

Customer Churn Prediction Lloyds Banking Group

Rounak Saha

2025-05-07

Introduction

Welcome to the Data Science & Analytics team at Lloyds Banking Group. As a new data science graduate, I have been entrusted with a critical project that could significantly impact our customer retention strategies. Li, our senior data scientist, has specifically chosen me to take over this project due to my strong analytical skills and enthusiasm for solving real-world business problems. This is an exciting opportunity for me to apply my knowledge and make a real impact within our team.

Context The project I am about to embark on is the “Customer Retention Enhancement through Predictive Analytics” initiative. This project arose from an urgent need to address declining retention rates within certain segments of our customer base. Over the past few months, we’ve noticed a worrying trend of increased customer churn, particularly among young professionals and small business owners. This poses a substantial threat to our market position and long-term profitability.

Our fictional client, SmartBank, a subsidiary of Lloyds, has reported that a substantial portion of their customer base is at risk of moving to competitors offering more personalised banking solutions. SmartBank has tasked our team with developing a predictive model to identify at-risk customers and propose targeted interventions to retain them.

Task 1

Introduction In this task, I will take the first critical steps toward building a predictive model for customer churn. Your work will involve gathering relevant data, conducting EDA, and preparing the data set for model development. These activities are foundational for ensuring the accuracy and reliability of your subsequent analysis and predictions.

Identify and gather data:

- Review the provided data sources and select those most relevant for predicting customer churn. Focus on key areas such as customer demographics, transaction history, and customer service interactions.
- Document your selection criteria and rationale for choosing each data set, ensuring that the data will provide meaningful insights into customer behaviour.

Perform EDA:

- Use statistical techniques and data visualisation tools to explore the data sets. Create visualisations such as histograms, scatter plots, and box plots to understand distributions, trends, and relationships between variables.

- Identify key features that may influence customer churn, paying special attention to patterns or anomalies that could be significant.

Clean and preprocess the data:

- Handle missing values by choosing appropriate methods such as imputation, removal, or flagging. Justify your chosen method based on the data and context.
- Detect and address outliers that could skew the analysis or predictions. Decide whether to cap, transform, or remove outliers based on their nature and potential impact.
- Standardise or normalise numerical features to ensure consistent scales across variables. This step is crucial for preparing the data for machine learning algorithms.
- Encode categorical variables using techniques like one-hot encoding to transform them into a numerical form appropriate for analysis.

Deliverable:

File submission: Submit a comprehensive report detailing your data gathering, EDA, and data cleaning processes. The report should include:

- A summary of the data sets selected and the rationale for their inclusion
- Visualisations and statistical summaries from the EDA
- A description of the data cleaning and preprocessing steps taken
- The cleaned and preprocessed data set ready for model building

Ensure that your report is clear, concise, and well-organised, as it will be a key component of the project's success, guiding future analysis and model development.

```
# install required packages
options(repos = c(CRAN = "https://cloud.r-project.org"))
install.packages("tidyverse")
install.packages("readxl", type = "source")
# install.packages("psych") ---- stats summary of the dataset
install.packages("skimr")          ## Detailed descriptive stats, includes numeric and categorical
install.packages("ggplot2")
install.packages("dplyr")
install.packages("fastDummies") ## for one-hot encoding
install.packages("data.table")
install.packages("e1071")
install.packages("corrplot") # plotting heatmap
install.packages("caTools")
install.packages("caret") # train control
install.packages("ROSE") # sampling imbalanced dataset
install.packages("smotefamily")
install.packages("rpart") #partitioning of the decision tree
install.packages("rpart.plot") #plot the decesion tree
install.packages("data.tree") #to get nice visual
install.packages("pROC") #RoC curve
install.packages("randomForest")
install.packages("xgboost")
```

```
# load the required libraries
library(tidyverse)
```

Load required packages and dataset

```
## Warning: package 'ggplot2' was built under R version 4.3.3
```

```
## -- Attaching core tidyverse packages ----- tidyverse 2.0.0 --
## v dplyr      1.1.4      v readr      2.1.5
## v forcats    1.0.0      v stringr   1.5.0
## v ggplot2    3.5.2      v tibble    3.2.1
## v lubridate  1.9.3      v tidyr     1.3.1
## v purrr      1.0.2
## -- Conflicts ----- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
library(readxl)
library(skimr)
#library(psych)
library(ggplot2)
library(dplyr)
library(fastDummies)
```

```
## Warning: package 'fastDummies' was built under R version 4.3.3
```

```
library(data.table)
```

```
## Warning: package 'data.table' was built under R version 4.3.3
```

```
##
## Attaching package: 'data.table'
##
## The following objects are masked from 'package:lubridate':
##
##   hour, isoweek, mday, minute, month, quarter, second, wday, week,
##   yday, year
##
## The following objects are masked from 'package:dplyr':
##
##   between, first, last
##
## The following object is masked from 'package:purrr':
##
##   transpose
```

```
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.3
```

```
library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 4.3.3
```

```
## corrplot 0.95 loaded
```

```
library(caTools)
```

```
## Warning: package 'caTools' was built under R version 4.3.3
```

```
library(caret)
```

```
## Warning: package 'caret' was built under R version 4.3.3
```

```
## Loading required package: lattice
```

```
##
```

```
## Attaching package: 'caret'
```

```
##
```

```
## The following object is masked from 'package:purrr':
```

```
##
```

```
## lift
```

```
library(ROSE)
```

```
## Warning: package 'ROSE' was built under R version 4.3.3
```

```
## Loaded ROSE 0.0-4
```

```
library(smotefamily)
```

```
library(rpart)
```

```
## Warning: package 'rpart' was built under R version 4.3.3
```

```
library(rpart.plot)
```

```
## Warning: package 'rpart.plot' was built under R version 4.3.3
```

```
library(data.tree)
```

```
## Warning: package 'data.tree' was built under R version 4.3.3
```

```
library(pROC)
```

```
## Warning: package 'pROC' was built under R version 4.3.3

## Type 'citation("pROC")' for a citation.
##
## Attaching package: 'pROC'
##
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```
library(randomForest)
```

```
## Warning: package 'randomForest' was built under R version 4.3.3

## randomForest 4.7-1.2
## Type rfNews() to see new features/changes/bug fixes.
##
## Attaching package: 'randomForest'
##
## The following object is masked from 'package:dplyr':
##
##     combine
##
## The following object is masked from 'package:ggplot2':
##
##     margin
```

```
library(xgboost)
```

```
## Warning: package 'xgboost' was built under R version 4.3.3

##
## Attaching package: 'xgboost'
##
## The following object is masked from 'package:dplyr':
##
##     slice
```

```
# We load the Excel workbook that contains multiple sheets of customer-related data
customer_demo <- read_xlsx("Customer_Churn_Data_Large.xlsx", sheet = 1, col_names = TRUE)
transaction_hist <- read_xlsx("Customer_Churn_Data_Large.xlsx", sheet = 2, col_names = TRUE)
customer_service <- read_xlsx("Customer_Churn_Data_Large.xlsx", sheet = 3, col_names = TRUE)
online_activity <- read_xlsx("Customer_Churn_Data_Large.xlsx", sheet = 4, col_names = TRUE)
churn_status <- read_xlsx("Customer_Churn_Data_Large.xlsx", sheet = 5, col_names = TRUE)
```

The dataset consists of multiple sheets from an Excel file. The above datasets were selected based on their relevance to Customer Churn prediction. All datasets have their respective data with the Customer ID which helps uniquely identify a customer.

- **Customer Demographics:** Contains features like Age, Gender, Marital Status and Income Level. These features are critical since demographic features often influence customer behaviour. There are 1000 samples in the dataset, one for each customer ID.
- **Transaction History:** Included features are: Transaction ID, Transaction Date, Amount Spent and Product Category. The Transaction ID was a redundant column since it is used to identify each transaction which has no effect on customer churn. There are 5054 samples in the dataset. Transaction patterns reveal spending behaviour which may affect churn.
- **Customer Service:** Include features like Interaction ID, Interaction Date, Interaction Type and Resolution Status. Customer service Interactions like Unresolved complaints could be a crucial indicator for customer churn. Samples: 1001
- **Online Activity:** This dataset contains the various online service usage of customers. Includes features like Last Login Date, Login Frequency and Service Usage. The online service usage patterns could signal a customer likely to churn. A high user may not be as likely to leave as their counterpart. Samples: 1000
- **Churn Status:** The target variable 'Churn Status' indicates whether a customer has churned (1) or not (0). 1000 unique samples in the dataset.

Exploratory Data Analysis In this stage, we conduct a thorough exploration of the dataset using both statistical summaries and visualizations. The goal is to understand variable distributions, detect anomalies, identify relationships between features, and uncover patterns that may influence customer churn.

Key steps include:

- Examining the distribution of numeric and categorical variables
- Identifying potential correlations or trends
- Visualizing churn rates across different customer segments
- Highlighting any unusual or missing data that requires attention

```
head(customer_demo)
```

Customer demographics

```
## # A tibble: 6 x 5
##   CustomerID Age Gender MaritalStatus IncomeLevel
##   <dbl> <dbl> <chr> <chr> <chr>
## 1         1   62 M      Single      Low
## 2         2   65 M      Married      Low
## 3         3   18 M      Single      Low
## 4         4   21 M      Widowed      Low
## 5         5   21 M      Divorced    Medium
## 6         6   57 F      Divorced    Medium
```

```
str(customer_demo)
```

```
## tibble [1,000 x 5] (S3: tbl_df/tbl/data.frame)
## $ CustomerID : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
## $ Age       : num [1:1000] 62 65 18 21 21 57 27 37 39 68 ...
```

```
## $ Gender      : chr [1:1000] "M" "M" "M" "M" ...
## $ MaritalStatus: chr [1:1000] "Single" "Married" "Single" "Widowed" ...
## $ IncomeLevel  : chr [1:1000] "Low" "Low" "Low" "Low" ...
```

```
skim(customer_demo)
```

Table 1: Data summary

Name	customer_demo
Number of rows	1000
Number of columns	5
Column type frequency:	
character	3
numeric	2
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
Gender	0	1	1	1	0	2	0
MaritalStatus	0	1	6	8	0	4	0
IncomeLevel	0	1	3	6	0	3	0

Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CustomerID	0	1	500.50	288.82	1	250.75	500.5	750.25	1000	
Age	0	1	43.27	15.24	18	30.00	43.0	56.00	69	

```
table(customer_demo$Gender)
```

```
##
##   F   M
## 513 487
```

```
table(customer_demo$MaritalStatus)
```

```
##
## Divorced Married Single Widowed
##    248    261    215    276
```

```
table(customer_demo$IncomeLevel)
```

```
##
##   High   Low Medium
##   349   325   326
```

The descriptive statistics of the dataset shows a well balanced demographics data.

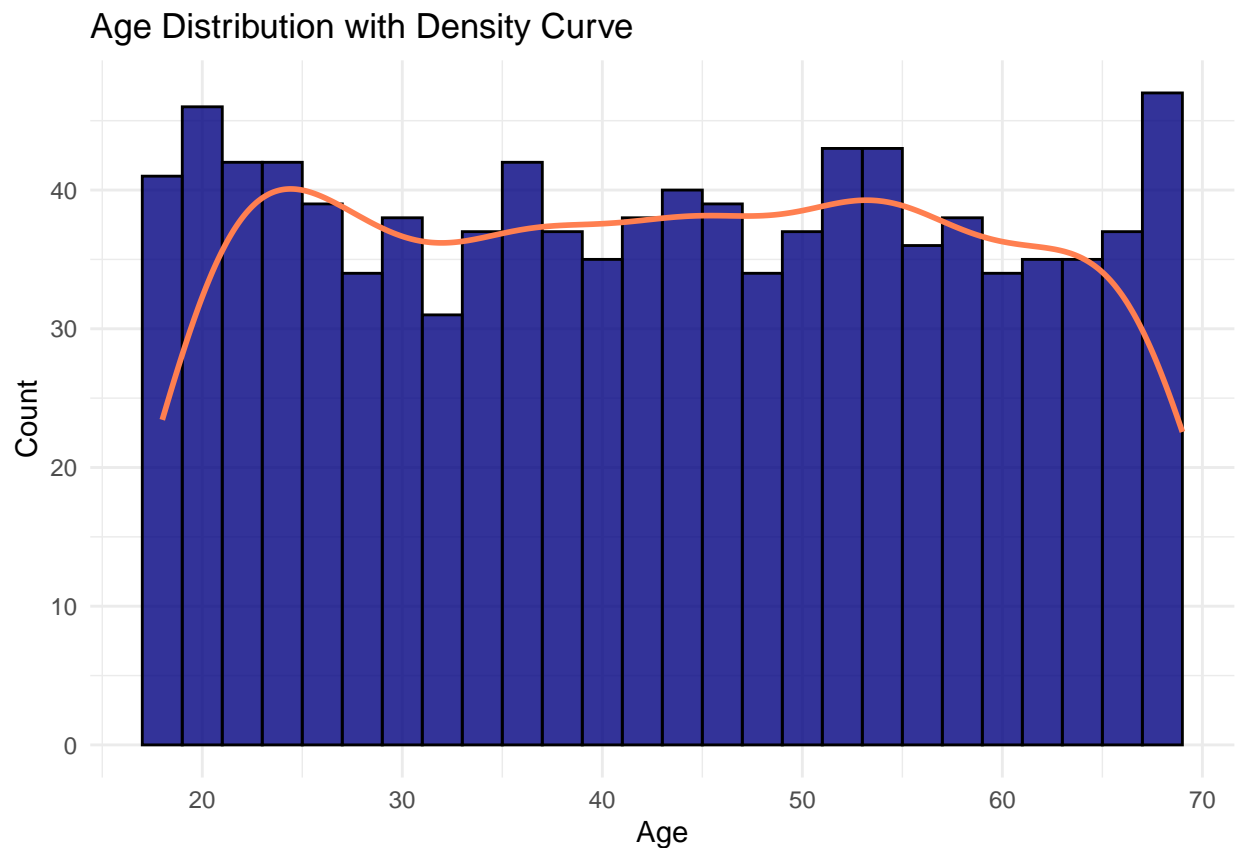
- Gender : Males and Females are equally distributed
- Age : Mean age is 43, and the customer base ranges from age 18 to 69.
- Marital status : There are 4 unique values, and the statistics suggest it is roughly equally distributed.
- Income level : Three unique levels for this feature, and equally distributed.

```
# checking for duplicates
sum(duplicated(customer_demo))
```

```
## [1] 0
```

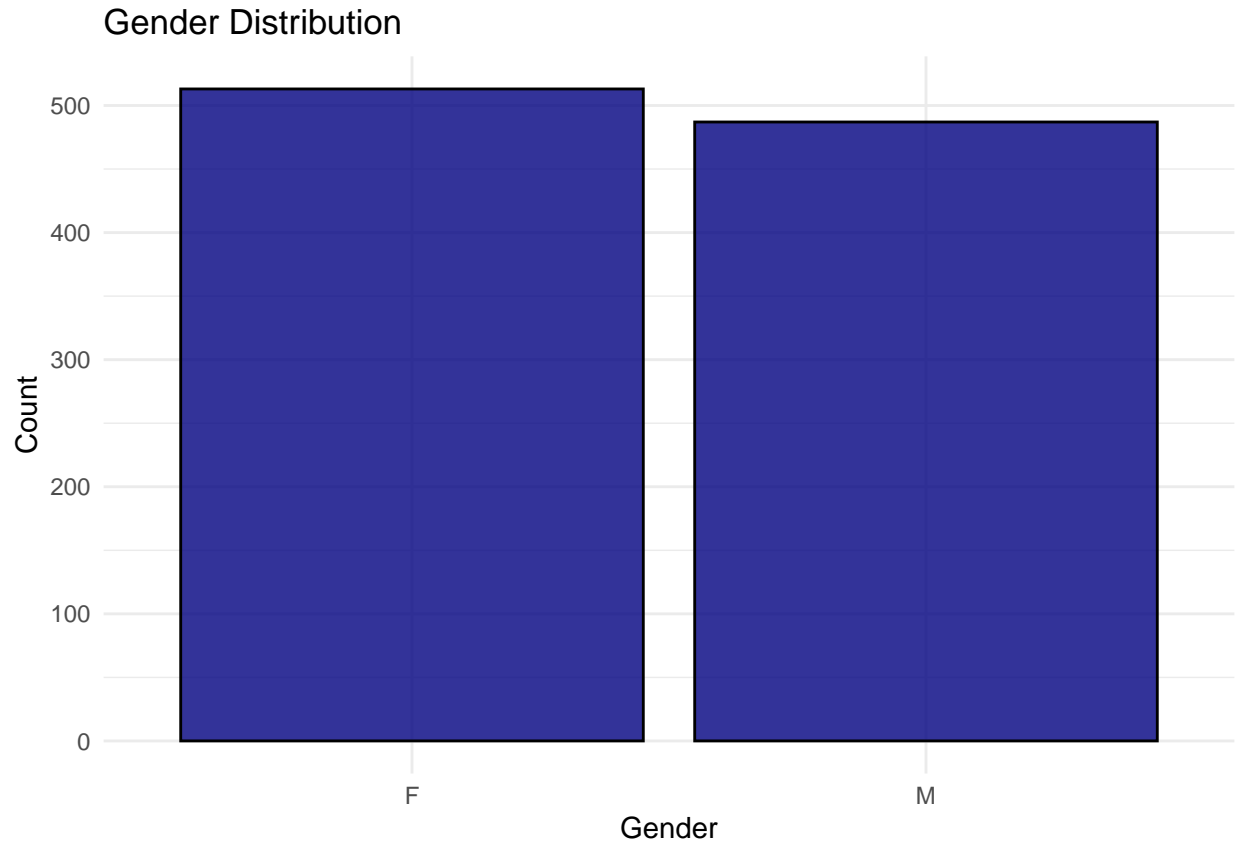
```
# there are no duplicate values and missing value
```

```
# plotting age distribution
ggplot(data = customer_demo, aes(x =Age)) +
  geom_histogram(binwidth= 2, fill = "navyblue", color = "black", alpha= 0.8) +
  geom_density(aes(y= ..count..*2), color= "coral", size = 1) +
  labs(title = "Age Distribution with Density Curve", x = "Age", y = "Count")+
  theme_minimal()
```

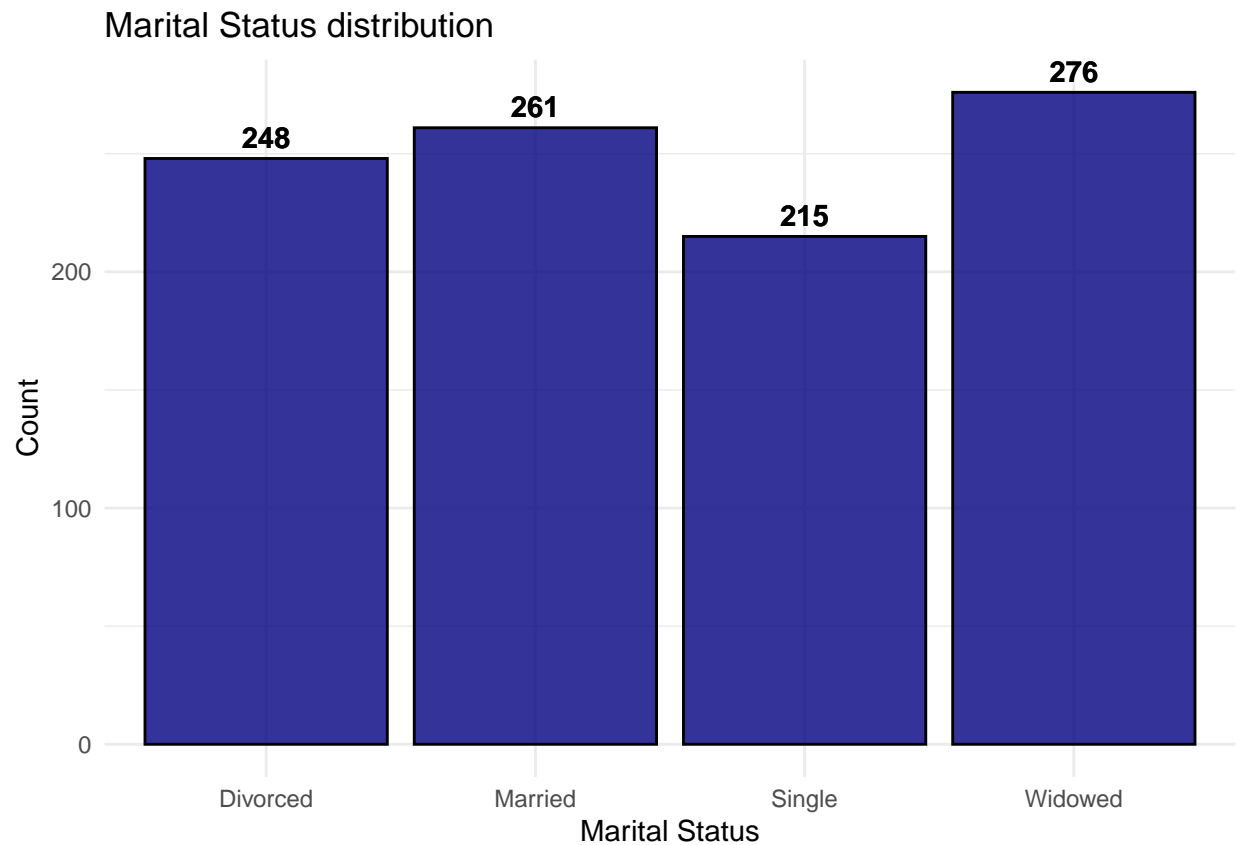


The spread of the Age column is uniform, the customer base is uniformly distributed across all age groups.


```
# Plotting Gender distribution
ggplot(data= customer_demo, aes(x= Gender)) +
  geom_bar(fill = "navyblue", color = "black", alpha = 0.8)+ theme_minimal()+
  labs( title = "Gender Distribution", x= "Gender", y= "Count")
```

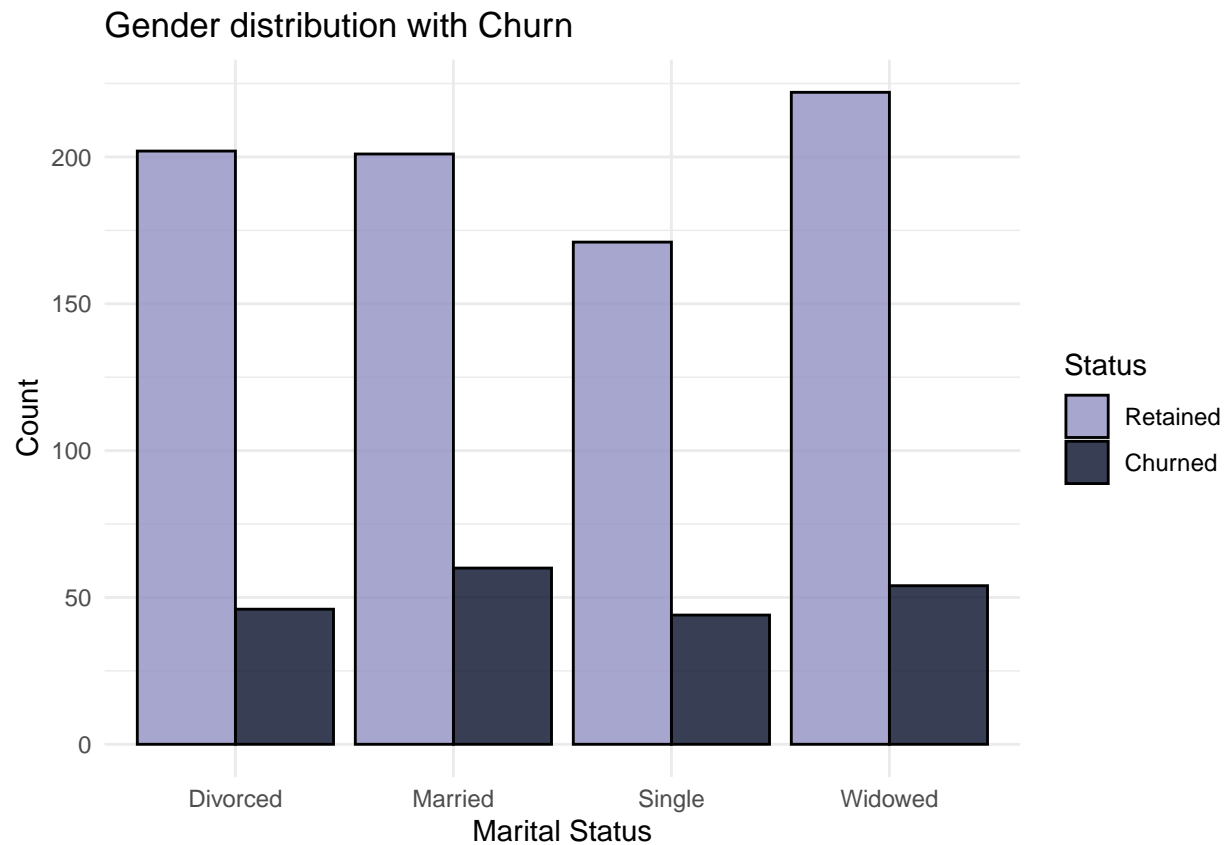


```
# Plotting Marital Status distribution
ggplot(data= customer_demo, aes(x= MaritalStatus)) +
  geom_bar(fill = "navyblue", color = "black", alpha = 0.8)+ theme_minimal()+
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, color = "black", fontface = "bold")+
  labs( title = "Marital Status distribution", x= "Marital Status", y= "Count")+
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, fontface = "bold")
```

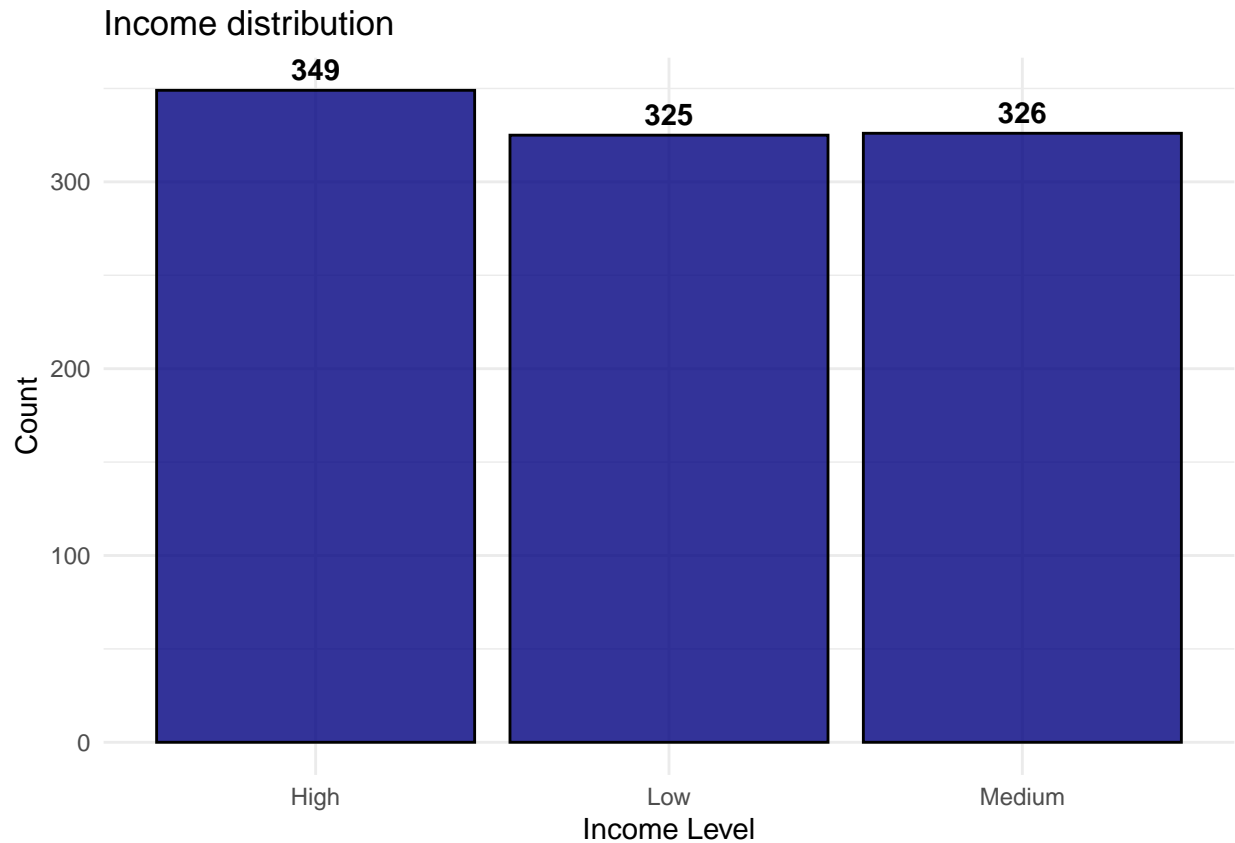


```
cust_demo_churn <- left_join(customer_demo, churn_status, by = "CustomerID")

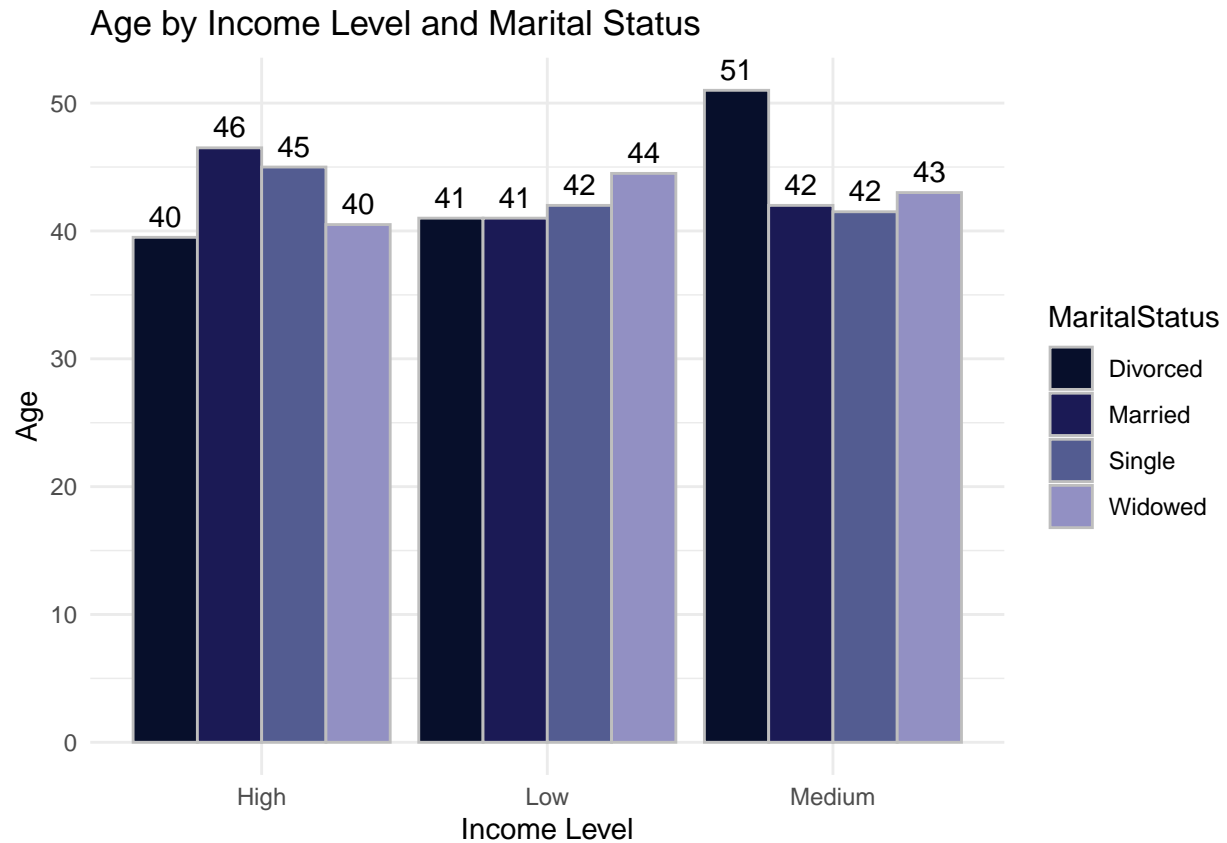
ggplot(data = cust_demo_churn,
       aes(x = factor(MaritalStatus), fill = factor(ChurnStatus, labels = c("Retained", "Churned")))) +
  geom_bar(color = "black", alpha = 0.8, position = "dodge") +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Gender distribution with Churn", x = "Marital Status", y = "Count", fill = "Status") +
  theme_minimal()
```



```
# Plotting income level distribution  
ggplot(data = customer_demo, aes(x= IncomeLevel))+  
  geom_bar(fill= "navyblue", color = "black", alpha= 0.8)+ theme_minimal()+  
  geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, fontface = "bold")+  
  labs(title = "Income distribution", x= "Income Level", y= "Count")
```



```
# plotting the median age of customers in each income level and marital status
customer_demo %>% group_by(IncomeLevel, MaritalStatus) %>%
  summarise(median_age = median(Age), .groups = 'drop') %>%
  ggplot(aes(x = IncomeLevel, y = median_age, fill = MaritalStatus)) +
  geom_bar(stat = "identity", position = "dodge", color = "grey") +
  labs(title = "Age by Income Level and Marital Status",
       x = "Income Level", y = "Age") +
  scale_fill_manual(values = c("Married" = "#1B1A55", "Single" = "#535C91", "Divorced" = "#070F2B", "Widowed" = "#A5A5A5")) +
  geom_text(aes(label = round(median_age, 0)), position = position_dodge(width = 0.9),
            vjust = -0.5) +
  theme_minimal()
```



The data is equally distributed across different categories

```
# Encoding the categorical variables
# Label encoding for Gender & Income Level
customer_demo$Gender <- as.numeric(factor(customer_demo$Gender,
                                          levels = unique(customer_demo$Gender)))-1
customer_demo$IncomeLevel <- as.numeric(factor(customer_demo$IncomeLevel,
                                              levels = c("Low", "Medium", "High"),
                                              labels = c(0,1,2)))-1

# one-hot encoding for Marital status
customer_demo <- dummy_cols(customer_demo, select_columns = "MaritalStatus",
                           remove_first_dummy = FALSE,
                           remove_selected_columns = TRUE)

head(customer_demo)
```

```
## # A tibble: 6 x 8
##   CustomerID Age Gender IncomeLevel MaritalStatus_Divorced
##   <dbl> <dbl> <dbl> <dbl> <int>
## 1     1    62     0     0         0
## 2     2    65     0     0         0
## 3     3    18     0     0         0
## 4     4    21     0     0         0
## 5     5    21     0     1         1
## 6     6    57     1     1         1
## # i 3 more variables: MaritalStatus_Married <int>, MaritalStatus_Single <int>,
## #   MaritalStatus_Widowed <int>
```

All categorical features has been converted to numerical values.

```
head(transaction_hist)
```

Transaction History

```
## # A tibble: 6 x 5
##   CustomerID TransactionID TransactionDate      AmountSpent ProductCategory
##   <dbl>          <dbl> <dtm>          <dbl> <chr>
## 1           1           7194 2022-03-27 00:00:00      416. Electronics
## 2           2           7250 2022-08-08 00:00:00       55.0 Clothing
## 3           2           9660 2022-07-25 00:00:00      198. Electronics
## 4           2           2998 2022-01-25 00:00:00      101. Furniture
## 5           2           1228 2022-07-24 00:00:00      397. Clothing
## 6           2           8903 2022-01-09 00:00:00      285. Electronics
```

```
str(transaction_hist)
```

```
## tibble [5,054 x 5] (S3: tbl_df/tbl/data.frame)
## $ CustomerID      : num [1:5054] 1 2 2 2 2 2 2 2 3 3 ...
## $ TransactionID   : num [1:5054] 7194 7250 9660 2998 1228 ...
## $ TransactionDate: POSIXct[1:5054], format: "2022-03-27" "2022-08-08" ...
## $ AmountSpent     : num [1:5054] 416 55 198 101 397 ...
## $ ProductCategory: chr [1:5054] "Electronics" "Clothing" "Electronics" "Furniture" ...
```

```
summary(transaction_hist)
```

```
##   CustomerID      TransactionID TransactionDate
## Min.   : 1.0      Min.   :1000      Min.   :2022-01-01 00:00:00.00
## 1st Qu.: 251.0    1st Qu.:3242      1st Qu.:2022-04-03 00:00:00.00
## Median : 506.0    Median :5530      Median :2022-07-01 00:00:00.00
## Mean   : 501.4     Mean   :5511      Mean   :2022-07-01 19:25:37.16
## 3rd Qu.: 749.0    3rd Qu.:7681      3rd Qu.:2022-09-29 00:00:00.00
## Max.   :1000.0    Max.   :9997      Max.   :2022-12-31 00:00:00.00
##   AmountSpent      ProductCategory
## Min.   : 5.18      Length:5054
## 1st Qu.:127.11     Class :character
## Median :250.53     Mode  :character
## Mean   :250.71
## 3rd Qu.:373.41
## Max.   :499.86
```

```
table(transaction_hist$ProductCategory)
```

```
##
##      Books      Clothing Electronics      Furniture      Groceries
##      1041         1000         1001          992         1020
```

```
# the product category variable is almost distributed equally
```

The descriptive statistics of the dataset shows a well balanced transaction data.

- Amount spent : Mean Amount spent by customers is \$250, and ranges from \$5 to \$500.
- Product category : Five unique levels for this feature, and equally distributed.

```
# check for null and duplicate values  
sum(is.na(transaction_hist))
```

```
## [1] 0
```

```
sum(duplicated(transaction_hist))
```

```
## [1] 0
```

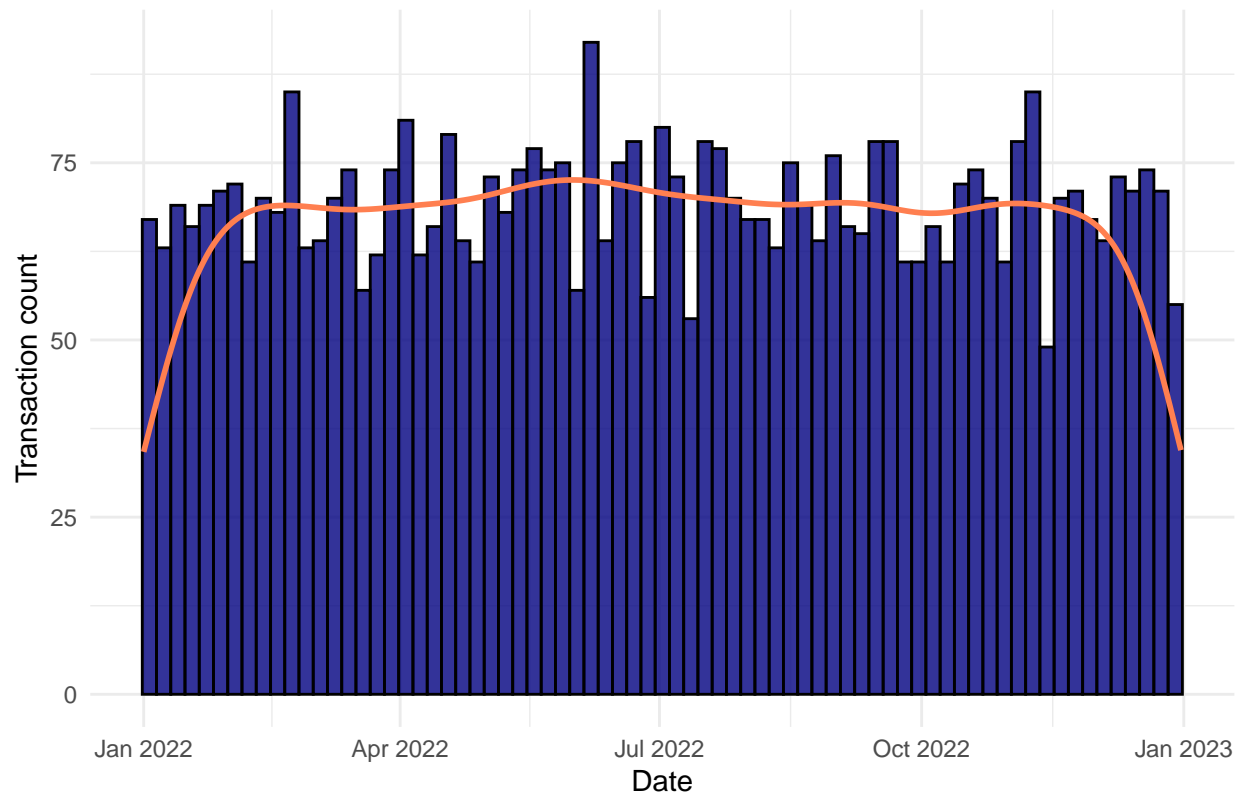
The dataset is clean without any null & duplicate values.

TransactionID is a redundant column. It represents an id to refer to the transaction which is irrelevant to customer churn, so it can be dropped.

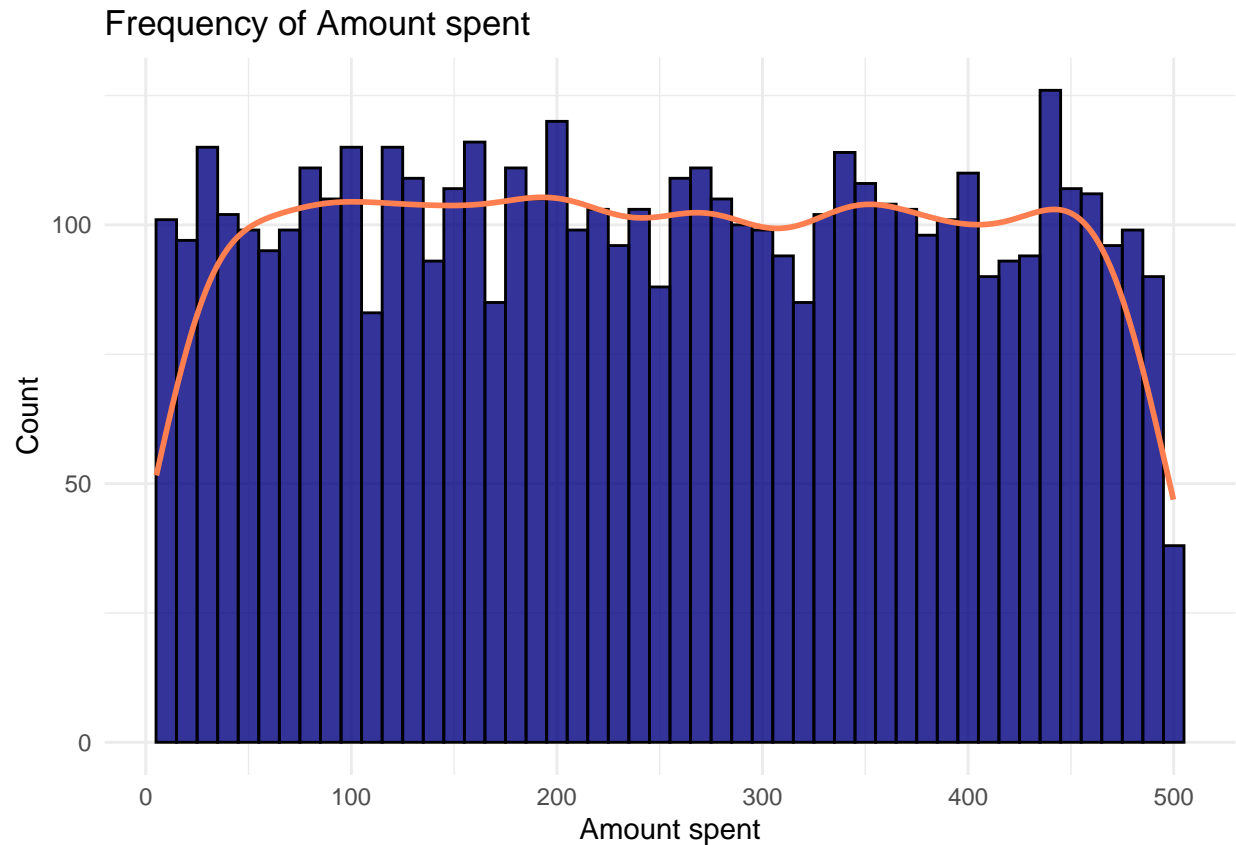
There are multiple transaction histories for a given customer, so feature engineering is required to aggregate multiple histories and create relevant features.

```
# Plotting transaction records by date  
ggplot(data = transaction_hist, aes(x= as.Date(TransactionDate))) +  
  geom_histogram(binwidth= 5, fill= "navyblue", color = "black", alpha= 0.8)+ theme_minimal()+  
  geom_density(aes(y= ..count..*5), color= "coral", size = 1)+  
  labs(title = "No of transactions by date", x= "Date", y = "Transaction count")
```

No of transactions by date

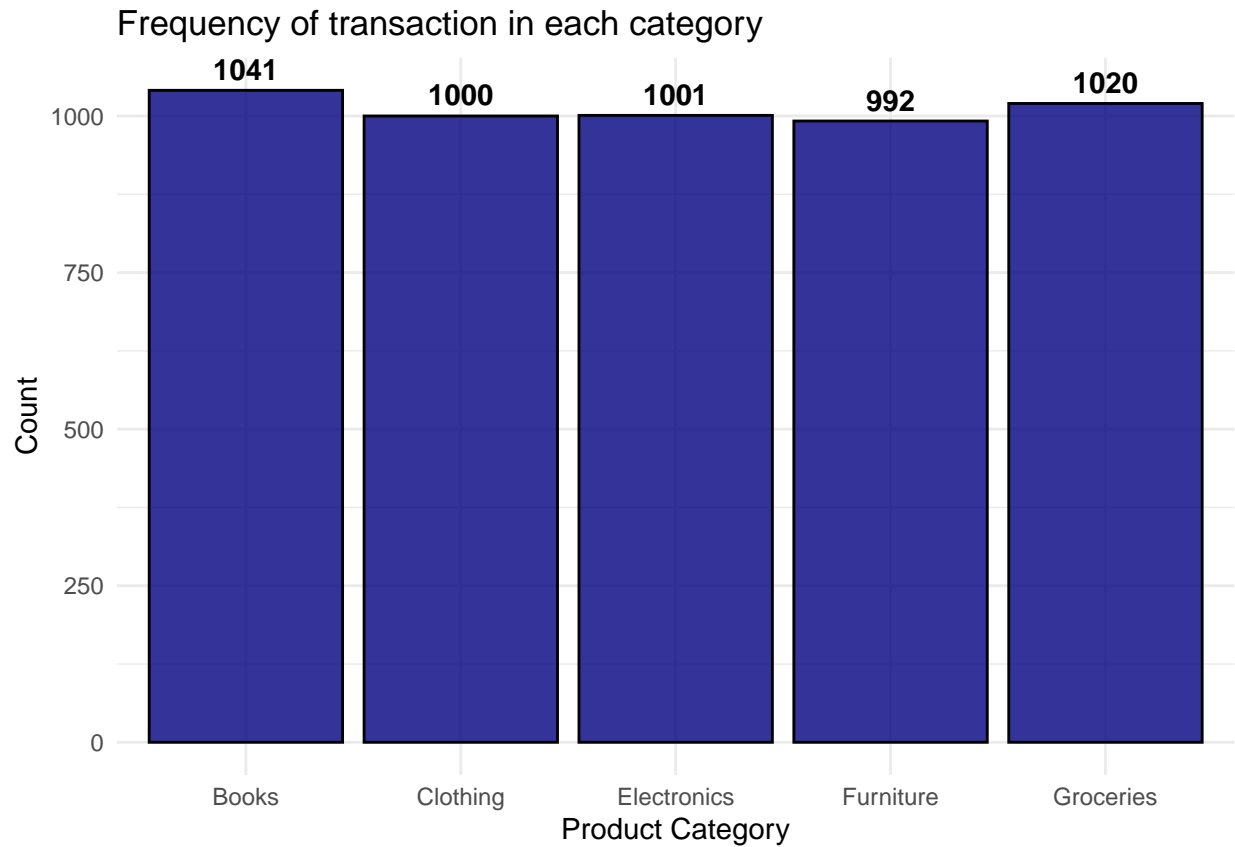


```
# Plotting amount spend
ggplot(data = transaction_hist, aes(x = AmountSpent)) +
  geom_histogram(binwidth = 10, fill= "navyblue", color= "black", alpha= 0.8)+ theme_minimal()+
  geom_density(aes(y= ..count..* 10), color= "coral", size= 1)+
  labs(title = "Frequency of Amount spent", x= "Amount spent", y = "Count")
```

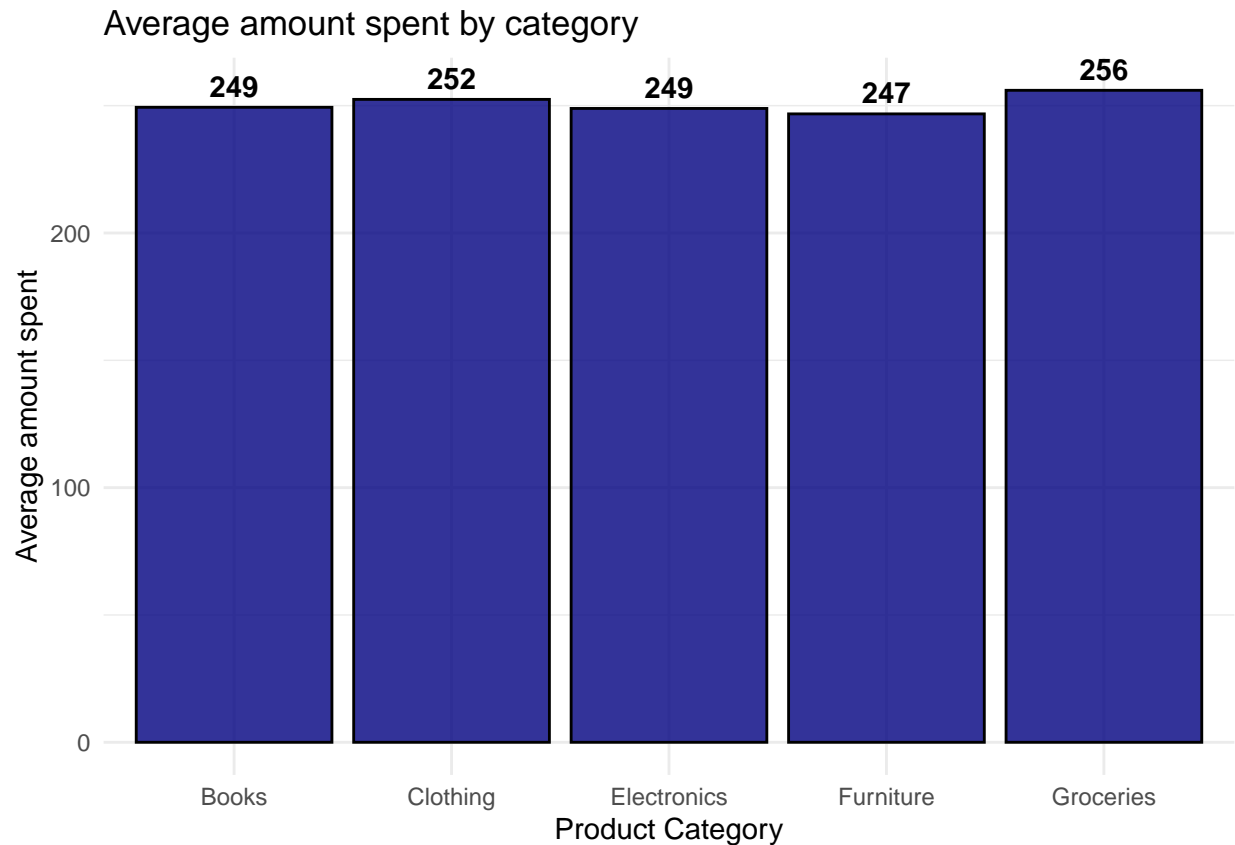



The Customer transactions and Amount spent are roughly uniformly distributed.

```
# Plotting frequency of transaction in each category
ggplot(data= transaction_hist, aes(x = ProductCategory))+
  geom_histogram(stat = "count", fill= "navyblue", color= "black", alpha= 0.8)+
  geom_text(stat= "count", aes(label= ..count..), vjust= -0.5, fontface= "bold")+
  labs(title = "Frequency of transaction in each category ", x= "Product Category", y = "Count")+
  theme_minimal()
```



```
# Mean amount spent for each category
transaction_hist %>% group_by(ProductCategory) %>%
  summarise(total_amt_spent = mean(AmountSpent)) %>%
  ggplot(aes(x= ProductCategory, y= total_amt_spent)) + geom_bar(stat = "identity",fill= "navyblue", col= "white") +
  labs(title = "Average amount spent by category ", x= "Product Category", y = "Average amount spent")+
  geom_text(aes(label = round(total_amt_spent,0)), vjust= -0.5, fontface= "bold")+
  theme_minimal()
```



```
# one-hot encoding for Product Category
transaction_hist <- dummy_cols(transaction_hist,
                                select_columns = "ProductCategory",
                                remove_first_dummy = FALSE,
                                remove_selected_columns = TRUE)

# dropping transaction ID column
transaction_hist <- transaction_hist[, -2]

# converting the transaction date to Transaction_frequency and grouping by customer ID
transaction_hist <- transaction_hist %>% group_by(CustomerID) %>% summarise(Total_amount_spent = sum(Amount),
                                                                              Transaction_frequency = n(),
                                                                              Books = sum(ProductCategory_Books),
                                                                              Clothing = sum(ProductCategory_Clothing),
                                                                              Electronics = sum(ProductCategory_Electronics),
                                                                              Furniture = sum(ProductCategory_Furniture),
                                                                              Groceries = sum(ProductCategory_Groceries))
```

The categorical features has been converted to numeric values.

```
head(customer_service)
```

Customer Service

```
## # A tibble: 6 x 5
##   CustomerID InteractionID InteractionDate      InteractionType ResolutionStatus
##       <dbl>         <dbl> <dtm>              <chr>          <chr>
## 1         1           6363 2022-03-31 00:00:00 Inquiry        Resolved
## 2         2           3329 2022-03-17 00:00:00 Inquiry        Resolved
## 3         3           9976 2022-08-24 00:00:00 Inquiry        Resolved
## 4         4           7354 2022-11-18 00:00:00 Inquiry        Resolved
## 5         4           5393 2022-07-03 00:00:00 Inquiry        Unresolved
## 6         6           2358 2022-05-05 00:00:00 Feedback       Resolved
```

```
str(customer_service)
```

```
## tibble [1,002 x 5] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID      : num [1:1002] 1 2 3 4 4 6 8 8 9 11 ...
##  $ InteractionID    : num [1:1002] 6363 3329 9976 7354 5393 ...
##  $ InteractionDate  : POSIXct[1:1002], format: "2022-03-31" "2022-03-17" ...
##  $ InteractionType  : chr [1:1002] "Inquiry" "Inquiry" "Inquiry" "Inquiry" ...
##  $ ResolutionStatus: chr [1:1002] "Resolved" "Resolved" "Resolved" "Resolved" ...
```

```
summary(customer_service)
```

```
##   CustomerID      InteractionID InteractionDate
## Min.   : 1.0      Min.   :2015      Min.   :2022-01-01 00:00:00.00
## 1st Qu.:238.2     1st Qu.:3992     1st Qu.:2022-04-07 00:00:00.00
## Median :474.5     Median :5912     Median :2022-07-02 12:00:00.00
## Mean   :485.2     Mean   :5953     Mean   :2022-07-02 19:28:22.99
## 3rd Qu.:735.8     3rd Qu.:7908     3rd Qu.:2022-09-30 00:00:00.00
## Max.   :995.0     Max.   :9997     Max.   :2022-12-30 00:00:00.00
## InteractionType   ResolutionStatus
## Length:1002      Length:1002
## Class :character  Class :character
## Mode  :character  Mode  :character
##
##
##
```

```
table(customer_service$InteractionType)
```

```
##
## Complaint Feedback Inquiry
##      335      360      307
```

```
table(customer_service$ResolutionStatus)
```

```
##
## Resolved Unresolved
##      523      479
```

```
# check null and duplicates
sum(is.na(customer_service))
```

```
## [1] 0
```

```
sum(duplicated(customer_service))
```

```
## [1] 0
```

The descriptive statistics of the dataset shows a well balanced customer service data.

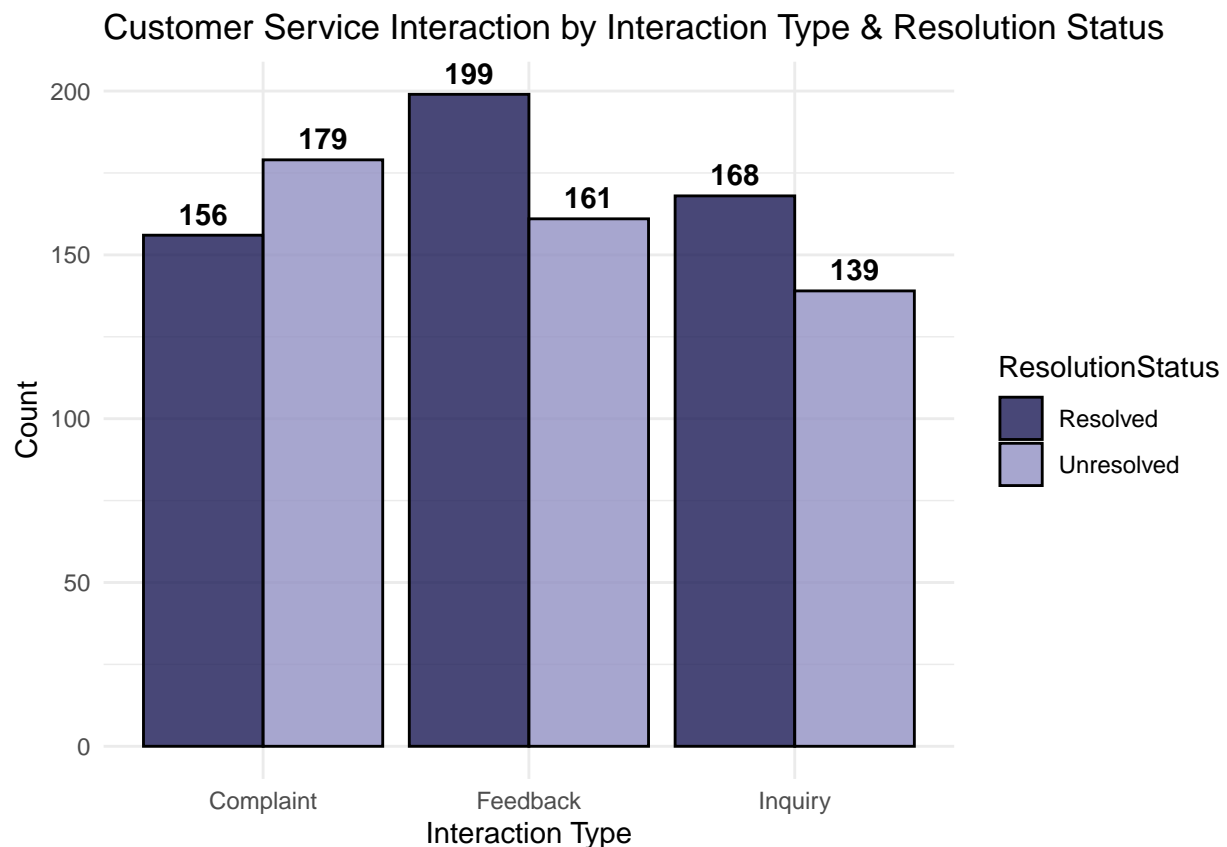
- InteractionType : Three unique levels for this feature, and equally distributed.
- Resolution Status : Two unique levels for this feature, and equally distributed.

InteractionID is a redundant column. It represents an id to refer to the interaction which is irrelevant to customer churn, so it can be dropped.

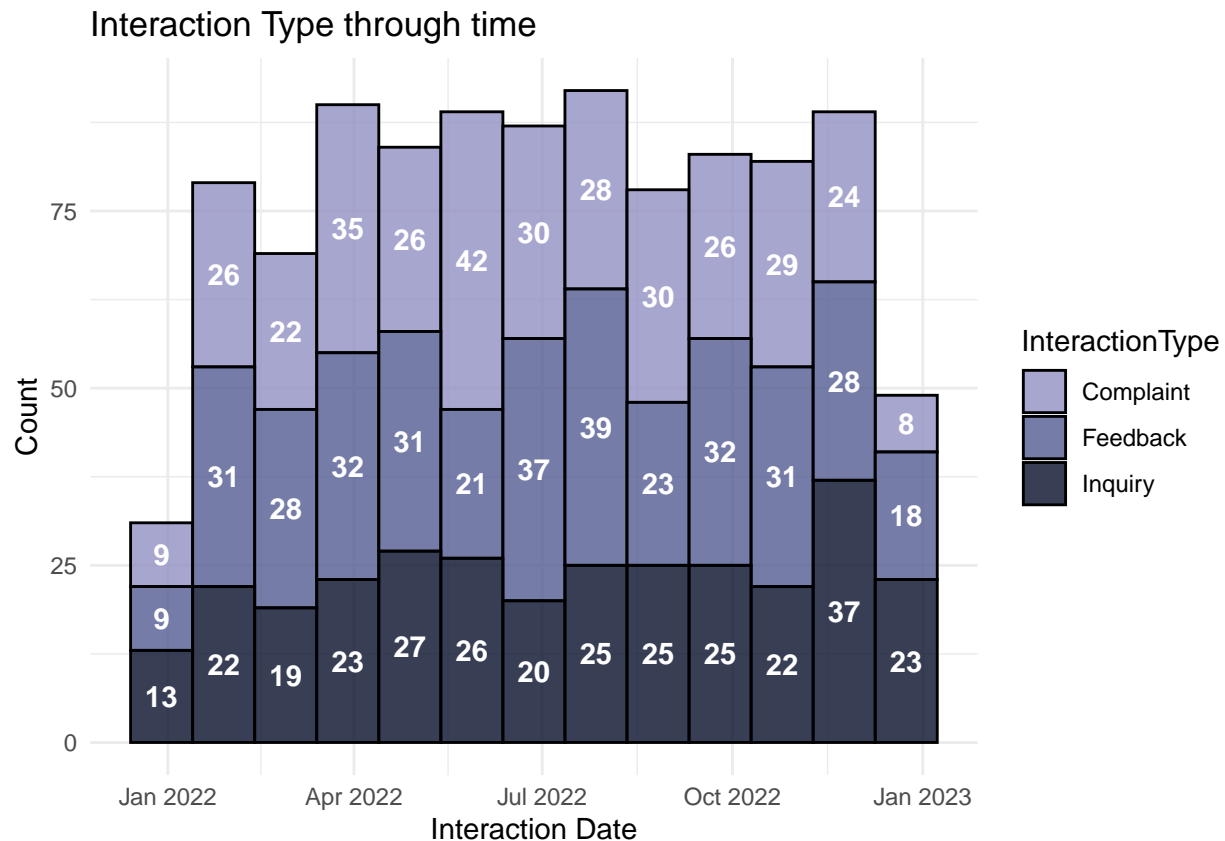
```
customer_service <- customer_service[,-2]
```

```
# Customer Service Interaction by Interaction Type & Resolution Status
```

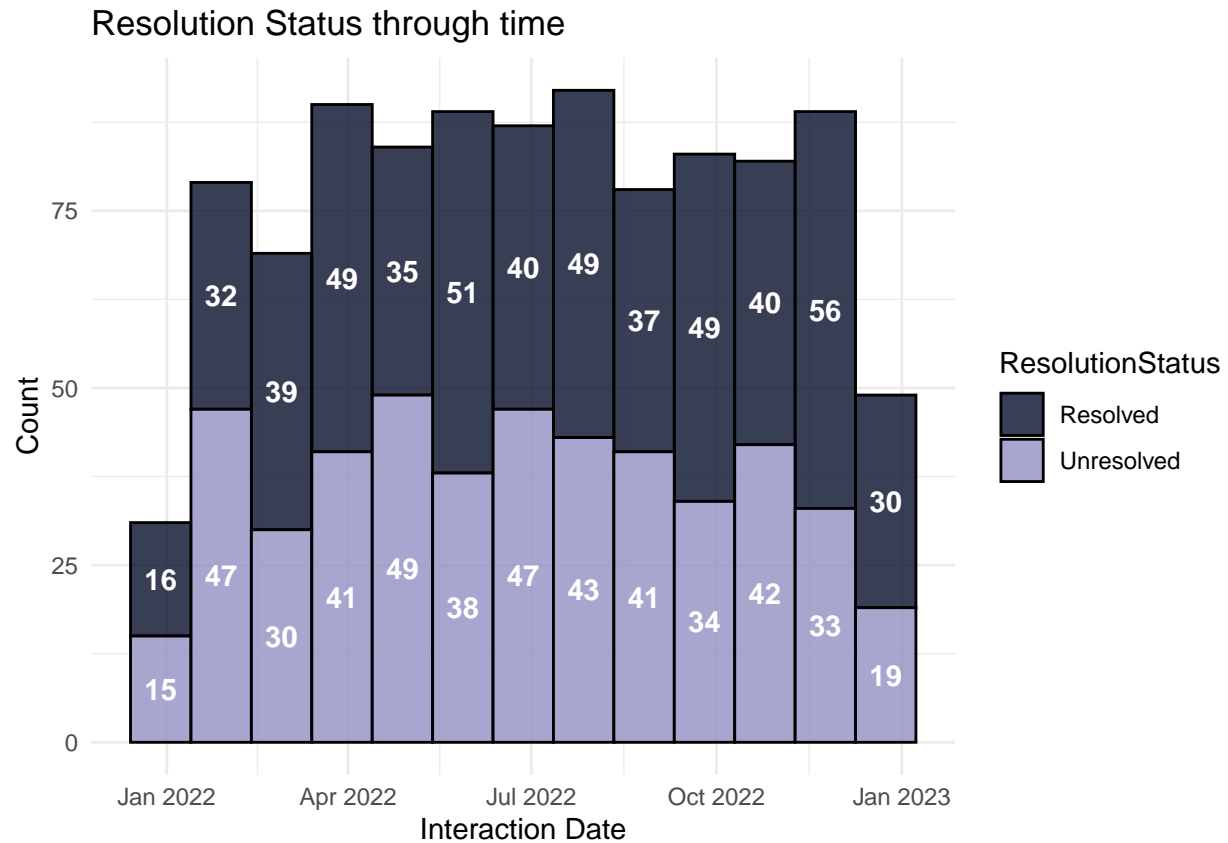
```
ggplot(data = customer_service, aes(x= InteractionType, fill = ResolutionStatus))+  
  geom_histogram(stat = "count",position = "dodge",color= "black", alpha= 0.8)+  
  scale_fill_manual(values = c("Resolved" = "#1B1A55", "Unresolved"= "#9290C3"))+  
  labs(title = "Customer Service Interaction by Interaction Type & Resolution Status", x= "Interaction Type", y= "Count") +  
  geom_text(stat= "count", aes(label = ..count..),position = position_dodge(width = 0.9), vjust= -0.5, color="black") +  
  theme_minimal()
```



```
# Plotting Interaction Type through time
ggplot(data = customer_service, aes(x = as.Date(InteractionDate), fill = InteractionType)) +
  geom_histogram(position = "stack", binwidth = 30, color = "black", alpha = 0.8) +
  scale_fill_manual(values = c("Feedback" = "#535C91", "Inquiry" = "#070F2B", "Complaint" = "#9290C3")) +
  geom_text(stat = "bin", binwidth = 30, aes(label = ..count.., group = InteractionType),
    position = position_stack(vjust = 0.5), color = "white", size = 4, fontface = "bold") +
  labs(title = "Interaction Type through time ", x = "Interaction Date", y = "Count") +
  theme_minimal()
```



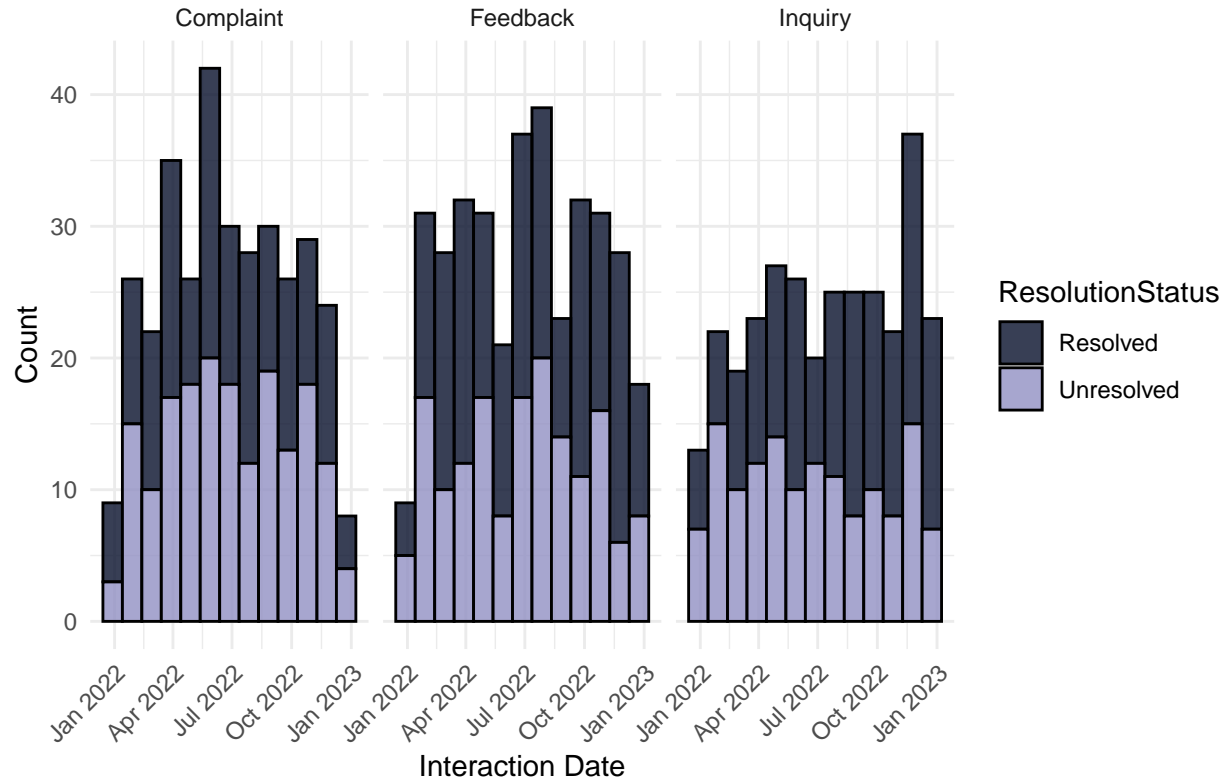
```
# Plotting Resolution Status through time
ggplot(data = customer_service, aes(x = as.Date(InteractionDate), fill = ResolutionStatus)) +
  geom_histogram(position = "stack", binwidth = 30, color = "black", alpha = 0.8) +
  scale_fill_manual(values = c("Resolved" = "#070F2B", "Unresolved" = "#9290C3")) +
  geom_text(stat = "bin", binwidth = 30, aes(label = ..count.., group = ResolutionStatus),
    position = position_stack(vjust = 0.5), color = "white", size = 4, fontface = "bold") +
  labs(title = "Resolution Status through time ", x = "Interaction Date", y = "Count") +
  theme_minimal()
```



The features and their relationship with each other are uniform. The interaction types and Resolution Status have stayed consistent over time.

```
# Plotting Interaction & Resolution Status through time
ggplot(data = customer_service, aes(x = as.Date(InteractionDate), fill = ResolutionStatus)) +
  geom_histogram(position = "stack", binwidth = 30, color = "black", alpha = 0.8) +
  scale_fill_manual(values = c("Resolved" = "#070F2B", "Unresolved" = "#9290C3")) +
  labs(title = "Interaction & Resolution Status through time",
       x = "Interaction Date", y = "Count") +
  theme_minimal() +
  facet_wrap(vars(InteractionType)) + theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

Interaction & Resolution Status through time



The complaints have reduced recently and enquiry resolution have gotten better over time which is shown by the increasing share of 'Resolved' status across categories.

The Unresolved Enquiries, particularly Complaints could be a significant factor in customer churn.

```
head(churn_status)
```

```
## # A tibble: 6 x 2
##   CustomerID ChurnStatus
##       <dbl>       <dbl>
## 1         1           0
## 2         2           1
## 3         3           0
## 4         4           0
## 5         5           0
## 6         6           0
```

```
# Plot churn status of customers who have unresolved complaints
df1<- unique(customer_service[customer_service$InteractionType == "Complaint" &
                             customer_service$ResolutionStatus == "Unresolved",
                             c(1, 3, 4)])
df2 <-left_join(df1, churn_status, by= "CustomerID")

ggplot(df2, aes(x = as.factor(ChurnStatus))) +
  geom_bar(fill = "navyblue", color = "black", alpha = 0.8) +
  theme_minimal() +
```



```
labs(title = "Churn status of customers with unresolved complaints", x = "Churn Status", y = "Count") +
geom_text(stat = "count", aes(label = ..count..), vjust = -0.5, fontface = "bold")
```



A direct correlation cannot be found between Unresolved complaints and Churn status.

```
# creating feature days since last interaction to the most recent date
max_date <- as.Date(max(customer_service$InteractionDate))
customer_service$InteractionDate <- as.Date(customer_service$InteractionDate)
customer_service$DaysSinceLastInteraction <- as.numeric(max_date - customer_service$InteractionDate)
head(customer_service)
```

```
## # A tibble: 6 x 5
##   CustomerID InteractionDate InteractionType ResolutionStatus
##       <dbl> <date>           <chr>           <chr>
## 1         1 2022-03-31      Inquiry        Resolved
## 2         2 2022-03-17      Inquiry        Resolved
## 3         3 2022-08-24      Inquiry        Resolved
## 4         4 2022-11-18      Inquiry        Resolved
## 5         4 2022-07-03      Inquiry        Unresolved
## 6         6 2022-05-05      Feedback       Resolved
## # i 1 more variable: DaysSinceLastInteraction <dbl>
```

```
# Encoding the categorical variables
# one-hot encoding for Interaction Type & Resolution Status
customer_service <- dummy_cols(customer_service, select_columns = "InteractionType",
```

```

remove_first_dummy = FALSE,
remove_selected_columns = TRUE)

customer_service <- dummy_cols(customer_service, select_columns = "ResolutionStatus",
                                remove_first_dummy = FALSE,
                                remove_selected_columns = TRUE)

# dropping InteractionDate column
customer_service<- customer_service[,-2]

# converting the Interaction date to DaysSinceLastLogin and grouping by customer ID
customer_service<- customer_service %>% group_by(CustomerID) %>% summarise(DaysSinceLastInteraction =ma
                                Resolved= sum(ResolutionStatus_Resolved),
                                Unresolved= sum(ResolutionStatus_Unresolved),
                                Complaint= sum(InteractionType_Complaint),
                                Feedback= sum(InteractionType_Feedback),
                                Inquiry= sum(InteractionType_Inquiry))

```

```
head(online_activity)
```

Online Activity

```
## # A tibble: 6 x 4
##   CustomerID LastLoginDate      LoginFrequency ServiceUsage
##   <dbl> <dtm>          <dbl> <chr>
## 1         1 2023-10-21 00:00:00          34 Mobile App
## 2         2 2023-12-05 00:00:00           5 Website
## 3         3 2023-11-15 00:00:00           3 Website
## 4         4 2023-08-25 00:00:00           2 Website
## 5         5 2023-10-27 00:00:00          41 Website
## 6         6 2023-09-17 00:00:00           2 Website

```

```
str(online_activity)
```

```
## tibble [1,000 x 4] (S3: tbl_df/tbl/data.frame)
## $ CustomerID : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
## $ LastLoginDate : POSIXct[1:1000], format: "2023-10-21" "2023-12-05" ...
## $ LoginFrequency: num [1:1000] 34 5 3 2 41 2 32 17 24 29 ...
## $ ServiceUsage : chr [1:1000] "Mobile App" "Website" "Website" "Website" ...

```

```
summary(online_activity)
```

```
##   CustomerID      LastLoginDate      LoginFrequency
## Min.   : 1.0    Min.   :2023-01-01 00:00:00   Min.   : 1.00
## 1st Qu.: 250.8  1st Qu.:2023-04-08 00:00:00   1st Qu.:13.75
## Median : 500.5  Median :2023-07-10 12:00:00   Median :27.00
## Mean   : 500.5  Mean   :2023-07-05 21:28:48   Mean   :25.91
## 3rd Qu.: 750.2  3rd Qu.:2023-10-01 06:00:00   3rd Qu.:38.00

```

```
## Max.      :1000.0    Max.      :2023-12-31 00:00:00    Max.      :49.00
## ServiceUsage
## Length:1000
## Class :character
## Mode  :character
##
##
##
```

```
table(online_activity$ServiceUsage)
```

```
##
##      Mobile App Online Banking      Website
##           342           349           309
```

```
# check null and duplicates
sum(is.na(online_activity))
```

```
## [1] 0
```

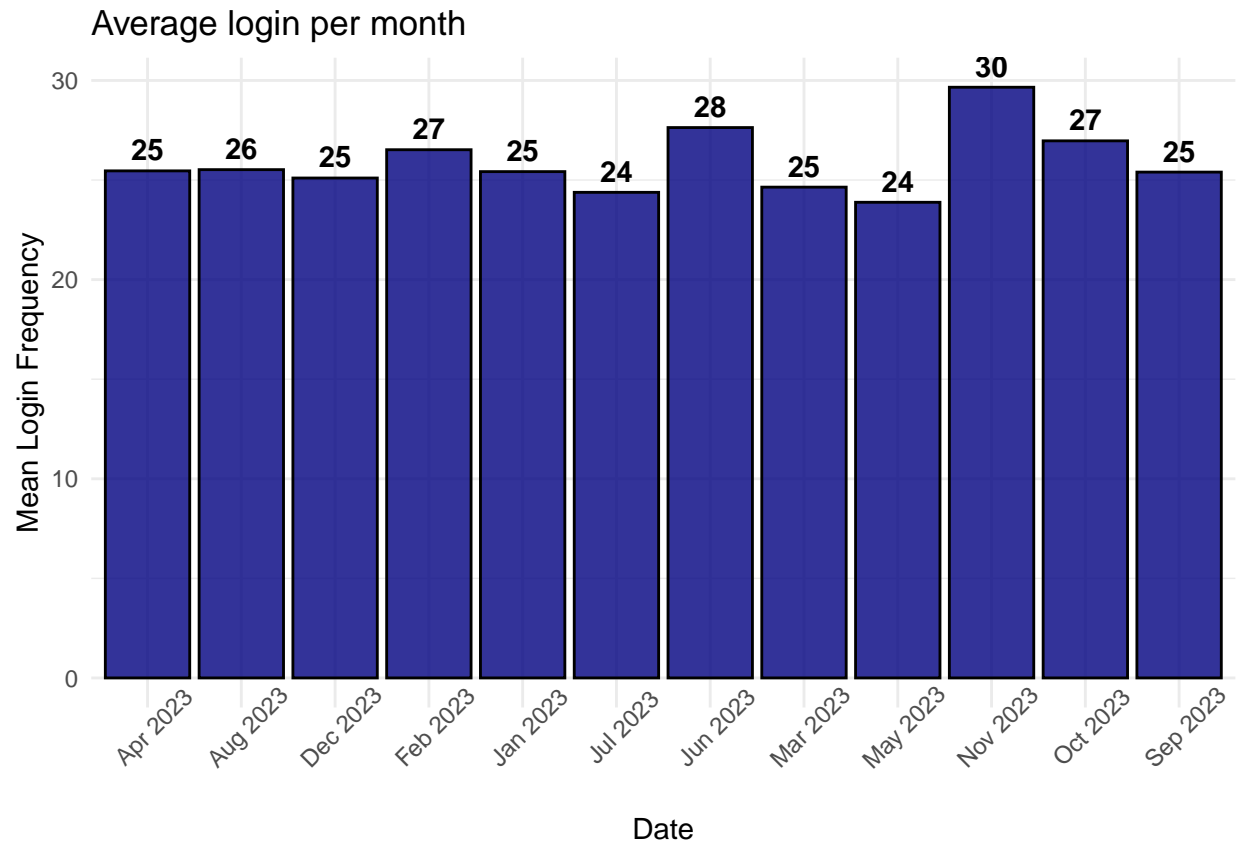
```
sum(duplicated(online_activity))
```

```
## [1] 0
```

The descriptive statistics of the dataset shows a well balanced customer activity data.

- Login Frequency: Customers have a mean login frequency of 25 times, with a min of 1 and max 49 throughout the time period
- Service Usage: There are 3 types of service customers avail, those are through mobile app, Online banking and website. And the data has roughly equal distribution

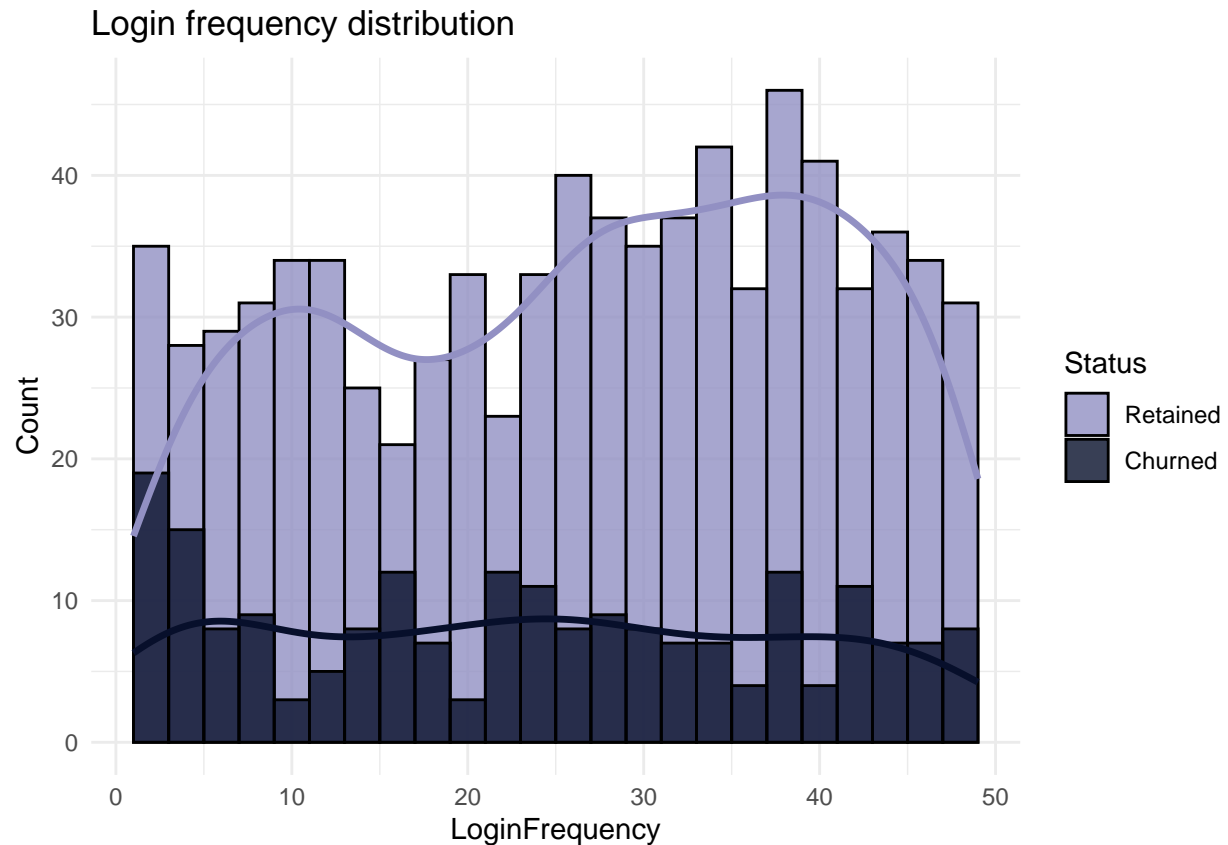
```
# Average login per month
online_activity %>%
  mutate(monthYear = floor_date>LastLoginDate, "month")) %>%
  group_by(monthYear) %>%
  summarise(mean = mean(LoginFrequency), .groups = "drop") %>%
  mutate(monthYear = format(monthYear, "%b %Y")) %>%
  ggplot(aes(x= monthYear, y= mean)) +
  geom_bar(stat = "identity", fill = "navyblue", color = "black", alpha = 0.8)+
  theme(axis.text.x = element_text(angle = 45))+
  labs(title = "Average login per month", x= "Date", y= "Mean Login Frequency")+
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45))+
  geom_text(aes(label = round(mean,0)), vjust = -0.5, fontface = "bold")
```



Average login Frequency is distributed equally over the year

```
df3<- left_join(online_activity, churn_status, by= "CustomerID")
```

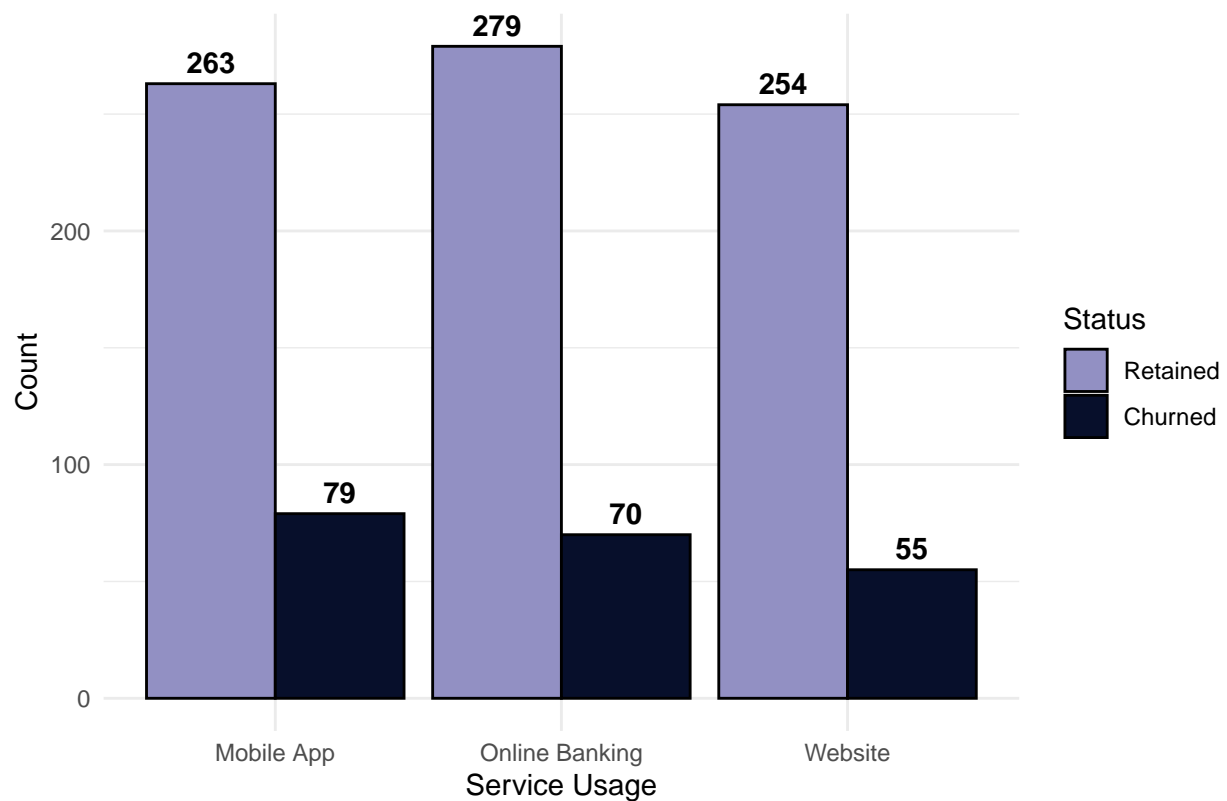
```
# Login frequency distribution with churn status
ggplot(data= df3, aes(x = LoginFrequency, fill = factor(ChurnStatus,labels = c("Retained","Churned")))) +
  geom_histogram(binwidth= 2, color= "black", alpha= 0.8,position = "identity")+
  geom_density(aes(y = ..count.. * 2,
    color = factor(ChurnStatus,labels = c("Retained", "Churned"))),
    size = 1.2, alpha = 0,show.legend = FALSE) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  scale_color_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Login frequency distribution", x = "LoginFrequency", y = "Count", fill = "Status")+
  theme_minimal()
```



The login frequencies of customers is spread across uniformly, but customers with very low login frequency tend to have churned more.

```
# Service Usage Pattern by Churn Status
ggplot(data = df3)+
  geom_histogram(aes(x= ServiceUsage,
    fill= factor(ChurnStatus,labels = c("Retained","Churned"))),
    stat = "count", position = "dodge", color= "black")+
  geom_text(stat = "count",aes(x = ServiceUsage, label= ..count.., group = factor(ChurnStatus)),
    position = position_dodge(width = 0.9),color = "black", size = 4,fontface= "bold", vjust= -0.5)+
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Service Usage Pattern by Churn Status", x= "Service Usage", y = "Count", fill= "Status")
  theme_minimal()
```

Service Usage Pattern by Churn Status



```
# creating feature days since last login to the most recent date
max_date <- as.Date(max(online_activity$LastLoginDate))
online_activity$LastLoginDate <- as.Date(online_activity$LastLoginDate)
online_activity$DaysSinceLastLogin <- as.numeric(max_date - online_activity$LastLoginDate)

#remove Last Login Date
online_activity <- online_activity[,-2]
```

The customer use all the banking services equally.

```
# one-hot encoding for Service Usage
online_activity <- dummy_cols(online_activity,
                              select_columns = "ServiceUsage",
                              remove_first_dummy = FALSE,
                              remove_selected_columns = TRUE)

head(online_activity)
```

```
## # A tibble: 6 x 6
##   CustomerID LoginFrequency DaysSinceLastLogin `ServiceUsage_Mobile App`
##       <dbl>         <dbl>         <dbl>                <int>
## 1         1             34             71                  1
## 2         2              5             26                  0
## 3         3              3             46                  0
## 4         4              2            128                  0
```

```
## 5          5          41          65          0
## 6          6          2         105          0
## # i 2 more variables: `ServiceUsage_Online Banking` <int>,
## #   ServiceUsage_Website <int>
```

Merging all the 5 datasets to see the patterns

```
str(customer_demo)
```

```
## tibble [1,000 x 8] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID      : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Age              : num [1:1000] 62 65 18 21 21 57 27 37 39 68 ...
##  $ Gender           : num [1:1000] 0 0 0 0 0 1 1 0 0 0 ...
##  $ IncomeLevel      : num [1:1000] 0 0 0 0 1 1 2 0 2 2 ...
##  $ MaritalStatus_Divorced: int [1:1000] 0 0 0 0 1 1 0 0 1 0 ...
##  $ MaritalStatus_Married : int [1:1000] 0 1 0 0 0 0 1 0 0 1 ...
##  $ MaritalStatus_Single  : int [1:1000] 1 0 1 0 0 0 0 1 0 0 ...
##  $ MaritalStatus_Widowed : int [1:1000] 0 0 0 1 0 0 0 0 0 0 ...
```

```
str(transaction_hist)
```

```
## tibble [1,000 x 8] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID      : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Total_amount_spent : num [1:1000] 416 1547 1703 917 2001 ...
##  $ Transaction_frequency: int [1:1000] 1 7 6 5 8 5 1 7 5 7 ...
##  $ Books            : int [1:1000] 0 0 1 0 0 1 1 2 0 0 ...
##  $ Clothing         : int [1:1000] 0 2 1 1 0 1 0 2 2 1 ...
##  $ Electronics      : int [1:1000] 1 3 0 2 3 1 0 1 1 1 ...
##  $ Furniture        : int [1:1000] 0 1 2 1 2 2 0 2 0 3 ...
##  $ Groceries        : int [1:1000] 0 1 2 1 3 0 0 0 2 2 ...
```

```
str(customer_service)
```

```
## tibble [668 x 7] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID      : num [1:668] 1 2 3 4 6 8 9 11 12 13 ...
##  $ DaysSinceLastInteraction: num [1:668] 274 288 128 180 239 245 137 338 352 156 ...
##  $ Resolved        : int [1:668] 1 1 1 1 1 0 1 1 1 1 ...
##  $ Unresolved      : int [1:668] 0 0 0 1 0 2 0 1 1 0 ...
##  $ Complaint       : int [1:668] 0 0 0 0 0 0 0 0 0 1 ...
##  $ Feedback        : int [1:668] 0 0 0 0 1 1 0 2 1 0 ...
##  $ Inquiry         : int [1:668] 1 1 1 2 0 1 1 0 1 0 ...
```

```
str(online_activity)
```

```
## tibble [1,000 x 6] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID      : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ LoginFrequency   : num [1:1000] 34 5 3 2 41 2 32 17 24 29 ...
##  $ DaysSinceLastLogin : num [1:1000] 71 26 46 128 65 105 358 253 187 352 ...
##  $ ServiceUsage_Mobile App : int [1:1000] 1 0 0 0 0 0 1 0 0 0 ...
##  $ ServiceUsage_Online Banking: int [1:1000] 0 0 0 0 0 0 0 1 0 1 ...
##  $ ServiceUsage_Website : int [1:1000] 0 1 1 1 1 1 0 0 1 0 ...
```

```
str(churn_status)
```

```
## tibble [1,000 x 2] (S3: tbl_df/tbl/data.frame)
## $ CustomerID : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
## $ ChurnStatus: num [1:1000] 0 1 0 0 0 0 0 1 0 1 ...
```

```
customer_churn_df <- customer_demo %>%
  left_join(transaction_hist, by = "CustomerID") %>%
  left_join(customer_service, by = "CustomerID") %>%
  left_join(online_activity, by = "CustomerID") %>%
  left_join(churn_status, by = "CustomerID")
```

```
sum(duplicated(customer_churn_df))
```

```
## [1] 0
```

```
colSums(is.na(customer_churn_df))
```

```
##           CustomerID           Age
##           0           0
##           Gender           IncomeLevel
##           0           0
## MaritalStatus_Divorced MaritalStatus_Married
##           0           0
## MaritalStatus_Single MaritalStatus_Widowed
##           0           0
## Total_amount_spent Transaction_frequency
##           0           0
##           Books           Clothing
##           0           0
##           Electronics           Furniture
##           0           0
##           Groceries DaysSinceLastInteraction
##           0           332
##           Resolved           Unresolved
##           332           332
##           Complaint           Feedback
##           332           332
##           Inquiry           LoginFrequency
##           332           0
## DaysSinceLastLogin ServiceUsage_Mobile App
##           0           0
## ServiceUsage_Online Banking ServiceUsage_Website
##           0           0
##           ChurnStatus
##           0
```

There are large amount of null values in some columns, removing all these column can result in significant amount of data loss so we have to impute the null values using appropriate method before sending it for model prediction

Best method for imputing columns like Resolved, Unresolved, Complaint, Feedback & Inquiry is to replace with 0 as the customer had never had any interaction


```
customer_churn_df$Resolved[is.na(customer_churn_df$Resolved)] = 0
customer_churn_df$Unresolved[is.na(customer_churn_df$Unresolved)] = 0
customer_churn_df$Complaint[is.na(customer_churn_df$Complaint)] = 0
customer_churn_df$Feedback[is.na(customer_churn_df$Feedback)] = 0
customer_churn_df$Inquiry[is.na(customer_churn_df$Inquiry)] = 0

customer_churn_df$DaysSinceLastInteraction[is.na(customer_churn_df$DaysSinceLastInteraction)] =
  median(customer_churn_df$DaysSinceLastInteraction, na.rm = TRUE)

sum(is.na(customer_churn_df))
```

```
## [1] 0
```

```
uniqueN(customer_churn_df$CustomerID)
```

```
## [1] 1000
```

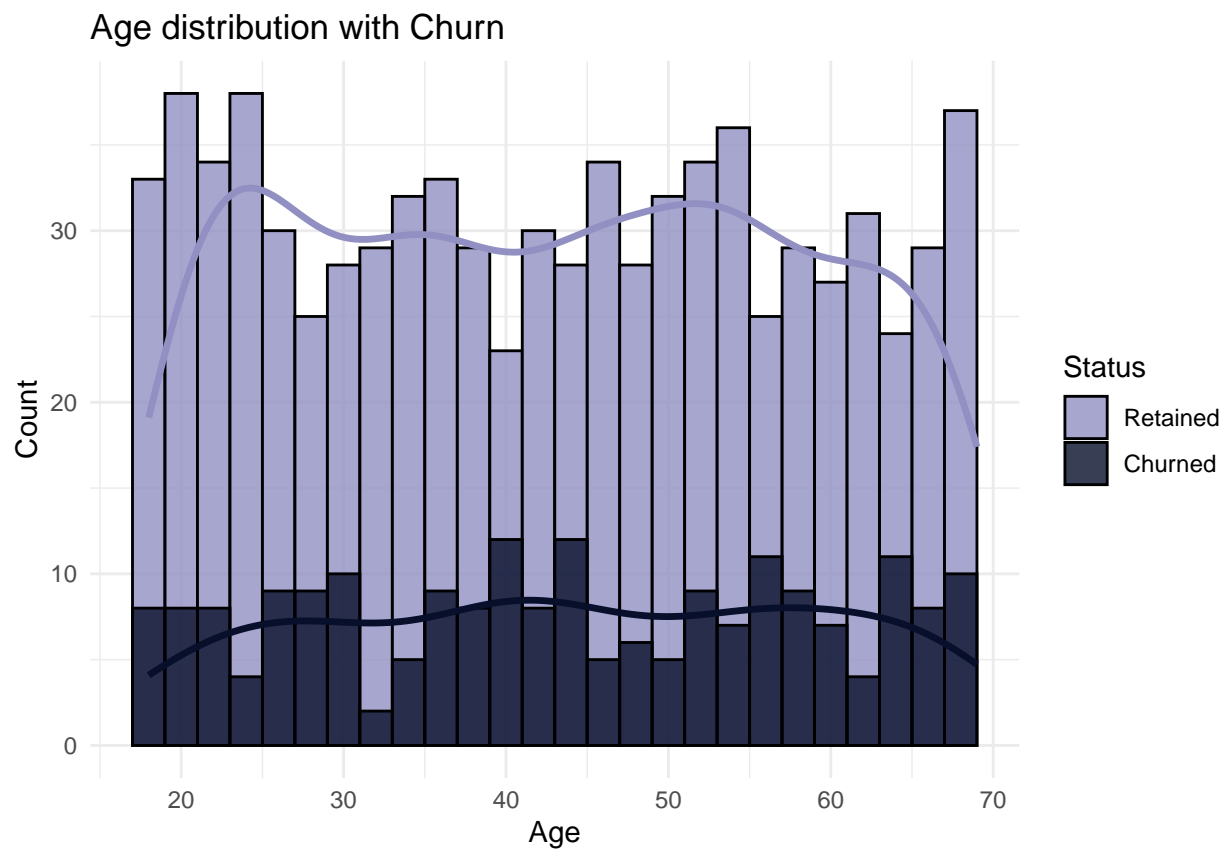
There are no duplicate customer ID in the dataset

```
str(customer_churn_df)
```

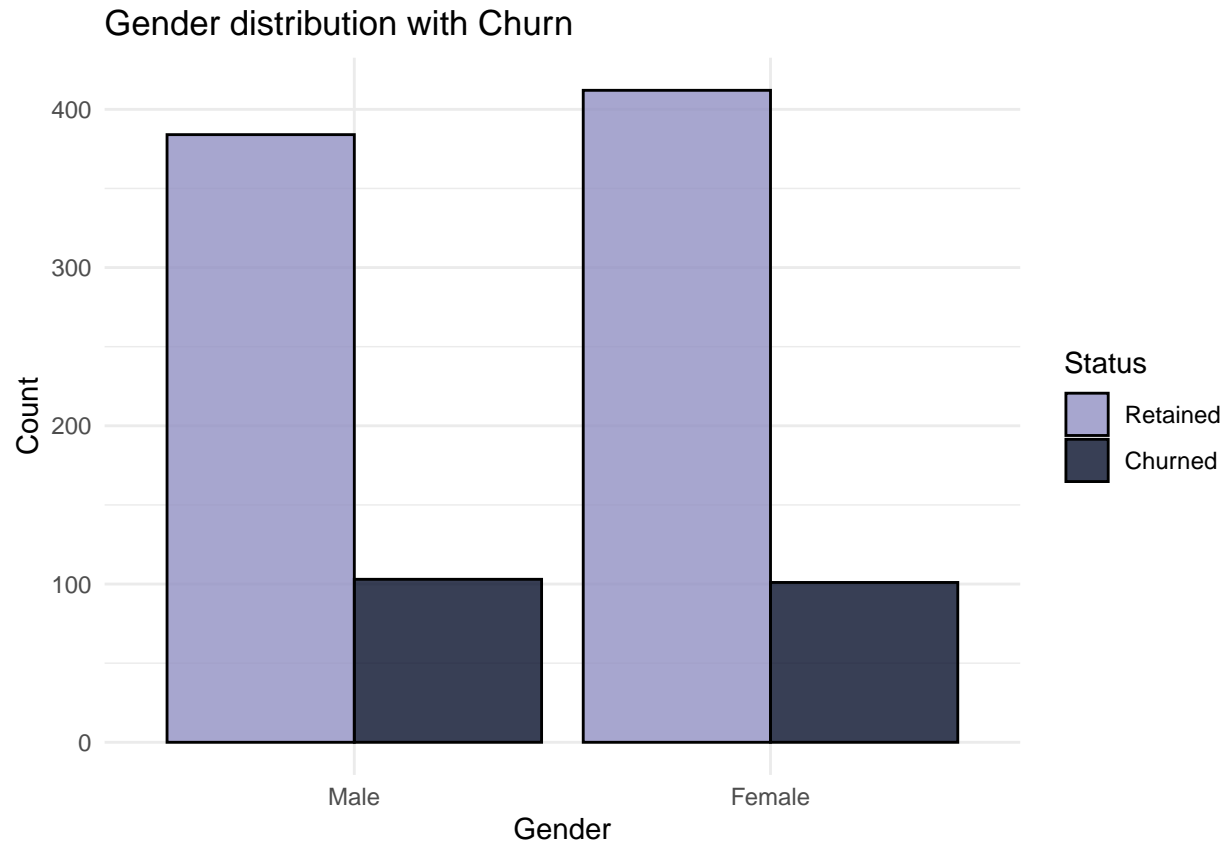
```
## tibble [1,000 x 27] (S3: tbl_df/tbl/data.frame)
## $ CustomerID      : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
## $ Age             : num [1:1000] 62 65 18 21 21 57 27 37 39 68 ...
## $ Gender          : num [1:1000] 0 0 0 0 0 1 1 0 0 0 ...
## $ IncomeLevel     : num [1:1000] 0 0 0 0 1 1 2 0 2 2 ...
## $ MaritalStatus_Divorced : int [1:1000] 0 0 0 0 1 1 0 0 1 0 ...
## $ MaritalStatus_Married : int [1:1000] 0 1 0 0 0 0 1 0 0 1 ...
## $ MaritalStatus_Single : int [1:1000] 1 0 1 0 0 0 0 1 0 0 ...
## $ MaritalStatus_Widowed : int [1:1000] 0 0 0 1 0 0 0 0 0 0 ...
## $ Total_amount_spent : num [1:1000] 416 1547 1703 917 2001 ...
## $ Transaction_frequency : int [1:1000] 1 7 6 5 8 5 1 7 5 7 ...
## $ Books           : int [1:1000] 0 0 1 0 0 1 1 2 0 0 ...
## $ Clothing        : int [1:1000] 0 2 1 1 0 1 0 2 2 1 ...
## $ Electronics     : int [1:1000] 1 3 0 2 3 1 0 1 1 1 ...
## $ Furniture       : int [1:1000] 0 1 2 1 2 2 0 2 0 3 ...
## $ Groceries       : int [1:1000] 0 1 2 1 3 0 0 0 2 2 ...
## $ DaysSinceLastInteraction : num [1:1000] 274 288 128 180 215 239 215 245 137 215 ...
## $ Resolved        : num [1:1000] 1 1 1 1 0 1 0 0 1 0 ...
## $ Unresolved      : num [1:1000] 0 0 0 1 0 0 0 2 0 0 ...
## $ Complaint       : num [1:1000] 0 0 0 0 0 0 0 0 0 0 ...
## $ Feedback        : num [1:1000] 0 0 0 0 0 1 0 1 0 0 ...
## $ Inquiry         : num [1:1000] 1 1 1 2 0 0 0 1 1 0 ...
## $ LoginFrequency  : num [1:1000] 34 5 3 2 41 2 32 17 24 29 ...
## $ DaysSinceLastLogin : num [1:1000] 71 26 46 128 65 105 358 253 187 352 ...
## $ ServiceUsage_Mobile App : int [1:1000] 1 0 0 0 0 0 1 0 0 0 ...
## $ ServiceUsage_Online Banking: int [1:1000] 0 0 0 0 0 0 0 1 0 1 ...
## $ ServiceUsage_Website : int [1:1000] 0 1 1 1 1 1 0 0 1 0 ...
## $ ChurnStatus     : num [1:1000] 0 1 0 0 0 0 0 1 0 1 ...
```

EDA on preprocessed data:

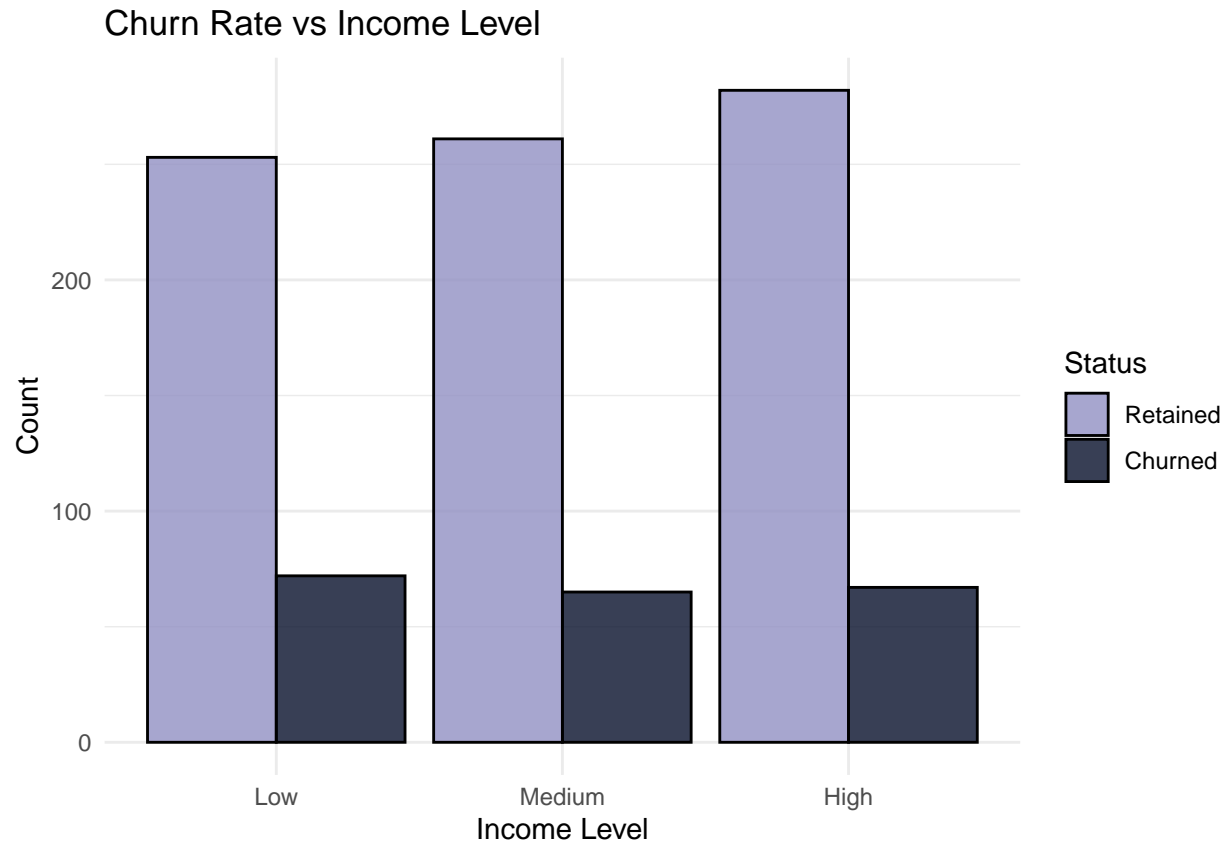
```
# Age distribution with Churn
ggplot(data= customer_churn_df, aes(x= Age, fill = factor(ChurnStatus,labels = c("Retained","Churned")))) +
  geom_histogram(binwidth=2, color= "black", alpha= 0.8,position = "identity")+
  geom_density(aes(y = ..count.. * 2,
                  color = factor(ChurnStatus,labels = c("Retained", "Churned"))),
              size = 1.2, alpha = 0,show.legend = FALSE) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  scale_color_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Age distribution with Churn", x= "Age", y = "Count", fill = "Status")+
  theme_minimal()
```



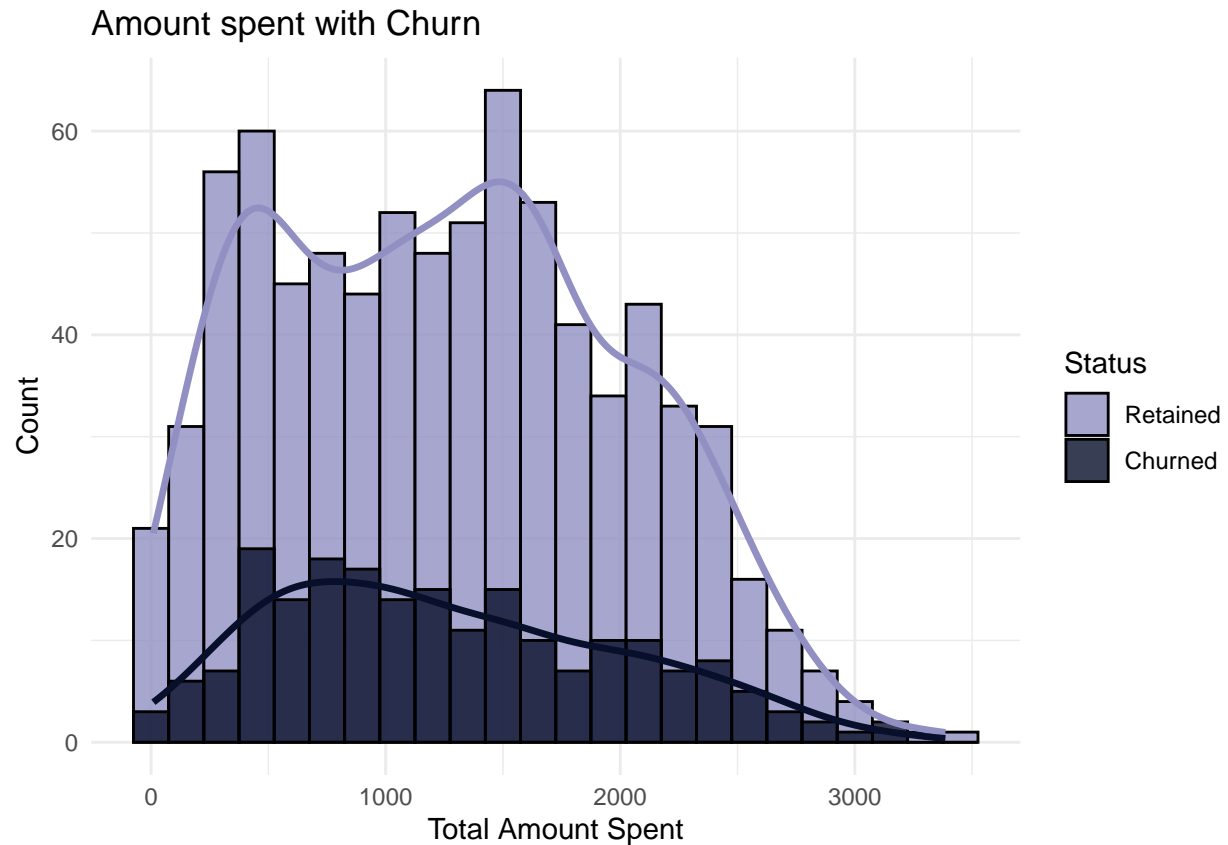
```
# Gender distribution with Churn
ggplot(data = customer_churn_df,
       aes(x = factor(Gender), fill = factor(ChurnStatus, labels =
                                             c("Retained", "Churned")))) +
  geom_bar(color = "black", alpha = 0.8, position = "dodge") +
  scale_x_discrete(labels = c("0" = "Male", "1" = "Female")) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Gender distribution with Churn", x = "Gender", y = "Count", fill = "Status") +
  theme_minimal()
```



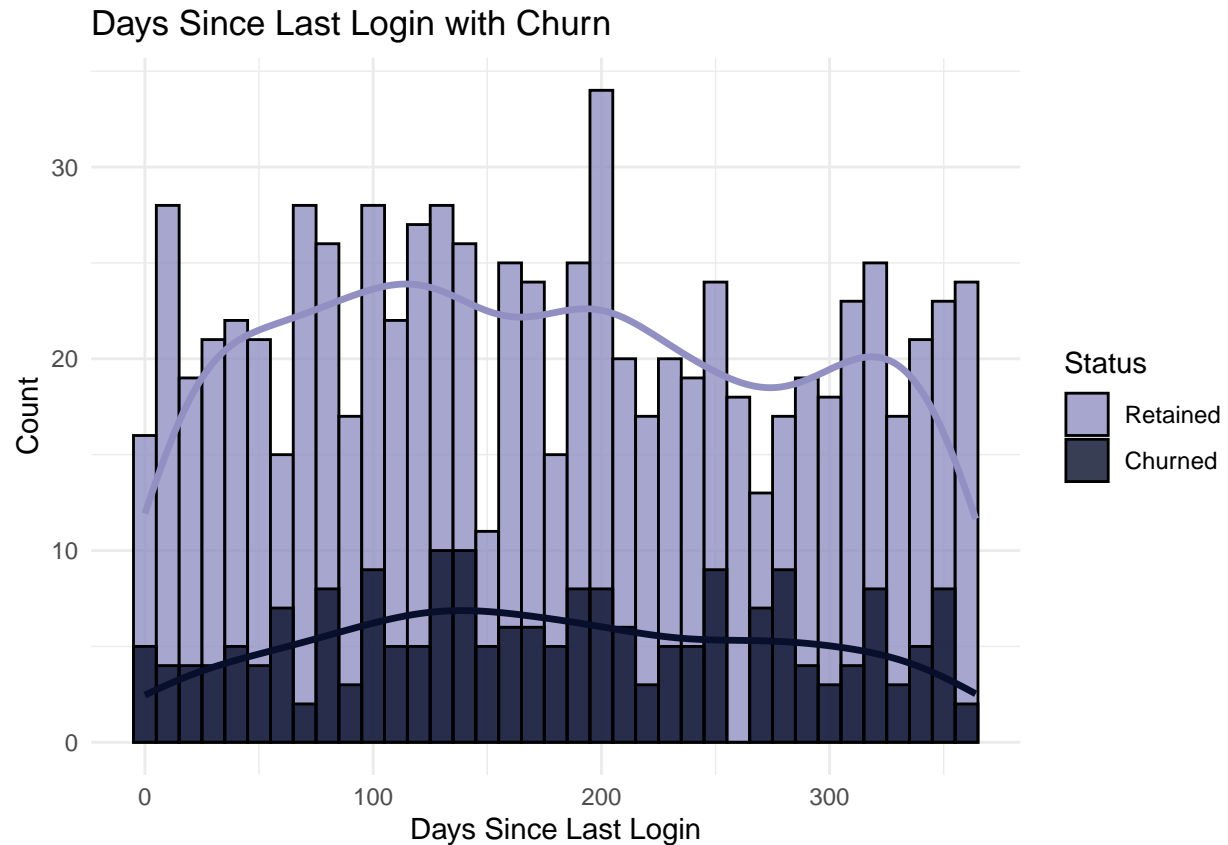
```
# Churn Rate vs Income Level
ggplot(data = customer_churn_df,
  aes(x = factor(IncomeLevel), fill = factor(ChurnStatus, labels = c("Retained", "Churned")))) +
  geom_bar(color = "black", alpha = 0.8, position = "dodge") +
  scale_x_discrete(labels = c("0" = "Low", "1" = "Medium", "2" = "High")) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Churn Rate vs Income Level", x = "Income Level", y = "Count", fill = "Status") +
  theme_minimal()
```



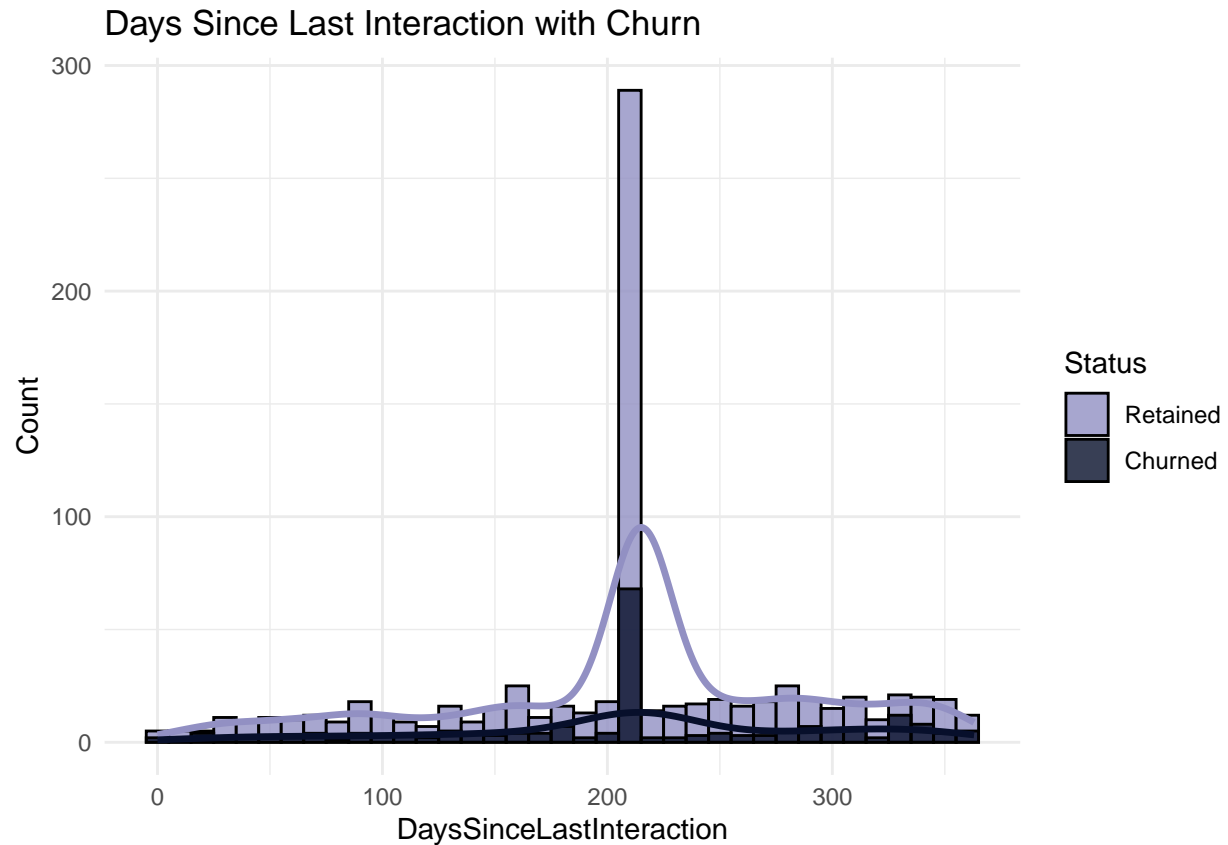
```
# Amount spent with Churn
ggplot(data= customer_churn_df, aes(x = Total_amount_spent, fill =
                                     factor(ChurnStatus,labels=
                                             c("Retained","Churned")))) +
  geom_histogram(binwidth= 150, color= "black", alpha= 0.8,position = "identity")+
  geom_density(aes(y = ..count.. * 150,
                  color = factor(ChurnStatus,labels = c("Retained", "Churned"))),
              size = 1.2, alpha = 0,show.legend = FALSE) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  scale_color_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Amount spent with Churn", x= "Total Amount Spent", y = "Count", fill = "Status")+
  theme_minimal()
```



```
# Days Since Last Login with Churn
ggplot(data= customer_churn_df, aes(x = DaysSinceLastLogin, fill = factor(ChurnStatus,labels = c("Retained", "Churned")),
  geom_histogram(binwidth= 10, color= "black", alpha= 0.8,position = "identity")+
  geom_density(aes(y = ..count.. * 10,
                    color = factor(ChurnStatus,labels = c("Retained", "Churned"))),
              size = 1.2, alpha = 0,show.legend = FALSE) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  scale_color_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Days Since Last Login with Churn",
        x= "Days Since Last Login", y = "Count", fill = "Status")+
  theme_minimal()
```



```
# Days Since Last Interaction with Churn
ggplot(data= customer_churn_df, aes(x = DaysSinceLastInteraction,fill=
                                     factor(ChurnStatus,labels=
                                             c("Retained","Churned"))))+
  geom_histogram(binwidth= 10, color= "black", alpha= 0.8,position = "identity")+
  geom_density(aes(y = ..count.. * 10,
                  color = factor(ChurnStatus,labels = c("Retained", "Churned"))),
              size = 1.2, alpha = 0,show.legend = FALSE) +
  scale_fill_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  scale_color_manual(values = c("Churned" = "#070F2B", "Retained" = "#9290C3")) +
  labs(title = "Days Since Last Interaction with Churn", x= "DaysSinceLastInteraction",
       y = "Count", fill = "Status")+
  theme_minimal()
```



```
str(customer_churn_df)
```

```
## tibble [1,000 x 27] (S3: tbl_df/tbl/data.frame)
##  $ CustomerID      : num [1:1000] 1 2 3 4 5 6 7 8 9 10 ...
##  $ Age              : num [1:1000] 62 65 18 21 21 57 27 37 39 68 ...
##  $ Gender            : num [1:1000] 0 0 0 0 0 1 1 0 0 0 ...
##  $ IncomeLevel       : num [1:1000] 0 0 0 0 1 1 2 0 2 2 ...
##  $ MaritalStatus_Divorced : int [1:1000] 0 0 0 0 1 1 0 0 1 0 ...
##  $ MaritalStatus_Married  : int [1:1000] 0 1 0 0 0 0 1 0 0 1 ...
##  $ MaritalStatus_Single   : int [1:1000] 1 0 1 0 0 0 0 1 0 0 ...
##  $ MaritalStatus_Widowed  : int [1:1000] 0 0 0 1 0 0 0 0 0 0 ...
##  $ Total_amount_spent     : num [1:1000] 416 1547 1703 917 2001 ...
##  $ Transaction_frequency : int [1:1000] 1 7 6 5 8 5 1 7 5 7 ...
##  $ Books                 : int [1:1000] 0 0 1 0 0 1 1 2 0 0 ...
##  $ Clothing              : int [1:1000] 0 2 1 1 0 1 0 2 2 1 ...
##  $ Electronics           : int [1:1000] 1 3 0 2 3 1 0 1 1 1 ...
##  $ Furniture             : int [1:1000] 0 1 2 1 2 2 0 2 0 3 ...
##  $ Groceries             : int [1:1000] 0 1 2 1 3 0 0 0 2 2 ...
##  $ DaysSinceLastInteraction : num [1:1000] 274 288 128 180 215 239 215 245 137 215 ...
##  $ Resolved              : num [1:1000] 1 1 1 1 0 1 0 0 1 0 ...
##  $ Unresolved            : num [1:1000] 0 0 0 1 0 0 0 2 0 0 ...
##  $ Complaint             : num [1:1000] 0 0 0 0 0 0 0 0 0 0 ...
##  $ Feedback              : num [1:1000] 0 0 0 0 0 1 0 1 0 0 ...
##  $ Inquiry               : num [1:1000] 1 1 1 2 0 0 0 1 1 0 ...
##  $ LoginFrequency        : num [1:1000] 34 5 3 2 41 2 32 17 24 29 ...
```

```
## $ DaysSinceLastLogin      : num [1:1000] 71 26 46 128 65 105 358 253 187 352 ...
## $ ServiceUsage_Mobile App : int [1:1000] 1 0 0 0 0 0 1 0 0 0 ...
## $ ServiceUsage_Online Banking: int [1:1000] 0 0 0 0 0 0 0 1 0 1 ...
## $ ServiceUsage_Website    : int [1:1000] 0 1 1 1 1 1 0 0 1 0 ...
## $ ChurnStatus              : num [1:1000] 0 1 0 0 0 0 0 1 0 1 ...
```

```
# remove custome ID column
customer_churn_df<- customer_churn_df[,-1]
```

```
# Plot heatmap
# convert all columns to numeric

i <- c(1:26)
customer_churn_df[, i] <- apply(customer_churn_df[, i], 2, function(x) as.numeric(x))
```

```
str(customer_churn_df)
```

```
## tibble [1,000 x 26] (S3: tbl_df/tbl/data.frame)
## $ Age                : num [1:1000] 62 65 18 21 21 57 27 37 39 68 ...
## $ Gender              : num [1:1000] 0 0 0 0 0 1 1 0 0 0 ...
## $ IncomeLevel         : num [1:1000] 0 0 0 0 1 1 2 0 2 2 ...
## $ MaritalStatus_Divorced : num [1:1000] 0 0 0 0 1 1 0 0 1 0 ...
## $ MaritalStatus_Married : num [1:1000] 0 1 0 0 0 0 1 0 0 1 ...
## $ MaritalStatus_Single  : num [1:1000] 1 0 1 0 0 0 0 1 0 0 ...
## $ MaritalStatus_Widowed : num [1:1000] 0 0 0 1 0 0 0 0 0 0 ...
## $ Total_amount_spent    : num [1:1000] 416 1547 1703 917 2001 ...
## $ Transaction_frequency : num [1:1000] 1 7 6 5 8 5 1 7 5 7 ...
## $ Books                : num [1:1000] 0 0 1 0 0 1 1 2 0 0 ...
## $ Clothing              : num [1:1000] 0 2 1 1 0 1 0 2 2 1 ...
## $ Electronics           : num [1:1000] 1 3 0 2 3 1 0 1 1 1 ...
## $ Furniture             : num [1:1000] 0 1 2 1 2 2 0 2 0 3 ...
## $ Groceries             : num [1:1000] 0 1 2 1 3 0 0 0 2 2 ...
## $ DaysSinceLastInteraction : num [1:1000] 274 288 128 180 215 239 215 245 137 215 ...
## $ Resolved              : num [1:1000] 1 1 1 1 0 1 0 0 1 0 ...
## $ Unresolved            : num [1:1000] 0 0 0 1 0 0 0 2 0 0 ...
## $ Complaint             : num [1:1000] 0 0 0 0 0 0 0 0 0 0 ...
## $ Feedback              : num [1:1000] 0 0 0 0 0 1 0 1 0 0 ...
## $ Inquiry               : num [1:1000] 1 1 1 2 0 0 0 1 1 0 ...
## $ LoginFrequency        : num [1:1000] 34 5 3 2 41 2 32 17 24 29 ...
## $ DaysSinceLastLogin    : num [1:1000] 71 26 46 128 65 105 358 253 187 352 ...
## $ ServiceUsage_Mobile App : num [1:1000] 1 0 0 0 0 0 1 0 0 0 ...
## $ ServiceUsage_Online Banking: num [1:1000] 0 0 0 0 0 0 0 1 0 1 ...
## $ ServiceUsage_Website   : num [1:1000] 0 1 1 1 1 1 0 0 1 0 ...
## $ ChurnStatus           : num [1:1000] 0 1 0 0 0 0 0 1 0 1 ...
```

```
customer_churn_df<- as.matrix(customer_churn_df)
class(customer_churn_df)
```

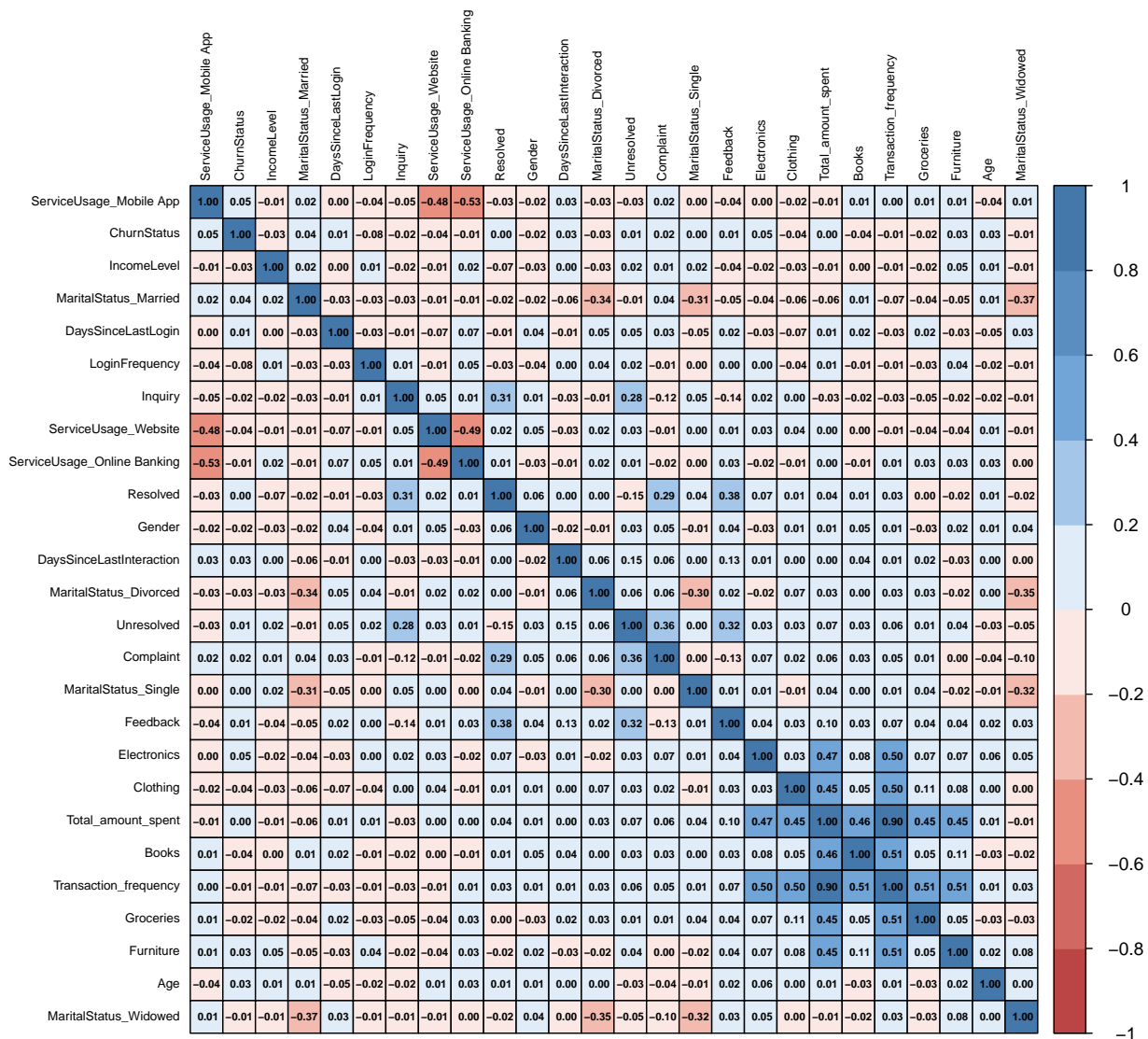
```
## [1] "matrix" "array"
```



```

# Compute correlation matrix
# Plot heatmap with correlation values
corrplot(cor(customer_churn_df),
  method = "color",          # color tiles
  #col = COL2('RdYlBu'),
  col <- colorRampPalette(c("#BB4444", "#EE9988", "#FFFFFF", "#77AADD", "#4477AA"))(10),
  type = "full",             # show upper triangle only
  addCoef.col = "black",     # add correlation coefficients
  tl.col = "black",          # text color
  number.cex = 0.5,          # size of correlation numbers
  tl.cex = 0.6,              # size of variable labels
  diag = TRUE,               # hide diagonal
  order = "AOE",
  outline= TRUE)

```



```
# Check skewness
customer_churn_df<- as.data.frame(customer_churn_df)

skew_values <- sapply(customer_churn_df, function(x) {
  if (is.numeric(x)) skewness(x, na.rm = TRUE) else NA
})
# View skewness of all numeric columns
print(skew_values)
```

```
##           Age           Gender
##      0.01312863      -0.05193958
##      IncomeLevel      MaritalStatus_Divorced
##      -0.04426777           1.16531680
```

```
##      MaritalStatus_Married      MaritalStatus_Single
##      1.08676003      1.38538037
##      MaritalStatus_Widowed      Total_amount_spent
##      1.00069660      0.26910140
##      Transaction_frequency      Books
##      -0.05877198      0.90981215
##      Clothing      Electronics
##      1.05494416      1.05360974
##      Furniture      Groceries
##      0.90203024      0.81424439
##      DaysSinceLastInteraction      Resolved
##      -0.42570791      0.80350613
##      Unresolved      Complaint
##      0.93027891      1.37650031
##      Feedback      Inquiry
##      1.37167397      1.42293149
##      LoginFrequency      DaysSinceLastLogin
##      -0.12749403      0.08008048
##      ServiceUsage_Mobile App ServiceUsage_Online Banking
##      0.66513426      0.63263318
##      ServiceUsage_Website      ChurnStatus
##      0.82545487      1.46689261
```

The data in the dataset is not scaled, we need to scale the data before training the model. Standardization is sufficient for logistic regression so we are not normalizing the data.

```
## Scaling the data
cols_to_scale <- c("Age",
                  "IncomeLevel",
                  "Total_amount_spent",
                  "Transaction_frequency",
                  "DaysSinceLastInteraction",
                  "LoginFrequency",
                  "DaysSinceLastLogin")

customer_churn_df[cols_to_scale] <- as.data.frame(scale(customer_churn_df[cols_to_scale],
                                                         center = TRUE, scale = TRUE))

head(customer_churn_df)
```

```
##      Age Gender IncomeLevel MaritalStatus_Divorced MaritalStatus_Married
## 1  1.2290131      0 -1.24720662      0      0
## 2  1.4258337      0 -1.24720662      0      1
## 3 -1.6576883      0 -1.24720662      0      0
## 4 -1.4608677      0 -1.24720662      0      0
## 5 -1.4608677      0 -0.02923141      1      0
## 6  0.9009789      1 -0.02923141      1      0
##      MaritalStatus_Single MaritalStatus_Widowed Total_amount_spent
## 1      1      0      -1.1516199
## 2      0      0      0.3795679
## 3      1      0      0.5901854
## 4      0      1     -0.4735847
## 5      0      0      0.9943474
## 6      0      0     -0.1391637
```

##	Transaction_frequency	Books	Clothing	Electronics	Furniture	Groceries
## 1	-1.55717487	0	0	1	0	0
## 2	0.74747467	0	2	3	1	1
## 3	0.36336641	1	1	0	2	2
## 4	-0.02074185	0	1	2	1	1
## 5	1.13158293	0	0	3	2	3
## 6	-0.02074185	1	1	1	2	0

##	DaysSinceLastInteraction	Resolved	Unresolved	Complaint	Feedback	Inquiry
## 1	0.78278685	1	0	0	0	1
## 2	0.95467950	1	0	0	0	1
## 3	-1.00980793	1	0	0	0	1
## 4	-0.37134952	1	1	0	0	2
## 5	0.05838211	0	0	0	0	0
## 6	0.35305523	1	0	0	1	0

##	LoginFrequency	DaysSinceLastLogin	ServiceUsage_Mobile	App
## 1	0.5754145	-1.0210463		1
## 2	-1.4877682	-1.4500373		0
## 3	-1.6300566	-1.2593747		0
## 4	-1.7012009	-0.4776577		0
## 5	1.0734242	-1.0782451		0
## 6	-1.7012009	-0.6969198		0

##	ServiceUsage_Online Banking	ServiceUsage_Website	ChurnStatus
## 1	0	0	0
## 2	0	1	1
## 3	0	1	0
## 4	0	1	0
## 5	0	1	0
## 6	0	1	0

Task 2: Building a Machine Learning Model

Introduction In this task, you will focus on developing a robust machine learning model to predict customer churn. Your objective is to select an appropriate algorithm, train and validate the model, and propose evaluation metrics that will help assess its performance. This task is pivotal for providing actionable insights that can inform business strategies at Lloyds Banking Group.

Instructions

Select an appropriate machine learning algorithm:

- Review the characteristics of the data set and the nature of the churn prediction problem.
- Consider algorithms such as logistic regression, decision trees, random forests, gradient boosting machines, or neural networks.
- Choose an algorithm that balances accuracy and interpretability, suitable for the business context.

Build and train the model:

- Use the preprocessed data set from Task 1 to train your chosen model.
- Implement techniques like cross-validation to ensure the model generalises well to unseen data.
- Perform hyperparameter tuning to optimise the model's performance.

Evaluate model performance:

- Select appropriate metrics to evaluate the model's performance, such as precision, recall, F1 score, ROC-AUC, and confusion matrix analysis.
- Consider the implications of each metric in the context of imbalanced data sets, ensuring that the evaluation provides a comprehensive view of the model's effectiveness.

Suggest ways to improve and utilise the model:

- Provide recommendations on how the model can be used by the business to identify at-risk customers and develop retention strategies.
- Discuss any potential improvements or adjustments to the model that could enhance its accuracy or applicability in different business scenarios.

Deliverable: Report submission: Compile a comprehensive report that includes:

- A detailed description of the selected algorithm and the rationale behind its choice.
- The trained model, along with performance metrics and evaluation results.
- Suggested ways to utilise the model's predictions for business decision-making and potential areas for improvement.

Ensure that your report is clear, concise, and well-organised, effectively communicating both the technical aspects of the model and its practical applications for the business. This report will be a critical tool for stakeholders to understand and leverage the predictive insights generated by your model.

```
# Duplicating the dataset
data<- customer_churn_df

# Convert the target variable as factor
data$ChurnStatus <- as.factor(data$ChurnStatus)
```

```
# logistic regression model

# Set seed for reproducibility
set.seed(123)

# Splitting the data
split <- sample.split(data$ChurnStatus, SplitRatio = 0.8)
train <- subset(data, split== TRUE)
test <- subset(data, split == FALSE)

table(train$ChurnStatus)
```

ML Model using Logistic Regression

```
##
##    0    1
## 637 163
```

```
table(test$ChurnStatus)
```

```
##
##    0    1
## 159  41
```

```
# Fitting test data into Logistic Regression model
model<- glm(ChurnStatus~., data = train, family = "binomial")
summary(model)
```

```
##
## Call:
## glm(formula = ChurnStatus ~ ., family = "binomial", data = train)
##
## Coefficients: (4 not defined because of singularities)
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.166084   0.503599  -4.301  1.7e-05 ***
## Age             0.059759   0.090389   0.661  0.50853
## Gender        -0.040578   0.181932  -0.223  0.82351
## IncomeLevel   -0.125840   0.089135  -1.412  0.15801
## MaritalStatus_Divorced -0.008744   0.252973  -0.035  0.97243
## MaritalStatus_Married  0.152765   0.246543   0.620  0.53550
## MaritalStatus_Single  0.006261   0.259807   0.024  0.98077
## MaritalStatus_Widowed      NA         NA      NA      NA
## Total_amount_spent  0.075632   0.213926   0.354  0.72368
## Transaction_frequency -0.375830   0.301873  -1.245  0.21314
## Books          -0.033969   0.131848  -0.258  0.79669
## Clothing       0.010272   0.132871   0.077  0.93838
## Electronics    0.278130   0.126885   2.192  0.02838 *
## Furniture      0.246185   0.131196   1.876  0.06059 .
## Groceries      NA         NA      NA      NA
## DaysSinceLastInteraction  0.136826   0.094319   1.451  0.14687
## Resolved      -0.046429   0.202599  -0.229  0.81874
## Unresolved     0.002130   0.205094   0.010  0.99171
## Complaint      0.223310   0.224860   0.993  0.32066
## Feedback       0.060785   0.220865   0.275  0.78315
## Inquiry        NA         NA      NA      NA
## LoginFrequency -0.271453   0.090636  -2.995  0.00274 **
## DaysSinceLastLogin  0.063324   0.090487   0.700  0.48404
## `ServiceUsage_Mobile App`  0.243196   0.227316   1.070  0.28468
## `ServiceUsage_Online Banking` 0.194365   0.227096   0.856  0.39207
## ServiceUsage_Website      NA         NA      NA      NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##    Null deviance: 808.89  on 799  degrees of freedom
## Residual deviance: 781.35  on 778  degrees of freedom
```

```
## AIC: 825.35
##
## Number of Fisher Scoring iterations: 4

# Predicting the test set
pred <- predict(model, test, type = "response")
pred_class <- as.factor(ifelse(pred > 0.5, "1", "0"))

pred_class <- factor(pred_class, levels = c("0", "1"))
test$ChurnStatus <- factor(test$ChurnStatus, levels = c("0", "1"))

confusionMatrix(pred_class, test$ChurnStatus, positive = "1")

## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##           0 158  41
##           1   1   0
##
##           Accuracy : 0.79
##           95% CI : (0.7269, 0.8443)
##       No Information Rate : 0.795
##       P-Value [Acc > NIR] : 0.6097
##
##           Kappa : -0.0099
##
##  Mcnemar's Test P-Value : 1.768e-09
##
##           Sensitivity : 0.0000
##           Specificity : 0.9937
##       Pos Pred Value : 0.0000
##       Neg Pred Value : 0.7940
##           Prevalence : 0.2050
##       Detection Rate : 0.0000
##   Detection Prevalence : 0.0050
##       Balanced Accuracy : 0.4969
##
##       'Positive' Class : 1
##
```

As there is high class imbalance the model predicted “0” for every test case, regardless of actual class. We can use methods like over-sampling, under sampling, ROSE, SMOTE methods to reduce this class imbalance

```
set.seed(123)

# Split data
split <- sample.split(data$ChurnStatus, SplitRatio = 0.8)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)

# Make names safe
names(train) <- make.names(names(train))
```

```

names(test) <- make.names(names(test))

# Ensure response is factor with valid level names
train$ChurnStatus <- factor(ifelse(train$ChurnStatus == "1", "yes", "no"))
test$ChurnStatus <- factor(ifelse(test$ChurnStatus == "1", "yes", "no"))

# Define a function to train, predict and evaluate
run_model <- function(sampled_data, test_data, method_name) {
  train.control <- trainControl(
    method = "cv",
    number = 5,
    classProbs = TRUE,
    summaryFunction = twoClassSummary
  )

  model <- train(
    ChurnStatus ~ .,
    data = sampled_data,
    method = "glm",
    family = "binomial",
    trControl = train.control,
    metric = "ROC"
  )

  pred <- predict(model, test_data, type = "prob")
  pred_class <- as.factor(ifelse(pred[, "yes"] > 0.5, "yes", "no"))

  cm <- confusionMatrix(pred_class, test_data$ChurnStatus, positive = "yes")
  cat("\n\n=====\\n")
  cat("Sampling Method:", method_name, "\\n")
  cat("=====\\n")
  print(table(sampled_data$ChurnStatus))
  print(cm)
}

# Apply different sampling methods and run the model

# Over-sampling
over_sampled <- ovun.sample(ChurnStatus ~ ., data = train, method = "over")$data
run_model(over_sampled, test, "Over-sampling")

##
##
## =====
## Sampling Method: Over-sampling
## =====
##
## no yes
## 637 631
## Confusion Matrix and Statistics
##
##           Reference
## Prediction no yes

```



```
##      no  88  26
##      yes 71  15
##
##              Accuracy : 0.515
##              95% CI : (0.4435, 0.5861)
##      No Information Rate : 0.795
##      P-Value [Acc > NIR] : 1
##
##              Kappa : -0.0573
##
##      McNemar's Test P-Value : 7.913e-06
##
##              Sensitivity : 0.3659
##              Specificity : 0.5535
##      Pos Pred Value : 0.1744
##      Neg Pred Value : 0.7719
##              Prevalence : 0.2050
##      Detection Rate : 0.0750
##      Detection Prevalence : 0.4300
##      Balanced Accuracy : 0.4597
##
##      'Positive' Class : yes
##
```

```
# Under-sampling
under_sampled <- ovun.sample(ChurnStatus ~ ., data = train, method = "under")$data
run_model(under_sampled, test, "Under-sampling")
```

```
##
##
## =====
## Sampling Method: Under-sampling
## =====
##
##  no yes
## 182 163
## Confusion Matrix and Statistics
##
##              Reference
## Prediction no yes
##      no  98  19
##      yes 61  22
##
##              Accuracy : 0.6
##              95% CI : (0.5285, 0.6685)
##      No Information Rate : 0.795
##      P-Value [Acc > NIR] : 1
##
##              Kappa : 0.1108
##
##      McNemar's Test P-Value : 4.563e-06
##
##              Sensitivity : 0.5366
##              Specificity : 0.6164
```

```
##          Pos Pred Value : 0.2651
##          Neg Pred Value : 0.8376
##          Prevalence : 0.2050
##          Detection Rate : 0.1100
##          Detection Prevalence : 0.4150
##          Balanced Accuracy : 0.5765
##
##          'Positive' Class : yes
##
```

```
# Both over & under-sampling
```

```
both_sampled <- ovun.sample(ChurnStatus ~ ., data = train, method = "both")$data
run_model(both_sampled, test, "Over + Under Sampling")
```

```
##
##
## =====
## Sampling Method: Over + Under Sampling
## =====
##
## no yes
## 376 424
## Confusion Matrix and Statistics
##
##          Reference
## Prediction no yes
##          no  80  20
##          yes 79  21
##
##          Accuracy : 0.505
##          95% CI : (0.4336, 0.5763)
##          No Information Rate : 0.795
##          P-Value [Acc > NIR] : 1
##
##          Kappa : 0.01
##
## Mcnemar's Test P-Value : 5.569e-09
##
##          Sensitivity : 0.5122
##          Specificity : 0.5031
##          Pos Pred Value : 0.2100
##          Neg Pred Value : 0.8000
##          Prevalence : 0.2050
##          Detection Rate : 0.1050
##          Detection Prevalence : 0.5000
##          Balanced Accuracy : 0.5077
##
##          'Positive' Class : yes
##
```

```
# ROSE sampling
```

```
rose_sampled <- ROSE(ChurnStatus ~ ., data = train)$data
run_model(rose_sampled, test, "ROSE Sampling")
```

```
##
##
## =====
## Sampling Method: ROSE Sampling
## =====
##
## no yes
## 398 402
## Confusion Matrix and Statistics
##
##           Reference
## Prediction no yes
##           no  90  27
##           yes  69  14
##
##           Accuracy : 0.52
##           95% CI : (0.4484, 0.591)
##           No Information Rate : 0.795
##           P-Value [Acc > NIR] : 1
##
##           Kappa : -0.067
##
## Mcnemar's Test P-Value : 2.857e-05
##
##           Sensitivity : 0.3415
##           Specificity : 0.5660
##           Pos Pred Value : 0.1687
##           Neg Pred Value : 0.7692
##           Prevalence : 0.2050
##           Detection Rate : 0.0700
##           Detection Prevalence : 0.4150
##           Balanced Accuracy : 0.4538
##
##           'Positive' Class : yes
##
```

```
# SMOTE
set.seed(121)
train <- subset(data, split== TRUE)
test  <- subset(data, split == FALSE)

names(train) <- make.names(names(train))
names(test)  <- make.names(names(test))

X1 <- train[, -which(names(train) == "ChurnStatus")] # all features
X2 <- as.numeric(as.character(train$ChurnStatus))    # convert factor to numeric (0/1)
train$ChurnStatus <- factor(train$ChurnStatus, levels = c("0", "1"))
smote_output <- SMOTE(X1, X2, K= 2)
SMOTE_sample_data <- smote_output$data

names(SMOTE_sample_data) <- make.names(names(SMOTE_sample_data))

# Ensure target is a factor with two levels: "0" (negative), "1" (positive)
SMOTE_sample_data$class <- factor(ifelse(SMOTE_sample_data$class == "1", "yes", "no"))
```

```

test$ChurnStatus <- factor(ifelse(test$ChurnStatus == "1", "yes", "no"))

# Train control with class probabilities and summary function for ROC
train.control <- trainControl(
  method = "cv",
  number = 5,
  classProbs = TRUE)

# Train logistic regression model using caret
model <- train(
  class ~ .,
  data = SMOTE_sample_data,
  method = "glm",
  family = "binomial",
  trControl = train.control,
  metric = "Kappa"
)

# Predict probabilities on test set
pred_probs <- predict(model, newdata = test, type = "prob")

# Convert probabilities to class predictions using threshold 0.5
pred_class <- as.factor(ifelse(pred_probs[, "yes"] > 0.5, "yes", "no"))

# Ensure test$ChurnStatus is a factor with same levels
# test$ChurnStatus <- factor(test$ChurnStatus, levels = c("0", "1"))

# Confusion matrix
confusionMatrix(pred_class, test$ChurnStatus, positive = "yes")

```

```

## Confusion Matrix and Statistics
##
##           Reference
## Prediction  no yes
##      no  121  29
##      yes   38  12
##
##               Accuracy : 0.665
##               95% CI : (0.595, 0.73)
##      No Information Rate : 0.795
##      P-Value [Acc > NIR] : 1.0000
##
##               Kappa : 0.0496
##
##  Mcnemar's Test P-Value : 0.3284
##
##               Sensitivity : 0.2927
##               Specificity : 0.7610
##      Pos Pred Value : 0.2400
##      Neg Pred Value : 0.8067
##      Prevalence : 0.2050
##      Detection Rate : 0.0600
##      Detection Prevalence : 0.2500

```

```
##          Balanced Accuracy : 0.5268
##
##          'Positive' Class : yes
##
```

```
# Decision Tree Model

set.seed(1234)

# Split into training and testing sets
split <- sample.split(data$ChurnStatus, SplitRatio = 0.8)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)

# Make column names safe
names(train) <- make.names(names(train))
names(test) <- make.names(names(test))

# Apply both over- and under-sampling
both_sampled <- ovun.sample(ChurnStatus ~ ., data = train, method = "both")$data

# Train decision tree classifier
tree_model <- rpart(ChurnStatus ~ ., data = both_sampled, method = "class",
  control = rpart.control(
    cp = 0.005,
    minsplit = 20,
    minbucket = 10,
    maxdepth = 5,
    xval = 10
  ))

# Get predicted probabilities for class "1"
tree_model_prob <- predict(tree_model, test, type = "prob")[, "1"]

# Convert probabilities to class predictions using 0.5 threshold
tree_model_prob_class <- as.factor(ifelse(tree_model_prob > 0.5, "1", "0"))

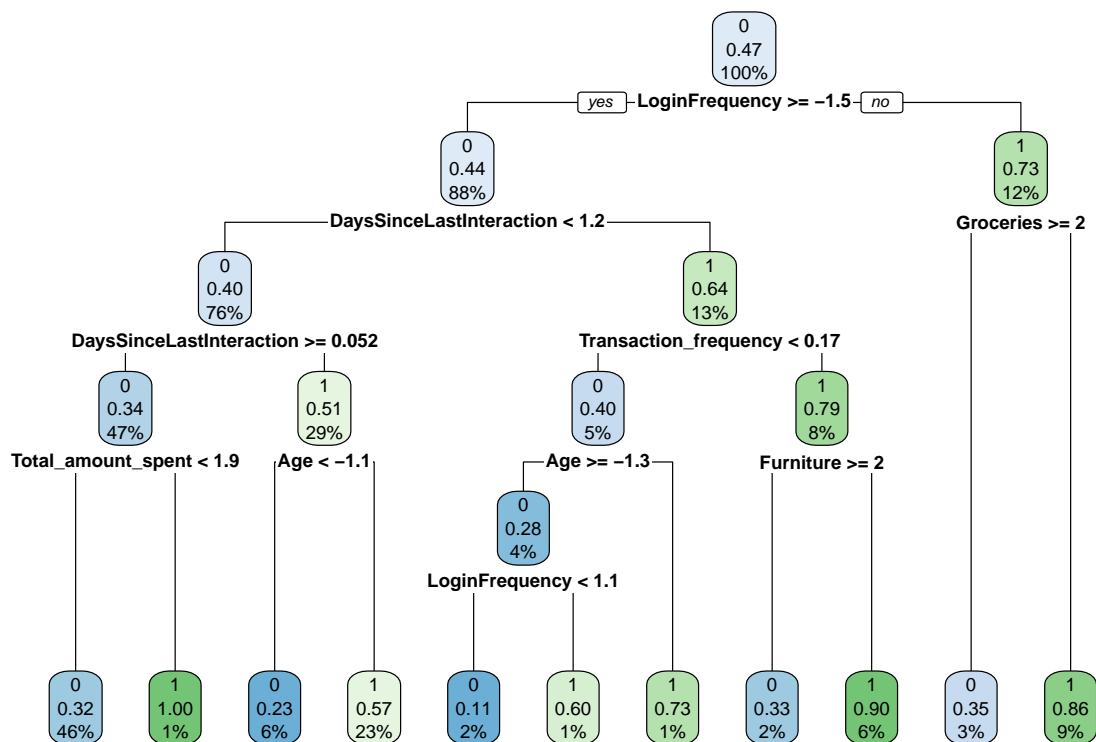
# Evaluate with confusion matrix
confusionMatrix(tree_model_prob_class, test$ChurnStatus, positive = "1")
```

ML Model using Decision Tree

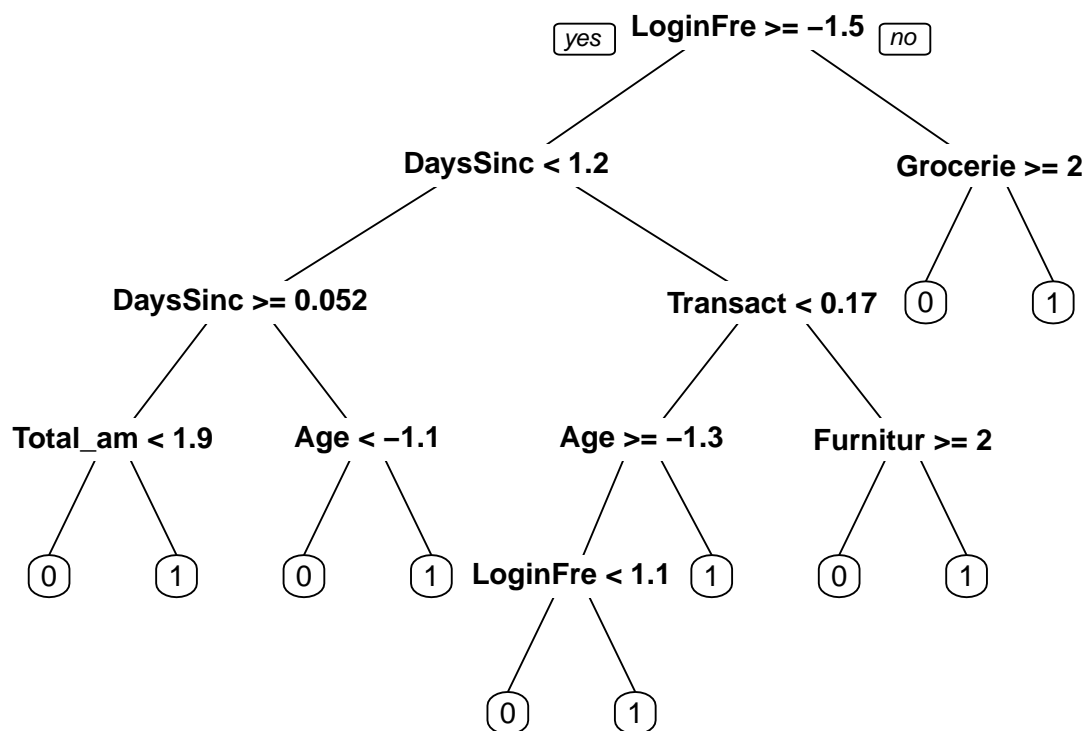
```
## Confusion Matrix and Statistics
##
##          Reference
## Prediction    0    1
##          0 100  20
##          1  59  21
##
##          Accuracy : 0.605
##          95% CI : (0.5336, 0.6732)
```

```
##      No Information Rate : 0.795
##      P-Value [Acc > NIR] : 1
##
##      Kappa : 0.1043
##
##      McNemar's Test P-Value : 1.909e-05
##
##      Sensitivity : 0.5122
##      Specificity : 0.6289
##      Pos Pred Value : 0.2625
##      Neg Pred Value : 0.8333
##      Prevalence : 0.2050
##      Detection Rate : 0.1050
##      Detection Prevalence : 0.4000
##      Balanced Accuracy : 0.5706
##
##      'Positive' Class : 1
##
```

```
# Visualize the tree
rpart.plot(tree_model)
```



```
prp(tree_model)
```



```

# Random Forest Model

set.seed(111)

# Split into training and testing sets
split <- sample.split(data$ChurnStatus, SplitRatio = 0.8)
train <- subset(data, split == TRUE)
test <- subset(data, split == FALSE)

# Make column names safe
names(train) <- make.names(names(train))
names(test) <- make.names(names(test))

train$ChurnStatus <- factor(ifelse(train$ChurnStatus == "1", "yes", "no"))
test$ChurnStatus <- factor(ifelse(test$ChurnStatus == "1", "yes", "no"))

# Apply both over- and under-sampling
under_sampled <- ovun.sample(ChurnStatus ~ ., data = train, method = "under")$data

control <- trainControl(method="cv", number=10, classProbs = TRUE,summaryFunction = twoClassSummary)
rf_model <- train(ChurnStatus~., data=under_sampled, method="rf",metric= "ROC", trControl=control)

rf_model

```

ML Model using Random Forest

```
## Random Forest
##
## 318 samples
## 25 predictor
## 2 classes: 'no', 'yes'
##
## No pre-processing
## Resampling: Cross-Validated (10 fold)
## Summary of sample sizes: 287, 287, 285, 287, 286, 286, ...
## Resampling results across tuning parameters:
##
## mtry  ROC      Sens      Spec
## 2     0.5344064 0.4758333 0.5463235
## 13    0.5734720 0.5287500 0.5577206
## 25    0.5931579 0.5625000 0.6000000
##
## ROC was used to select the optimal model using the largest value.
## The final value used for the model was mtry = 25.

rf_model_pred <- predict(rf_model, test, type = "prob")
rf_model_pred_class <- as.factor(ifelse(rf_model_pred[, "yes"] > 0.53, "yes", "no"))

confusionMatrix(rf_model_pred_class, test$ChurnStatus, positive = "yes")

## Confusion Matrix and Statistics
##
##              Reference
## Prediction no yes
##      no  92  19
##      yes  67  22
##
##              Accuracy : 0.57
##              95% CI : (0.4983, 0.6396)
##      No Information Rate : 0.795
##      P-Value [Acc > NIR] : 1
##
##              Kappa : 0.0803
##
##      McNemar's Test P-Value : 4.017e-07
##
##              Sensitivity : 0.5366
##              Specificity : 0.5786
##              Pos Pred Value : 0.2472
##              Neg Pred Value : 0.8288
##              Prevalence : 0.2050
##              Detection Rate : 0.1100
##      Detection Prevalence : 0.4450
##              Balanced Accuracy : 0.5576
##
##      'Positive' Class : yes
##
```


[illegible]

[illegible]


```
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
## [00:08:13] WARNING: src/c_api/c_api.cc:935: `ntree_limit` is deprecated, use `iteration_range` instead
```

```
sink()
unlink(tmp) # optional: remove temp file

xgb_model_pred <- predict(xgb_model, test, type = "prob")
xgb_model_pred_class <- as.factor(ifelse(xgb_model_pred[, "yes"] > 0.51, "yes", "no"))

confusionMatrix(xgb_model_pred_class, test$ChurnStatus, positive = "yes")
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction no yes
##           no  94  15
##           yes  65  26
##
##           Accuracy : 0.6
##           95% CI : (0.5285, 0.6685)
##           No Information Rate : 0.795
##           P-Value [Acc > NIR] : 1
##
##           Kappa : 0.1551
##
## Mcnemar's Test P-Value : 4.293e-08
##
##           Sensitivity : 0.6341
##           Specificity : 0.5912
##           Pos Pred Value : 0.2857
##           Neg Pred Value : 0.8624
##           Prevalence : 0.2050
##           Detection Rate : 0.1300
##           Detection Prevalence : 0.4550
##           Balanced Accuracy : 0.6127
##
##           'Positive' Class : yes
##
```

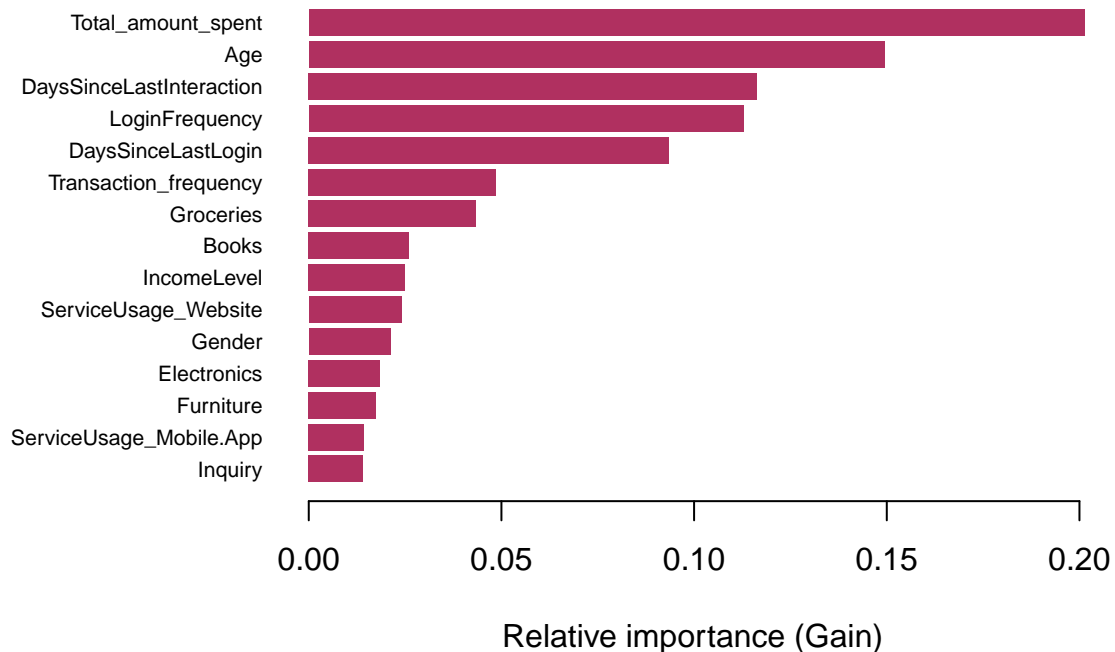
```
# Extract raw xgb.Booster from caret model
booster_model <- xgb_model$finalModel

feature_importance <- xgb.importance(model = booster_model)
print(feature_importance)
```

```
##           Feature           Gain           Cover           Frequency
```

	<char>	<num>	<num>	<num>
## 1:	Total_amount_spent	0.201274914	0.1984067415	0.186885246
## 2:	Age	0.149370455	0.1122520615	0.127868852
## 3:	DaysSinceLastInteraction	0.116336958	0.1512057455	0.131147541
## 4:	LoginFrequency	0.112788943	0.1079146891	0.101639344
## 5:	DaysSinceLastLogin	0.093447621	0.1213746172	0.124590164
## 6:	Transaction_frequency	0.048561288	0.0528356085	0.042622951
## 7:	Groceries	0.043381901	0.0521035451	0.045901639
## 8:	Books	0.026008836	0.0157982563	0.026229508
## 9:	IncomeLevel	0.024929280	0.0332139113	0.022950820
## 10:	ServiceUsage_Website	0.024085363	0.0251149860	0.022950820
## 11:	Gender	0.021164693	0.0223896856	0.016393443
## 12:	Electronics	0.018494956	0.0117234884	0.019672131
## 13:	Furniture	0.017418612	0.0109805645	0.019672131
## 14:	ServiceUsage_Mobile.App	0.014364001	0.0030803679	0.016393443
## 15:	Inquiry	0.014078786	0.0082638086	0.013114754
## 16:	MaritalStatus_Married	0.013035210	0.0099130106	0.013114754
## 17:	MaritalStatus_Divorced	0.012913335	0.0099823429	0.006557377
## 18:	Clothing	0.011497347	0.0066348682	0.009836066
## 19:	ServiceUsage_Online.Banking	0.010973340	0.0038207237	0.013114754
## 20:	MaritalStatus_Widowed	0.006798349	0.0107773695	0.009836066
## 21:	Feedback	0.006736266	0.0172935494	0.009836066
## 22:	MaritalStatus_Single	0.004678737	0.0050491849	0.006557377
## 23:	Unresolved	0.004373489	0.0068290392	0.006557377
## 24:	Complaint	0.001939639	0.0022691244	0.003278689
## 25:	Resolved	0.001347680	0.0007727103	0.003278689
##	Feature	Gain	Cover	Frequency

```
xgb.plot.importance(feature_importance, xlab = "Relative importance (Gain)", col= "maroon", top_n = 15)
```



```
# ROC curve and AUC
```

```
par(pty = "s")
roc_obj <- roc(test$ChurnStatus, xgb_model_pred[, "yes"], plot= TRUE,
  legacy.axes = TRUE, percent = TRUE,
  xlab= "False Positive percentage",
  ylab = "True Positive Percentage", col="#102E50",lwd = 2,
  print.auc = TRUE,print.auc.y = 65, print.auc.x = 27)
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
plot.roc(test$ChurnStatus, rf_model_pred[, "yes"],
  percent = TRUE,col="#B2A5FF",lwd = 2, print.auc = TRUE, add= TRUE, print.auc.y = 57,
  print.auc.x = 27)
```

```
## Setting levels: control = no, case = yes
```

```
## Setting direction: controls < cases
```

```
plot.roc(test$ChurnStatus, tree_model_prob,percent = TRUE,col="#48A6A7",lwd = 2, print.auc = TRUE, add=
```

```
## Setting levels: control = no, case = yes
```

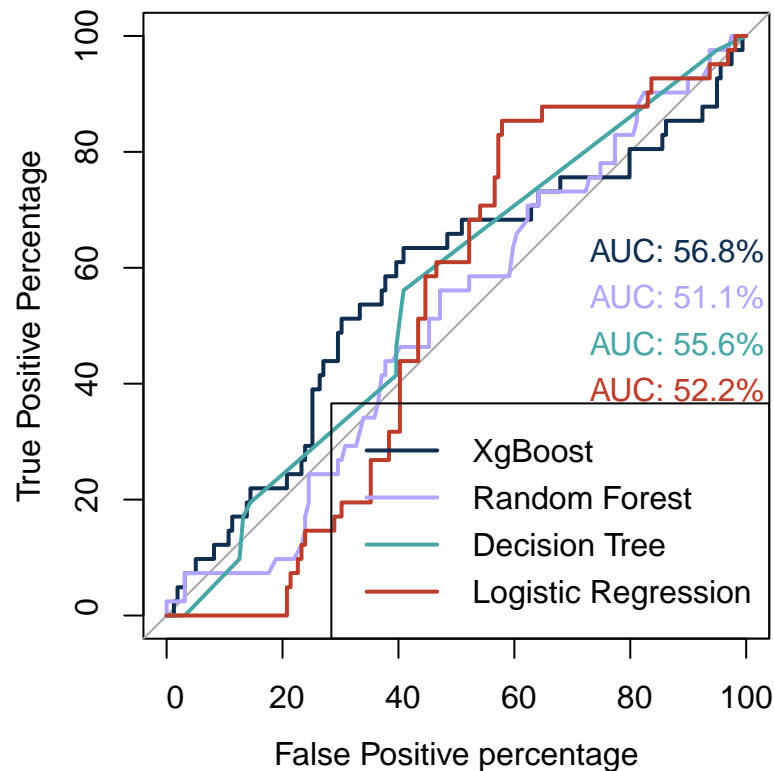
```
## Setting direction: controls < cases
```

```
plot.roc(test$ChurnStatus,pred_probs[,"yes"],percent = TRUE,col="#BE3D2A",lwd = 2, print.auc = TRUE, ad

## Setting levels: control = no, case = yes

## Setting direction: controls > cases

legend("bottomright", legend = c("XgBoost","Random Forest", "Decision Tree", "Logistic Regression"),
      col = c("#102E50", "#B2A5FF", "#48A6A7", "#BE3D2A"), lwd= 2)
```



Model Selection: Among the four models evaluated — Logistic Regression, Decision Tree, Random Forest, and XGBoost — the XGBoost model was selected as the most effective. While Logistic Regression achieved the highest overall accuracy (66.5%), it performed poorly in identifying the positive class, with a sensitivity of only 29.27%. In contrast, XGBoost achieved a significantly higher sensitivity (63.41%), the highest balanced accuracy (61.27%), and the highest ROC-AUC score, indicating superior discriminatory power between the classes. Additionally, XGBoost yielded the highest Cohen’s Kappa (0.1551), reflecting better agreement beyond chance. Therefore, due to its superior performance in handling class imbalance, distinguishing between classes, and effectively detecting the minority class, XGBoost is the most suitable model for this classification task.