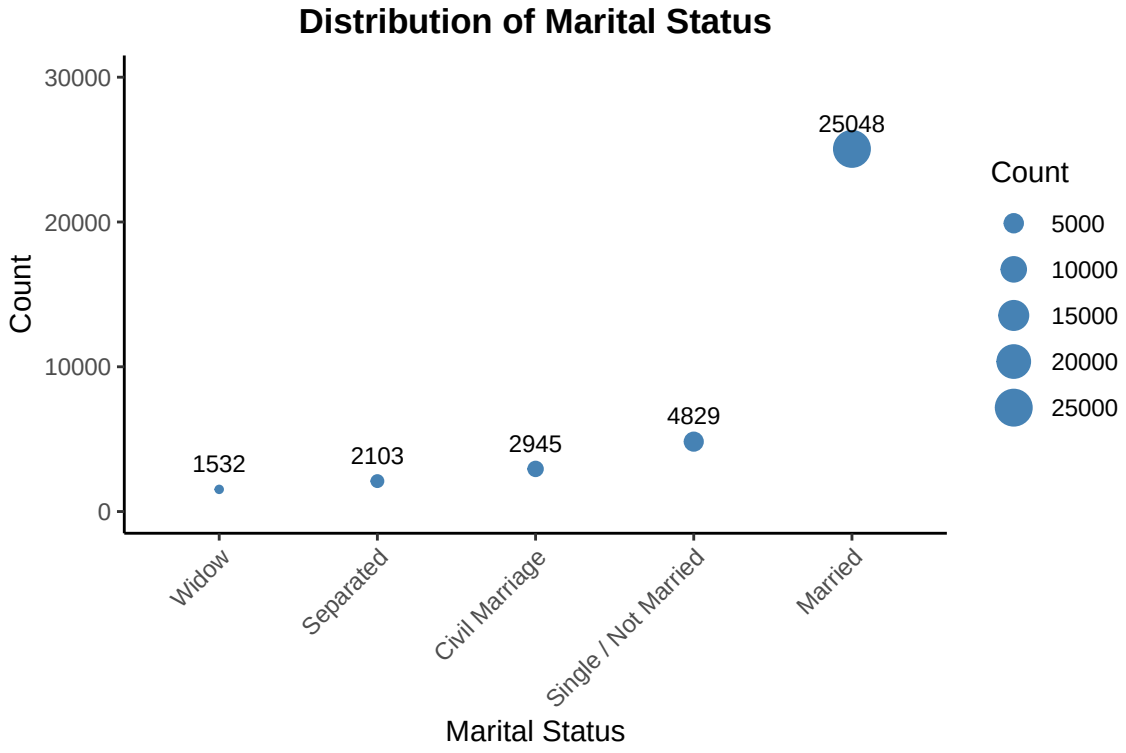


Figure 4: Distribution of Marital Status of Customers

```r
analyze_character_column(df, "Marital.status")
```

```
## Column: Marital.status
## Number of unique values: 5
## Value counts:
##                Category Count Percentage
## 1               Married 25048      68.71
## 2 Single / Not Married  4829      13.25
## 3       Civil Marriage  2945       8.08
## 4            Separated  2103       5.77
## 5                Widow  1532       4.20
```

More three-fifths of customers are married. The next largest group are those who are not married, followed by those who are separated or widowed.

```r
#install.packages("treemap")
library(ggplot2)
library(treemap)

dwelling <- table(df$Dwelling)
dwelling_df <- as.data.frame(dwelling)
names(dwelling_df) <- c("Dwelling", "Count")
treemap(dwelling_df,
        index = "Dwelling",
        vSize = "Count",
        title = "Distribution of Dwelling",
        palette = "Set2")
```

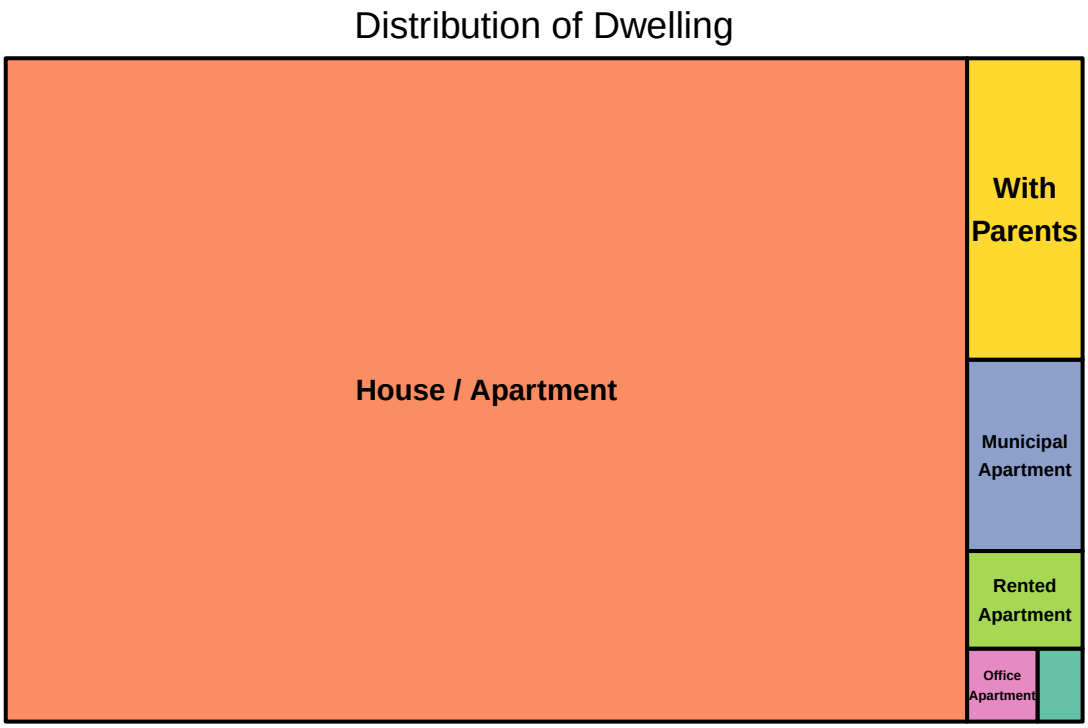


Figure 5: Distribution of Customer Dwelling

```r
analyze_character_column(df, "Dwelling")
```

```
## Column: Dwelling
## Number of unique values: 6
## Value counts:
##              Category Count Percentage
## 1   House / Apartment 32548      89.28
## 2        With Parents  1776       4.87
## 3 Municipal Apartment  1128       3.09
## 4    Rented Apartment   575       1.58
## 5    Office Apartment   262       0.72
## 6     Co-Op Apartment   168       0.46
```

The majority of customers live in a home or apartment, and they make up more than 89% of the dataset. The rest of the dwellings (a small minority) consist of living with parents, municipal apartments, rented apartments, office apartments, and co-op apartments.

```r
#install.packages("treemap")
library(ggplot2)
library(treemap)

job_title <- table(df$Job.title)
job_title_df <- as.data.frame(job_title)
names(job_title_df) <- c("Job.title", "Count")
treemap(job_title_df,
        index = "Job.title",
        vSize = "Count",
        title = "Distribution of Job.title",
        palette = "Set2")
```

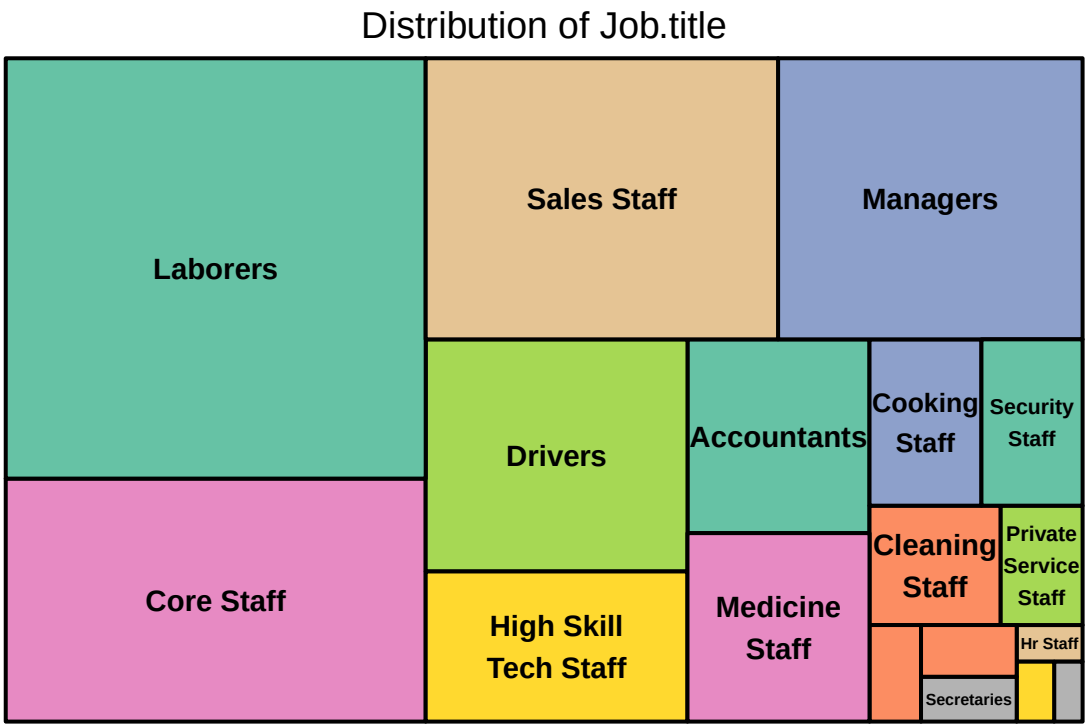


Figure 6: Distribution of Customer Dwelling

```r
analyze_character_column(df, "Job.title")
```

```
## Column: Job.title
## Number of unique values: 18
## Value counts:
##                  Category Count Percentage
## 1                Laborers  6211      24.71
## 2             Core Staff  3591      14.29
## 3            Sales Staff  3485      13.87
## 4               Managers  3012      11.98
## 5                Drivers  2138       8.51
## 6  High Skill Tech Staff  1383       5.50
## 7            Accountants  1241       4.94
## 8         Medicine Staff  1207       4.80
## 9          Cooking Staff   655       2.61
## 10        Security Staff   592       2.36
## 11        Cleaning Staff   551       2.19
## 12 Private Service Staff   344       1.37
```

```
## 13    Low-Skill Laborers   175       0.70
## 14  Waiters/Barmen Staff   174       0.69
## 15           Secretaries   151       0.60
## 16              Hr Staff    85       0.34
## 17         Realty Agents    79       0.31
## 18               It Staff    60       0.24
```

'Laborers', 'Core Staff' and 'Sales Staff' are the most common customer jobs, collectively representing over half of the dataset. 'Managers', 'Drivers', and 'High Skill Tech Staff' also contribute significantly to the job distribution.

#### 3.2.3 Binary Features

```
logical_columns <- df %>% select_if(is.logical)
names(logical_columns)
```

```
## [1] "Has.a.car"        "Has.a.property"   "Has.a.work.phone" "Has.a.phone"
## [5] "Has.an.email"     "Is.high.risk"
```

Above are the logical columns explored in this section.

```r
plot_df <- data.frame(
  col_name = c("Has.a.car",
               "Has.a.property", "Has.a.work.phone",
               "Has.a.phone", "Has.an.email"),
  true = c(sum(df$Has.a.car),
           sum(df$Has.a.property), sum(df$Has.a.work.phone),
           sum(df$Has.a.phone), sum(df$Has.an.email)),
  false = c(nrow(df) - sum(df$Has.a.car),
            nrow(df) - sum(df$Has.a.property), nrow(df) - sum(df$Has.a.work.phone),
            nrow(df) - sum(df$Has.a.phone), nrow(df) - sum(df$Has.an.email))
)

plot_df_long <- tidyr::pivot_longer(plot_df, cols = c("true", "false"),
                                    names_to = "value", values_to = "count")

ggplot(plot_df_long, aes(x = col_name, y = count / nrow(df) * 100, fill = value)) +
  geom_bar(stat = "identity") +
  labs(x = "Binary Features", y = "Percentage", fill = NULL) +
  geom_text(aes(label = count), position = position_stack(vjust = 0.5), size = 3) +
  theme_bw() +
  theme(legend.position = "bottom")
```

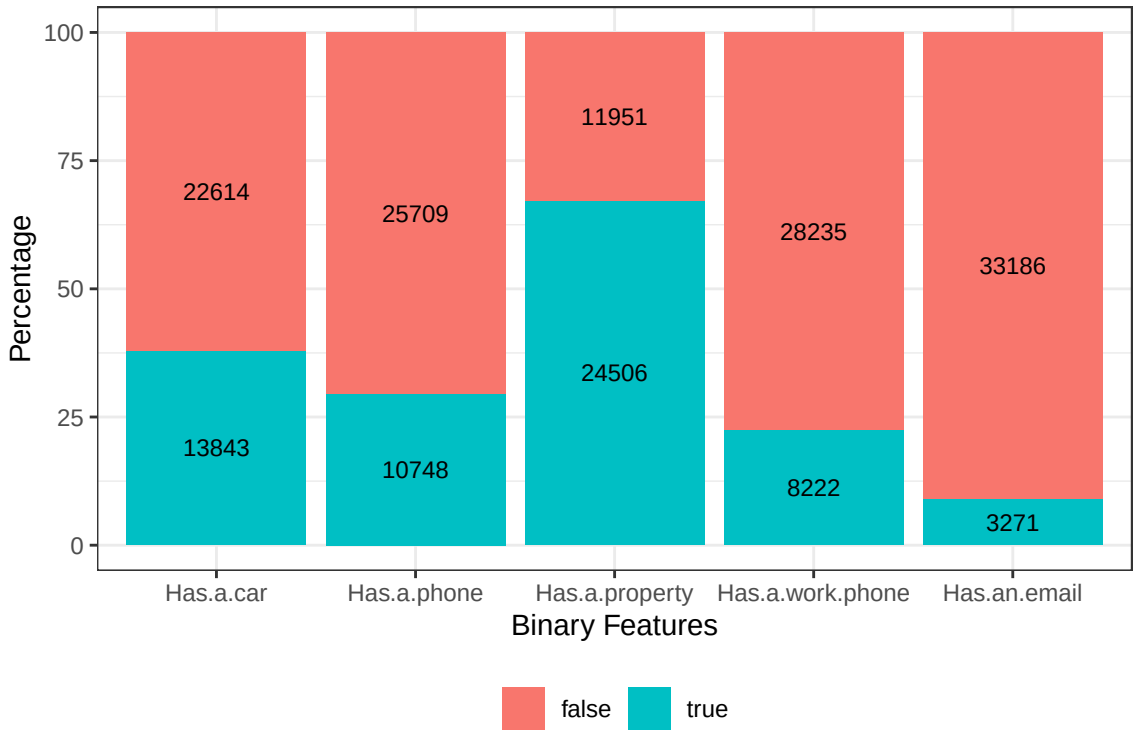


Figure 7: Distribution of Five Binary Features

More than three-fifths of the customer do not have cars. Customer with cars make up a smaller but still significant group within the dataset.

There are significantly more customers who owns property as compared with those who do not.

Surprisingly, most customers do not have neither a work phone nor a personal phone. This could mean customer did not include their phone contact details during creation of their accounts.

Only around 9% of the customers provided their email address. Rest of the customers either did not provide it or do not have an email.

```
#install.packages("lessR")
library(lessR)

PieChart(Is.high.risk, data = df,
         hole = 0.4,
         fill = 'blues',
         color = "black",
         lwd = 1.5,
         main = "Distribution of Is.high.risk")
```

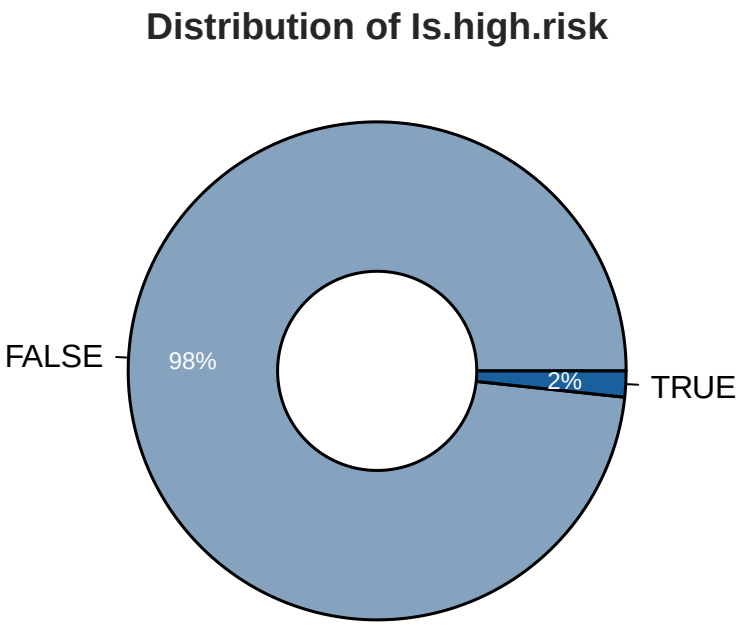


Figure 8: Distribution of Customer Risk

Almost all customers are classified as low risk in the dataset. High Risk customer makes up around 1.69% in the dataset.

#### 3.2.4 Numerical Features

```
numeric_columns <- df %>% select_if(function(x) is.integer(x) || is.numeric(x))
names(numeric_columns)

## [1] "Children.count"              "Income"
## [3] "Customer.relative.age"      "Employment.relative.length"
## [5] "Family.member.count"        "Account.relative.age"
```

Above features are explored in this section. The three columns: 'Customer.relative.age', 'Employment.relative.length', and 'Account.relative.age' are measured by subtracting the actual value from a reference point value. So, they can contain negative values.

```
ggplot(df, aes_string(x = "Children.count")) +
  geom_bar(aes(fill = cut(Children.count, breaks = c(-Inf, 0, 1, 2, 3, 4, 5, Inf))),
           show.legend = FALSE) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, size=3) +
  labs(title = paste0("Distribution of Children.count"),
       x = "Number of Children",
       y = "Count") +
  theme_classic()+
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  scale_x_continuous(limits = c(-1, 6), breaks = seq(0, 5, by = 1)) +
  scale_fill_brewer(palette = "Set2")
```

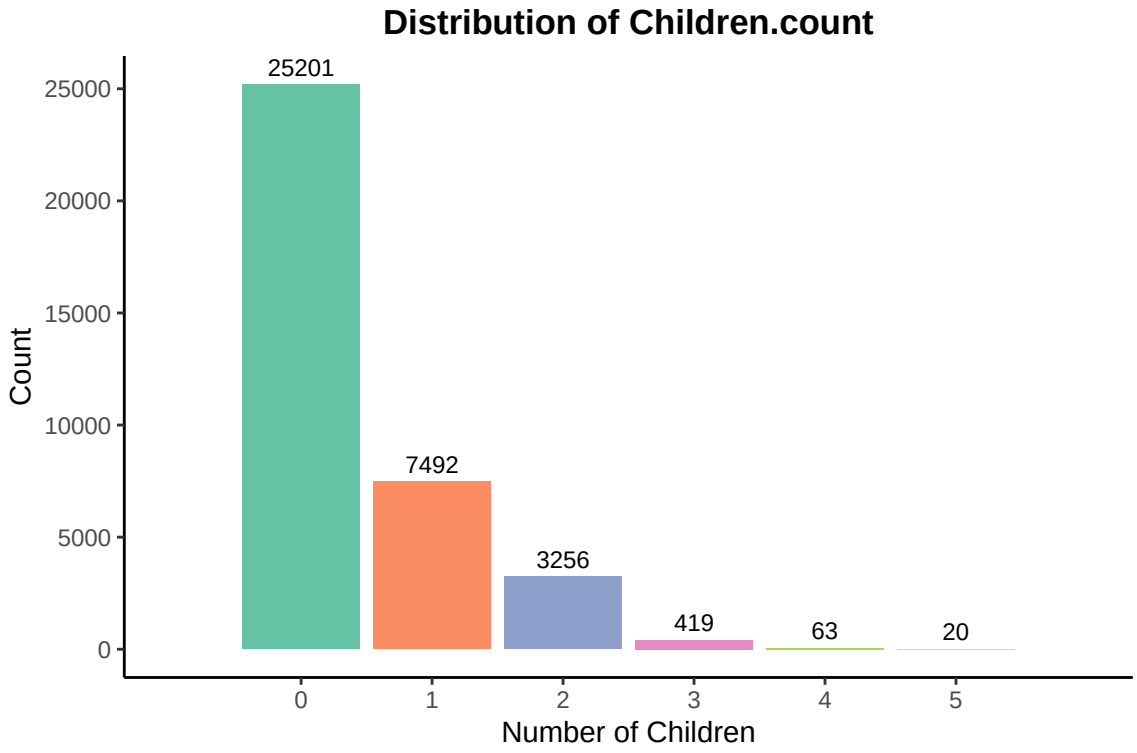


Figure 9: Distribution of Childred Customers Have

```
table(df$Children.count)

##
##     0     1     2     3     4     5     7    14    19
## 25201  7492  3256   419    63    20     2     3     1
```

Most customers have no children. There is a single customer with highest number of children of 19. Majority of the customers have either no children, 1 child or 2 children.

```r
ggplot(df, aes_string(x = "Family.member.count")) +
  geom_bar(
    aes(fill = cut(Family.member.count, breaks = c(-Inf, 1, 2, 3, 4, 5, 6, 7, Inf))),
    show.legend = FALSE) +
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, size = 3) +
  labs(title = paste0("Distribution of Family.member.count"),
       x = "Number of Family Members",
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold")) +
  scale_x_continuous(limits = c(0, 8), breaks = seq(1, 7, by = 1)) +
  scale_fill_brewer(palette = "Set2")
```

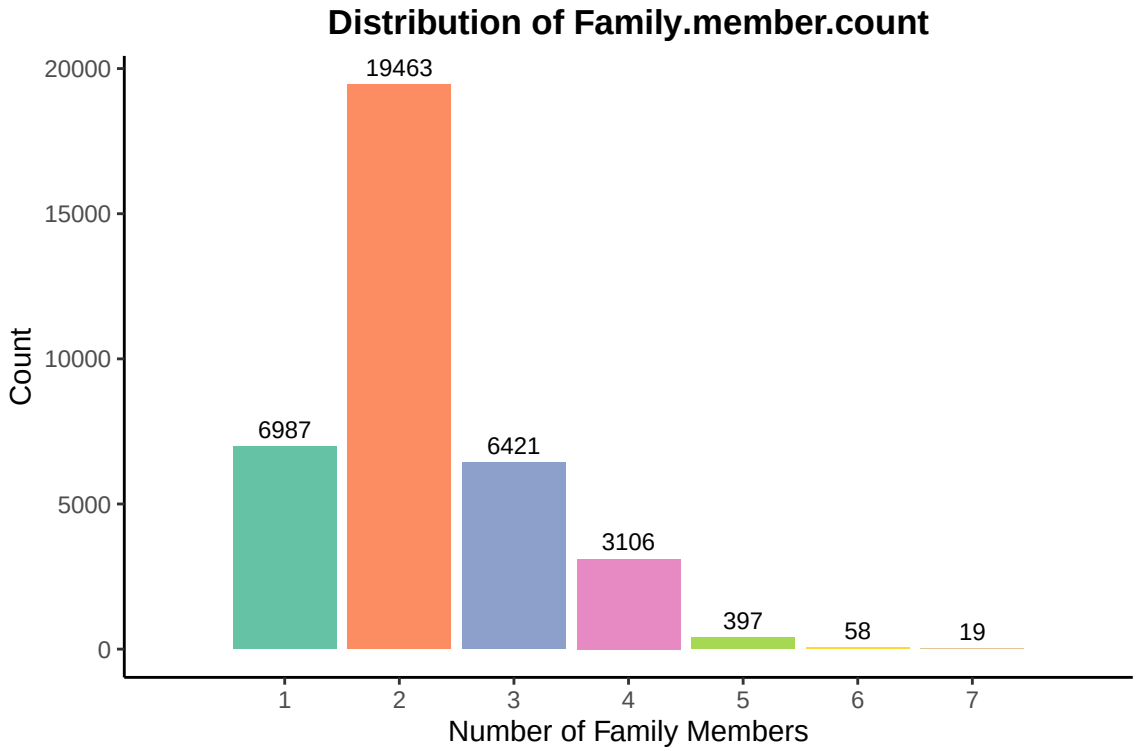


Figure 10: Distribution of Number of Customer Family Members

The majority of customers report having 2 family members, followed by those reporting 1 or 3 family members with nearly equal proportions between the two categories.

```r
ggplot(df, aes_string(x = "Customer.relative.age")) +
  geom_histogram(aes(fill = ..count..), bins = 60, show.legend = FALSE) +
  labs(title = paste0("Distribution of Customer.relative.age"),
       x = "Relative Age",
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        panel.grid.major.y = element_line(linetype = "dashed", color = "gray")) +
  scale_y_continuous(limits = c(0, 1000)) +
  scale_fill_gradient(low = "#132c45", high = "#54a7e7")
```

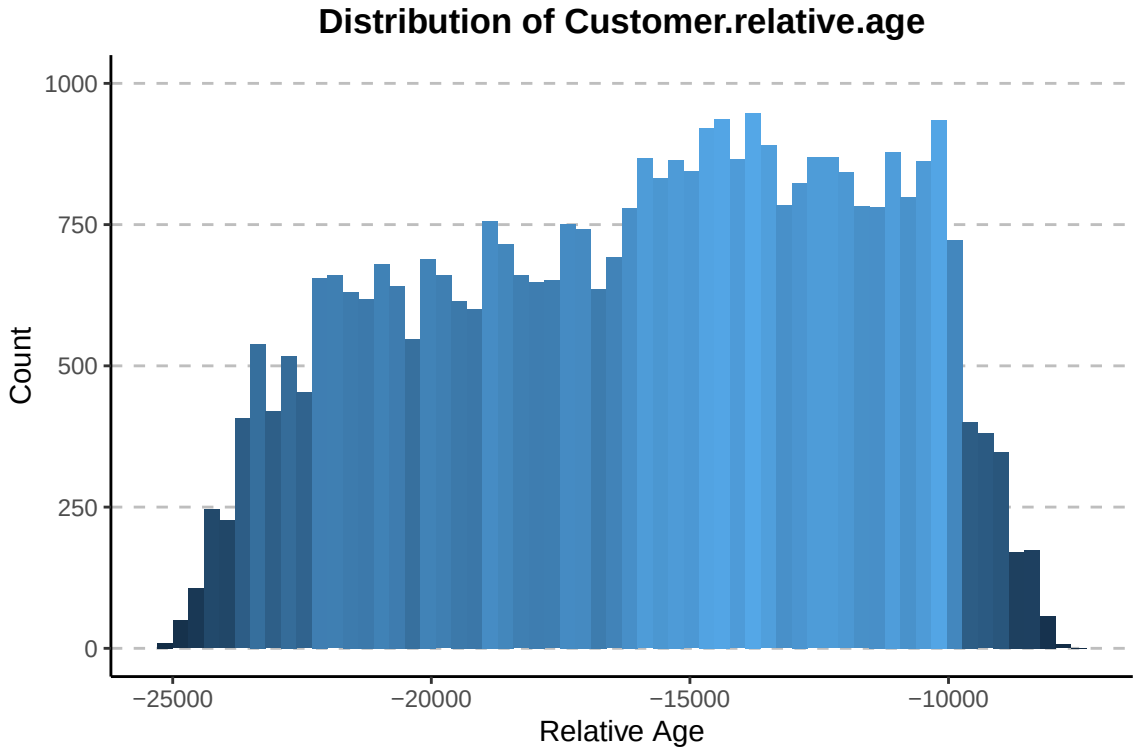


Figure 11: Distribution of Relative Age of Customers

When comparing the ages of customers, a higher proportion falls into the younger age group.

```r
ggplot(df, aes_string(x = "Account.relative.age")) +
  geom_histogram(aes(fill = ..count..), bins = 30, show.legend = FALSE) +
  labs(title = paste0("Distribution of Account.relative.age"),
       x = "Relative Account Age",
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        panel.grid.major.y = element_line(linetype = "dashed", color = "gray")) +
  scale_y_continuous(limits = c(0, 2000)) +
  scale_fill_gradient(low = "#132c45", high = "#54a7e7")
```

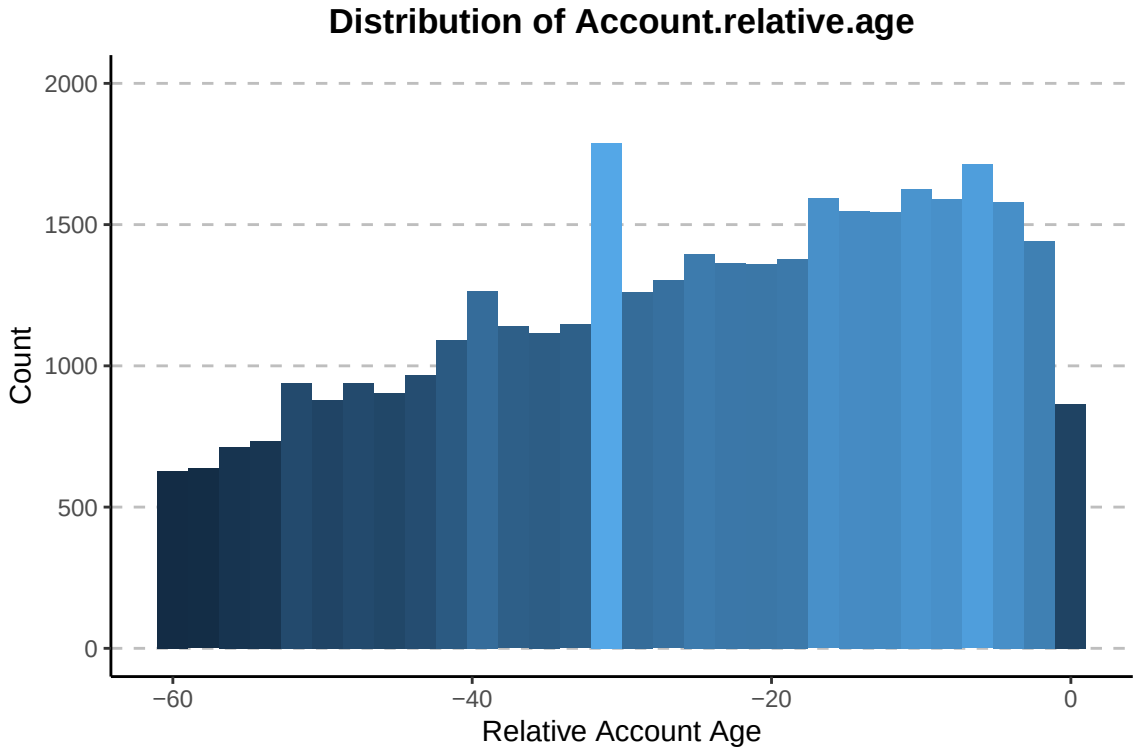


Figure 12: Distribution of Relative Age of Cutomer's Account

Similar trend is observed in the distribution of relative age of accounts where comparatively more proportion of accounts seem to created recently.

```r
#install.packages('patchwork')
library(patchwork)
library(ggplot2)

df_neg <- df[df$Employment.relative.length <= 0, ]
df_pos <- df[df$Employment.relative.length > 0, ]

bins <- 30

p1 <- ggplot(df_neg, aes(x = Employment.relative.length)) +
  geom_histogram(aes(fill = ..count..), bins = bins, show.legend = FALSE) +
  labs(title = "Employment Period",
       x = "Relative Duration",
       y = "Count") +
  theme_classic() +
  theme(
```

```
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.major.y = element_line(linetype = "dashed", color = "gray")) +
  scale_y_continuous(limits = c(0, 6500)) +
  scale_fill_gradient(low = "#132c45", high = "#54a7e7")

p2_start = 360000
p2 <- ggplot(df_pos, aes(x = Employment.relative.length)) +
  geom_histogram(aes(fill = ..count..), bins = bins, show.legend = FALSE) +
  labs(title = "Unemployment Period",
       x = "Relative Duration",
       y = "Count") +
  theme_classic() +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    panel.grid.major.y = element_line(linetype = "dashed", color = "gray"))  +
  scale_x_continuous(limits = c(p2_start, p2_start + 15000)) +
  scale_y_continuous(limits = c(0, 6500)) +
  scale_fill_gradient(low = "#132c45", high = "#54a7e7")

p1 + p2
```

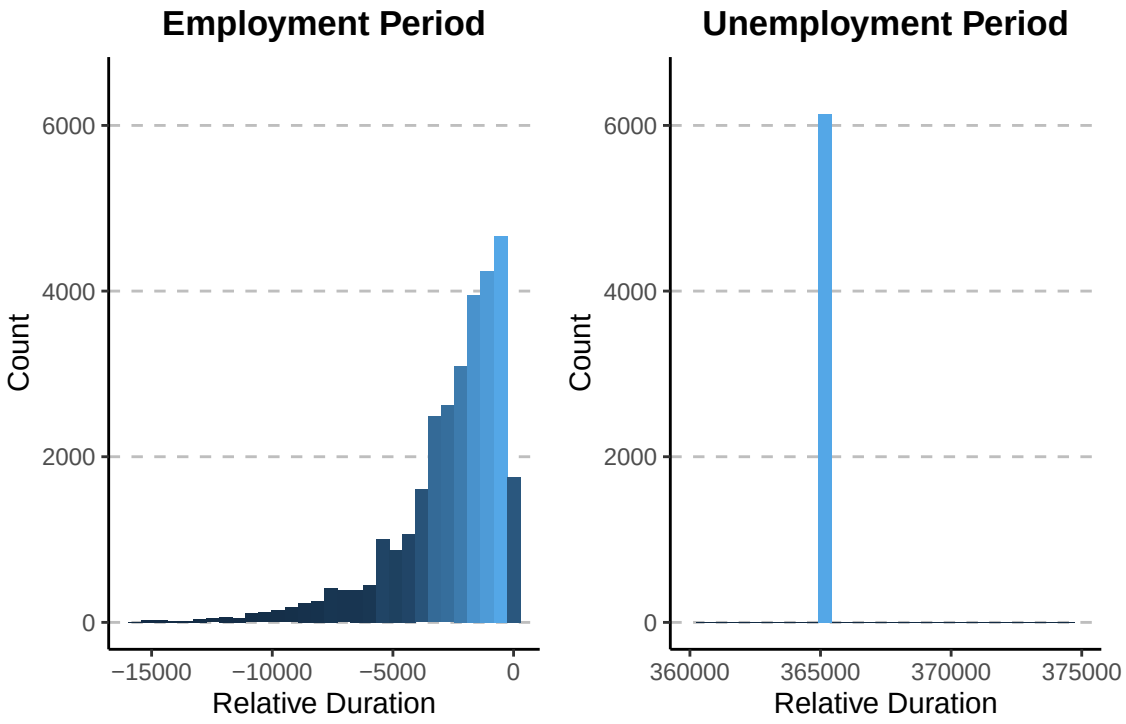


Figure 13: Distribution of Relative Employment Length

In the 'Employment.relative.length' column, negative values mean employment periods. Significantly higher proportion of customers are employed.

```r
ggplot(df, aes_string(x = "Income")) +
  geom_histogram(aes(fill = ..count..), bins = 60, show.legend = FALSE) +
  labs(title = paste0("Distribution of Income"),
       x = "Income",
       y = "Count") +
  theme_classic() +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        panel.background = element_blank(),
        panel.grid.major.y = element_line(linetype = "dashed", color = "gray"))  +
  scale_x_continuous(labels = function(x) format(x, scientific = FALSE)) +
  scale_fill_gradient(low = "#132c45", high = "#54a7e7")
```

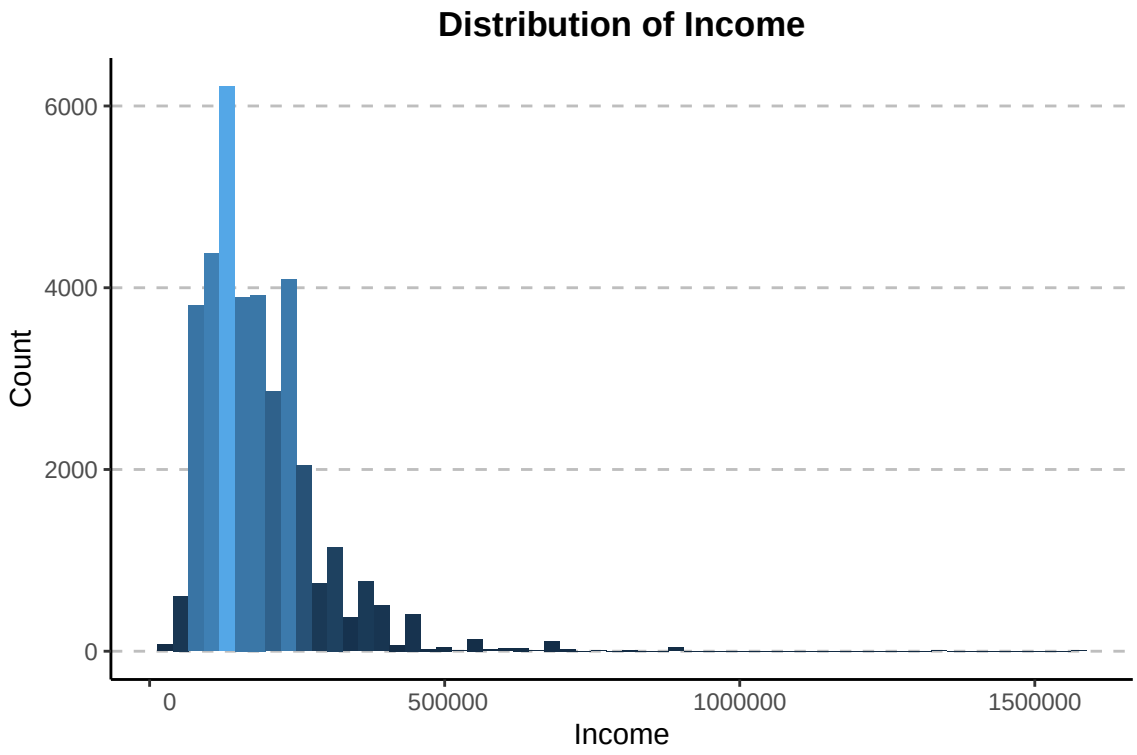


Figure 14: Distribution of Income of Customers

# 4 Experimental Results

The experimental results to the previously stated non-trivial research questions were obtained using multivariate analysis and predictive modeling. It required grouping, pivoting or transforming data which were then analyzed conducted using high quality graphs made using ggplot2.

## 4.1 What factors greatly contribute to risk status of customers for credit card?

The approach to obtaining solution in this section includes creating a XGBoost classifer model and visualizing its feature importance. The target feature is 'Is.high.risk'.

```
# install.packages("reshape2")
# install.packages("ggplot2")
library(reshape2)
library(ggplot2)

numeric_cols <- sapply(
  df, function(x) is.numeric(x) | is.integer(x) | is.logical(x)
)
df_numeric <- df[, numeric_cols]
cormat <- round(cor(df_numeric),2)

upper_tri <- cormat
upper_tri[lower.tri(cormat)] <- NA

melted_cormat <- melt(upper_tri, na.rm = TRUE)

ggplot(data = melted_cormat, aes(Var2, Var1, fill = value))+
  geom_tile(color = "white")+
  scale_fill_gradient2(low = "blue", high = "red", mid = "white",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal()+
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1))+
  coord_fixed() +
  geom_text(aes(Var2, Var1, label = value), color = "black", size = 4) +
  theme(
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    panel.grid.major = element_blank(),
    panel.border = element_blank(),
    panel.background = element_blank(),
    axis.ticks = element_blank(),
    legend.justification = c(1, 0),
    legend.position = c(0.6, 0.7),
    legend.direction = "horizontal")+
  guides(fill = guide_colorbar(barwidth = 7, barheight = 1,
                               title.position = "top", title.hjust = 0.5))
```

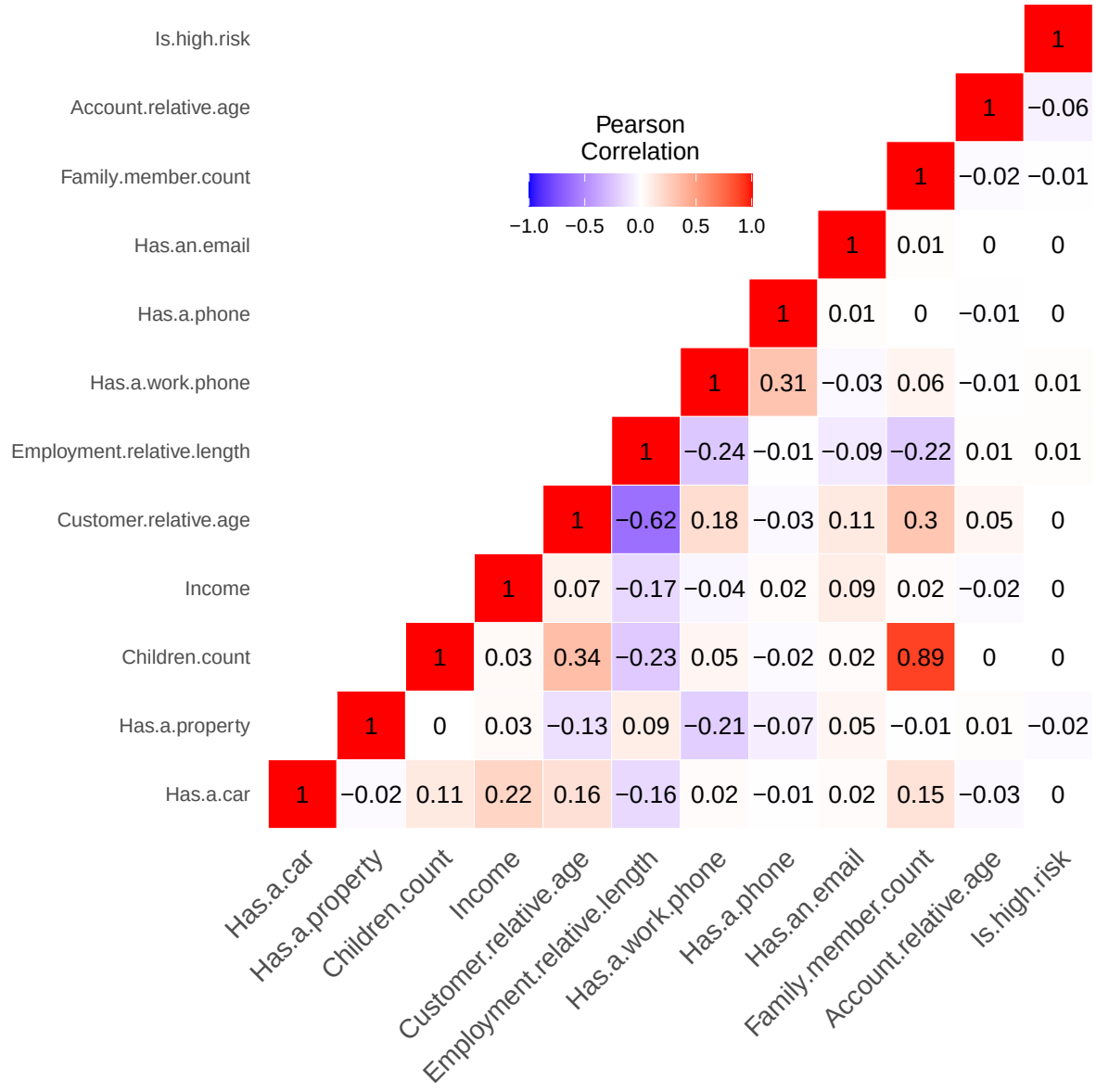


Figure 15: Correlation Heatmap

No features seem to be correlated with the target feature 'Is.high.risk'.

```
# install.packages("xgboost")
# install.packages("SHAPforxgboost")
# install.packages("caret")
library(xgboost)
library(SHAPforxgboost)
library(caret)
```

Above necessary libraries were imported.

```
df_clean <- na.omit(df)

features <- df_clean[, !(names(df_clean) %in% c("Is.high.risk"))] # Exclude target co
target <- df_clean$Is.high.risk
```

```r
# Train-Test Split

set.seed(555) # for reproducibility
train_size <- floor(0.8 * nrow(features))
train_indices <- sample(seq_len(nrow(features)), size = train_size)

X_train <- features[train_indices, ]
preprocessed_data <- dummyVars(" ~ .", data = X_train)
X_train <- data.frame(predict(preprocessed_data, newdata = X_train))
X_train = as.matrix(X_train)

y_train <- target[train_indices]
y_train <- as.numeric(y_train)

X_test <- features[-train_indices, ]
preprocessed_data <- dummyVars(" ~ .", data = X_test)
X_test <- data.frame(predict(preprocessed_data, newdata = X_test))
X_test = as.matrix(X_test)

y_test <- target[-train_indices]
y_test <- as.numeric(y_test)

dim(X_train)

## [1] 20107    57

dim(X_test)

## [1] 5027   57
```

The wrangled dataframe was split into train-test split of 0.8 and then one-hot-encoded. 80% was used for training and 20% for testing .

```r
dtrain <- xgb.DMatrix(data = X_train, label=y_train)
dtest <- xgb.DMatrix(data = X_test, label=y_test)
watchlist <- list(train=dtrain, test=dtest)
```

The train-test splits were converted to DMatrix objects which are used to efficiently store the data in a format optimized for training and evaluating an XGBoost model.

```r
model <- xgb.train(
  data=dtrain,
  lambda = 5,
  alpha=5,
  max.depth=4,
  eta=0.3,
  subsample=0.8,
  colsample_bytree=0.9,
  nthread = 3,
  nrounds=1000,
  scale_pos_weight = 30,
  watchlist=watchlist,
  objective = "binary:logistic",
  eval_metric = "logloss",
  verbose = 0
)
```

```r
y_pred = predict(model, X_test)
y_pred <- ifelse(y_pred > 0.5, 1, 0)
y_test_factor <- factor(y_test, levels = c(0, 1))
y_pred_factor <- factor(y_pred, levels = c(0, 1))
confusionMatrix(y_test_factor, y_pred_factor)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 4888   62
##          1   47   30
##
##                Accuracy : 0.9783
##                  95% CI : (0.9739, 0.9822)
##     No Information Rate : 0.9817
##     P-Value [Acc > NIR] : 0.9644
##
##                   Kappa : 0.3441
##
##  Mcnemar's Test P-Value : 0.1799
##
##             Sensitivity : 0.9905
##             Specificity : 0.3261
##          Pos Pred Value : 0.9875
##          Neg Pred Value : 0.3896
##              Prevalence : 0.9817
##          Detection Rate : 0.9723
##    Detection Prevalence : 0.9847
##       Balanced Accuracy : 0.6583
##
##        'Positive' Class : 0
##
```

The hyperparameters of the model was set manually through trial and error resulting in a model that has an accuracy of almost 98% and a high number of true negatives (4888) but relatively fewer true positives (30) in test dataset.

```r
# install.packages("Ckmeans.1d.dp")
library(Ckmeans.1d.dp)
importance_matrix <- xgb.importance(colnames(X_train), model = model)
xgb.ggplt <- xgb.ggplot.importance(
  importance_matrix = importance_matrix, top_n = 5
)
xgb.ggplt +
  theme_classic() +
  theme(legend.position = "bottom") +
  labs(title = "Important Features for Classifying Credit Card Risk",
       y = "Feature Importance Score (XGBoost)") +
  theme(plot.title = element_text(face = "bold"))
```

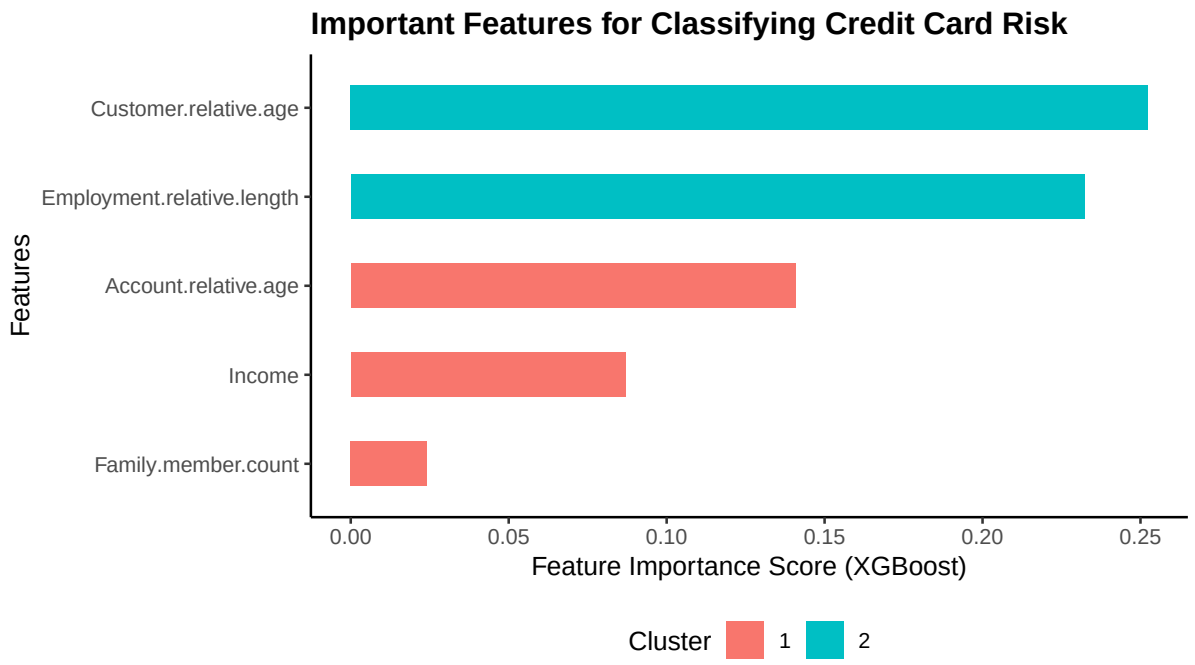


Figure 16: Top Five Important Features for Classifying Customer Risk

#### 4.1.1 Analysis and answer

Employment tenure and customer age demonstrated the strongest impact on risk status, followed by account age. How much a customer earn, and to a lesser extent how many family members they have, also contributes. This suggests that stability and longevity in personal and financial aspects are more predictive of credit card risk. Given all these points, financial stability and longevity in personal and financial aspects are key predictors of credit card risk, with employment length and customer age exhibiting the most pronounced influence when classifying customers into high risk or low risk.

### 4.2 Which professions have a higher chance of being classified as high risk?

```
library(ggplot2)

job_counts <- table(df$Job.title, df$Is.high.risk)
job_counts_df <- as.data.frame(job_counts)
job_counts_df <- job_counts_df %>% pivot_wider(
  names_from = Var2, values_from = Freq
)

colnames(job_counts_df) <- c("Job.title", "Low.risk", "High.risk")

job_counts_df <- job_counts_df %>%
  mutate(
    Low.risk.perct = round((Low.risk / (Low.risk + High.risk)) * 100, 3),
    High.risk.perct = round((High.risk / (Low.risk + High.risk)) * 100, 3)
  )

job_counts_df <- job_counts_df %>%
  arrange(desc(High.risk.perct))
```

```r
job_counts_df <- job_counts_df %>%
  mutate(Risk = case_when(
    High.risk.perct > 4 ~ "Very High Risk",
    High.risk.perct < 1 ~ "Low Risk",
    TRUE ~ "Moderate Risk"
  ))

risk_colors <- c(
  "Very High Risk" = "#f3661f",
  "Low Risk" = "#50bb6d",
  "Moderate Risk" = "#1470bb"
)

ggplot(
  job_counts_df,
  aes(x = High.risk.perct, y = reorder(Job.title, High.risk.perct), fill = Risk)
  ) +
  geom_bar(stat = "identity", width = 0.7, color = NA) +
  geom_text(aes(label = paste0(round(High.risk.perct, 1), "%")),
            hjust = -0.5,
            vjust = 0.4,
            color = "black",
            size = 3) +
  geom_vline(xintercept = 0, color = "black", linetype = "solid") +
  scale_fill_identity(
    name = "",
    guide = guide_legend(title.position = "top", title.hjust = 0.5)) +
  labs(
    title = "Examining High-Risk Job Titles in Credit Card Applications",
    subtitle = "(Percentage of High Risk Individuals)",
  ) +
  theme_minimal() +
  theme(
    plot.title = element_text(size = 14, face = "bold", hjust = 0),
    plot.subtitle = element_text(size = 12, hjust = 0, margin = margin(b = 15)),
    panel.grid.major.x = element_line(linetype = "dashed", color = "gray"),
    panel.grid.major.y = element_blank(),
    panel.grid.minor.x = element_blank(),
    panel.grid.minor.y = element_blank(),
    axis.text.y = element_text(size = 10, face = "bold"),
    axis.title = element_text(size = 12),
    axis.text.x = element_blank(),
    axis.title.x = element_blank(),
    axis.title.y = element_blank(),
    legend.position = "bottom"
  ) +
  scale_fill_manual(values = risk_colors, name = "")
```

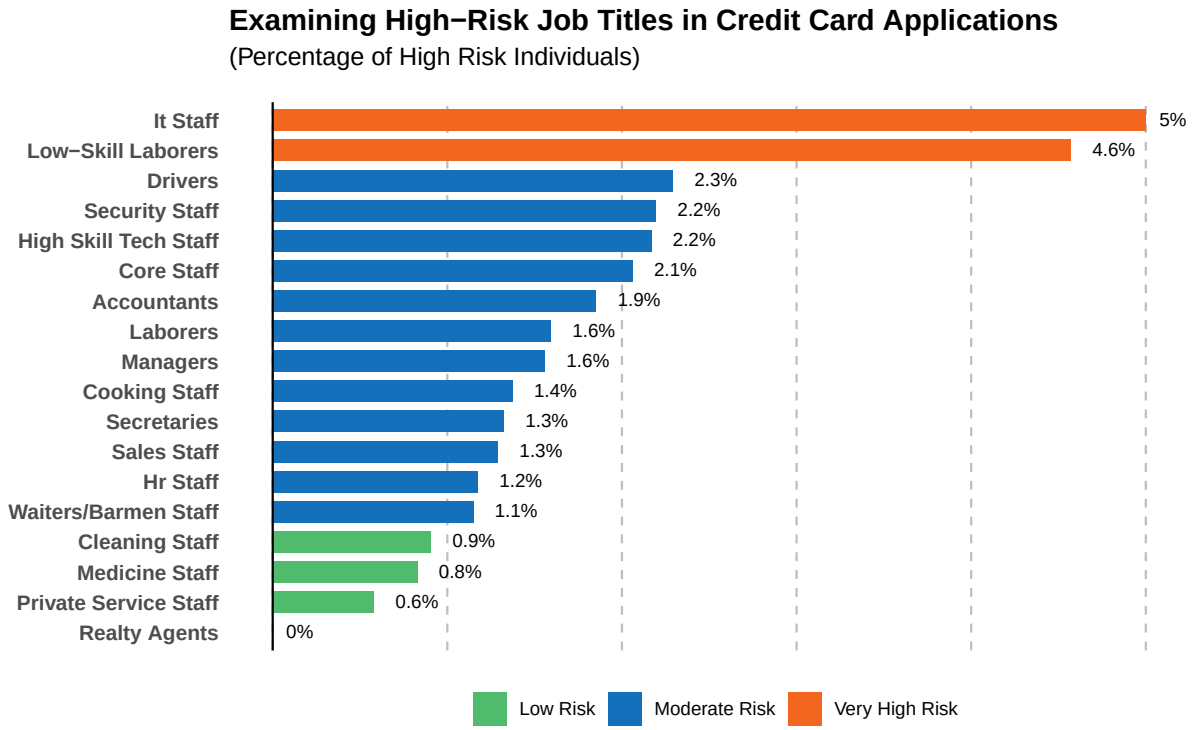


Figure 17: High-Risk Jobs

```
#install.packages("ggridges")
library(ggridges)
library(ggplot2)

df_clean <- na.omit(df)

df_clean$Risk <-  with(
  df_clean,
  ifelse(
    Job.title %in% c('It Staff', 'Low-Skill Laborers'),
    'Very High Risk',
    ifelse(
      Job.title %in% c(
        'Cleaning Staff', 'Medicine Staff',
        'Private Service Staff', 'Realty Agents'),
      'Low Risk',
      'Moderate Risk')
    )
  )

risk_colors <- c(
  "Very High Risk" = "#f3661f",
  "Low Risk" = "#50bb6d",
  "Moderate Risk" = "#1470bb"
)

ggplot(df_clean, aes(x = Income, y = Job.title, fill = Risk)) +
  geom_density_ridges() +
  theme_ridges() +
  theme(legend.position = "bottom") +
```

```
  scale_fill_manual(values = risk_colors, name = "") +
  theme_ridges() +
  theme(legend.position = "bottom",
        plot.title = element_text(hjust = 0.5)) +
  scale_x_continuous(limits = c(-1, 1000000)) +
  labs(title = "Income Distribution by Job Title Risk",
       x = "Income",
       y = "Job Title")
```

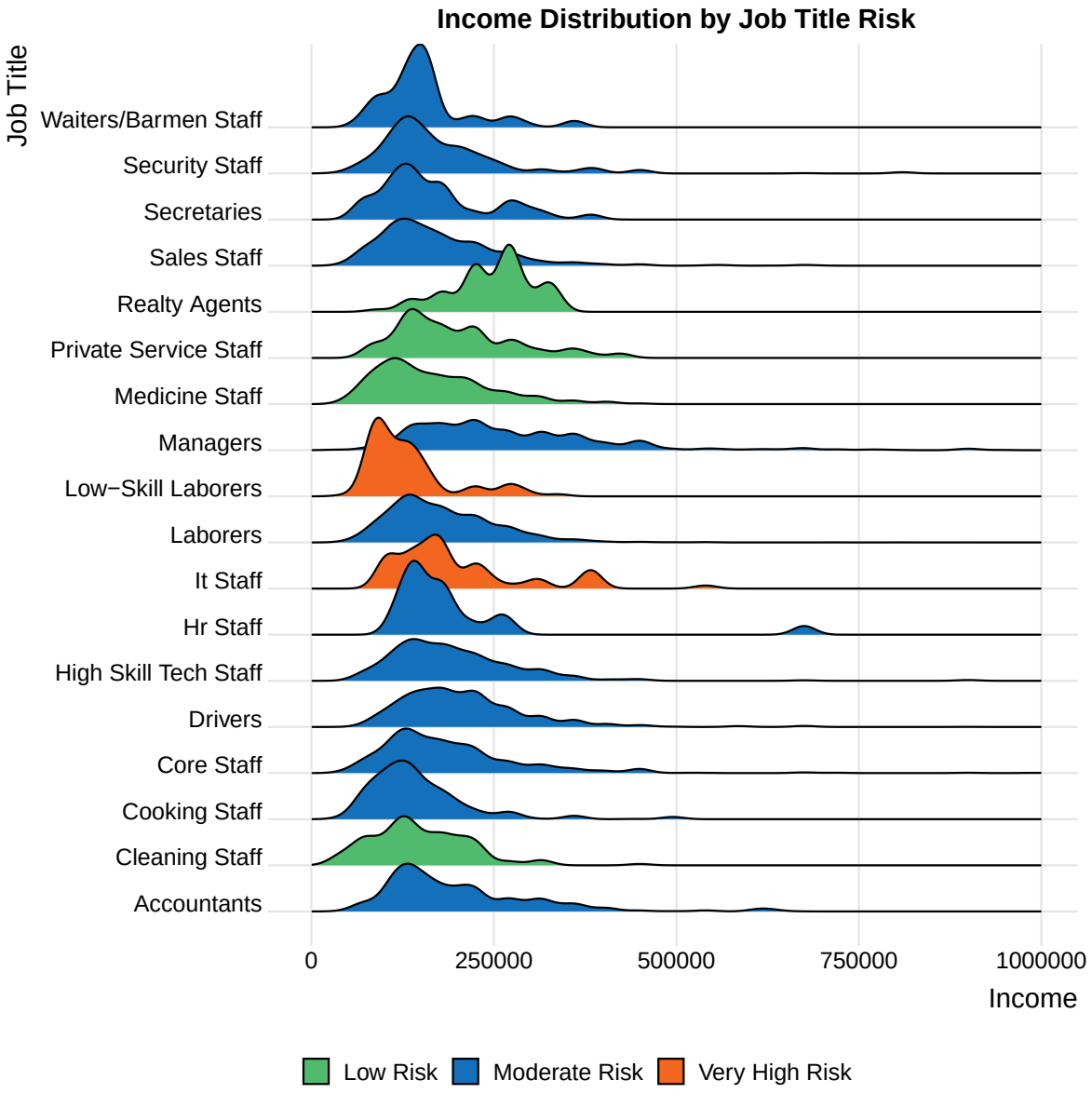


Figure 18: Income by Job Title Risks

#### 4.2.1 Analysis and answer

'It Staff' and 'Low-Skill Laborers' job are at risk the most as compared with others. On the other hand, 'Cleaning Staff', 'Medicine Staff', 'Private Service Staff' and 'Realty Agents' jobs have low risk associated, with 'Realty Agents' job being practically risk free. Moderate risk jobs has a wider range of incomes than the other two risk classes.

It includes both high-income earners and those with lower incomes. In light of these information, 'IT Staff' (4.6%) and 'Low-Skill Laborers' (5%) includes a disproportionately high percentage of high-risk individuals, significantly higher than the average proportion of 1.7% across moderate risk job titles and less than 1% across low risk job titles.

Upon investigating their income pattern, it can be observed that jobs with highly skewed income distribution were high risk. Moderate risk jobs follow a similar normal distribution, while low risk jobs often have a more concentrated income range

### 4.3 How does employment duration and status influence income levels among customers?

```
percentile_90 <- quantile(df$Income, probs = 0.9)

t2.rect1 <- data.frame(
  xmin = min(log(abs(df$Employment.relative.length))),
  xmax = max(log(abs(df$Employment.relative.length))),
  ymin = percentile_90,
  ymax = max(df$Income)
)

ggplot(df,
       aes(x = log(abs(Employment.relative.length)), y = Income,
           shape = Employment.status, color = Employment.status)) +
  geom_rect(data = t2.rect1,
            aes(xmin = xmin, xmax = xmax, ymin = ymin, ymax = ymax),
            fill = "#50bb6d", alpha = 0.2, inherit.aes = FALSE) +
  geom_point(size = 0.6, alpha = 0.5) +
  theme_classic() +
  labs(x = "Log of Absolute Employment Relative Length",
       y = "Income",
       title = "Relationship between Employment Duration and Income") +
  geom_hline(
    yintercept = percentile_90, linetype = "dashed", color = "black") +
  annotate("text",
           x = max(log(abs(df$Employment.relative.length))),
           y = percentile_90, label = "90th Percentile",
           vjust = 1, hjust = 1.25, color = "black") +
  annotate("text",
           x = mean(log(abs(df$Employment.relative.length))),
           y = mean(c(percentile_90, max(df$Income))),
           label = "High Income", color = "black", size = 5,
           hjust = 3.5, vjust = -8) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        legend.position = "bottom",
        legend.key.size = unit(1.2, "cm")) +
  guides(color = guide_legend(override.aes = list(size = 4)))
```


Figure 19: Employment Duration and Income

#### 4.3.1 Analysis and answer

The findings suggests that employment status and duration are indeed important contributors to observed earnings differentials. The majority of high-income individuals (defined as those earning above the 90th percentile) are concentrated in the 'Commercial Associate' category. Other high income customers have employment status either 'State Servant' or 'Working'.

Furthermore, early years in career seem pretty key to broader earning opportunity later on, especially for those in 'Commercial Associate' roles. For those in 'Working' status, employment duration is associated with income in a non-linear way with income increasing at first but showing a potential plateau or decline after a certain point.

### 4.4 Does asset ownership significantly impact income levels, and does this impact differ between high-risk and low-risk customers?

```
ggplot(df, aes(x = Has.a.car, y = Income, color = Is.high.risk)) +
  geom_boxplot() +
  facet_wrap(~ Has.a.property,
             labeller = as_labeller(
               c("TRUE" = "Has Property", "FALSE" = "No Property")
             )
  ) +
  labs(
    x = "Has Car",
    y = "Income",
    title = "Income Distribution by Risk Status, Property, and Car Ownership"
  ) +
  theme_classic() +
  theme(
```

```
    legend.position = "bottom",
    legend.title = element_blank(),
    plot.title = element_text(
      hjust = 0.5, face = "bold", margin = margin(b = 10))) +
  scale_color_manual(
    values = c("red", "blue"), labels = c("High Risk", "Low Risk"))
```

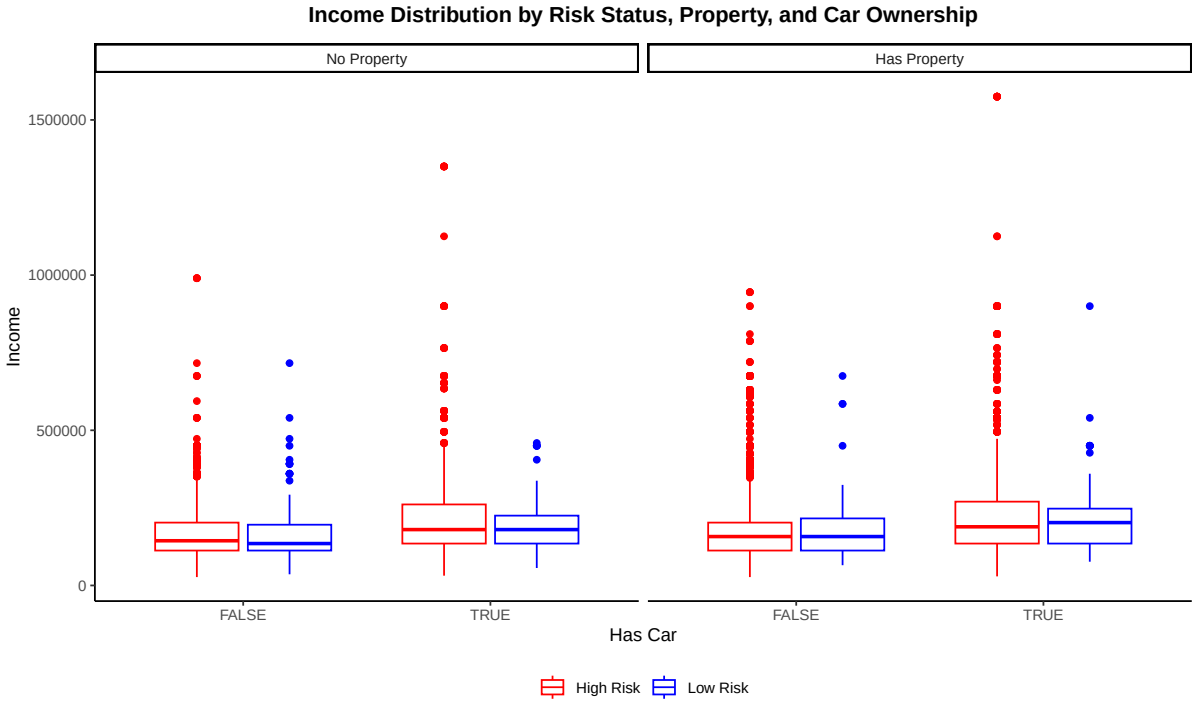


Figure 20: Feature Importance for Classifying Customer Risk

### 4.4.1 Analysis and answer

The results show a positive association between asset ownership and income. Customers who own property tend to have higher incomes than those without. In the same way car ownership indicates a higher income. These findings suggest that asset accumulation plays a role in influencing earning potential.

The study also identified a unique pattern for high-risk customers: higher and a highly skewed distribution of income. This means that even though people that are high-risk tend to earn more in total, their income levels are also more spread out with a good number of them having very high incomes.

To sum up, the results show a positive association between asset ownership and income, and high-risk customers exhibited a higher and more skewed income distribution, highlighting the need for targeted financial products that address income variability.

## 4.5 Does higher education level mean high income?

```
unique(df$Education.level)

## [1] "Secondary / Secondary Special" "Higher Education"
## [3] "Incomplete Higher"             "Lower Secondary"
```

```
## [5] "Academic Degree"

education_order <- c("Lower Secondary", "Secondary / Secondary Special",
                     "Incomplete Higher", "Higher Education", "Academic Degree")

median_income <- df %>%
  group_by(Education.level) %>%
  summarize(Median.Income = median(Income))

ggplot(median_income,
       aes(
         x = factor(Education.level, levels = education_order),
         y = Median.Income)) +
  geom_segment(aes(x = factor(Education.level, levels = education_order),
                   xend = factor(Education.level, levels = education_order),
                   y = 0, yend = Median.Income),
               color = "steelblue", linetype = "dashed", alpha = 0.5) +
  geom_point(aes(size = Median.Income), color = "#f3661f") +
  geom_text(aes(label = Median.Income), vjust = -1, size = 3) +
  theme_classic() +
  labs(title = "Median Income by Education Level",
       x = "Education Level (Increasing)",
       y = "Median Income") +
  scale_x_discrete(labels = c("Lower Secondary", "Secondary",
                              "Incomplete Higher", "Higher Education",
                              "Academic Degree")) +
  scale_y_continuous(labels = scales::comma,
                     breaks = seq(0, 300000, by = 50000),
                     limits = c(0, 300000)) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold", margin = margin(b = 10)),
    axis.title.x = element_text(margin = margin(t = 10)))
```

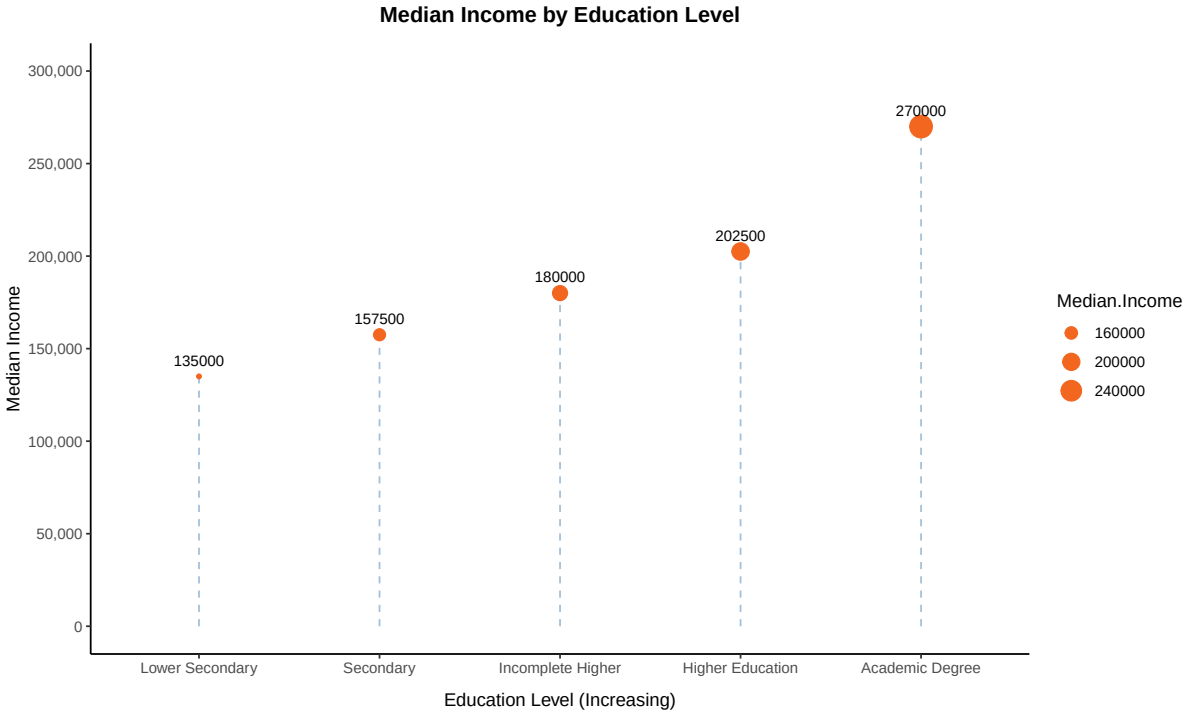


Figure 21: Median Income by Customer Education Level

### 4.5.1 Analysis and answer

We see a positive correlation between education level and median income.It shows a strong upwards trend with an increase in median income with every higher level of educational attainment, from Lower Secondary to Academic Degree. Those having an Academic Degree make nearly the double the median income of those with Lower Secondary education. In other words, in the simplest of terms, it illustrates even more clearly that a higher educational attainment leads to a higher ability to earn, and supports the notion that education plays a crucial role in socioeconomic advancement.

# 5 Conclusion

This report investigated the correlation of demographic and financials characteristics with the costumer credit risk using a comprehensive dataset named 'Credit Card Eligibility Dataset'. The analysis was targeted to understanding the influence of employment status, education level, asset ownership, and income on credit risk as well as major factors effecting credit risk. Through data visualization techniques and predictive modeling, this study reveals significant correlations and trends, providing key insights on the intricate relationship between these variables and creditworthiness.

This analysis of the dataset highlights the significant role of several factors in determining customer credit risk. The report's findings to the non-trivial research questions include:

1. **Q:** What factors greatly contribute to risk status of customers for credit card?
   **A:** Employment tenure and customer age have the strongest impact on risk status, followed by account age, income, and family member count respectively.

2. **Q:** Which professions have a higher chance of being classified as high risk?
   **A:** Jobs in the categories 'IT Staff' and 'Low-Skill Laborers' have much higher odds of being deemed high-risk.

3. **Q:** How does employment duration and status influence income levels among customers?
   **A:** Customers in mid-stage of career is the highest earning segment, with majority of high earners being 'Commercial Associate' and 'Working' status customers.

4. **Q:** Does asset ownership significantly impact income levels, and does this impact differ between high-risk and low-risk customers?
   **A:** Yes, asset ownership significantly impacts income levels. High-risk customers with more assets tend to have higher income levels compared to low-risk customers with similar assets.

5. **Q:** Does higher education level mean high income?
   **A:** Yes, higher education levels are strongly associated with higher median income.

Similarly, the report highlights key findings about the customer distribution, including but not limited to:

1. Customer distribution show a predominant presence of males, married individuals, and customers with education levels typically up to Secondary/Secondary Special.
2. Most customers have 2 members in their family
3. Most customers have 0 children living.
4. Most customers live in 'House / Appartment'

The dataset can be further explored to obtain additional findings. There is ample scope for asking more questions and performing deeper analysis to answer them using the dataset explored in this report.

## References

Inastrilla, C. R. A. (2023), Data visualization in the information society, *in* 'Seminars in Medical Writing and Education', Vol. 2, pp. 25–25.

Qin, X., Luo, Y., Tang, N. & Li, G. (2020), 'Making data visualization more efficient and effective: a survey', *The VLDB Journal* **29**(1), 93–117.

Shakeel, H. M., Iram, S., Al-Aqrabi, H., Alsboui, T. & Hill, R. (2022), 'A comprehensive state-of-the-art survey on data visualization tools: Research developments, challenges and future domain specific visualization framework', *IEEE Access* **10**, 96581–96601.