EIE3005 STATISTICAL COMPUTING

SEMESTER 2 2022/2023

THE PREDICTION ON CHURNING CREDIT CARD CUSTOMER

GROUP ASSIGNMENT PROJECT

PAPER

(25%)

GROUP CHERRY

NAME	MATRIC ID
XU HAO NING	S2022492
MUHAMMAD ARIEF MUQRISH	17202862
DIVYAH THANGADURAI	U2003491
KIRTINI SURESH	U2003379
SUN WENBIN	S2007796
MUHAMMAD AFNAN DARWISY	17124877

Part 1.0: Introduction

Because of the enormous amount of suppliers of services, particularly banks, available globally, the industry is ever-changing and highly adversarial. among the most significant issues for this industry is the shift in client behaviour. Clients are at the heart of all companies, particularly customer-dependent organisations such as the financial services industry, which accepts savings, makes investments, and lending. Thus, in an ever-changing and highly adversarial market surroundings, client attrition, or churn of customers, is a common issue encountered by credit card businesses. Client churn is described as the loss of a client to an opponent, which results in earnings losses. Maintaining existing customers becomes crucial for long-term achievement and sustainable profitability in a business where obtaining new consumers may be expensive and taking time. Furthermore, customer churn not just results in revenue loss, but it also has an impact on credit card businesses' entire brand reputation and market position.

It is critical to recognise clients who will probably to go to a rival bank in order to handle churn. As a result, using the credit card customer information data collected from online sources, this study hopes to analyse the factors influencing the loss rate of credit card customers through data cleansing, exploratory data analysis and visualisation, developing Logistic regression models, and other steps.

Initial, we will do Data cleansing on the collected information in order to standardize the data for later analysis. Following data cleansing, we will first gain a better understanding of the dataset by investigating five questions. The first Logistic regression model is then built using one dependent variable and all independent variables from the cleaned dataset. We will model and make predictions again after identifying variables that do not have a substantial impact on customer churn rate using the P-value.

1.1: Objectives:

(1) To analyse the percentage of churning and existing customers

(2) To determine the number of customers with different levels of education

(3) To analyze the churn rate of customers holding different types of credit cards

(4) To identify a significant difference between the Line of credit of loss customers and existing customers

(5) To observe variables that significantly affect the churn rate of credit card customers

Part 2.0: Data cleaning

2.1 Load packages and dataset

In this section, we load some necessary packages and dataset, and name the dataset **"data"**.

```
library(tidyverse)
library(tidyr)
library(dplyr)
library(lubridate)
library(janitor)
library(ISLR)
library(tree)
data <- read_csv("BankChurners.csv")
```

2.2 Check the data structure

The **str()** function below shows the structure of the information on each column's class, length, and content for the analysis. Based on the result, there are a total of 10127 observations and 23 variables.

```
str(data)

## spc_tbl_ [10,127 × 23] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ CLIENTNUM                                                               :
num [1:10127] 7.69e+08 8.19e+08 7.14e+08 7.70e+08 7.09e+08 ...
##  $ Attrition_Flag                                                       :
chr [1:10127] "Existing Customer" "Existing Customer" "Existing Customer"
"Existing Customer" ...
##  $ Customer_Age                                                            :
num [1:10127] 45 49 51 40 40 44 51 32 37 48 ...
##  $ Gender                                                                  :
chr [1:10127] "M" "F" "M" "F" ...
##  $ Dependent_count                                                         :
num [1:10127] 3 5 3 4 3 2 4 0 3 2 ...
##  $ Education_Level                                                         :
chr [1:10127] "High School" "Graduate" "Graduate" "High School" ...
##  $ Marital_Status                                                          :
chr [1:10127] "Married" "Single" "Married" "Unknown" ...
##  $ Income_Category
: chr [1:10127] "$60K - $80K" "Less than $40K" "$80K - $120K" "Less than
$40K" ...
##                     $          Card_Category
: chr [1:10127] "Blue" "Blue" "Blue" "Blue" ...
##  $ Months_on_book
: num [1:10127] 39 44 36 34 21 36 46 27 36 36 ...
##  $ Total_Relationship_Count                                                :
num [1:10127] 5 6 4 3 5 3 6 2 5 6 ...
##  $ Months_Inactive_12_mon                                                  :
num [1:10127] 1 1 1 4 1 1 1 2 2 3 ...
##  $ Contacts_Count_12_mon
```

```
: num [1:10127] 3 2 0 1 0 2 3 2 0 3 ...
##  $ Credit_Limit                                                              :
num [1:10127] 12691 8256 3418 3313 4716 ...
##  $ Total_Revolving_Bal                                                       :
num [1:10127] 777 864 0 2517 0 ...
##  $ Avg_Open_To_Buy                                                           :
num [1:10127] 11914 7392 3418 796 4716 ...
##   $ Total_Amt_Chng_Q4_Q1                                                  :
num [1:10127] 1.33 1.54 2.59 1.4 2.17 ...
##  $ Total_Trans_Amt                                                           :
num [1:10127] 1144 1291 1887 1171 816 ...
##                     $          Total_Trans_Ct
: num [1:10127] 42 33 20 20 28 24 31 36 24 32 ...
##   $ Total_Ct_Chng_Q4_Q1                  : num
[1:10127] 1.62 3.71 2.33 2.33 2.5 ...
##  $ Avg_Utilization_Ratio                                                     :
num [1:10127] 0.061 0.105 0 0.76 0 0.311 0.066 0.048 0.113 0.144 ...
##  $
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dep
endent_count_Education_Level_Months_Inactive_12_mon_1: num [1:10127] 9.34e-05
5.69e-05 2.11e-05 1.34e-04 2.17e-05 ...
##  $
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dep
endent_count_Education_Level_Months_Inactive_12_mon_2: num [1:10127] 1 1 1 1
1 ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   CLIENTNUM = col_double(),
##   ..   Attrition_Flag = col_character(),
##   ..   Customer_Age = col_double(),
##   ..   Gender = col_character(),
##   ..   Dependent_count = col_double(),
##   ..   Education_Level = col_character(), ##
..   Marital_Status = col_character(),
##   ..   Income_Category = col_character(),
##   ..   Card_Category = col_character(),
##   ..   Months_on_book = col_double(),
##   ..   Total_Relationship_Count = col_double(),
##   ..   Months_Inactive_12_mon = col_double(),
##   ..   Contacts_Count_12_mon = col_double(),
##   ..   Credit_Limit = col_double(),
##   ..   Total_Revolving_Bal = col_double(),
##   ..   Avg_Open_To_Buy = col_double(),
##   ..   Total_Amt_Chng_Q4_Q1 = col_double(),
##   ..   Total_Trans_Amt = col_double(),
##   ..   Total_Trans_Ct = col_double(),
##   ..   Total_Ct_Chng_Q4_Q1 = col_double(), ##
..   Avg_Utilization_Ratio = col_double(), ##
..
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dep
```

```
endent_count_Education_Level_Months_Inactive_12_mon_1 = col_double(), ##
..
Naive_Bayes_Classifier_Attrition_Flag_Card_Category_Contacts_Count_12_mon_Dep
endent_count_Education_Level_Months_Inactive_12_mon_2 = col_double() ##   .. )
##  - attr(*, "problems")=<externalptr>
```

2.3 Select data

CLIENTNUM indicates customer number information, which is not useful for EDA and modeling, so we remove it from the dataset by using **selected()** function. The last two variables are also not relevant to the logistic regression model, so they are also removed from the dataset.

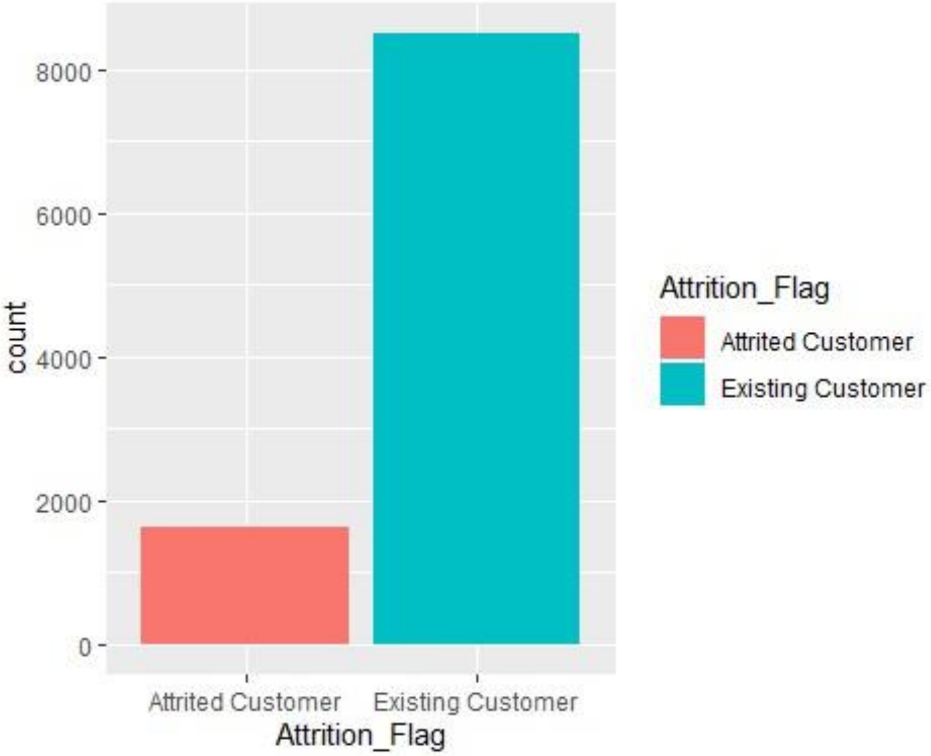
```
selected<- select(data,c(-1,-22,-23))
```

2.4 Filter the row

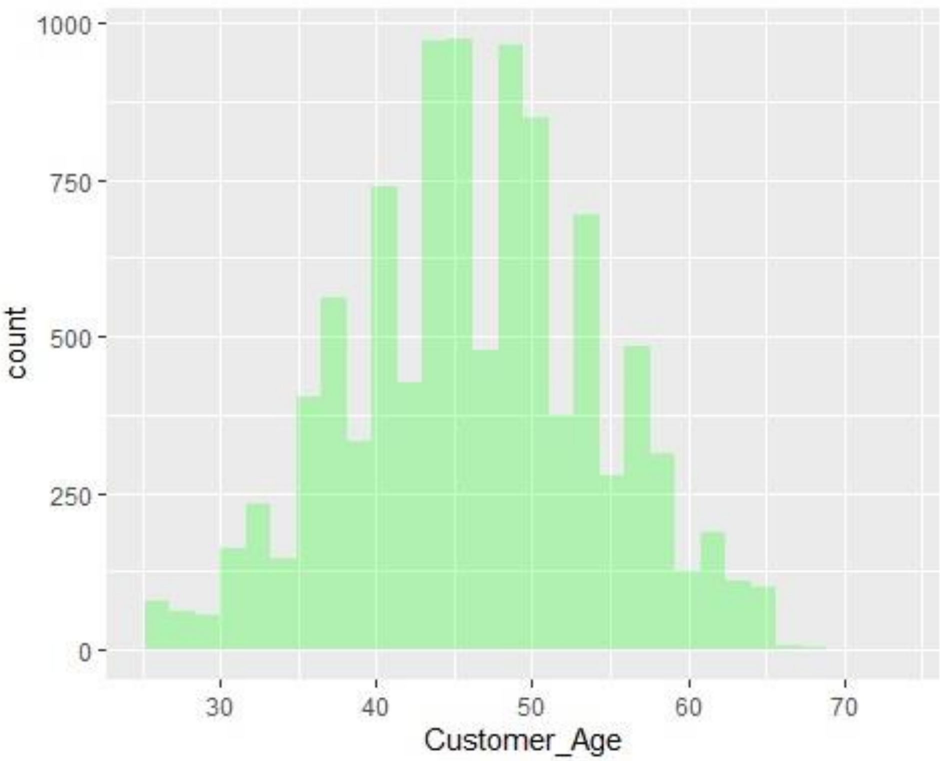
we use the “ggplot2” package to visualize the variables for checking whether there has any problems with the data, such as unusual values and missing values.

For **categorical variables** and **numerical variables with a relatively small number of values**, we can visualize them using bar charts. For **numerical variables with more values**, we use histograms for visualization. According to this principle, for the purpose of data cleaning, we then perform preliminary visualization of each of the nine selected variables.

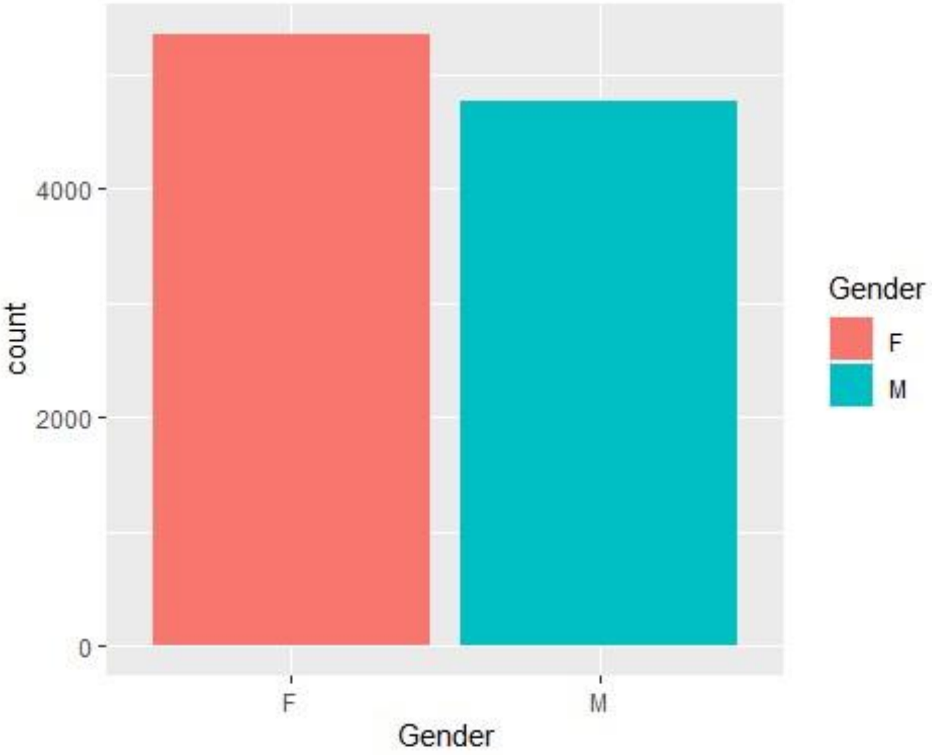
```
par(mfrow=c(4,5), no.readonly = FALSE)
ggplot(selected)+geom_bar(aes(x=Attrition_Flag,fill=Attrition_Flag))
```



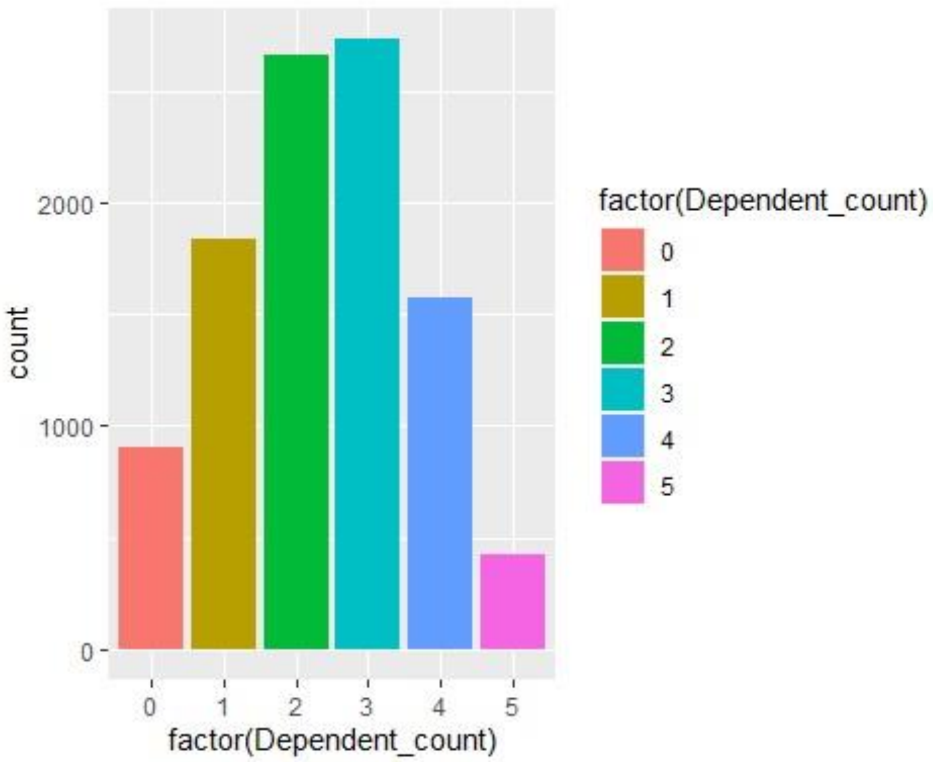
```
ggplot(selected) +   geom_histogram(aes(x = Customer_Age),
fill='green',alpha = 0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



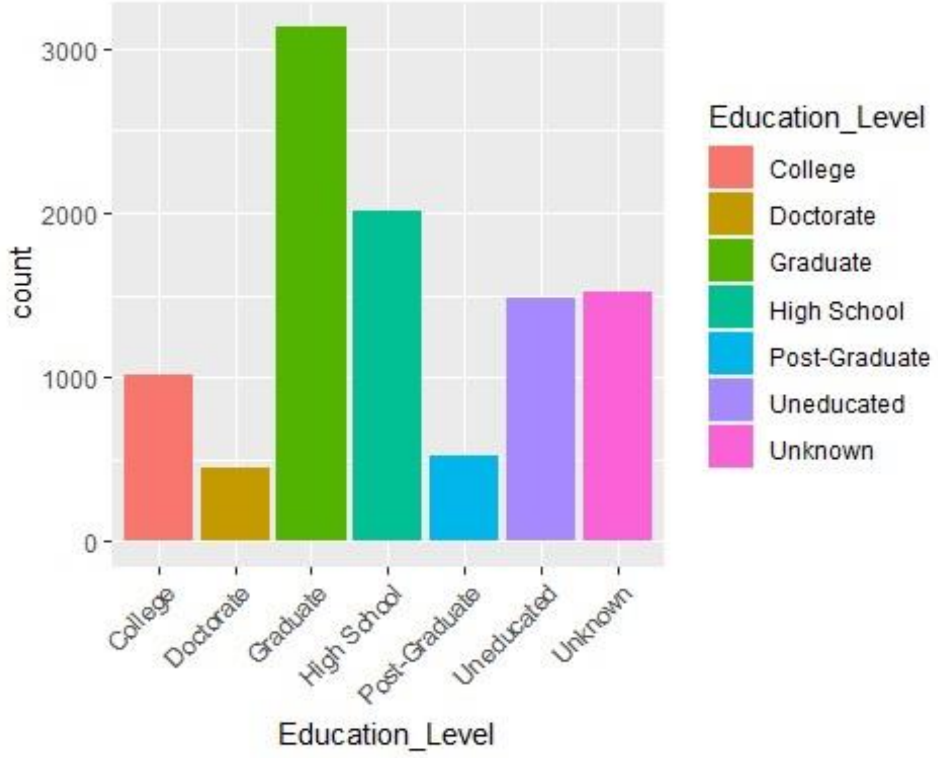
```
ggplot(selected)+geom_bar(aes(x=Gender,fill=Gender))
```



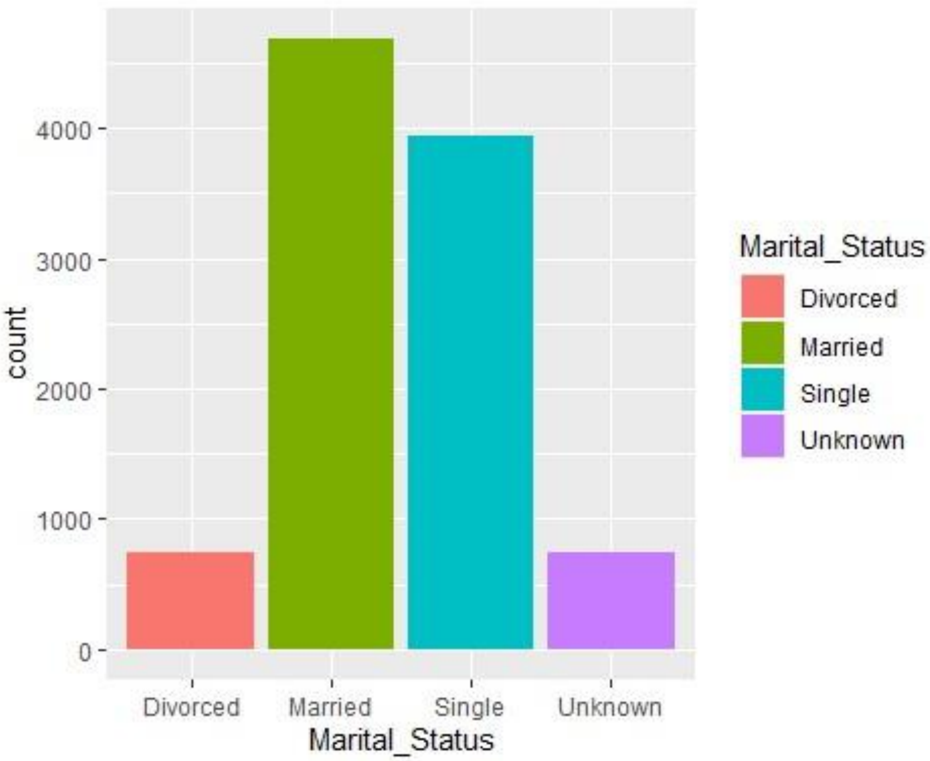
```
ggplot(selected)+geom_bar(aes(x=factor(Dependent_count),fill=factor(Dependent
_count)))
```



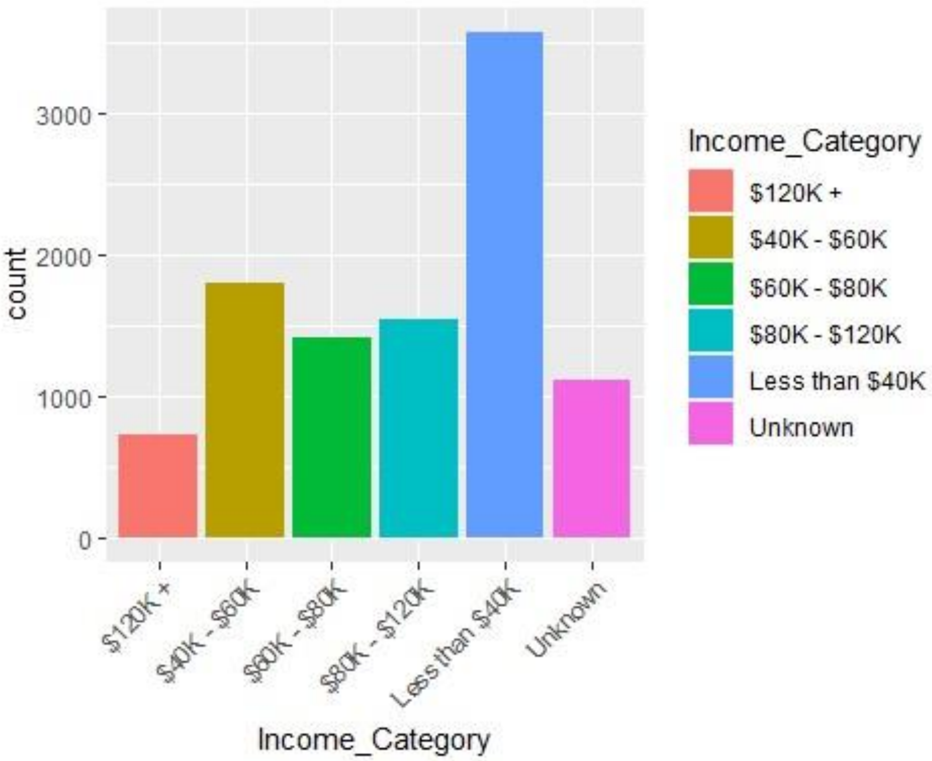
```
ggplot(selected)+geom_bar(aes(x=Education_Level,fill=Education_Level))+
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



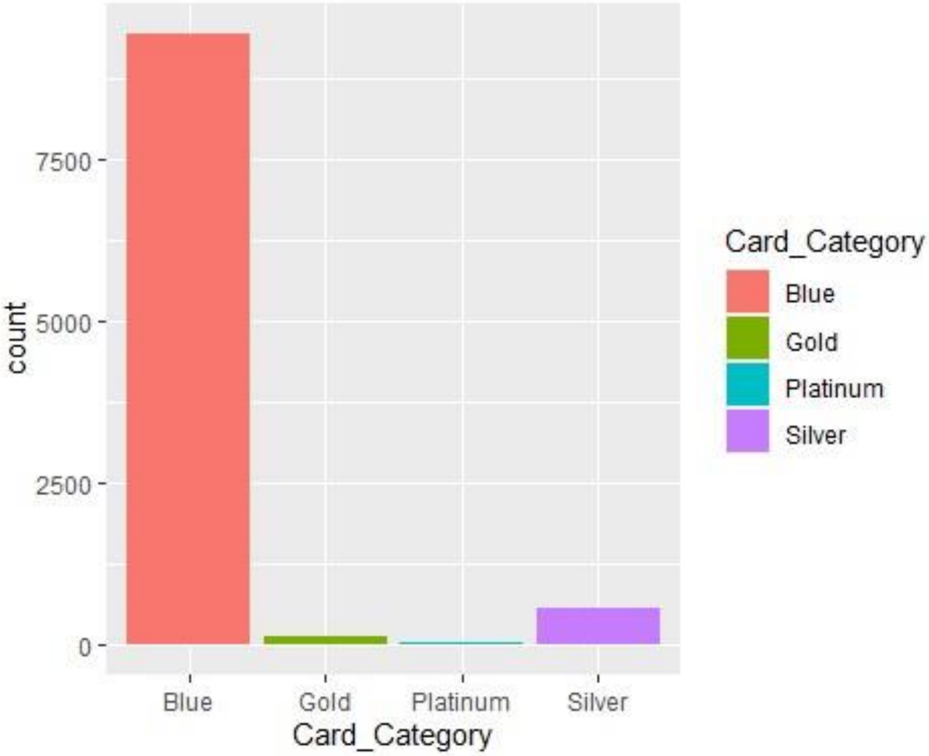
```
ggplot(selected)+geom_bar(aes(x=Marital_Status,fill=Marital_Status))
```



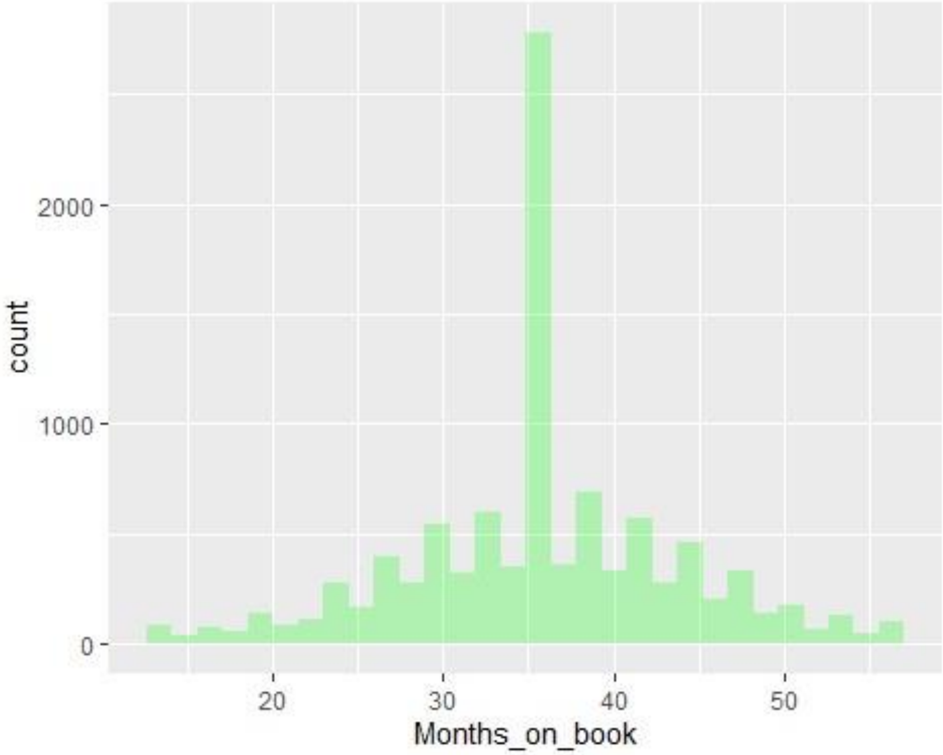
```
ggplot(selected)+geom_bar(aes(x=Income_Category,fill=Income_Category))+
theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



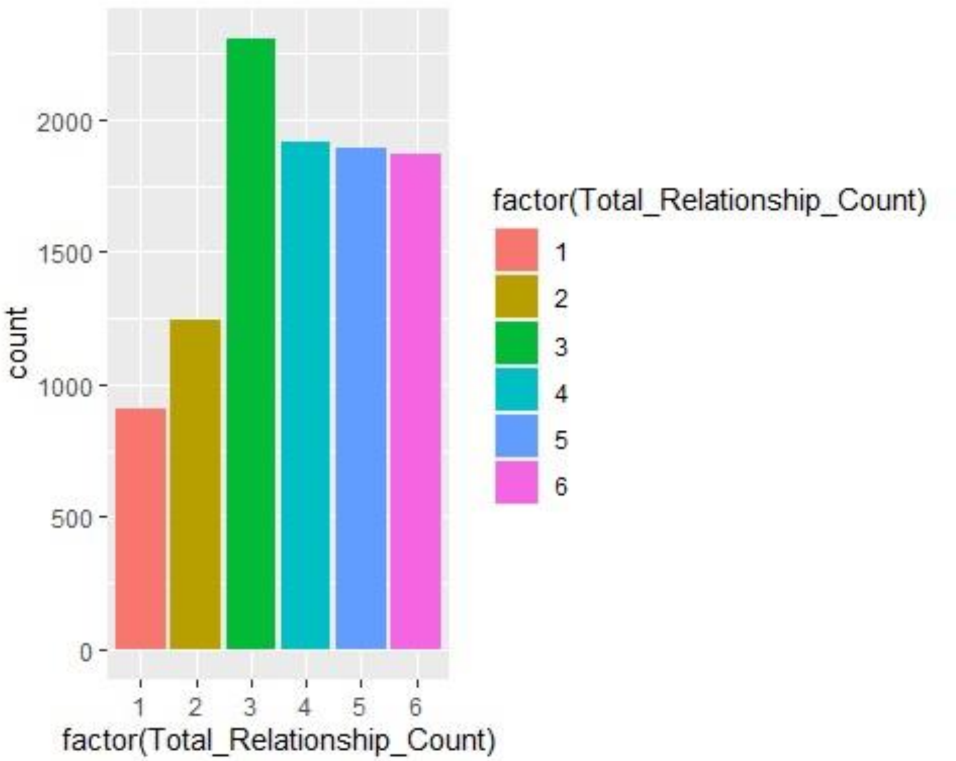
```
ggplot(selected)+geom_bar(aes(x=Card_Category,fill=Card_Category))
```



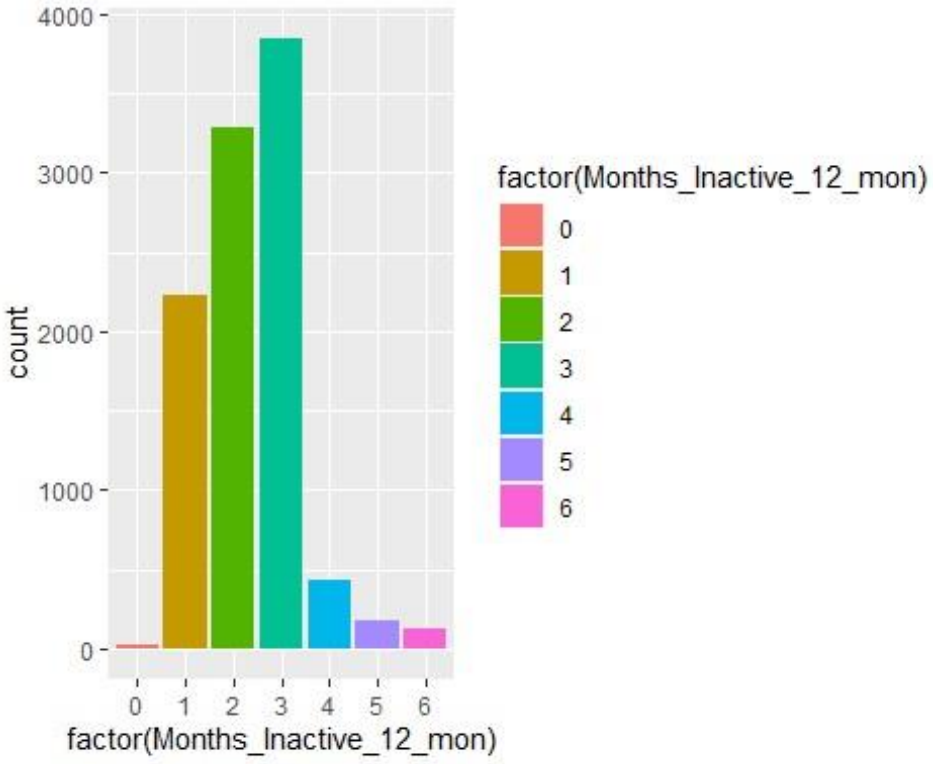
```
ggplot(selected)+geom_histogram(aes(x=Months_on_book),fill='green',alpha =
0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



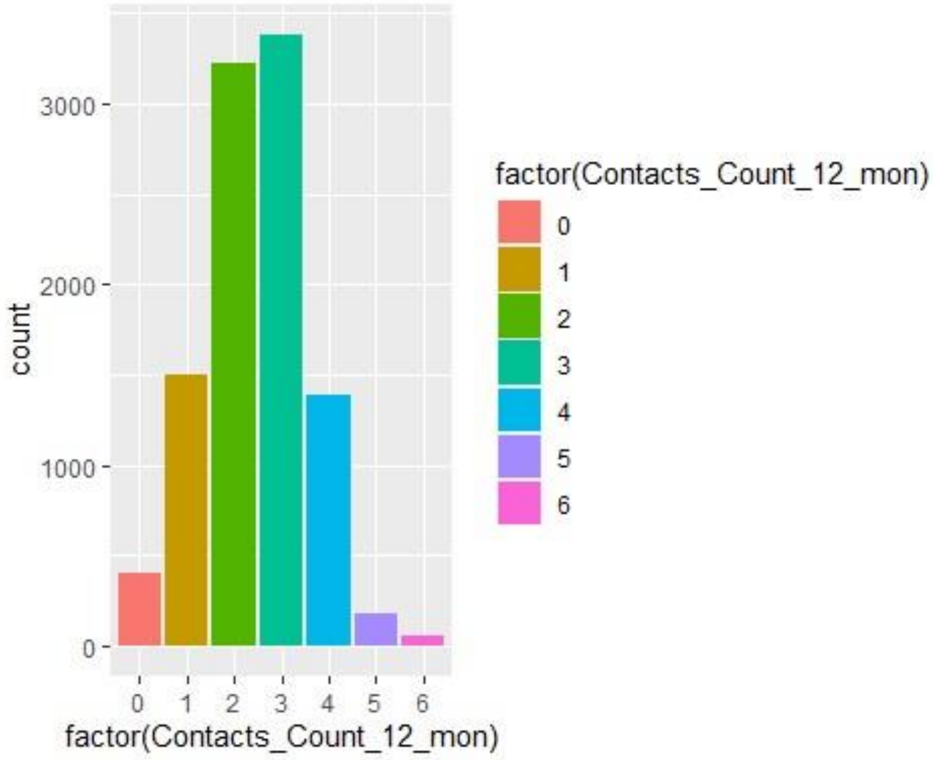
```
ggplot(selected)+geom_bar(aes(x=factor(Total_Relationship_Count),fill=factor(
Total_Relationship_Count)))
```



```
ggplot(selected)+geom_bar(aes(x=factor(Months_Inactive_12_mon),fill=factor(Mo
nths_Inactive_12_mon)))
```

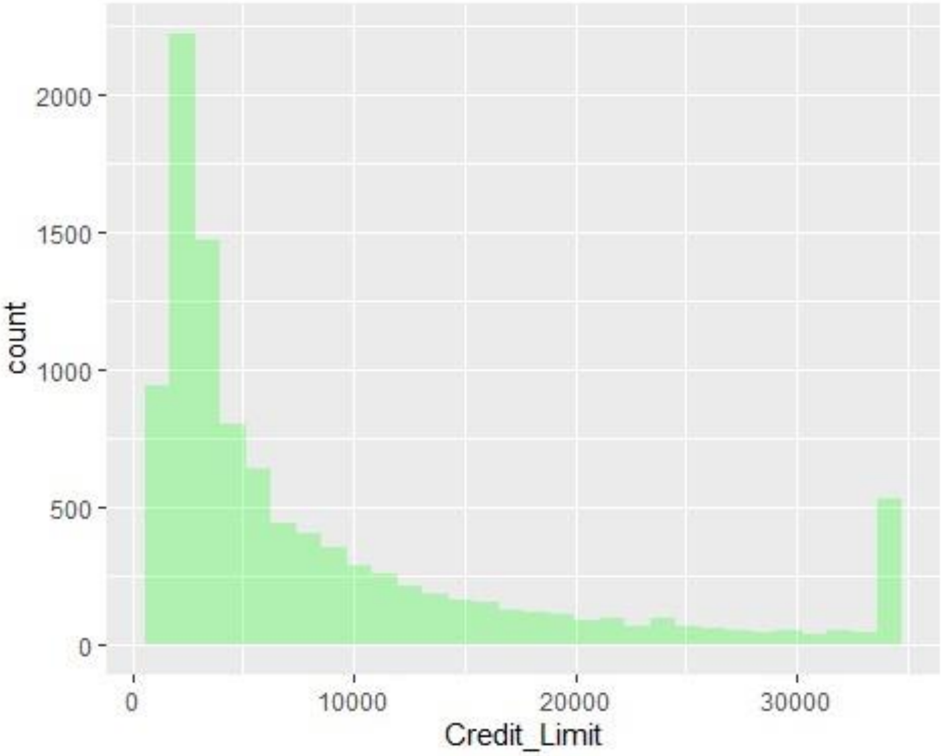


```
ggplot(selected)+geom_bar(aes(x=factor(Contacts_Count_12_mon),fill=factor(Con
tacts_Count_12_mon)))
```



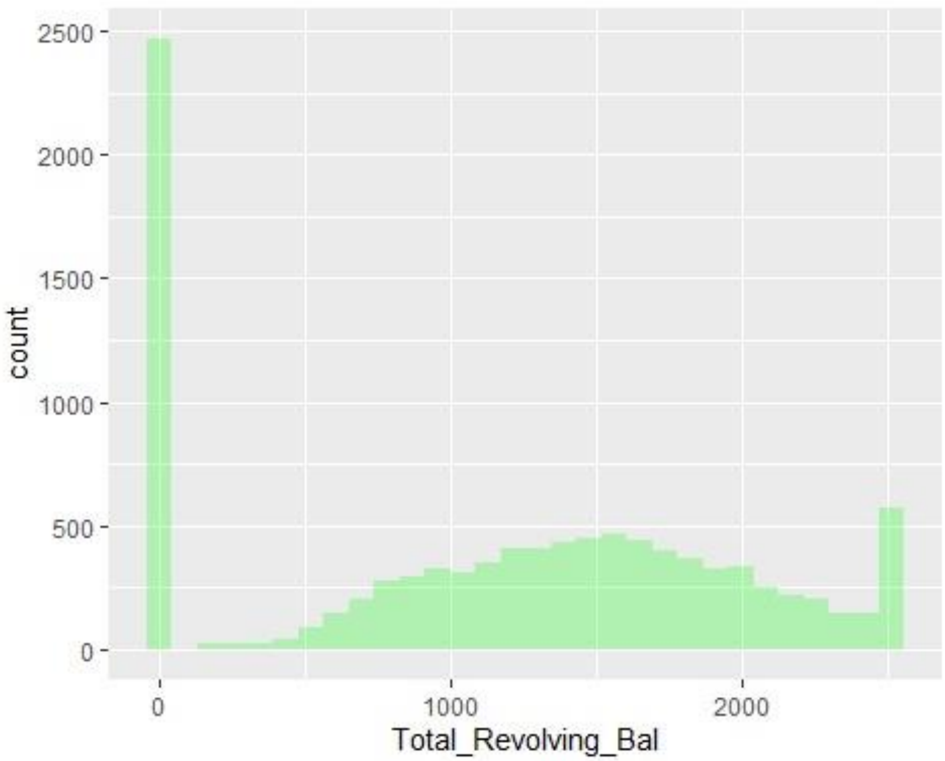
```
ggplot(selected)+geom_histogram(aes(x=Credit_Limit),fill='green',alpha =
0.25)
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

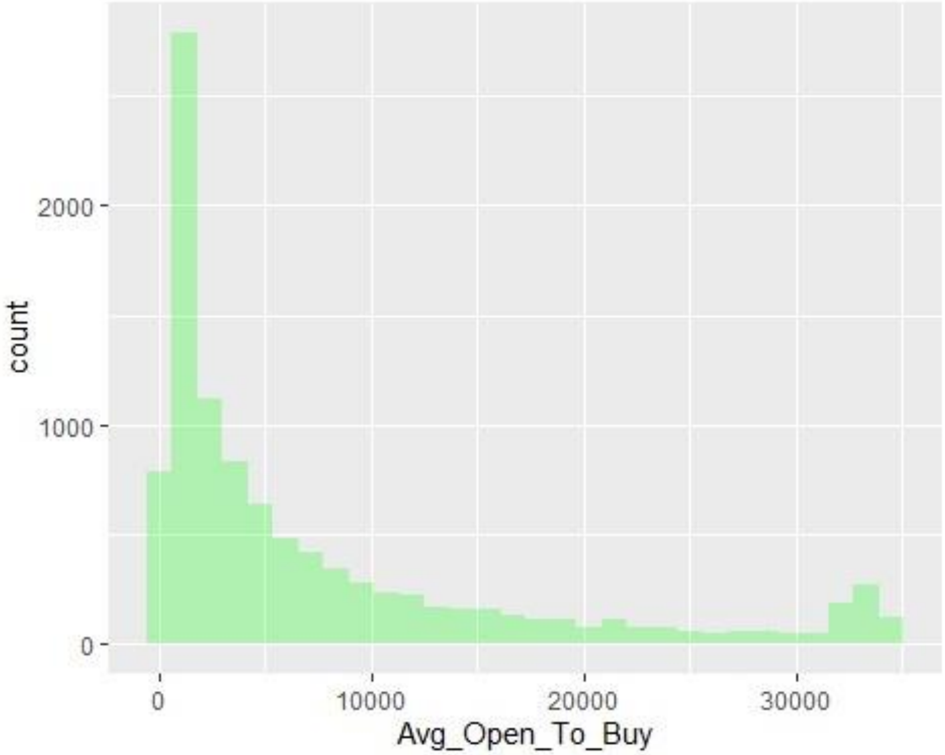


```r
ggplot(selected)+geom_histogram(aes(x=Total_Revolving_Bal),fill='green',alpha = 0.25)
```

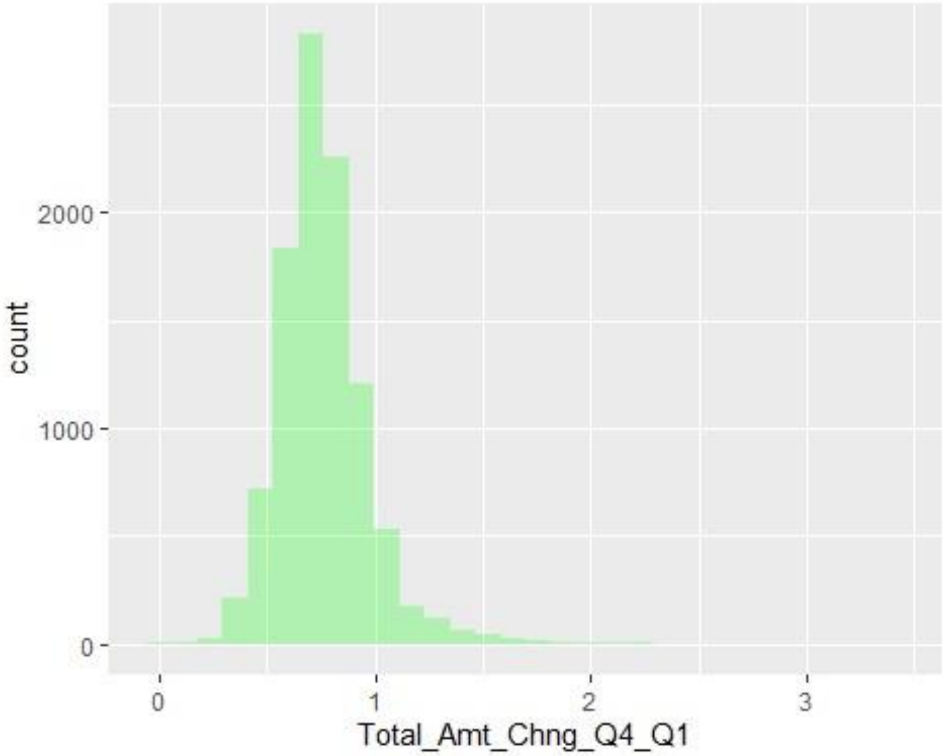
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



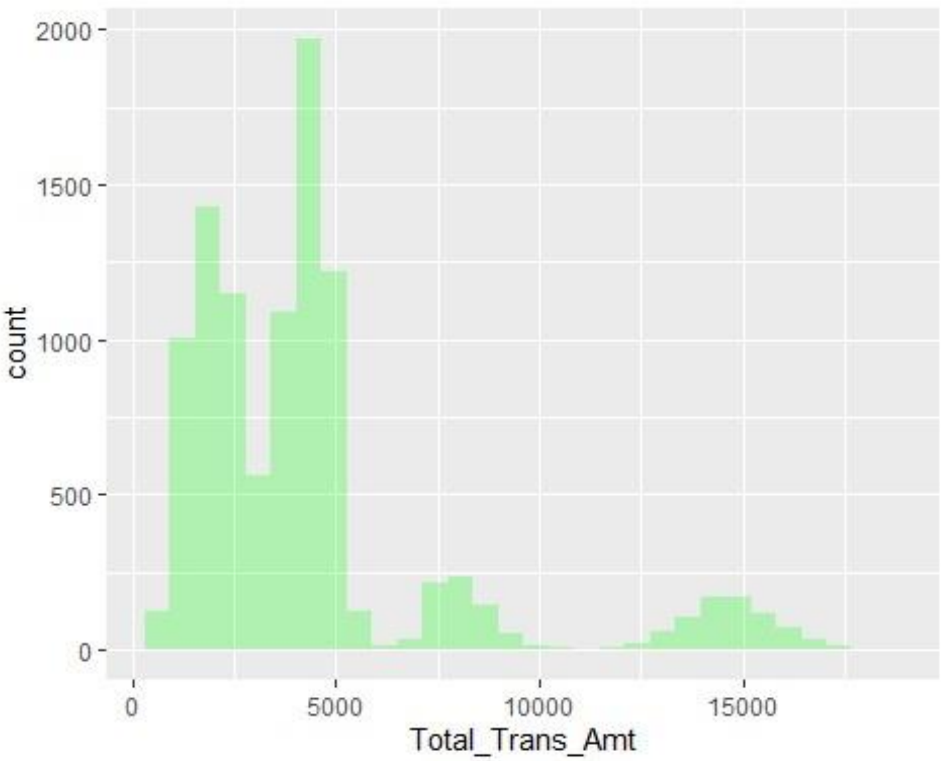
```
ggplot(selected)+geom_histogram(aes(x=Avg_Open_To_Buy),fill='green',alpha =
0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(selected)+geom_histogram(aes(x=Total_Amt_Chng_Q4_Q1),fill='green',alpha = 0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

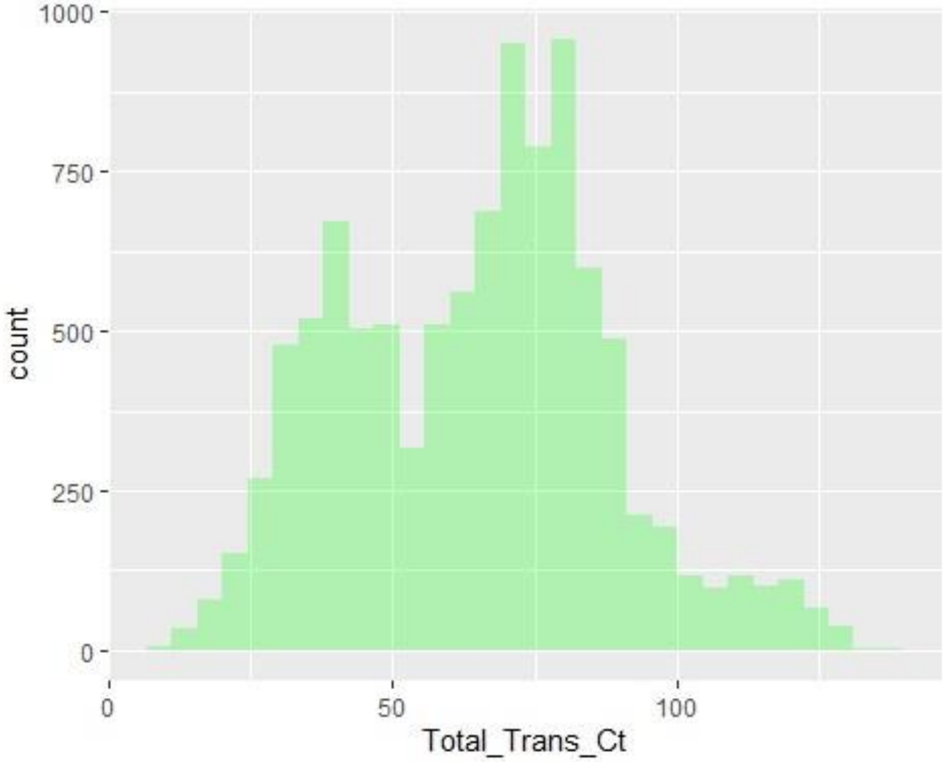


```
ggplot(selected)+geom_histogram(aes(x=Total_Trans_Amt),fill='green',alpha = 0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```

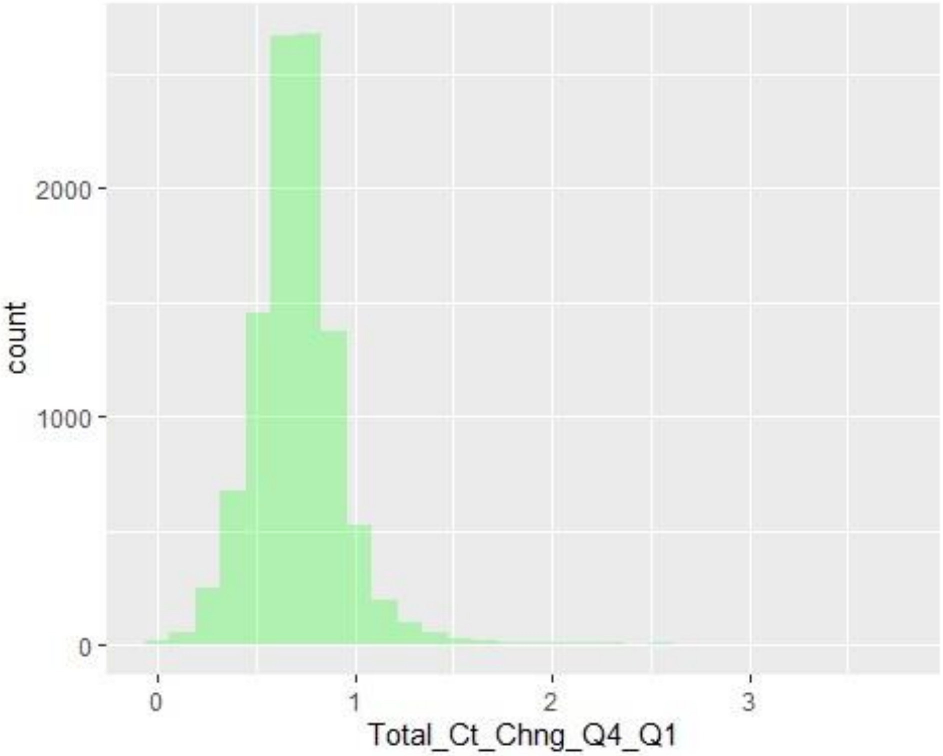


```
ggplot(selected)+geom_histogram(aes(x=Total_Trans_Ct),fill='green',alpha =
0.25)
```

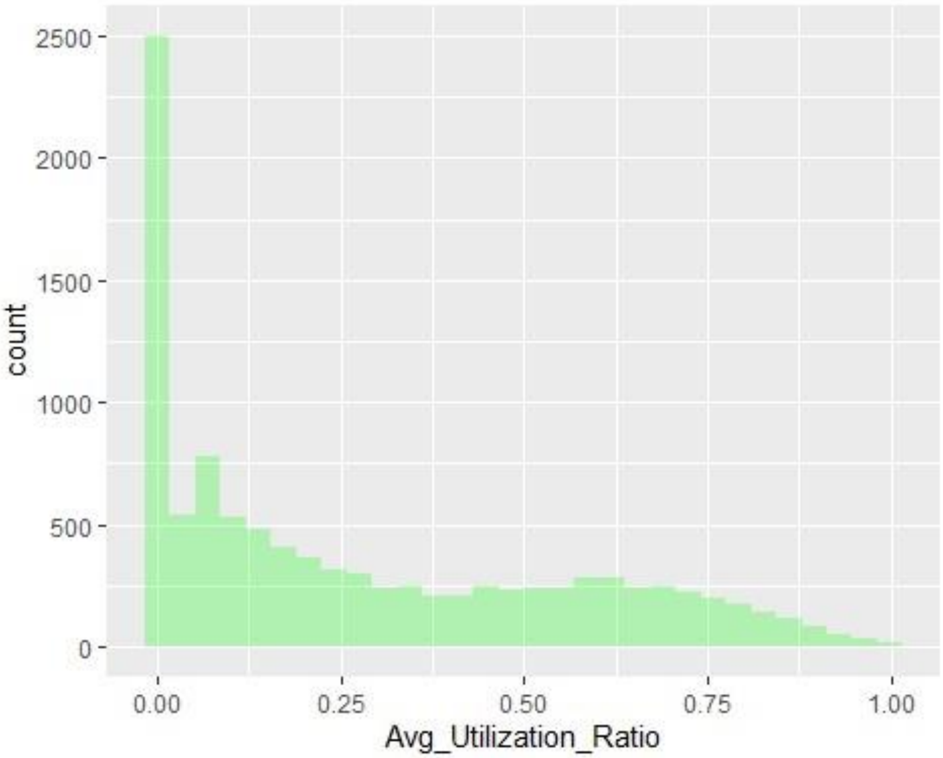
```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(selected)+geom_histogram(aes(x=Total_Ct_Chng_Q4_Q1),fill='green',alpha
= 0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
ggplot(selected)+geom_histogram(aes(x=Avg_Utilization_Ratio),fill='green',alp
ha = 0.25)

## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Based on the above graphs, we did not find any significant problems with the selected variables.It is worth mentioning that for some “Unknow” value in the Income_Category variable, Education_Level variable and Marital_Status variable, We would not do anything with it. Because in all observations, the number of “Unknow” is not very small, so if we directly delete the observations whose Income_Category / Education_Level / marital status is “Unknow”, **a lot of useful other information will be lost**.

In addition, combined with reality, we believe that credit card customers who are unwilling to disclose their income/education level/marital status may have some special reasons when they are surveyed, and these reasons may potentially affect their loyalty. Therefore, we include Unknow as an independent category in subsequent modeling, rather than NA/NaN or delete it.

Next, we use the **summary()** function to see the max/min values and quartiles of the numeric variables for double check and **sum(is.na(selected))** to see if NA is present in the data.

```
summary(selected)
```

```
##  Attrition_Flag      Customer_Age      Gender           Dependent_count
##  Length:10127       Min.   :26.00   Length:10127       Min.   :0.000  
##  Class :character   1st Qu.:41.00   Class :character   1st Qu.:1.000  
##  Mode  :character   Median :46.00   Mode  :character   Median :2.000  
##                     Mean   :46.33                      Mean   :2.346  
##                     3rd Qu.:52.00                      3rd Qu.:3.000  
##                     Max.   :73.00                      Max.   :5.000  
##  Education_Level    Marital_Status     Income_Category    Card_Category
##  Length:10127       Length:10127       Length:10127       Length:10127
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##  Months_on_book  Total_Relationship_Count Months_Inactive_12_mon
##  Min.   :13.00   Min.   :1.000            Min.   :0.000
##  1st Qu.:31.00   1st Qu.:3.000            1st Qu.:2.000
##  Median :36.00   Median :4.000            Median :2.000
##  Mean   :35.93   Mean   :3.813            Mean   :2.341
##  3rd Qu.:40.00   3rd Qu.:5.000            3rd Qu.:3.000
##  Max.   :56.00   Max.   :6.000            Max.   :6.000
##  Contacts_Count_12_mon  Credit_Limit   Total_Revolving_Bal Avg_Open_To_Buy
##  Min.   :0.000         Min.   : 1438   Min.   :   0        Min.   :    3
##  1st Qu.:2.000         1st Qu.: 2555   1st Qu.: 359        1st Qu.: 1324
##  Median :2.000         Median : 4549   Median :1276        Median : 3474
##  Mean   :2.455         Mean   : 8632   Mean   :1163        Mean   : 7469
##  3rd Qu.:3.000         3rd Qu.:11068   3rd Qu.:1784        3rd Qu.: 9859
##  Max.   :6.000         Max.   :34516   Max.   :2517        Max.   :34516
##  Total_Amt_Chng_Q4_Q1 Total_Trans_Amt Total_Trans_Ct   Total_Ct_Chng_Q4_Q1
##  Min.   :0.0000       Min.   :  510   Min.   : 10.00   Min.   :0.0000
##  1st Qu.:0.6310       1st Qu.: 2156   1st Qu.: 45.00   1st Qu.:0.5820
##  Median :0.7360       Median : 3899   Median : 67.00   Median :0.7020
##  Mean   :0.7599       Mean   : 4404   Mean   : 64.86   Mean   :0.7122
##  3rd Qu.:0.8590       3rd Qu.: 4741   3rd Qu.: 81.00   3rd Qu.:0.8180
##  Max.   :3.3970       Max.   :18484   Max.   :139.00   Max.   :3.7140
##  Avg_Utilization_Ratio
##  Min.   :0.0000
##  1st Qu.:0.0230
##  Median :0.1760
##  Mean   :0.2749
##   3rd  Qu.:0.5030
##   Max.    :0.9990
sum(is.na(selected))
## [1] 0
```

With the summary() function, we did not find any exceptions for the maximum/minimum/quantile values of each variable, and there is no NA in the data.

2.5 Processing Categorical variable

Next, we need to convert the Categorical variable in the dataset **from “character” to “factor”**, so as to facilitate subsequent operations to establish a mathematical model.

```
selected$Attrition_Flag=factor(selected$Attrition_Flag,levels=c('Existing
Customer','Attrited Customer'))
selected$Gender=factor(selected$Gender,levels=c('M','F'))
selected$Education_Level=factor(selected$Education_Level,levels=c('College','
Doctorate','Graduate','High School','Post-Graduate','Uneducated','Unknown'))
selected$Marital_Status=factor(selected$Marital_Status,levels=c('Divorced','M
arried','Single','Unknown'))
selected$Income_Category=factor(selected$Income_Category,levels=c('$120K
+','$40K - $60K','$60K - $80K','$80K - $120K','Less than $40K','Unknown'))
selected$Card_Category=factor(selected$Card_Category,levels=c('Blue','Gold','
Platinum','Silver'))
```

2.6 The summary of the new dataset `str(selected)`

```
## tibble [10,127 × 20] (S3: tbl_df/tbl/data.frame)
##  $ Attrition_Flag          : Factor w/ 2 levels "Existing Customer",..: 1
1 1 1 1 1 1 1 1 1 ...
##  $ Customer_Age            : num [1:10127] 45 49 51 40 40 44 51 32 37 48
...
##  $ Gender                  : Factor w/ 2 levels "M","F": 1 2 1 2 1 1 1 1 1
1 ...
##  $ Dependent_count         : num [1:10127] 3 5 3 4 3 2 4 0 3 2 ...
##  $ Education_Level         : Factor w/ 7 levels "College","Doctorate",..:
4 3 3 4 6 3 7 4 6 3 ...
##  $ Marital_Status          : Factor w/ 4 levels "Divorced","Married",..: 2
3 2 4 2 2 2 4 3 3 ...
##  $ Income_Category         : Factor w/ 6 levels "$120K +","$40K -
$60K",..: 3 5 4 5 3 2 1 3 3 4 ...
##  $ Card_Category           : Factor w/ 4 levels "Blue","Gold",..: 1 1 1 1
1 1 2 4 1 1 ...
##  $ Months_on_book          : num [1:10127] 39 44 36 34 21 36 46 27 36 36
...
##  $ Total_Relationship_Count: num [1:10127] 5 6 4 3 5 3 6 2 5 6 ...
##  $ Months_Inactive_12_mon  : num [1:10127] 1 1 1 4 1 1 1 2 2 3 ...
##  $ Contacts_Count_12_mon   : num [1:10127] 3 2 0 1 0 2 3 2 0 3 ...
##  $ Credit_Limit            : num [1:10127] 12691 8256 3418 3313 4716 ...
##  $ Total_Revolving_Bal     : num [1:10127] 777 864 0 2517 0 ...
##  $ Avg_Open_To_Buy         : num [1:10127] 11914 7392 3418 796 4716 ...
##  $ Total_Amt_Chng_Q4_Q1    : num [1:10127] 1.33 1.54 2.59 1.4 2.17 ...
##  $ Total_Trans_Amt         : num [1:10127] 1144 1291 1887 1171 816 ... ##
$ Total_Trans_Ct          : num [1:10127] 42 33 20 20 28 24 31 36 24 32 ...
##  $ Total_Ct_Chng_Q4_Q1     : num [1:10127] 1.62 3.71 2.33 2.33 2.5 ... ##
$ Avg_Utilization_Ratio   : num [1:10127] 0.061 0.105 0 0.76 0 0.311 0.066
0.048 0.113 0.144 ...
```

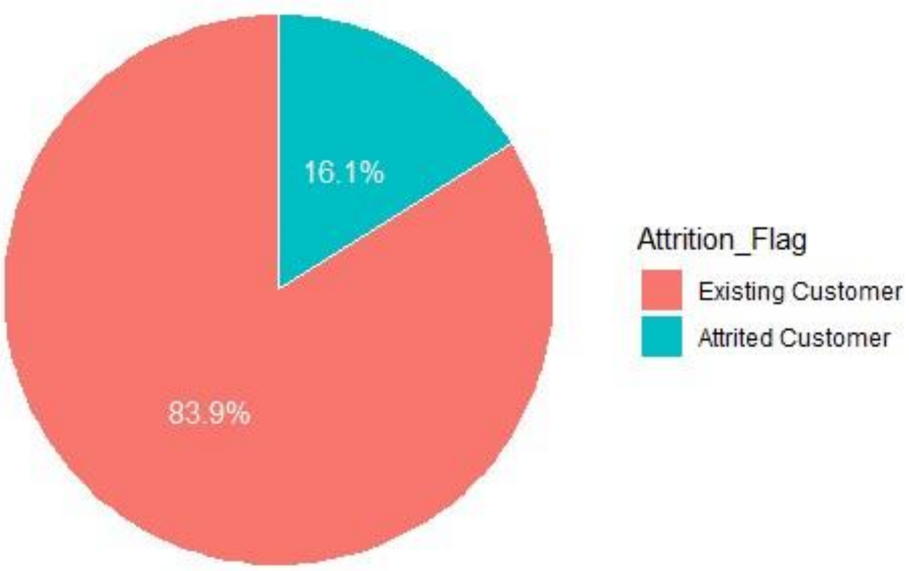
Part 3.0: Exploratory Data Analysis with visulization

Next, we would solve Objective1 to Objective4 by using Exploratory Data Analysis with visulization.

3.1 To analyse the percentage of churning and existing customers.

As the only dependent variable in the data set, we will build a logistic regression model to predict it. Before that, we want to know the distribution of the values of this binary variable.

```
attrition_counts <- count(selected, Attrition_Flag)

ggplot(attrition_counts, aes(x = "", y = n, fill = Attrition_Flag)) +
geom_bar(stat = "identity", width = 1, color = "white") +coord_polar("y",
start = 0) +labs(fill = "Attrition_Flag") +theme_void() + geom_text(aes(label
= paste0(sprintf("%.1f", n/sum(n) * 100), "%")),position =
position_stack(vjust = 0.5),color = "white",size = 4)
```

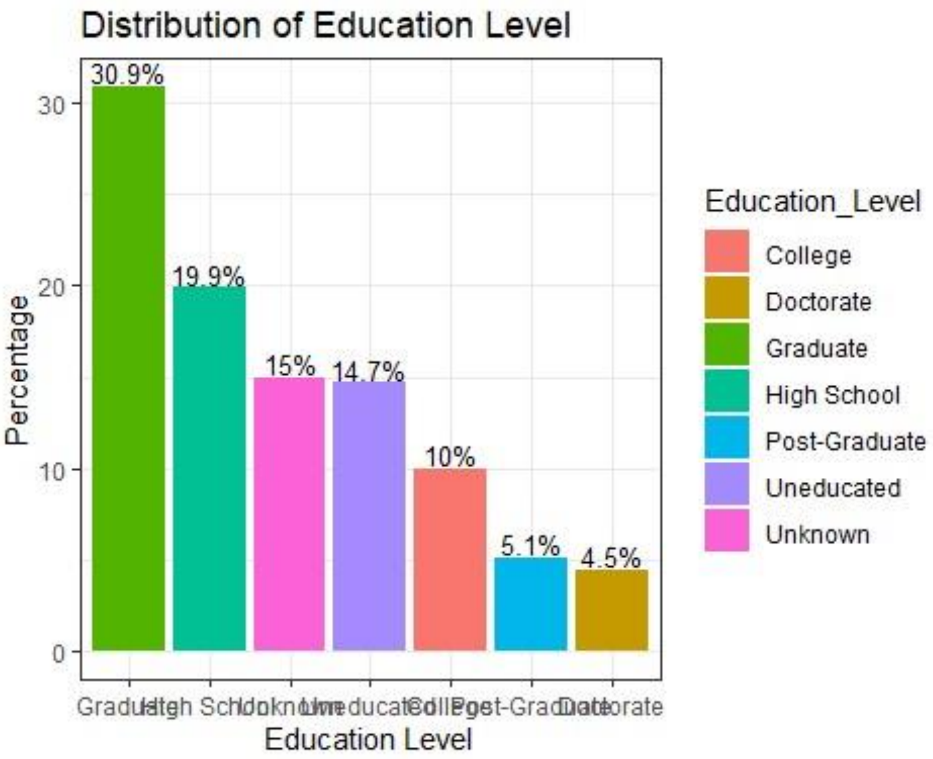


According to the pie chart above, we can see that 16.1% of customers have lost and 83.9% of customers are existing. Although a small percentage of customers are churning, it still signifies that the bank's credit card business has growth potential. By studying the factors influencing churn rate and implementing strategies to attract new customers and retain existing ones, the bank can continue to expand its market share and gain a competitive advantage in the highly competitive credit card industry.

3.2 To determine the number of customers with different levels of education.

Now we understand the distribution of the dependent variable. Next, we would like to know the distribution of customers' education levels.

```
education_counts <- selected %>% count(Education_Level) %>% mutate(percentage
= n / sum(n) * 100)

ggplot(education_counts, aes(x = reorder(Education_Level, -percentage), y =
percentage, fill = Education_Level)) +geom_col() + geom_text(aes(label =
paste0(round(percentage, 1), "%")), vjust = -0.15, size = 3.5) + labs(x =
"Education Level", y = "Percentage", title = "Distribution of Education
Level") + theme_bw()
```



Based on the pie chart above, we can see that among all customers, the number of customers with bachelor's and high school degrees is the highest, accounting for 30.9% and 19.9% of the total number, respectively. On the other hand, the clients with the fewest proportion of people are those who have obtained doctoral degrees, which is in line with our common sense. It is worth mentioning that 15% of clients' educational information is unknown, which may be due to errors in statistical data or clients' unwillingness to disclose their educational information.

Business suggestions:

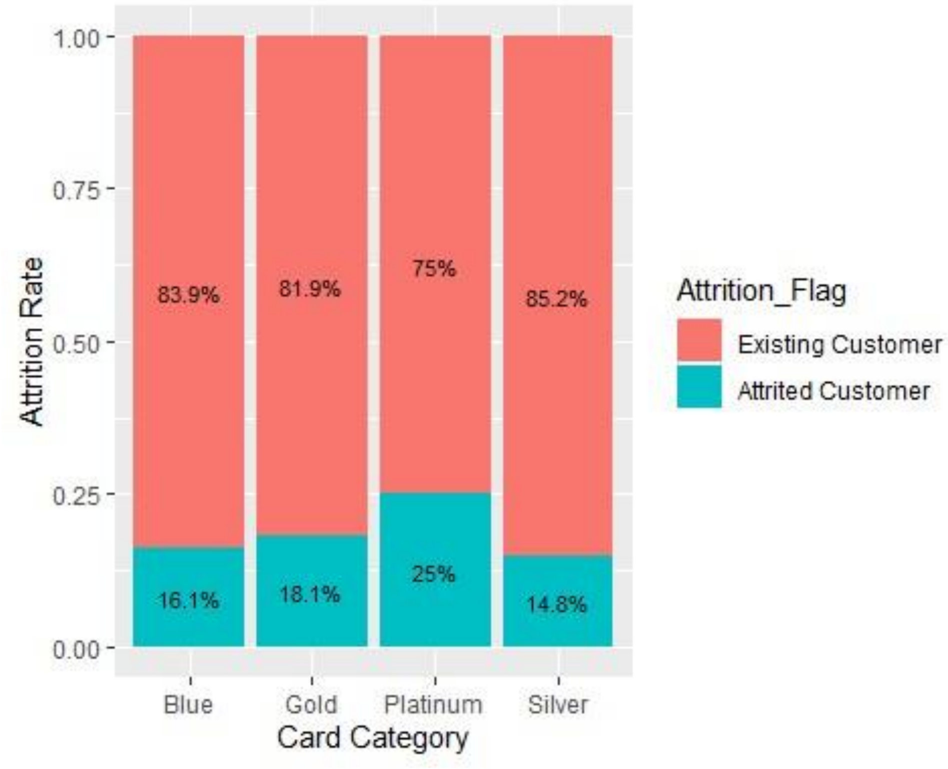
1.For the largest group of customers with a bachelor's degree, the bank can provide more targeted products and services. For example, offering special education loans and startup loans to support their academic or business pursuits. The bank can also organize targeted and personalized activities based on the preferences of customers with a bachelor’s degree.

2.The significant proportion of customers with a high school education and the 14.7% of customers who have not received any education indicate that the bank should not overlook the needs of customers with lower educational backgrounds in their future operations. The bank should provide simplified and easily understandable products and services to ensure that these customers can fully comprehend and utilize credit card products.

3.The unknown educational information of some customers can pose challenges to data analysis and business decision-making. The bank can take measures such as actively collecting educational information or providing incentives to increase customers’ willingness to disclose their educational information, thus obtaining a more accurate and comprehensive customer profile.

3.3 To analyze the churn rate of customers holding different types of credit cards.

```
  group_by(count(selected,Card_Category, Attrition_Flag),Card_Category) %>%
mutate(Attrition_Rate = n / sum(n) * 100) %>%

  ggplot(aes(x = Card_Category, y = Attrition_Rate, fill = Attrition_Flag)) +
geom_col(position = "fill") +   geom_text(aes(label =
paste0(round(Attrition_Rate, 1), "%")),position =  position_fill(vjust =
0.5),size = 3) +labs(x = "Card Category", y = "Attrition Rate")
```



As shown in the chart above, the highest percentage of customer churn was for customers whose credit card was Platinum, at 25%. This is followed by customers with Gold cards. Silver card holders had the lowest churn rate at 14.8%, which is consistent with our intuition that higher-rated credit card customers tend to be more "loyal".

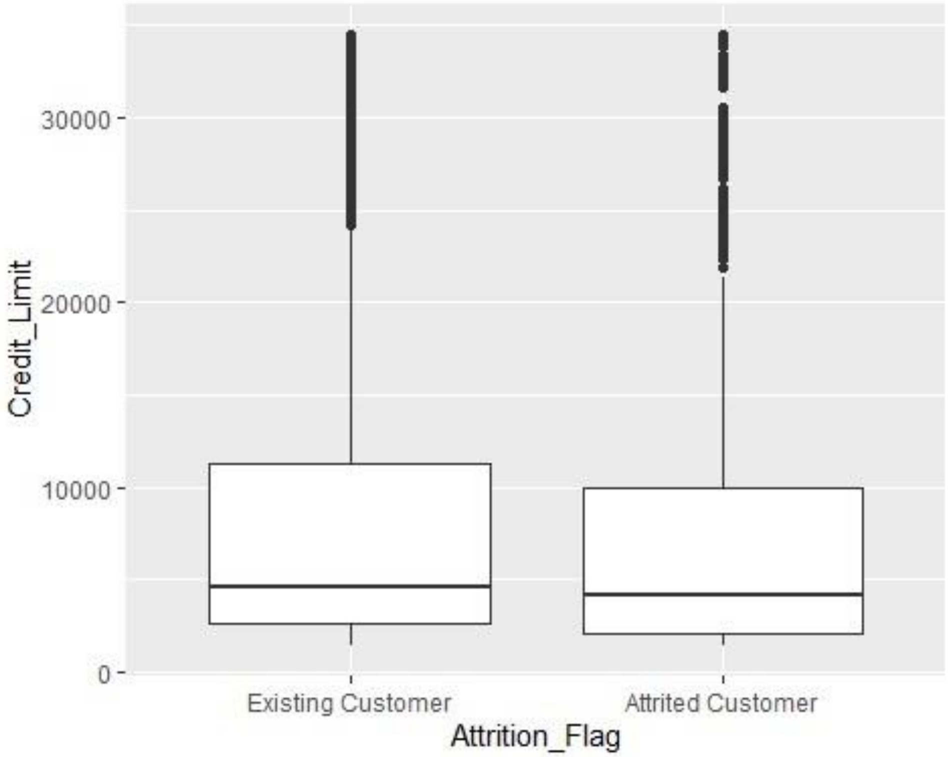
Business suggestions:

1.Focus on Platinum and Gold cardholders: The churn rates for Platinum and Gold cardholders are the highest, at 25% and 18.1% respectively. This indicates that the bank needs to pay special attention to these customer segments and take measures to reduce their churn rates. Consider offering more value-added services, personalized customer support, and more competitive incentives to retain these customers.

2.Maintain satisfaction of Silver cardholders: Silver cardholders have the lowest churn rate at 14.8%. The bank should continue to provide services that align with the expectations and needs of this high-value customer segment. By engaging in regular communication and offering customized benefits, the bank can maintain their satisfaction and loyalty.

3.4 To identify a significant difference between the Line of credit of loss customers and existing customers.

```
ggplot(selected)+geom_boxplot(aes(x=Attrition_Flag,y=Credit_Limit))
```



According to the box chart, we found that the Line of credit of lost customers and non lost customers did not appear to be significantly different. To provide a more accurate answer to this question, we will proceed with hypothesis testing.

H0:There is no significant difference between the Line of credit of lost customers and existing customers.

H1: There is a significant difference between the Line of credit of lost customers and existing customers.

```
churn_customers <- selected[selected$Attrition_Flag == "Attrited Customer", ]
existing_customers <- selected[selected$Attrition_Flag == "Existing
Customer", ] churn_credit_limit <-
mean(churn_customers$Credit_Limit) existing_credit_limit <-
mean(existing_customers$Credit_Limit)

t_test_result <-
t.test(churn_customers$Credit_Limit,
existing_customers$Credit_Limit) print(t_test_result)

##
##  Welch Two Sample t-test
##
```

```
## data:  churn_customers$Credit_Limit and existing_customers$Credit_Limit
## t = -2.401, df = 2290.4, p-value = 0.01643
## alternative hypothesis: true difference in means is not equal to 0 ##
95  percent  confidence  interval: ##    -1073.4010    -108.2751 ##  sample
estimates:
## mean of x mean of y
##  8136.039  8726.878
```

We conducted independent sample t-tests through **t. test()** function. The p-value is 0.01643, which is less than the usual significance level of 0.05. Therefore, we can reject the original assumption that there is no significant difference between the Line of credit of lost customers and that of existing customers. This means that there is a statistically significant difference between the Line of credit of lost customers and existing customers in the analyzed dataset.

Business suggestions:

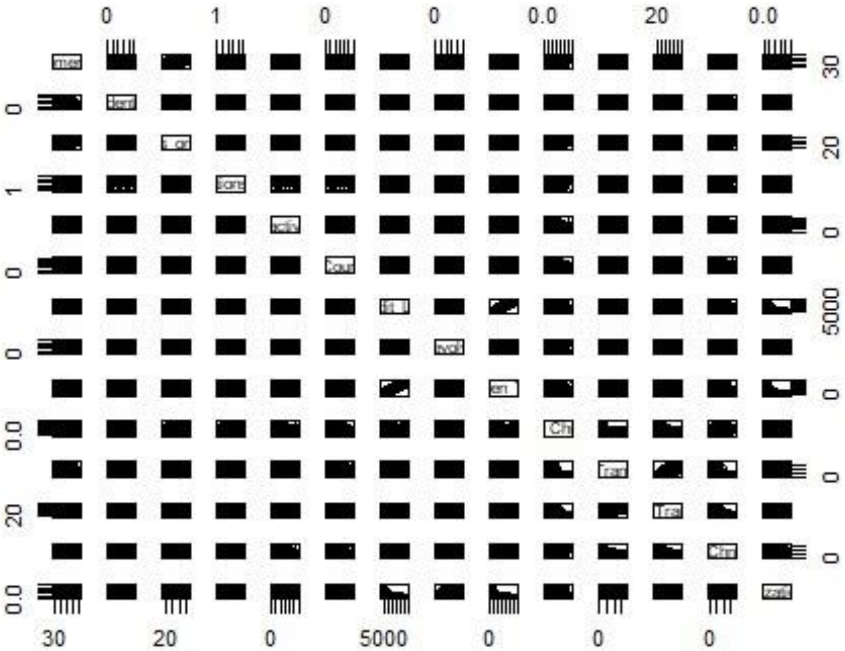
1.Customers with lower credit limits are more prone to churn, which may indicate a higher demand for credit limits. Therefore, the bank can improve its credit risk algorithms to increase customers' credit limits while ensuring the safety of funds, in order to better meet their financial needs.

2.Enhance customized products and services. For customers who are unable to have their credit limits increased, the bank can design customized services to meet their specific needs, such as credit repair solutions or programs to improve their credit scores, shortterm credit limit enhancements, and so on.

Part 4.0: Modelling & Interpretation

4.1 Correlation of Independent variables

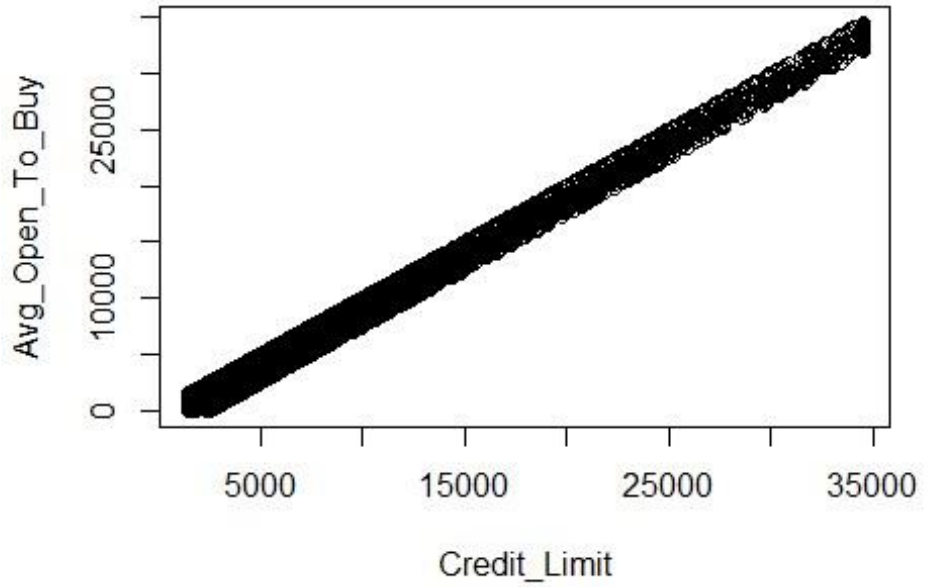
Correlation between the Independent variable, to check any variables are correlated each



other:

The correlation graph reveals a substantial positive relationship among the average opento-buy credit line over the previous 12 months and the credit card credit limit. This means that when the open-to-buy credit line expands, so does the credit limit on the credit card.

Here below we can see the relationship between credit limit and average open to buy more



clearly:

The high positive association among the average open-to-buy credit line over the past 12 months and the credit card credit limit implies that both of these factors tend to change in tandem. The credit limit and open-to-buy credit line on a credit card are each decided by the cardholder's eligibility and risk analysis. Credit card companies are more inclined to grant greater limits on credit and bigger open-to-buy credit lines when the client shows a strong credit record, appropriate financial behaviour, and a small risk of defaulting.

4.2 Logistic Regression Model

4.2.1 Regression Model including all independent variables

```
##
## Call:
## glm(formula = Attrition_Flag ~ ., family = binomial, data = selected)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1157  -0.3618  -0.1703  -0.0673   3.5351   ##
## Coefficients: (1 not defined because of singularities)
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)                    5.830e+00  4.418e-01  13.194  < 2e-16 ***
## Customer_Age                  -6.131e-03  7.711e-03  -0.795 0.426532
## GenderF                        8.938e-01  1.455e-01   6.142 8.16e-10 ***
## Dependent_count                1.358e-01  2.998e-02   4.530 5.89e-06 ***
## Education_LevelDoctorate       3.689e-01  2.081e-01   1.773 0.076218 .
## Education_LevelGraduate       -5.798e-03  1.396e-01  -0.042 0.966864
## Education_LevelHigh School     1.026e-02  1.488e-01   0.069 0.945016
## Education_LevelPost-Graduate   3.112e-01  2.050e-01   1.518 0.128952
## Education_LevelUneducated      6.955e-02  1.573e-01   0.442 0.658477
## Education_LevelUnknown         1.329e-01  1.554e-01   0.855 0.392310
## Marital_StatusMarried         -4.994e-01  1.544e-01  -3.234 0.001219 **
## Marital_StatusSingle           1.081e-01  1.549e-01   0.698 0.485248
## Marital_StatusUnknown          4.528e-02  1.962e-01   0.231 0.817467
## Income_Category$40K - $60K    -9.083e-01  2.026e-01  -4.484 7.33e-06 ***
## Income_Category$60K - $80K    -6.405e-01  1.791e-01  -3.576 0.000349 ***
## Income_Category$80K - $120K   -2.983e-01  1.663e-01  -1.794 0.072811 .
## Income_CategoryLess than $40K -7.702e-01  2.190e-01  -3.516 0.000438 ***
## Income_CategoryUnknown        -8.321e-01  2.322e-01  -3.584 0.000338 ***
## Card_CategoryGold              1.066e+00  3.521e-01   3.026 0.002475 **
## Card_CategoryPlatinum          9.816e-01  6.813e-01   1.441 0.149654
## Card_CategorySilver            4.502e-01  1.962e-01   2.294 0.021778 *
## Months_on_book                -4.685e-03  7.673e-03  -0.611 0.541484
## Total_Relationship_Count      -4.493e-01  2.750e-02 -16.338  < 2e-16 ***
## Months_Inactive_12_mon         5.078e-01  3.793e-02  13.387  < 2e-16 ***
## Contacts_Count_12_mon          5.133e-01  3.655e-02  14.044  < 2e-16 ***
## Credit_Limit                  -1.971e-05  6.860e-06  -2.873 0.004064 **
## Total_Revolving_Bal           -9.321e-04  7.207e-05 -12.934  < 2e-16 ***
## Avg_Open_To_Buy                       NA         NA      NA       NA
## Total_Amt_Chng_Q4_Q1          -4.262e-01  1.878e-01  -2.269 0.023253 *
## Total_Trans_Amt                4.855e-04  2.295e-05  21.154  < 2e-16 ***
## Total_Trans_Ct                -1.192e-01  3.731e-03 -31.944  < 2e-16 ***
## Total_Ct_Chng_Q4_Q1           -2.798e+00  1.889e-01 -14.813  < 2e-16 ***
## Avg_Utilization_Ratio        -1.253e-01  2.470e-01  -0.507 0.612020    ##
---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 ##
## (Dispersion parameter for binomial family taken to be 1)
##
```

```
##     Null deviance: 8927.2  on 10126  degrees of freedom ## Residual
deviance: 4710.6  on 10095  degrees of freedom
## AIC: 4774.6
##
## Number of Fisher Scoring iterations: 6
```

In this framework, the Attrition Flag (DV)' elements are categorised as existing customers(1) and churning customers(2). According to this result, the value of the coefficient of Number of dependents is 0.0138, indicating that the independent variable has a greater effect on churned consumers than on existing customers. For each rise in the number of dependents for clients, the likelihood of a client being a churned customer increases by 0.0138, where opposed to an existing customer which is statistically substantial.

The coefficient of 10.66 for the gold card group indicates that clients with a gold card are more likely to be churned than clients with a blue card (which acts as the reference group in this example). A positive coefficient implies a higher likelihood of being a churned client as when compared with the reference group. As a result, we may conclude that having a gold card is associated with being a churned client.

The -27.98 coefficient represents the effect of a change in transactions count on becoming a churning client vs a maintained customer. Because the coefficient is negative, clients who have increased their transaction count over this period are less likely to churn than customers who have decreased or decreased their transaction count. This research implies that increased engagement or activity, as seen by a rise in transactions, may be a barrier against churn. Customers who regularly use their credit card services and have a positive change in their transaction count are more unlikely to abandon the service.

The variables “client education level”, “average utilization ratio”, “month on book”, and “marital status” seem to have p-values larger than 0.05 in your regression model, suggesting that they're not statistically significant. This signifies that there's inadequate proof to reject the null hypothesis, implying that these factors have no effect on the churning customer. The variable “average open to buy” being reported as “NA” in the model of regression indicates that it is substantial correlation with the credit limit variable. So, we do exclude those insignificant variables and correlated variable from the regression model.

4.2.2 Regression Model exclude few independent variables

Exclude insignificant variables as well as correlated variable:

```
##
## Call:
## glm(formula = Attrition_Flag ~ Gender + Dependent_count + Income_Category
+
##     Card_Category + Total_Relationship_Count + Months_Inactive_12_mon +
##     Contacts_Count_12_mon + Credit_Limit + Total_Revolving_Bal +
##     Total_Amt_Chng_Q4_Q1 + Total_Trans_Amt + Total_Trans_Ct +
##     Total_Ct_Chng_Q4_Q1, family = binomial, data = selected) ##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max   ##
-3.0469  -0.3671  -0.1755  -0.0710   3.5002
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     5.052e+00  3.139e-01  16.095  < 2e-16 ***
## GenderF                         8.658e-01  1.443e-01   6.002 1.95e-09 ***
## Dependent_count                 1.400e-01  2.944e-02   4.757 1.97e-06 ***
## Income_Category$40K - $60K    -8.205e-01  2.002e-01  -4.099 4.15e-05 ***
## Income_Category$60K - $80K    -5.723e-01  1.772e-01  -3.230  0.00124 **
## Income_Category$80K - $120K   -2.604e-01  1.645e-01  -1.583  0.11334
## Income_CategoryLess than $40K -6.871e-01  2.165e-01  -3.174  0.00150 **
## Income_CategoryUnknown        -7.299e-01  2.298e-01  -3.176  0.00149 **
## Card_CategoryGold               1.028e+00  3.518e-01   2.921  0.00349 **
## Card_CategoryPlatinum           1.039e+00  6.827e-01   1.522  0.12812
## Card_CategorySilver             4.767e-01  1.938e-01   2.460  0.01390 *
## Total_Relationship_Count      -4.523e-01  2.730e-02 -16.571  < 2e-16 ***
## Months_Inactive_12_mon          4.961e-01  3.723e-02  13.326  < 2e-16 ***
## Contacts_Count_12_mon           5.079e-01  3.613e-02  14.056  < 2e-16 ***
## Credit_Limit                  -1.603e-05  6.204e-06  -2.583  0.00978 **
## Total_Revolving_Bal           -9.820e-04  4.606e-05 -21.319  < 2e-16 ***
## Total_Amt_Chng_Q4_Q1          -4.148e-01  1.834e-01  -2.262  0.02371 *
## Total_Trans_Amt                 4.728e-04  2.250e-05  21.010  < 2e-16 ***
## Total_Trans_Ct                -1.149e-01  3.627e-03 -31.673  < 2e-16 ***
## Total_Ct_Chng_Q4_Q1           -2.820e+00  1.874e-01 -15.048  < 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: 8927.2  on 10126  degrees of freedom
## Residual deviance: 4783.1  on 10107  degrees of freedom ##
AIC: 4823.1
##
## Number of Fisher Scoring iterations: 6
```

The greatest coefficient is -28.20 for fluctuations in changes of total number of transactions. A huge negative value indicates that a drop in changes of the total number of transactions is related with a considerable rise in the likelihood of being a churning client. This means that a decrease in the change of total number of transactions is significantly linked to an increased chance of customer turnover.The greatest coefficient is -2.820e+00 for change in total number of transactions. A big negative value indicates that a drop in the total number of transactions is related with a considerable rise in the likelihood of being a churning client. This means that a decrease in the total number of transactions is significantly linked to an increased chance of customer turnover. Total transaction count has a -0.1149 coefficient; a lower total number of transactions is associated with a reduced likelihood of a customer churning or quitting the organisation. Total revolving balance has a -0.0009820 coefficient, which suggests that larger revolving balances or credit card debt are connected with a decreased risk of customers switching.

Total transaction count and total revolving balance variables have an essential part in determining client churn by considering their coefficients within the setting of the churning customer. Clients are less likely to churn or leave a firm if the overall transaction count is lower and the revolving balance is larger.

PART 5.0: Resampling Methods

5.1 Fitting Classification Tree

5.1.1 Adding “Churn” variable in our dataset

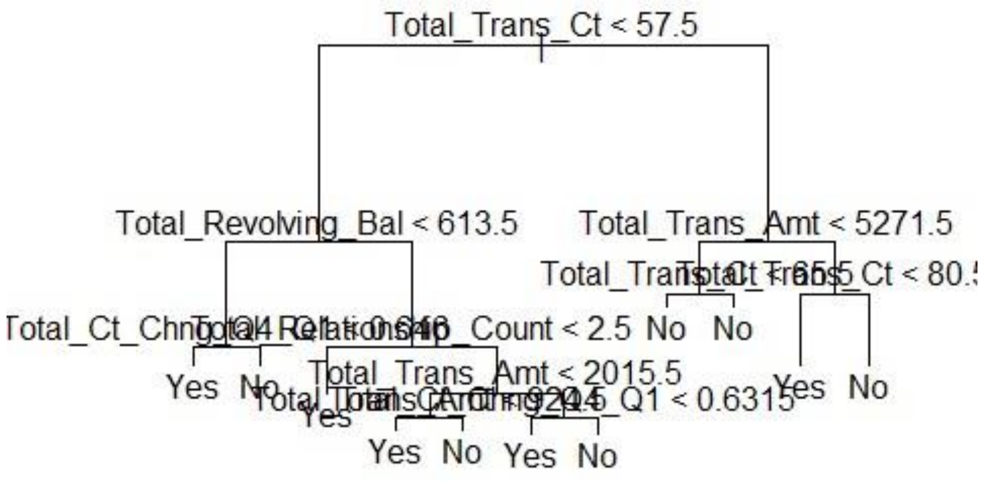
From the logistic regression, we going to predict how many % that we truly predict the churn customer

```
##  Factor w/ 2 levels "Existing Customer",..: 1 1 1 1 1 1 1 1 1 1 ...

## Attrition_Flag
## Existing Customer Attrited Customer
##              8500              1627

## Churn
##   No  Yes
## 8500 1627

## 'data.frame':    10127 obs. of  21 variables:
##  $ Attrition_Flag          : Factor w/ 2 levels "Existing Customer",..: 1 1
1 1 1 1 1 1 1 1 ...
##  $ Customer_Age            : num  45 49 51 40 40 44 51 32 37 48 ...
##  $ Gender                  : Factor w/ 2 levels "M","F": 1 2 1 2 1 1 1 1 1
1 ...
##  $ Dependent_count         : num  3 5 3 4 3 2 4 0 3 2 ...
##  $ Education_Level         : Factor w/ 7 levels "College","Doctorate",..:
4 3 3 4 6 3 7 4 6 3 ...
##  $ Marital_Status          : Factor w/ 4 levels "Divorced","Married",..: 2
3 2 4 2 2 2 4 3 3 ...
##  $ Income_Category         : Factor w/ 6 levels "$120K +","$40K - $60K",..:
3 5 4 5 3 2 1 3 3 4 ...
##  $ Card_Category           : Factor w/ 4 levels "Blue","Gold",..: 1 1 1 1
1 1 2 4 1 1 ...
##  $ Months_on_book          : num  39 44 36 34 21 36 46 27 36 36 ...
##  $ Total_Relationship_Count: num  5 6 4 3 5 3 6 2 5 6 ...
##  $ Months_Inactive_12_mon  : num  1 1 1 4 1 1 1 2 2 3 ...
##  $ Contacts_Count_12_mon   : num  3 2 0 1 0 2 3 2 0 3 ...
##  $ Credit_Limit            : num  12691 8256 3418 3313 4716 ...
##  $ Total_Revolving_Bal     : num  777 864 0 2517 0 ...
##  $ Avg_Open_To_Buy         : num  11914 7392 3418 796 4716 ...
##  $ Total_Amt_Chng_Q4_Q1    : num  1.33 1.54 2.59 1.4 2.17 ...
##  $ Total_Trans_Amt         : num  1144 1291 1887 1171 816 ...
##  $ Total_Trans_Ct          : num  42 33 20 20 28 24 31 36 24 32 ...
##  $ Total_Ct_Chng_Q4_Q1     : num  1.62 3.71 2.33 2.33 2.5 ...
##  $ Avg_Utilization_Ratio   : num  0.061 0.105 0 0.76 0 0.311 0.066 0.048
0.113 0.144 ...
##  $ Churn                   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1
1 1 1 ...
```

We can see from the table above that the attrition flag variable is recoded to the Churn variable, and if the customer is churning, we recode it as Yes under the Churn variable, and if the customer is still existing, we recognise it as No under the Churn variable.

5.1.2 Decision Tree

```
##
## Classification tree:
## tree(formula = Churn ~ . - Attrition_Flag, data = selected) ## Variables
actually used in tree construction:
## [1] "Total_Trans_Ct"           "Total_Revolving_Bal"      ## [3]
"Total_Ct_Chng_Q4_Q1"      "Total_Relationship_Count"
## [5] "Total_Trans_Amt"
## Number of terminal nodes:  11
## Residual mean deviance:  0.392 = 3966 / 10120
## Misclassification error rate: 0.08097 = 820 / 10127
```



```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 10127 8927.00 No ( 0.8393404 0.1606596 )
##    2) Total_Trans_Ct < 57.5 3736 4890.00 No ( 0.6381156 0.3618844 )
##      4) Total_Revolving_Bal < 613.5 1210 1454.00 Yes ( 0.2884298 0.7115702
)
##        8) Total_Ct_Chng_Q4_Q1 < 0.646 854  803.00 Yes ( 0.1791569
0.8208431 ) *
##        9) Total_Ct_Chng_Q4_Q1 > 0.646 356  489.90 No ( 0.5505618 0.4494382
) *
##      5) Total_Revolving_Bal > 613.5 2526 2488.00 No ( 0.8056215 0.1943785
)
```

```
##       10) Total_Relationship_Count < 2.5 224  235.30 Yes ( 0.2187500
0.7812500 ) *
##       11) Total_Relationship_Count > 2.5 2302 1842.00 No ( 0.8627281
0.1372719 )
##         22) Total_Trans_Amt < 2015.5 1550  703.60 No ( 0.9400000 0.0600000
)
##           44) Total_Trans_Amt < 924.5 23   17.81 Yes ( 0.1304348 0.8695652
) *
##           45) Total_Trans_Amt > 924.5 1527  586.40 No ( 0.9521938
0.0478062 ) *
##         23) Total_Trans_Amt > 2015.5 752  914.30 No ( 0.7034574 0.2965426
)
##           46) Total_Ct_Chng_Q4_Q1 < 0.6315 315  435.80 Yes ( 0.4730159
0.5269841 ) *
##           47) Total_Ct_Chng_Q4_Q1 > 0.6315 437  338.40 No ( 0.8695652
0.1304348 ) *
##    3) Total_Trans_Ct > 57.5 6391 2268.00 No ( 0.9569707 0.0430293 )
##      6) Total_Trans_Amt < 5271.5 4831  491.70 No ( 0.9910992 0.0089008 )
##       12) Total_Trans_Ct < 65.5 959  332.50 No ( 0.9582899 0.0417101 ) *
##       13) Total_Trans_Ct > 65.5 3872   48.98 No ( 0.9992252 0.0007748 ) *
##      7) Total_Trans_Amt > 5271.5 1560 1312.00 No ( 0.8512821 0.1487179 )
##       14) Total_Trans_Ct < 80.5 303  387.10 Yes ( 0.3366337 0.6633663 ) *
##       15) Total_Trans_Ct > 80.5 1257  290.80 No ( 0.9753381 0.0246619 ) *
```

The tree of classification study reveals the characteristics that have the greatest influence on customer attrition. Total transaction count, total revolving balance, total change in transaction count, total number of credit card products owned by clients, and the total amount of transactions are the variables used to build the tree. The structure of the tree is made up of 11 terminal nodes that represent various client categories or groupings. The tree is developed, with each split depending on a different variable and threshold. The splits are obtained by identifying the factors and thresholds that result in the best prediction of the outcome (churn or no churn). The complete dataset of 10,127 records is taken into account. The predominant class is “No churn” (83.93%), implying that the vast majority of the consumers in the sample did not churn. We understand that the initial split is based on total transaction count, indicating that this variable has a significant influence on forecasting customer attrition. Customers having a low transaction count, a large revolving balance, and a low transaction count change. This node’s majority group is “Existing customer” (82.08%), suggesting that these consumers are more unlikely to leave. At the same time, clients with fewer transactions, a greater outstanding load, and fewer credit cards are more likely to leave the credit card service (21.87%). The classification tree highlights critical characteristics in forecasting customer turnover as total transaction count, revolving balance, change in transaction count, and total relationship count. The tree illustrates how these factors and their respective thresholds divide the data set into various sections, each with a different chance of churning.

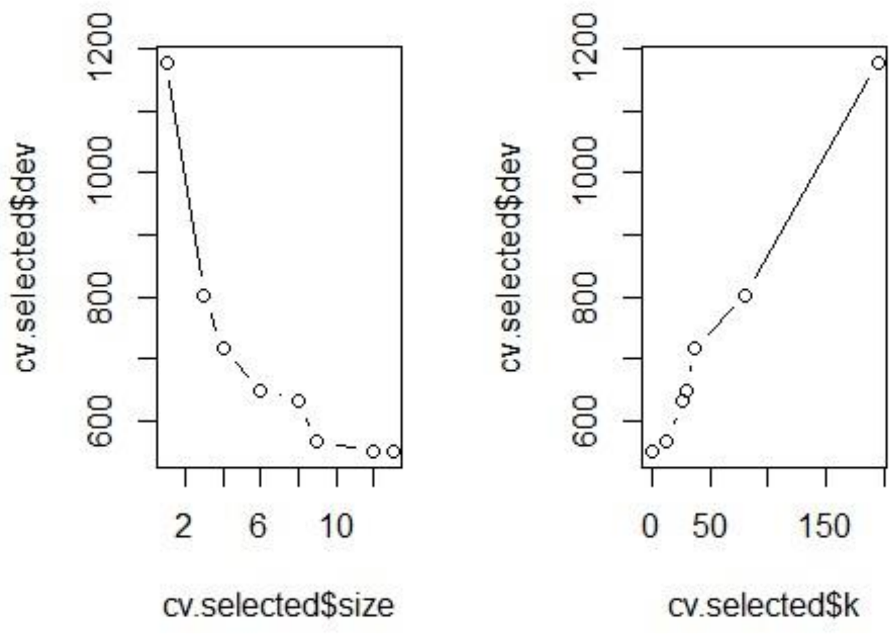
5.1.3 Training & Test set

```
## [1] 10127

##           Churn.test
## tree.pred   No  Yes
##        No  2482  109
##        Yes  106  341

## [1] 0.9292298

## [1] 0.07077024
```

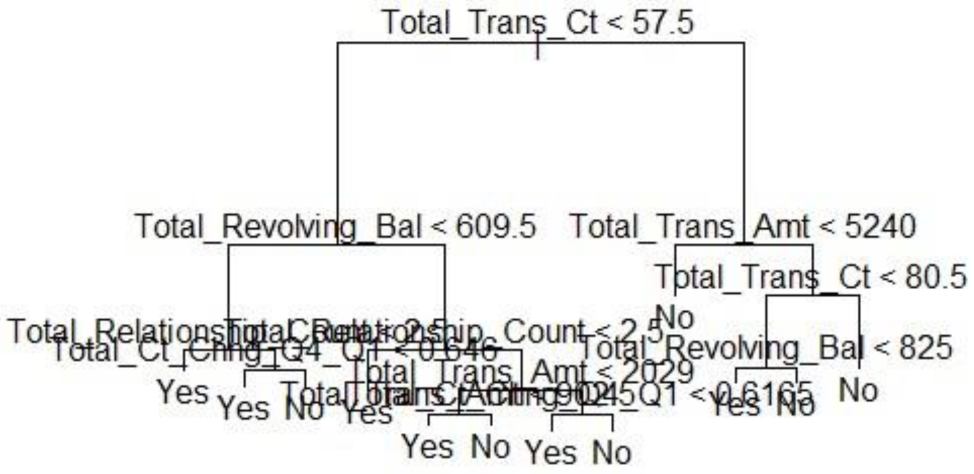
From the details that were supplied, it seems that the classification tree model was tested on a test set of 10,127 entries. The model anticipated “Yes” for churned consumers, while the actual number of churned customers was 315. The model projected “No” for nonchurned consumers, however there were 2479 real non-churned customers. The model forecasted “Yes” for non-churned consumers, whereas the actual number of non-churned clients was 145. Those are referred to as Type I mistakes. The algorithm predicted “No” for churned customers, yet there were 99 lost customers. These are referred to as Type II mistakes. The model’s precision is around 91.97%, implying that it precisely forecast the churn status of roughly 91.97% of the clients in the test dataset. The model incorrectly classified the churn state of roughly 8.03% of the clients, suggesting that the model incorrectly classified the churn situation for roughly 8.03% of the clients.

5.1.4 Cross Validation

```
## [1] "size"   "dev"    "k"      "method"

## $size
## [1] 13 12  9  8  6  4  3  1
##
## $dev
## [1]  551  551  566  632  649  718  803 1177 ##
## $k
## [1]      -Inf   0.00000  12.33333  27.00000  30.50000  37.00000  81.00000
## [8] 194.00000
##
## $method
## [1] "misclass"
##
## attr(,"class")
## [1] "prune"         "tree.sequence"
```



According on the data that was delivered, a pruning procedure was undertaken on the classification tree model. Pruning is a strategy for reducing decision tree complexity by deleting unneeded branches or nodes, resulting in a smaller and more interpretable tree. The result shows that sizes 13 and 12 have the lowest deviation, with lower deviance values indicating a model with a superior fit.


Total_Trans_Ct < 57.5
Total_Revolving_Bal < 609.5
Total_Trans_Amt < 5240
Total_Trans_Ct < 80.5
No
Total_Relationship_Count < 2.5
Total_Relationship_Count < 2.5
Total_Ct_Chng_Q4_Q1 < 0.646
Total_Revolving_Bal < 825
Total_Trans_Amt < 2029
Total_Trans_Amt < 902.5
Total_Ct_Chng_Q4_Q1 < 0.6165
Yes
Yes
No
Yes
Yes
No
No
Yes
No
Yes
No

```
## node), split, n, deviance, yval, (yprob)
##       * denotes terminal node
##
##  1) root 7089 6374.000 No ( 0.833968 0.166032 )
##    2) Total_Trans_Ct < 57.5 2619 3466.000 No ( 0.624666 0.375334 )
##      4) Total_Revolving_Bal < 609.5 880 1043.000 Yes ( 0.279545 0.720455 )
##        8) Total_Relationship_Count < 2.5 214   47.450 Yes ( 0.023364
0.976636 ) *
##        9) Total_Relationship_Count > 2.5 666  871.800 Yes ( 0.361862
0.638138 )
##         18) Total_Ct_Chng_Q4_Q1 < 0.646 459  498.500 Yes ( 0.233115
0.766885 ) *
##         19) Total_Ct_Chng_Q4_Q1 > 0.646 207  268.700 No ( 0.647343
0.352657 ) *
##      5) Total_Revolving_Bal > 609.5 1739 1744.000 No ( 0.799310 0.200690 )
##       10) Total_Relationship_Count < 2.5 153  167.000 Yes ( 0.235294
0.764706 ) *
##       11) Total_Relationship_Count > 2.5 1586 1320.000 No ( 0.853720
0.146280 )
##         22) Total_Trans_Amt < 2029 1074  533.500 No ( 0.932030 0.067970 )
##           44) Total_Trans_Amt < 902.5 15    7.348 Yes ( 0.066667 0.933333
) *
##           45) Total_Trans_Amt > 902.5 1059  455.400 No ( 0.944287 0.055713
) *
##         23) Total_Trans_Amt > 2029 512  634.400 No ( 0.689453 0.310547 )
##           46) Total_Ct_Chng_Q4_Q1 < 0.6165 214  294.000 Yes ( 0.443925
0.556075 ) *
##           47) Total_Ct_Chng_Q4_Q1 > 0.6165 298  235.000 No ( 0.865772
0.134228 ) *
##    3) Total_Trans_Ct > 57.5 4470 1597.000 No ( 0.956600 0.043400 )
##      6) Total_Trans_Amt < 5240 3366  342.900 No ( 0.991087 0.008913 ) *
##      7) Total_Trans_Amt > 5240 1104  927.800 No ( 0.851449 0.148551 )
##       14) Total_Trans_Ct < 80.5 212  267.500 Yes ( 0.325472 0.674528 )
##         28) Total_Revolving_Bal < 825 111   40.770 Yes ( 0.045045 0.954955
) *
##         29) Total_Revolving_Bal > 825 101  132.700 No ( 0.633663 0.366337
) *
##       15) Total_Trans_Ct > 80.5 892  199.000 No ( 0.976457 0.023543 ) *

##          Churn.test
## tree.pred   No  Yes
##       No  2482  109
##       Yes  106  341

## [1] 0.9292298

## [1] 0.07077024
```

The tree contains 12 nodes that are terminal. The incorrect classification error rate is 0.07307439, showing that the pruning tree correctly forecasts customer turnover for the vast majority of clients. The first division is based on the total transactions count, with a limit of 57.5. The overall transaction amount is greater, and the largest class is determined by the total transaction count and total revolving balance. The erroneous classification matrix in the trimmed tree displays the expected and actual test data results. The model predicts 2473 “No” (existing customer) and 343 “Yes” (churned customer) situations correctly. The pruning tree’s total precision is 92.69%.

PART 6.0: Conclusion

Finally, the data shows some crucial insights about credit card user attrition. We discovered that around 16.07% (1627 consumers) (Attrited consumers) opted to discontinue or terminate their credit card services. Customers with bachelor’s degrees outnumber those with high school diplomas. Customers with PhD degrees had the lowest share, as predicted. Clients with higher-rated credit cards are also more devoted and less inclined to churn, according to the findings. According to the research, the advantages, awards, or features connected with higher-tier credit cards may help with customer loyalty and involvement. If it involves churning, customers with larger or lesser line of credit levels may display various behaviors and inclinations. This knowledge can help credit card firms identify the link between limit of credit and client attrition. It may aid in the development of customized customer churn management tactics, such as providing customized credit limits or incentives to consumers at risk of turnover depending on their line of credit. The overall transaction count is thought to be the most significant predictor of attrition. consumers with a smaller total transaction count have a greater probability to be nonchurned, and additional variables such as revolving balance and variations in transaction count further narrow the classification of churned and non-churned consumers.