Determinants of Customer Churn in E-commerce: An Econometric Analysis

Trang Hoang 476929

2025-04-28

Contents

1. Abstract

The aim of this paper is to identify and quantify the factors that influence customer churn in e-commerce from an econometric perspective. Customer churn remains a critical challenge for business profitability, as retaining existing customers is often more cost-effective than acquiring new ones. In a competitive landscape marked by evolving consumer expectations, understanding the drivers of churn is more important than ever.

Using a synthetic e-commerce dataset, we employ a binary logistic regression model to examine the impact of customer characteristics such as shopping behavior, engagement metrics, demographics, satisfaction levels, and product preferences on churn likelihood. Contrary to initial expectations, higher satisfaction scores were associated with a slight increase in churn probability, suggesting complex dynamics at play. Meanwhile, longer tenure and recent purchase activity significantly reduced the risk of churn, while complaints and greater delivery distance increased it.

Notably, customers with more registered devices or a preference for niche product categories were more likely to churn, while those favoring mainstream categories like laptops or mobile phones were less likely to leave. The results confirm that both behavioral and demographic variables play a significant role in customer retention. These findings provide practical guidance for firms to improve complaint handling, segment at-risk

customers more effectively, and tailor engagement strategies. Furthermore, the study builds a foundation for applying advanced predictive models such as machine learning to refine churn prediction and enhance customer relationship management.

2. Introduction

In recent years, e-commerce has developed into a fast-growing industry in which competition between companies is becoming increasingly fierce. One of the biggest challenges for e-commerce platforms is customer churn. When customers stop shopping or switch to another platform, this not only reduces revenue, but also wastes investment costs for marketing, advertising and customer care. According to a Harvard Business Review report, the cost of acquiring a new customer can be 5 to 25 times higher than the cost of retaining an existing customer. In addition, research by Bain & Company shows that a 5% increase in customer retention can increase profits by 25 to 95%. These figures underline the economic importance of analyzing customer churn behavior in the e-commerce sector.

In this context, understanding the factors that influence churn behavior is a top priority for companies in order to optimize their customer retention strategies. Although many previous studies have applied statistical methods and machine learning models to predict customer churn, there is still a lack of in-depth analysis of the causal relationship between customer behavior characteristics and churn probability based on traditional econometric approaches.

This study adopts a binary logit regression model to quantify the relationship between customer characteristics and churn probability. The variables analyzed include time spent on the platform, purchase behavior, engagement, payment method, satisfaction level and complaint occurrence. By applying this methodology, the study not only identifies the key factors influencing churn, but also provides clear quantitative evidence to support effective customer management strategies.

It is expected that the results of this study will provide insights for the development of more advanced customer churn prediction models in the future and assist managers in making strategic decisions about marketing, customer care and the optimization of company resources.

3. Literature Review

Customer churn is a pressing issue in the e-commerce sector, where acquiring new customers is generally more expensive than retaining existing ones. Understanding the underlying drivers of churn enables businesses to allocate marketing resources more efficiently and develop effective retention strategies. A growing body of research has explored behavioral, demographic, experiential, and preference-based factors influencing customer churn using both traditional econometric models and modern machine learning approaches.

3.1. Customer Behavior and Engagement

Customer behavior and engagement are among the most consistently cited predictors of churn. Li (2022) applied a Random Forest model to e-commerce churn prediction and found that OrderCount, DaySinceLastOrder, and Complain were among the most influential predictors. Similarly, Berger and Kompan (2019) emphasized behavioral metrics such as HourSpendOnApp and NumberOfDeviceRegistered, arguing that decreased engagement with the platform often precedes churn.

Bhattacharya (2021) also supported this view, showing that declining frequency of interactions and lower transaction volume significantly raised churn risk among e-commerce customers. The study highlighted that users with sporadic usage patterns are more susceptible to disengagement. These results underscore the importance of continuous engagement in mitigating churn.

3.2. Demographic Factors

Demographic characteristics, such as gender, age, and tenure, are also commonly linked to churn behavior. In Li's (2022) study, female users and customers with shorter tenure periods were more likely to churn. Similarly, Berger and Kompan (2019) found gender to be a significant factor, while Liu and Wang (2010) identified educational attainment as a key demographic influencing churn in the service sector—suggesting that more educated users may have higher service expectations and a lower tolerance for dissatisfaction.

Adding to this, Ahmad et al. (2019) showed that age group and marital status play a significant role in customer retention, with younger and single customers demonstrating a higher probability of switching platforms. These insights are especially relevant when designing personalized retention strategies.

3.3. Product Preferences

Product preferences, while often overlooked, are crucial in understanding churn. Berger and Kompan (2019) found that customers with narrow or niche product interests were more prone to churn when platforms failed to meet their expectations. In a similar study, Dahiya and Bhatia (2020) analyzed product-level transaction data and concluded that customers who primarily purchased electronics or single-category goods exhibited more volatile loyalty patterns, particularly when competitors offered better alternatives.

This aligns with the current study's findings, where customers who preferred mainstream product categories (e.g., laptops or mobile phones) were significantly less likely to churn than those in "Other" categories.

3.4. Platform Experience and Satisfaction

Customer experience—including satisfaction levels and service-related issues—plays a fundamental role in determining churn. Li (2022) found that SatisfactionScore and Complain strongly influence churn likelihood, echoing earlier findings by Mittal and Kamakura (2001), who showed that customer satisfaction and complaint handling quality directly impact retention and brand loyalty.

Moreover, Jaiswal and Niraj (2011) emphasized that satisfaction must be interpreted in context; not all satisfied customers are loyal. Factors such as perceived switching cost, emotional attachment, and service recovery also mediate the relationship between satisfaction and churn.

4. Data

The dataset used for this analysis is Ecommerce Customer Churn Analysis and Prediction.cvs, which contains information about customer behavior including: demographics, frequency, usages and attitudes, loyalty, complain provided by Kaggle (https://www.kaggle.com/datasets/ankitverma2010/ecommerce-customer-churn-analysis-and-prediction)

Dataset contains 17 variables:

Demographic Factors

- Gender: The gender of the customer.
- MaritalStatus: The marital status of the customer.
- CityTier: The tier of the customer's city.
- NumberOfAddress: The number of addresses the customer has registered.
- WarehouseToHome: The distance between the warehouse and the customer's home

Customer Behavior and Engagement

- HourSpendOnApp: The number of hours the customer spends on the app.
- NumberOfDeviceRegistered: The number of devices registered by the customer.
- OrderCount: The total number of orders made by the customer.
- DaySinceLastOrder: The number of days since the customer's last order.
- CouponUsed: The number of coupons used by the customer.
- OrderAmountHikeFromlastYear: The increase in the order amount compared to the previous year.
- CashbackAmount: The amount of cashback the customer has received.
- PreferredLoginDevice: The device the customer prefers to use when logging in.
- PreferredPaymentMode: The customer's preferred mode of payment.
- PreferedOrderCat: The type of products the customer prefers to order.

Experience with the Platform

- SatisfactionScore: The satisfaction score that the customer has given to the service.
- Complain: Whether the customer has made a complaint or not.

4.1. Data preparation

```
# Packages & Libraries
if (!require("pacman")) install.packages("pacman")
```

```
## Loading required package: pacman
```

```
pacman::p_load(GGally,corrplot,car,lmtest,modelsummary,margins,BaylorEdPsych,ResourceSelection,caret,pR
```

```
# Import dataset
```

```
data<-read.csv("E Commerce Dataset.csv", sep=";", dec=".",header=TRUE)
```

```
# Checking data
head(data)
```

```
##   CustomerID Churn Tenure PreferredLoginDevice CityTier WarehouseToHome
## 1      50001     1      4         Mobile Phone        3               6
## 2      50002     1     NA                Phone        1               8
## 3      50003     1     NA                Phone        1              30
## 4      50004     1      0                Phone        3              15
## 5      50005     1      0                Phone        1              12
## 6      50006     1      0             Computer        1              22
##   PreferredPaymentMode Gender HourSpendOnApp NumberOfDeviceRegistered
## 1           Debit Card Female              3                         3
## 2                 UPI   Male              3                         4
## 3           Debit Card   Male              2                         4
## 4           Debit Card   Male              2                         4
## 5                   CC   Male             NA                         3
## 6           Debit Card Female              3                         5
##     PreferedOrderCat SatisfactionScore MaritalStatus NumberOfAddress Complain
## 1 Laptop & Accessory                2        Single               9        1
## 2             Mobile                3        Single               7        1
```

```
## 3             Mobile                 3        Single                6        1
## 4 Laptop & Accessory                 5        Single                8        0
## 5             Mobile                 5        Single                3        0
## 6       Mobile Phone                 5        Single                2        1
##   OrderAmountHikeFromlastYear CouponUsed OrderCount DaySinceLastOrder
## 1                          11          1          1                 5
## 2                          15          0          1                 0
## 3                          14          0          1                 3
## 4                          23          0          1                 3
## 5                          11          1          1                 3
## 6                          22          4          6                 7
##   CashbackAmount
## 1            160
## 2            121
## 3            120
## 4            134
## 5            130
## 6            139
```

```
dim(data)
```

```
## [1] 5630   20
```

```
summary(data)
```

```
##    CustomerID        Churn            Tenure      PreferredLoginDevice
##  Min.   :50001   Min.   :0.0000   Min.   : 0.00   Length:5630
##  1st Qu.:51408   1st Qu.:0.0000   1st Qu.: 2.00   Class :character
##  Median :52816   Median :0.0000   Median : 9.00   Mode  :character
##  Mean   :52816   Mean   :0.1684   Mean   :10.19
##  3rd Qu.:54223   3rd Qu.:0.0000   3rd Qu.:16.00
##  Max.   :55630   Max.   :1.0000   Max.   :61.00
##                                   NA's   :264
##     CityTier     WarehouseToHome  PreferredPaymentMode    Gender
##  Min.   :1.000   Min.   :  5.00   Length:5630          Length:5630
##  1st Qu.:1.000   1st Qu.:  9.00   Class :character     Class :character
##  Median :1.000   Median : 14.00   Mode  :character     Mode  :character
##  Mean   :1.655   Mean   : 15.64
##  3rd Qu.:3.000   3rd Qu.: 20.00
##  Max.   :3.000   Max.   :127.00
##                  NA's   :251
##  HourSpendOnApp  NumberOfDeviceRegistered PreferedOrderCat   SatisfactionScore
##  Min.   :0.000   Min.   :1.000            Length:5630        Min.   :1.000
##  1st Qu.:2.000   1st Qu.:3.000            Class :character   1st Qu.:2.000
##  Median :3.000   Median :4.000            Mode  :character   Median :3.000
##  Mean   :2.932   Mean   :3.689                               Mean   :3.067
##  3rd Qu.:3.000   3rd Qu.:4.000                               3rd Qu.:4.000
##  Max.   :5.000   Max.   :6.000                               Max.   :5.000
##  NA's   :255
##  MaritalStatus      NumberOfAddress     Complain
##  Length:5630        Min.   : 1.000   Min.   :0.0000
##  Class :character   1st Qu.: 2.000   1st Qu.:0.0000
##  Mode  :character   Median : 3.000   Median :0.0000
```

```
##                     Mean   : 4.214   Mean   :0.2849
##                     3rd Qu.: 6.000   3rd Qu.:1.0000
##                     Max.   :22.000   Max.   :1.0000
##
##  OrderAmountHikeFromlastYear   CouponUsed       OrderCount
##  Min.   :11.00               Min.   : 0.000   Min.   : 1.000
##  1st Qu.:13.00               1st Qu.: 1.000   1st Qu.: 1.000
##  Median :15.00               Median : 1.000   Median : 2.000
##  Mean   :15.71               Mean   : 1.751   Mean   : 3.008
##  3rd Qu.:18.00               3rd Qu.: 2.000   3rd Qu.: 3.000
##  Max.   :26.00               Max.   :16.000   Max.   :16.000
##  NA's   :265                 NA's   :256      NA's   :258
##  DaySinceLastOrder CashbackAmount
##  Min.   : 0.000    Min.   :  0.0
##  1st Qu.: 2.000    1st Qu.:146.0
##  Median : 3.000    Median :163.0
##  Mean   : 4.543    Mean   :177.2
##  3rd Qu.: 7.000    3rd Qu.:196.0
##  Max.   :46.000    Max.   :325.0
##  NA's   :307
```

There are many NA values in the dataset, so I use complete.case to remove rows with NA value

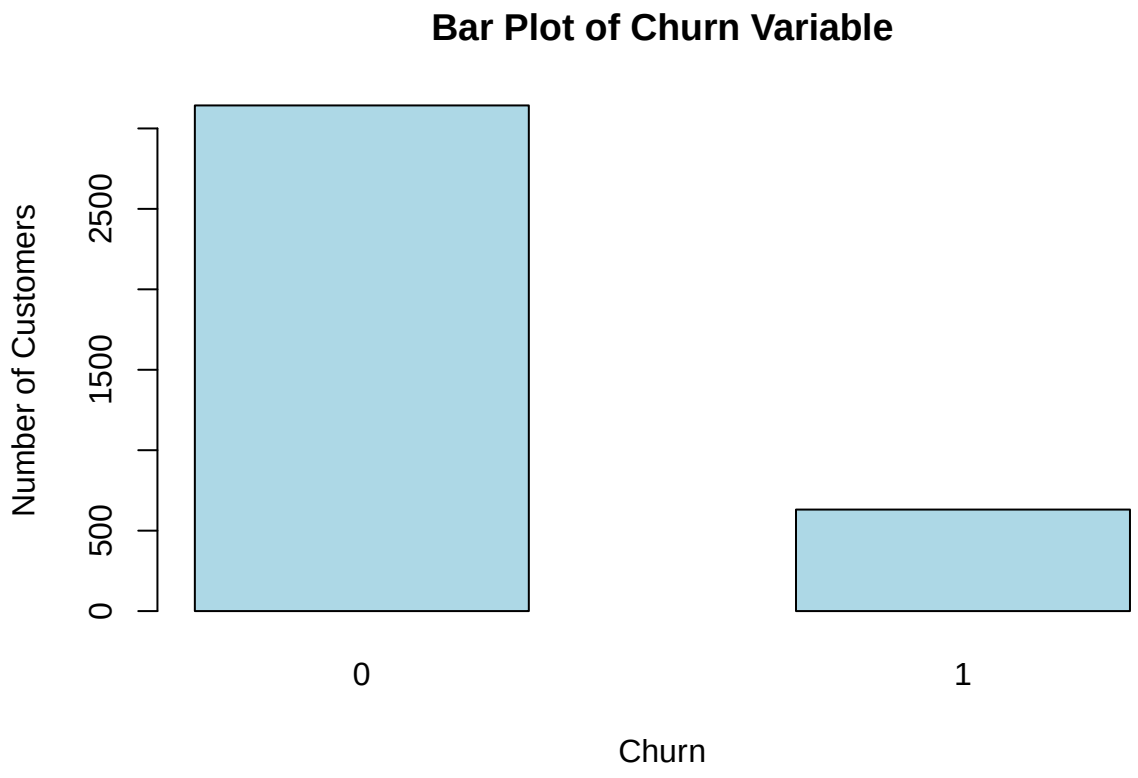
```
data <- data[complete.cases(data), ]
any(is.na(data))
```

```
## [1] FALSE
```

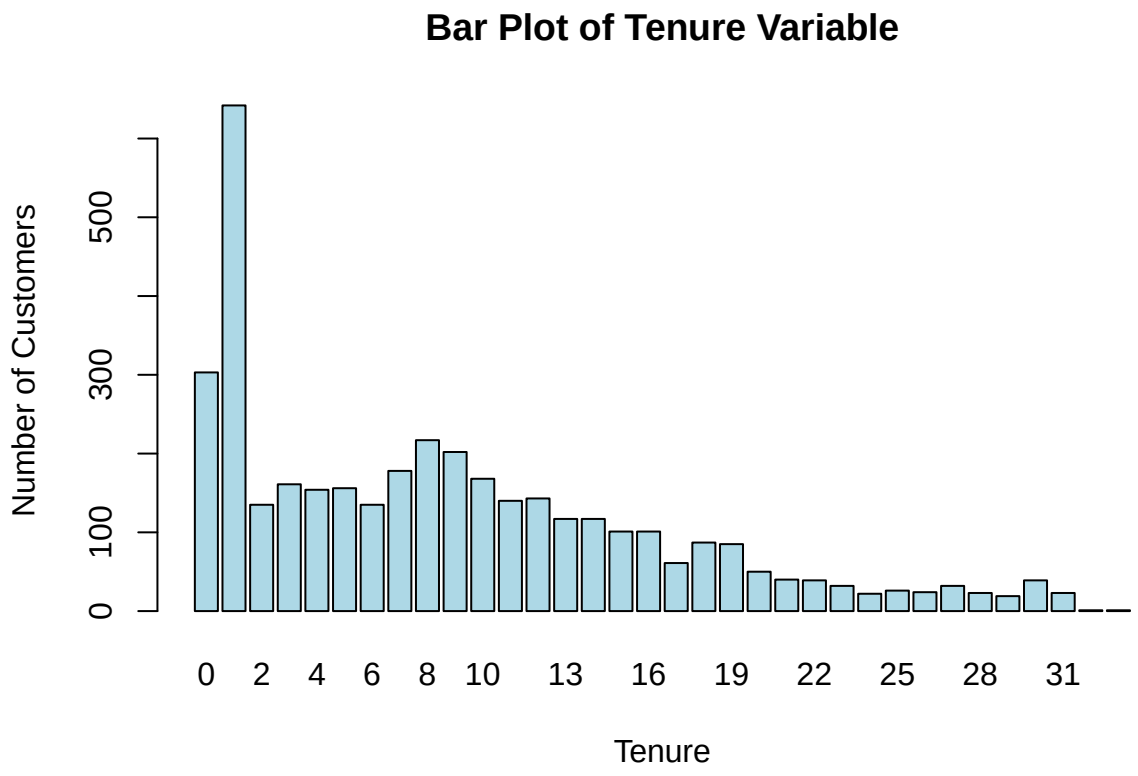
4.2. Data processing

```
# Remove ID column (non-meaning variable)
data <- data[, -1]
```

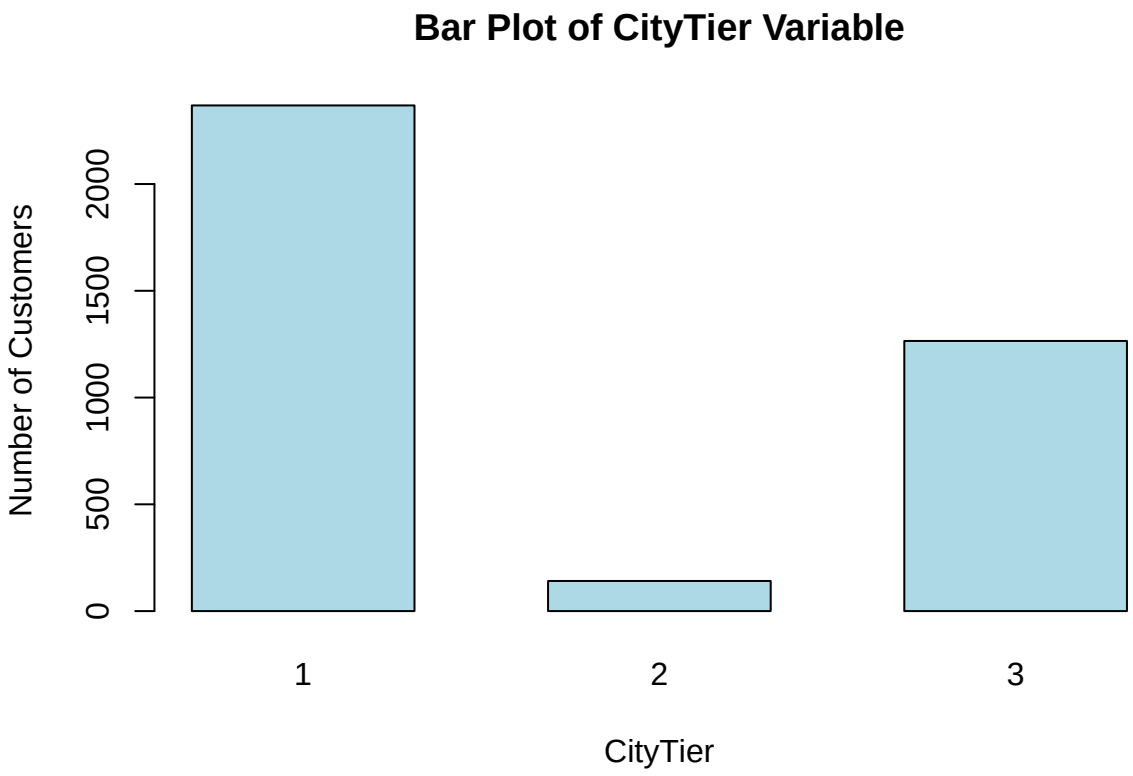
```
# Checking distribution of numeric data
barplot(table(data$Churn),
        col = "lightblue",
        xlab = "Churn",
        ylab = "Number of Customers",
        main = "Bar Plot of Churn Variable",
        space = 0.8)
```



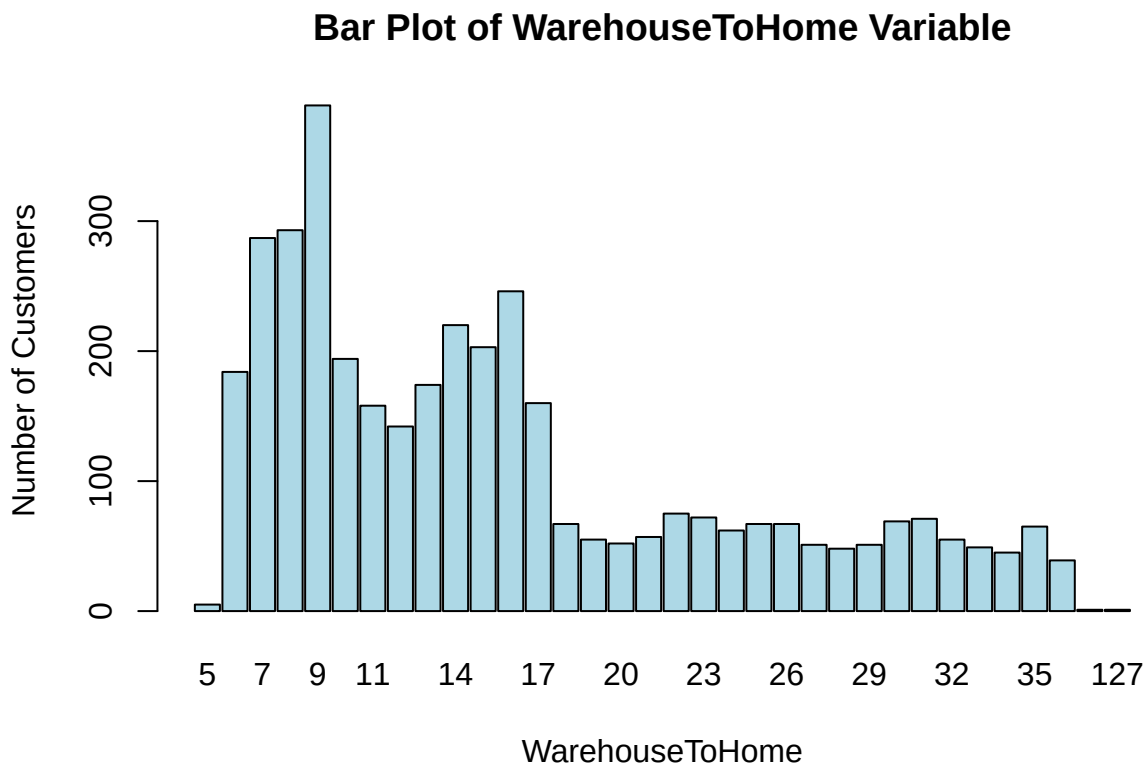

```r
barplot(table(data$Tenure),
        col = "lightblue",
        xlab = "Tenure",
        ylab = "Number of Customers",
        main = "Bar Plot of Tenure Variable",
        space = 0.2)
```



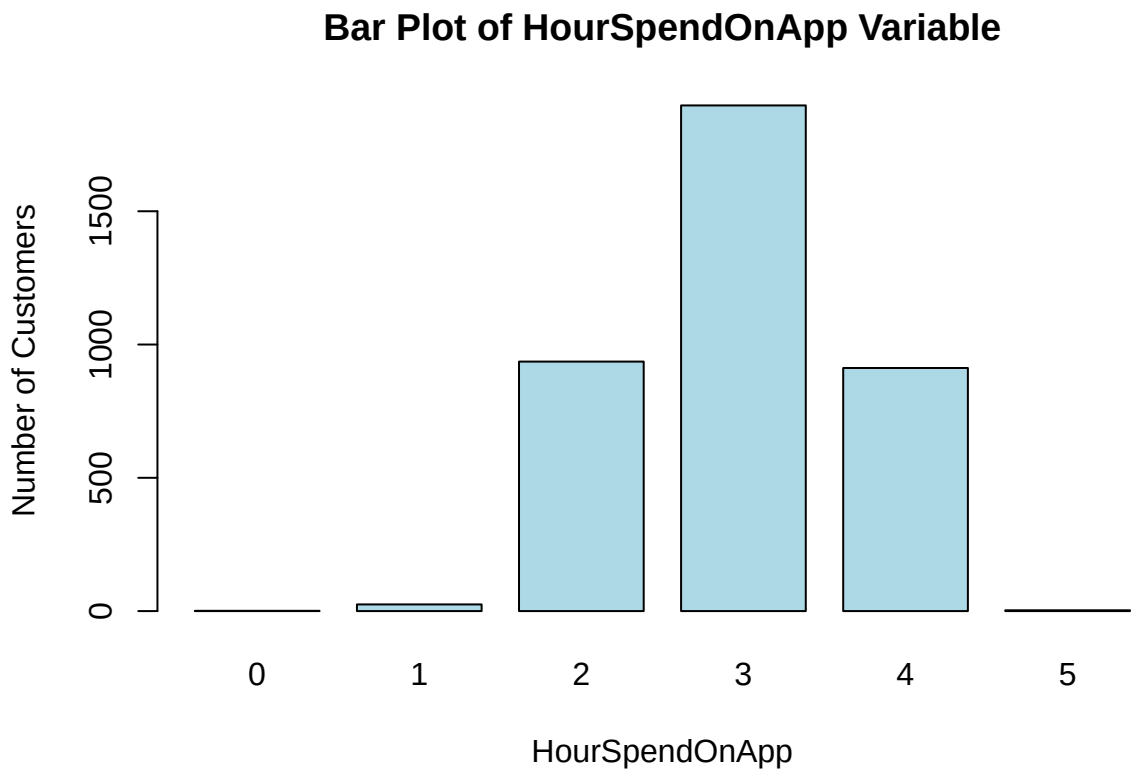

```
barplot(table(data$CityTier),
        col = "lightblue",
        xlab = "CityTier",
        ylab = "Number of Customers",
        main = "Bar Plot of CityTier Variable",
        space = 0.6)
```



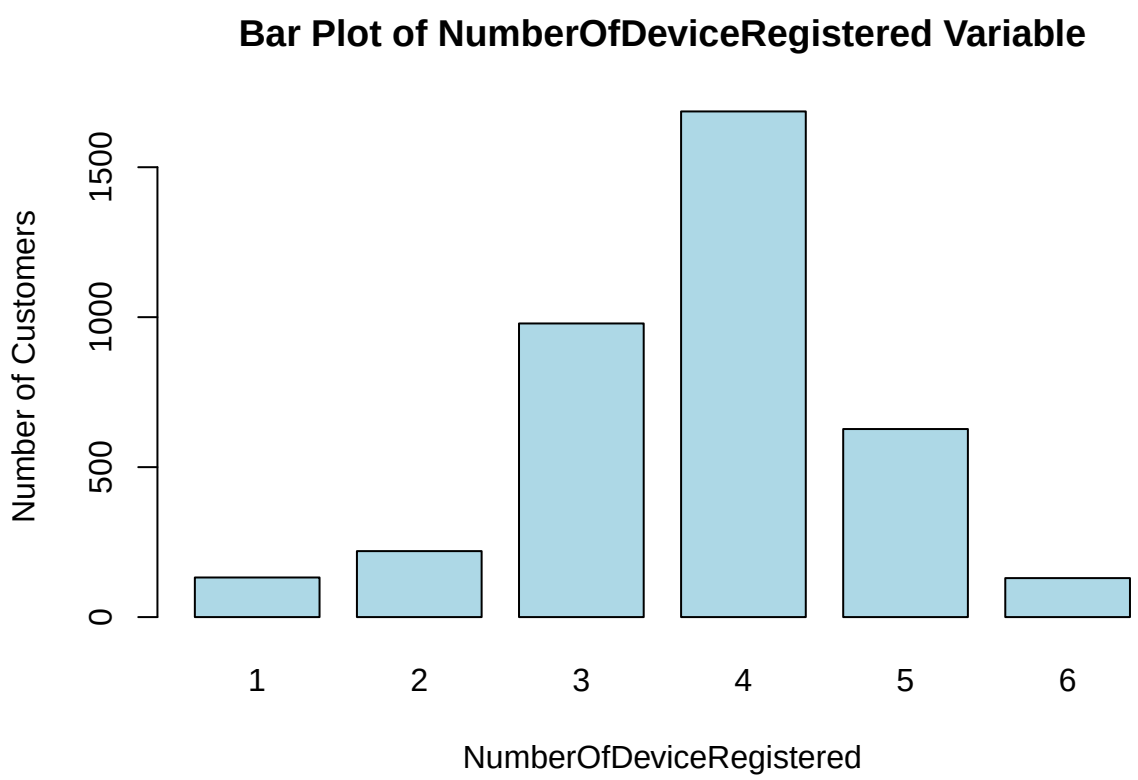

```
barplot(table(data$WarehouseToHome),
        col = "lightblue",
        xlab = "WarehouseToHome",
        ylab = "Number of Customers",
        main = "Bar Plot of WarehouseToHome Variable",
        space = 0.1)
```




```r
barplot(table(data$HourSpendOnApp),
        col = "lightblue",
        xlab = "HourSpendOnApp",
        ylab = "Number of Customers",
        main = "Bar Plot of HourSpendOnApp Variable",
        space = 0.3)
```

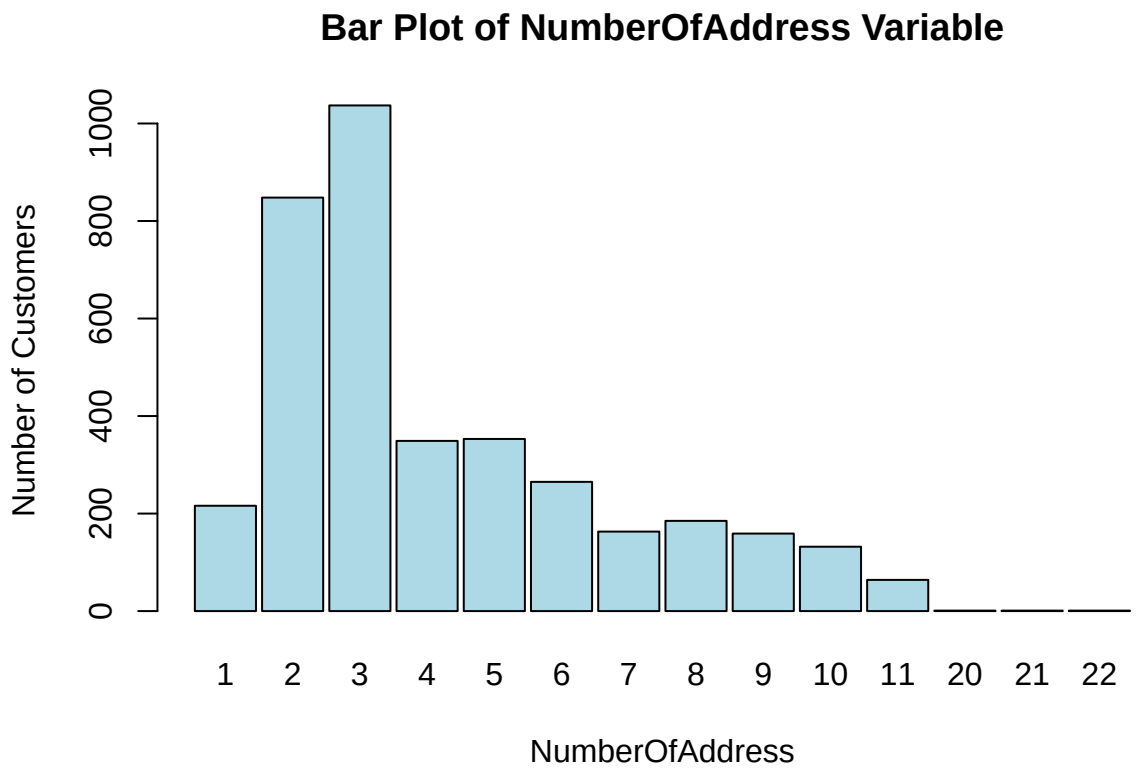



```r
barplot(table(data$NumberOfDeviceRegistered),
        col = "lightblue",
        xlab = "NumberOfDeviceRegistered",
        ylab = "Number of Customers",
        main = "Bar Plot of NumberOfDeviceRegistered Variable",
        space = 0.3)
```

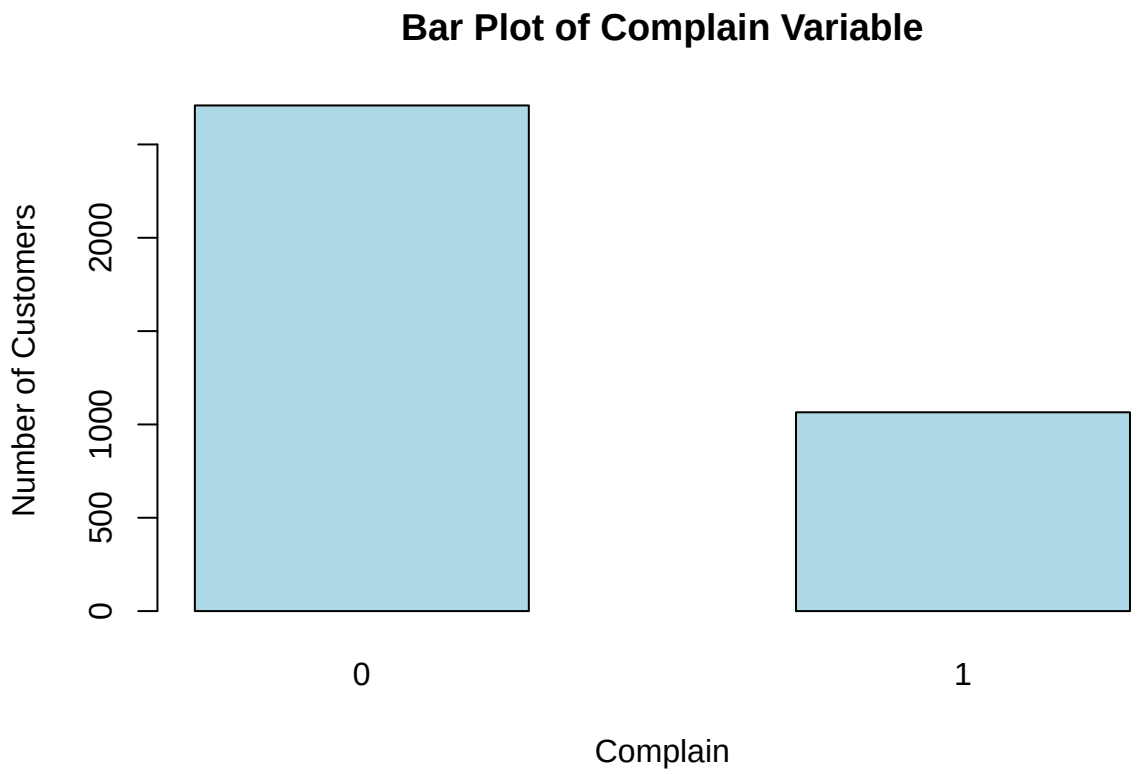



```
barplot(table(data$SatisfactionScore),
        col = "lightblue",
        xlab = "SatisfactionScore",
        ylab = "Number of Customers",
        main = "Bar Plot of SatisfactionScore Variable",
        space = 0.3)
```

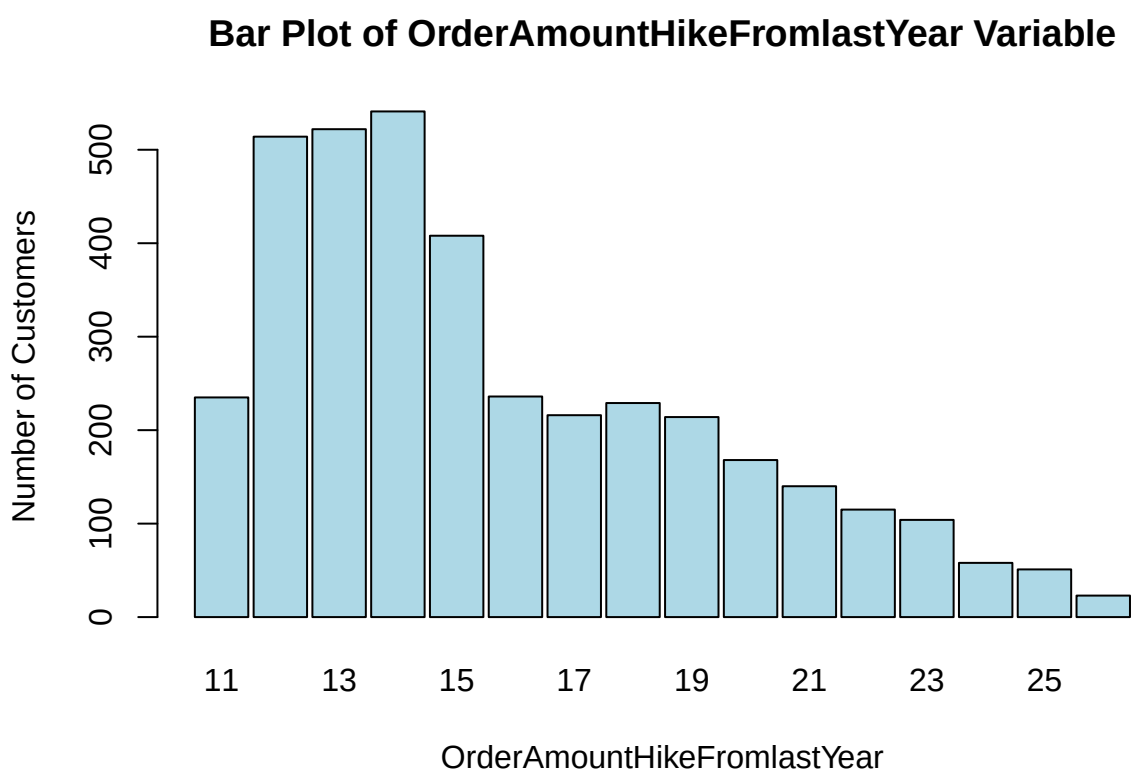

```r
barplot(table(data$NumberOfAddress),
        col = "lightblue",
        xlab = "NumberOfAddress",
        ylab = "Number of Customers",
        main = "Bar Plot of NumberOfAddress Variable",
        space = 0.1)
```



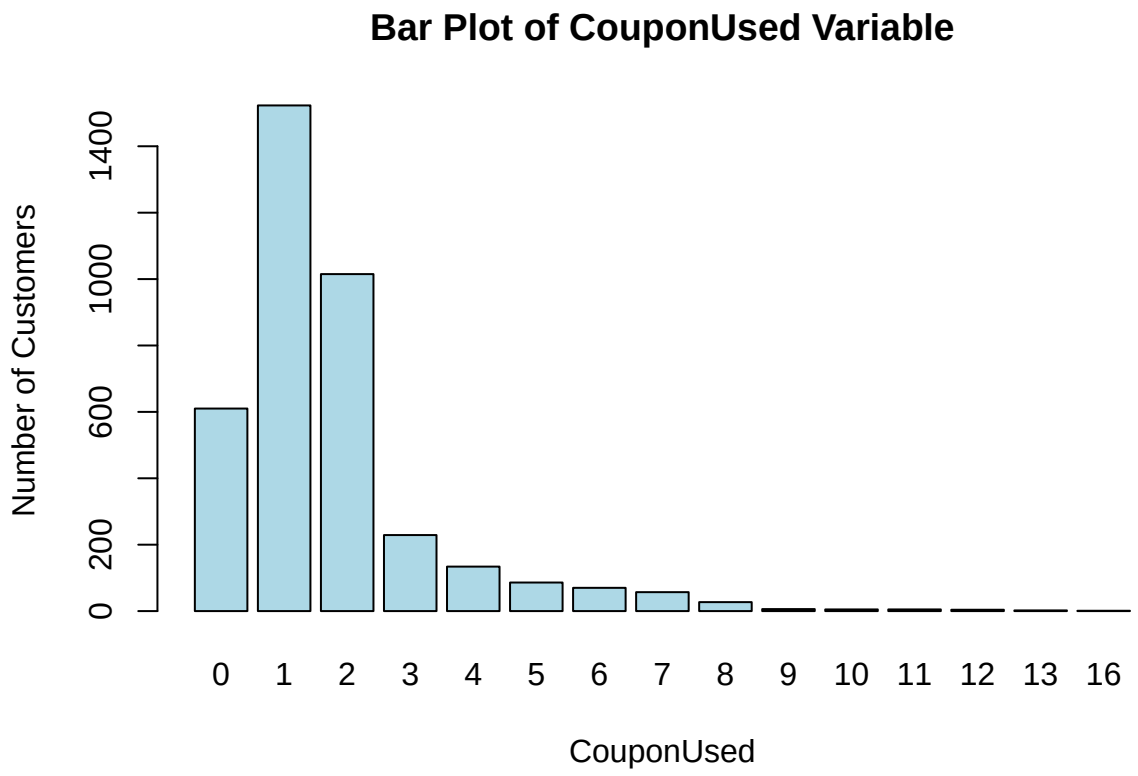

```
barplot(table(data$Complain),
        col = "lightblue",
        xlab = "Complain",
        ylab = "Number of Customers",
        main = "Bar Plot of Complain Variable",
        space = 0.8)
```



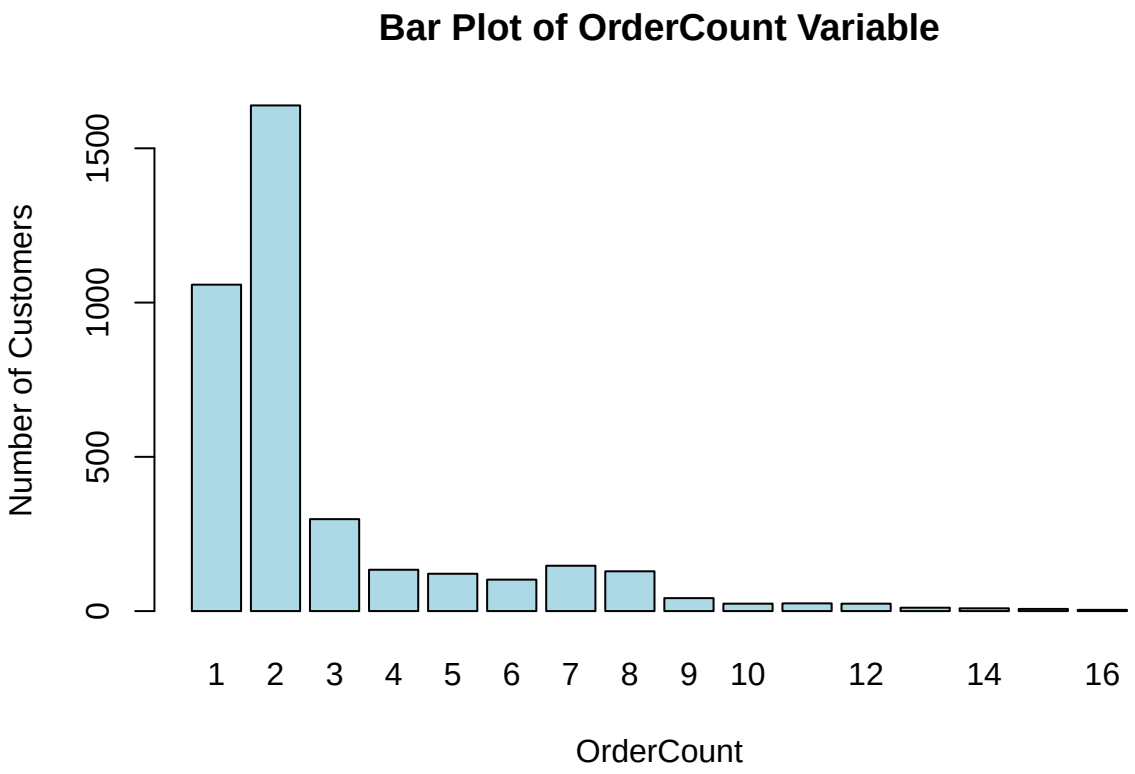

```
barplot(table(data$OrderAmountHikeFromlastYear),
        col = "lightblue",
        xlab = "OrderAmountHikeFromlastYear",
        ylab = "Number of Customers",
        main = "Bar Plot of OrderAmountHikeFromlastYear Variable",
        space = 0.1)
```



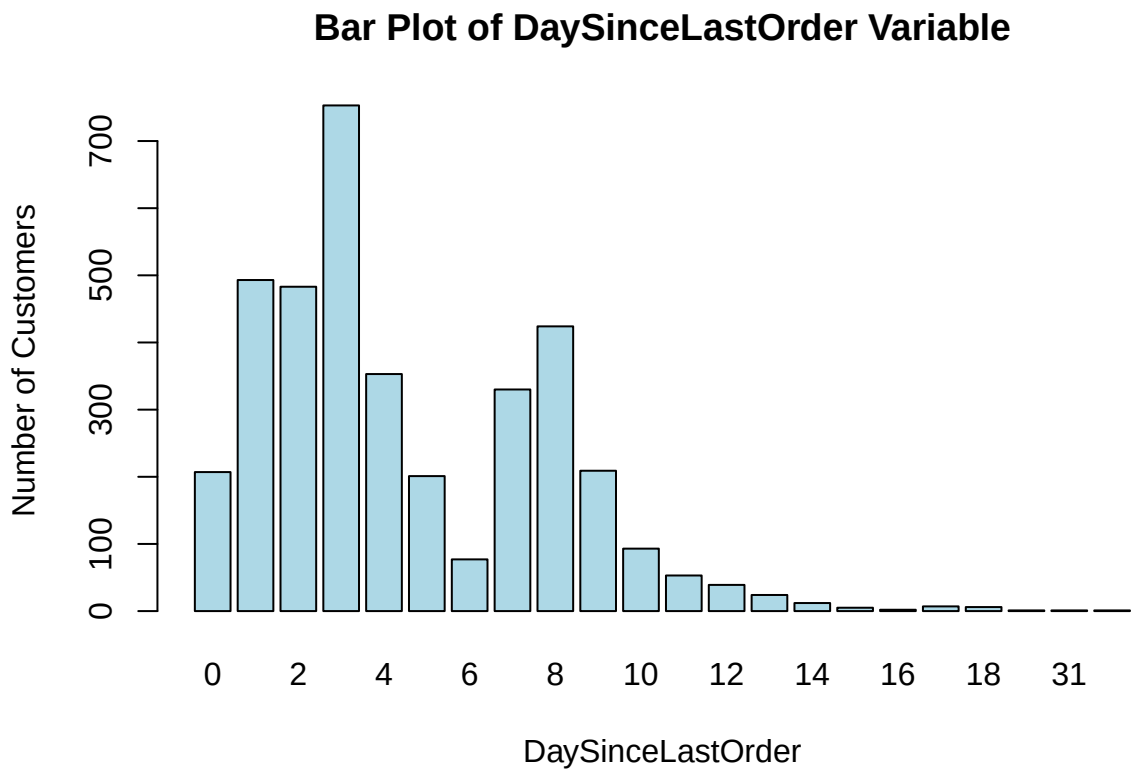

```r
barplot(table(data$CouponUsed),
        col = "lightblue",
        xlab = "CouponUsed",
        ylab = "Number of Customers",
        main = "Bar Plot of CouponUsed Variable",
        space = 0.2)
```



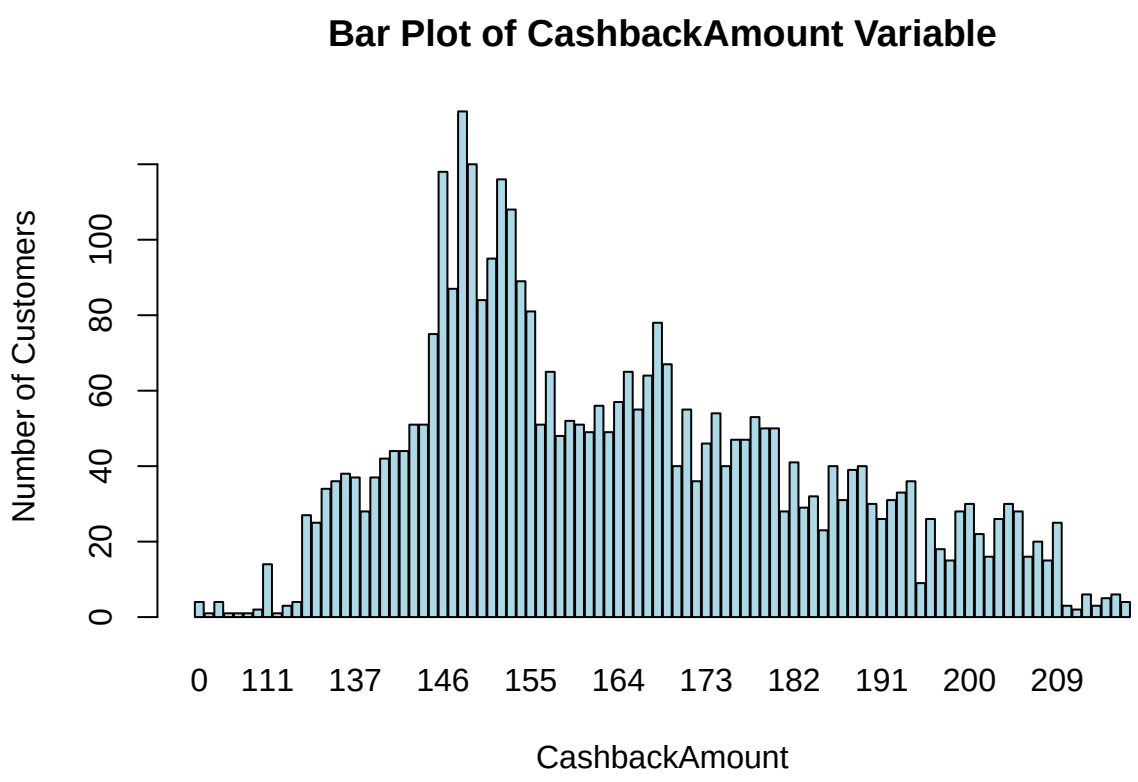

```
barplot(table(data$OrderCount),
        col = "lightblue",
        xlab = "OrderCount",
        ylab = "Number of Customers",
        main = "Bar Plot of OrderCount Variable",
        space = 0.2)
```



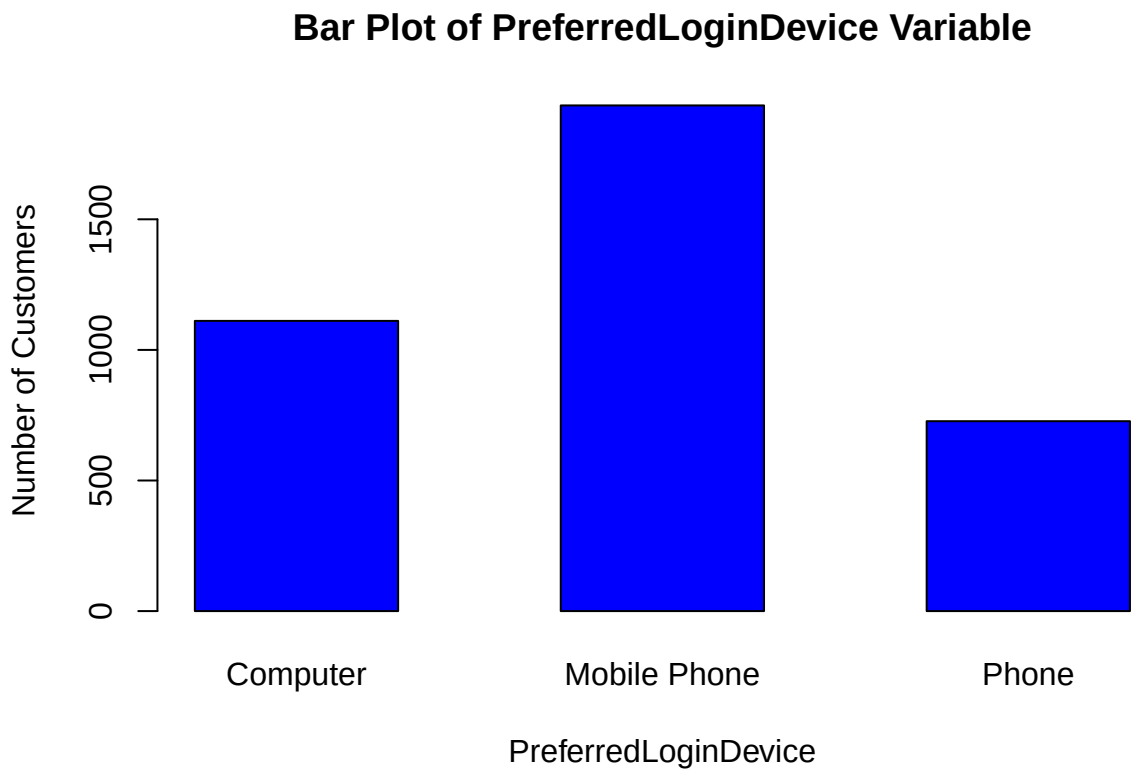

```
barplot(table(data$DaySinceLastOrder),
        col = "lightblue",
        xlab = "DaySinceLastOrder",
        ylab = "Number of Customers",
        main = "Bar Plot of DaySinceLastOrder Variable",
        space = 0.2)
```



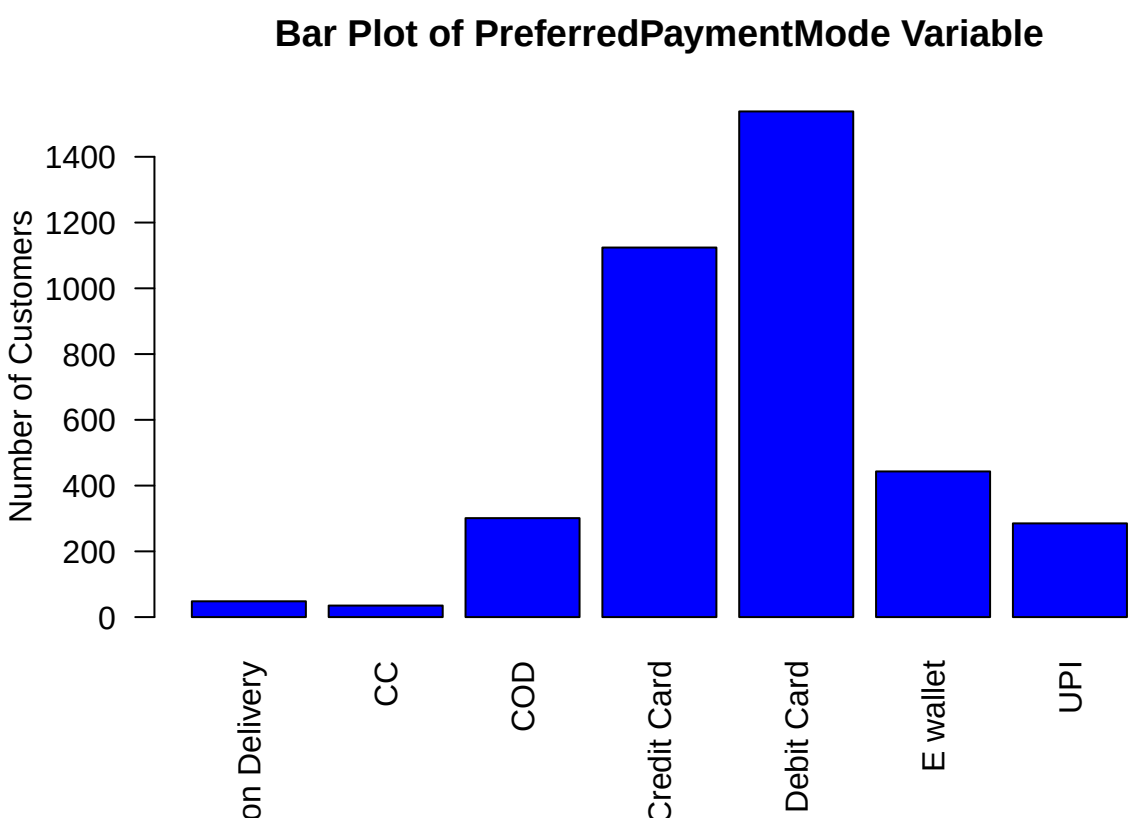

```r
barplot(table(data$CashbackAmount),
        col = "lightblue",
        xlab = "CashbackAmount",
        ylab = "Number of Customers",
        main = "Bar Plot of CashbackAmount Variable",
        space = 0.1)
```



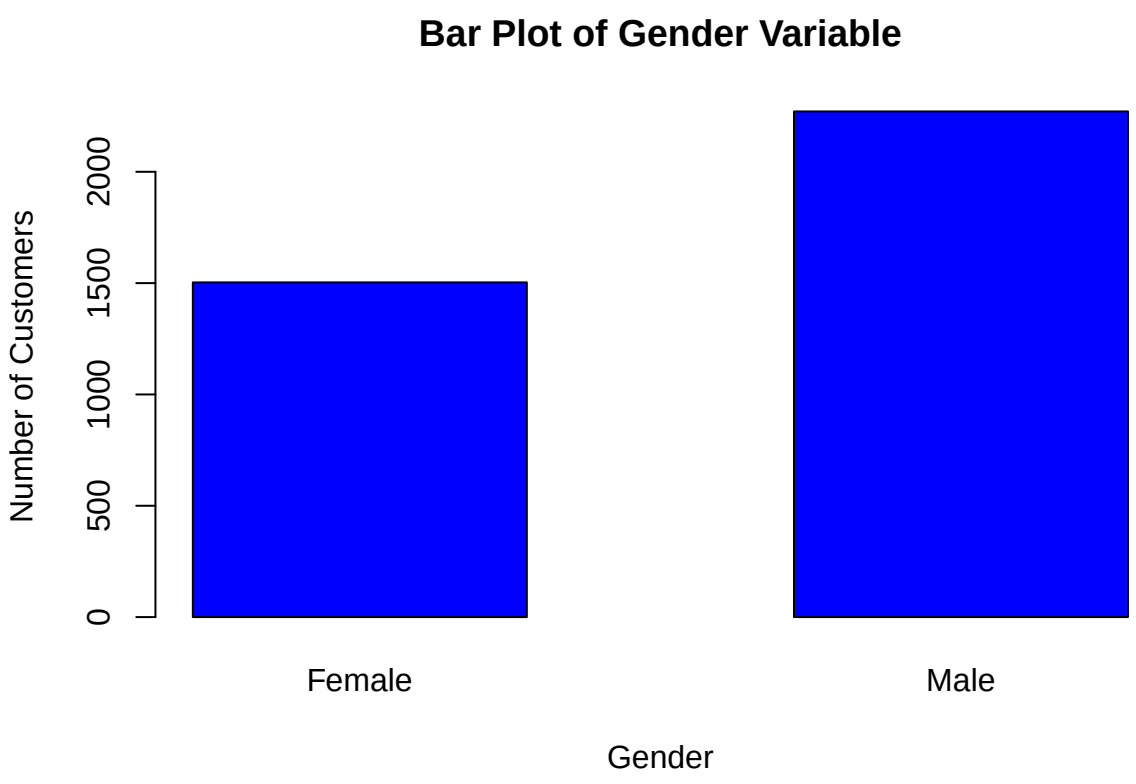

```r
# Checking distribution of character data

barplot(table(data$PreferredLoginDevice),
        col = "blue",
        xlab = "PreferredLoginDevice",
        ylab = "Number of Customers",
        main = "Bar Plot of PreferredLoginDevice Variable",
        space = 0.8)
```



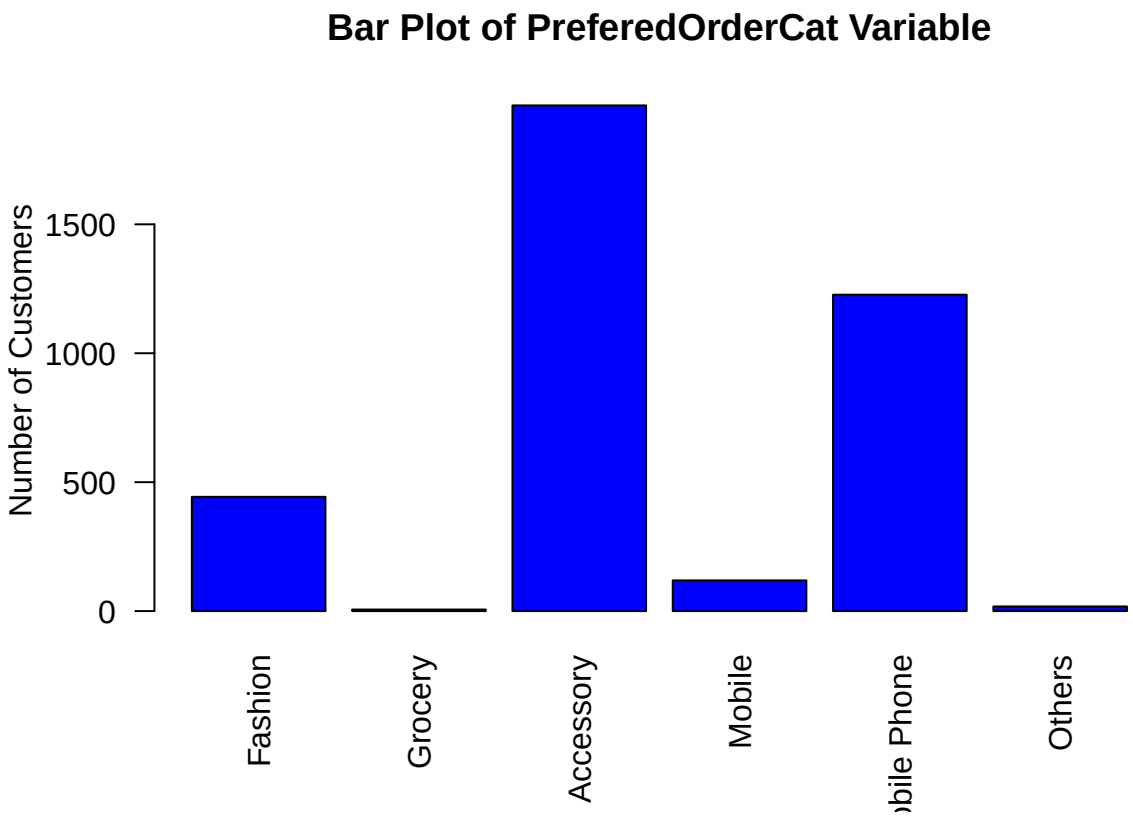

```r
barplot(table(data$PreferredPaymentMode),
        col = "blue",
        ylab = "Number of Customers",
        main = "Bar Plot of PreferredPaymentMode Variable",
        space = 0.2,
        las = 2,
        cex.names = 1)
```



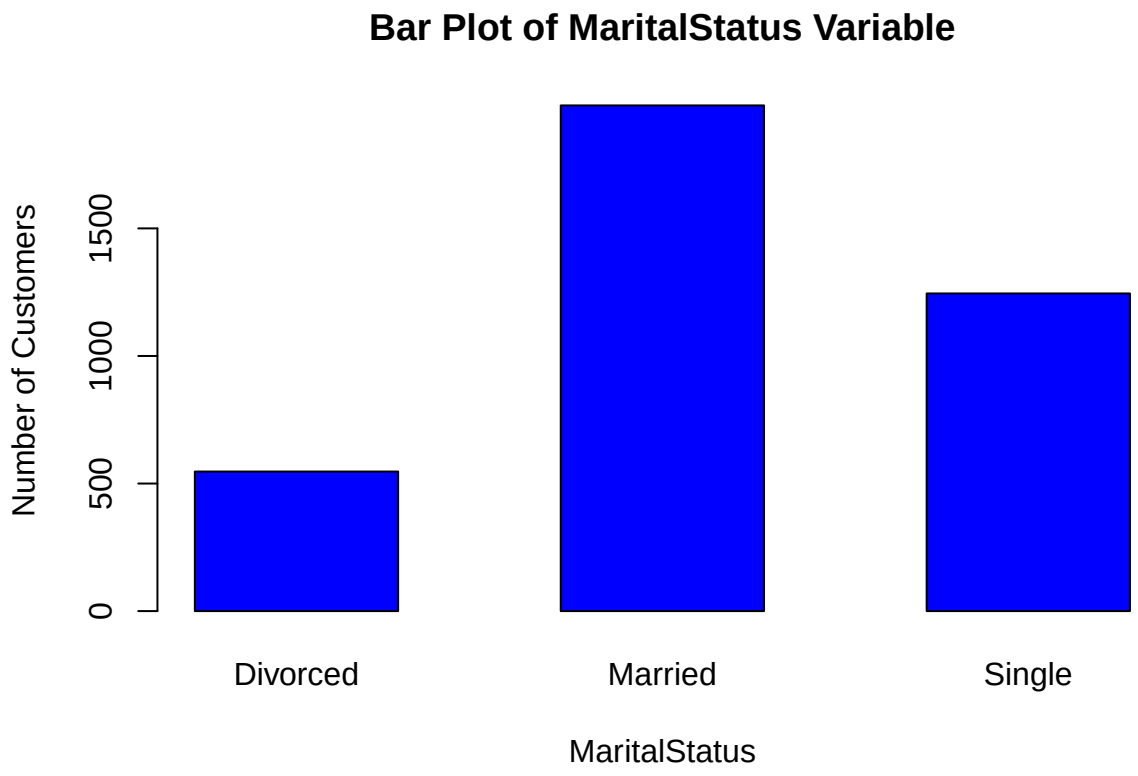

```
barplot(table(data$Gender),
        col = "blue",
        xlab = "Gender",
        ylab = "Number of Customers",
        main = "Bar Plot of Gender Variable",
        space = 0.8)
```



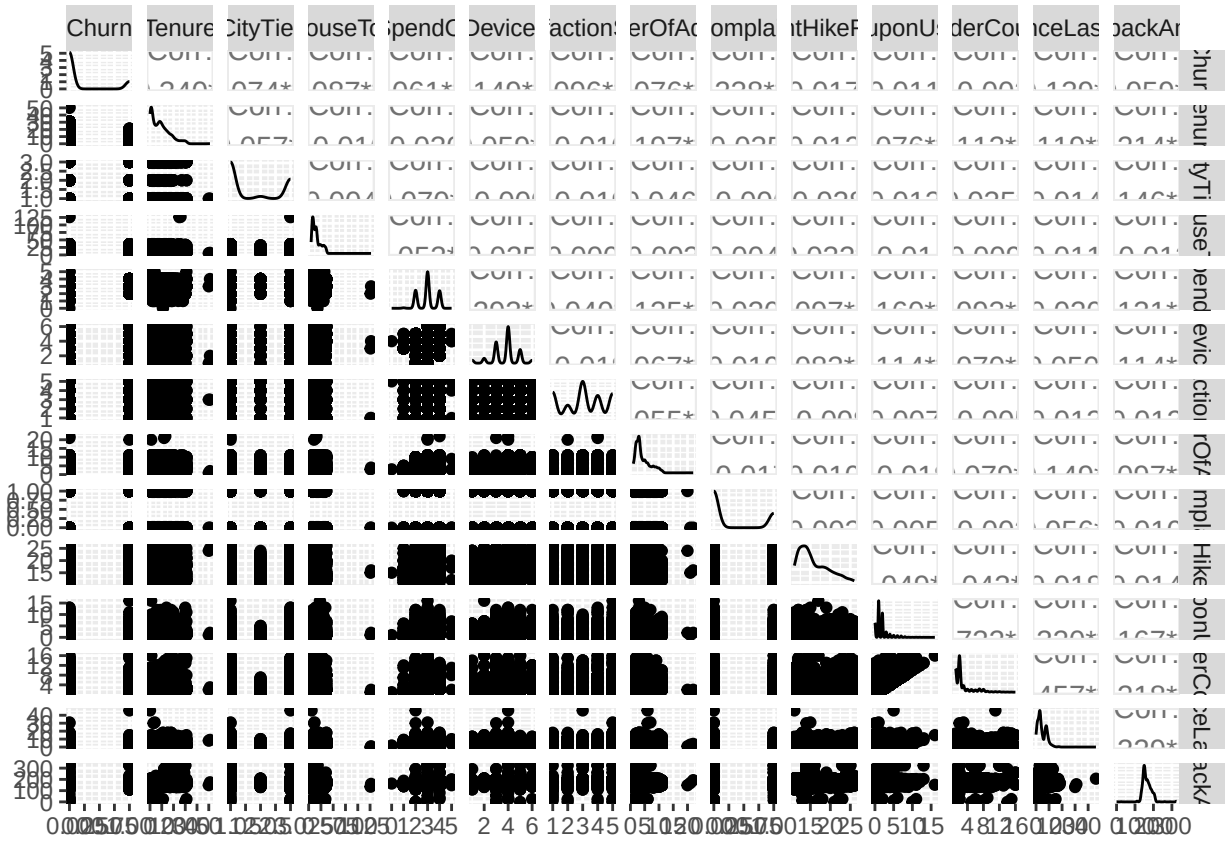

```r
barplot(table(data$PreferedOrderCat),
        col = "blue",
        ylab = "Number of Customers",
        main = "Bar Plot of PreferedOrderCat Variable",
        space = 0.2,
        las = 2,
        cex.names = 1)
```



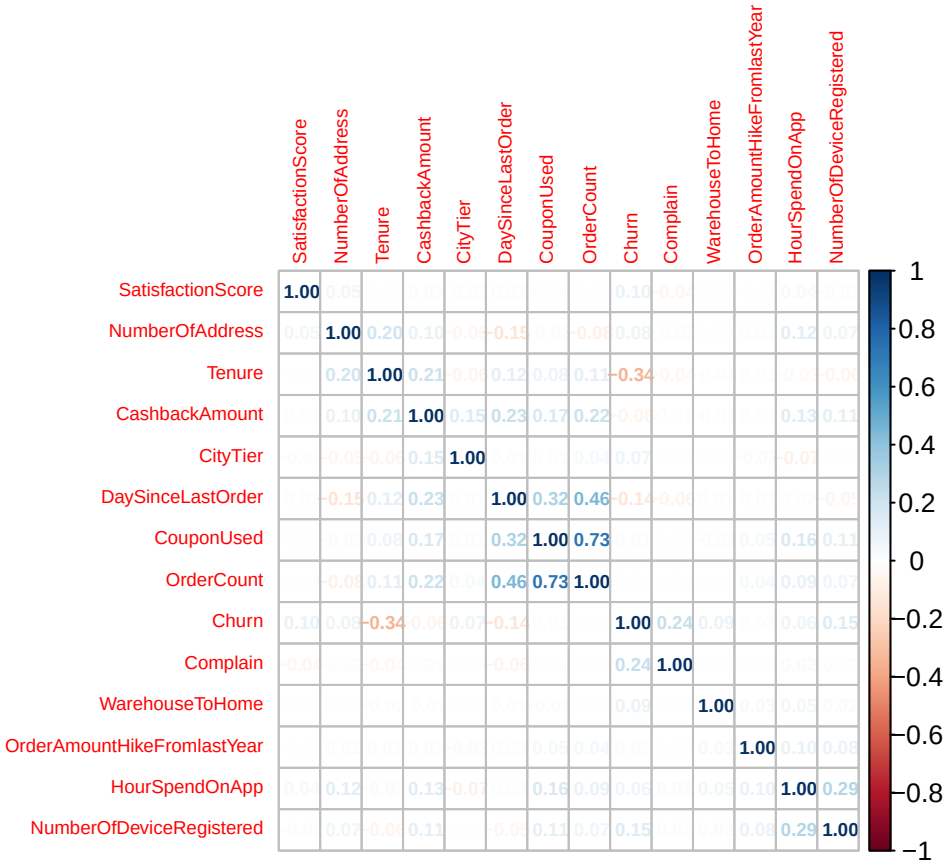

```
barplot(table(data$MaritalStatus),
        col = "blue",
        xlab = "MaritalStatus",
        ylab = "Number of Customers",
        main = "Bar Plot of MaritalStatus Variable",
        space = 0.8)
```




```
# Check the relationship among numeric variables
numeric_data <- data[, sapply(data, is.numeric)]
ggpairs(numeric_data)
```




```
corrplot(cor(numeric_data),
         method = "number",
         number.cex = 0.6,
         tl.cex = 0.6,
         order = "hclust")
```



This heatmap shows the correlation matrix between various customer behavior and profile features.

Correlation values:

- Close to 1 → strong positive relationship.
- Close to -1 → strong negative relationship.
- Close to 0 → no linear relationship.

Color scheme: Dark blue indicates strong positive correlation; red indicates strong negative correlation.

Key Correlations

Strong Positive Correlations:

- CouponUsed and OrderCount: 0.73. Customers who use more coupons tend to place more orders.
- OrderCount and DaySinceLastOrder: 0.46. Customers who place more orders tend to have more days since their last order. This is a little counterintuitive — it might suggest that high-frequency buyers could also experience recent inactivity (possible loyalty fatigue or seasonal effects). Needs deeper analysis
- CouponUsed and DaySinceLastOrder: 0.32. Customers who use more coupons also show longer gaps since the last order. Possibly because promotions attract buyers during campaigns, but they are not consistent buyers otherwise
- HourSpendOnApp and NumberOfDeviceRegistered; 0.29. Customers who spend more time on the app tend to register more devices. Likely reflects tech-savvy, engaged customers who access the service from multiple devices

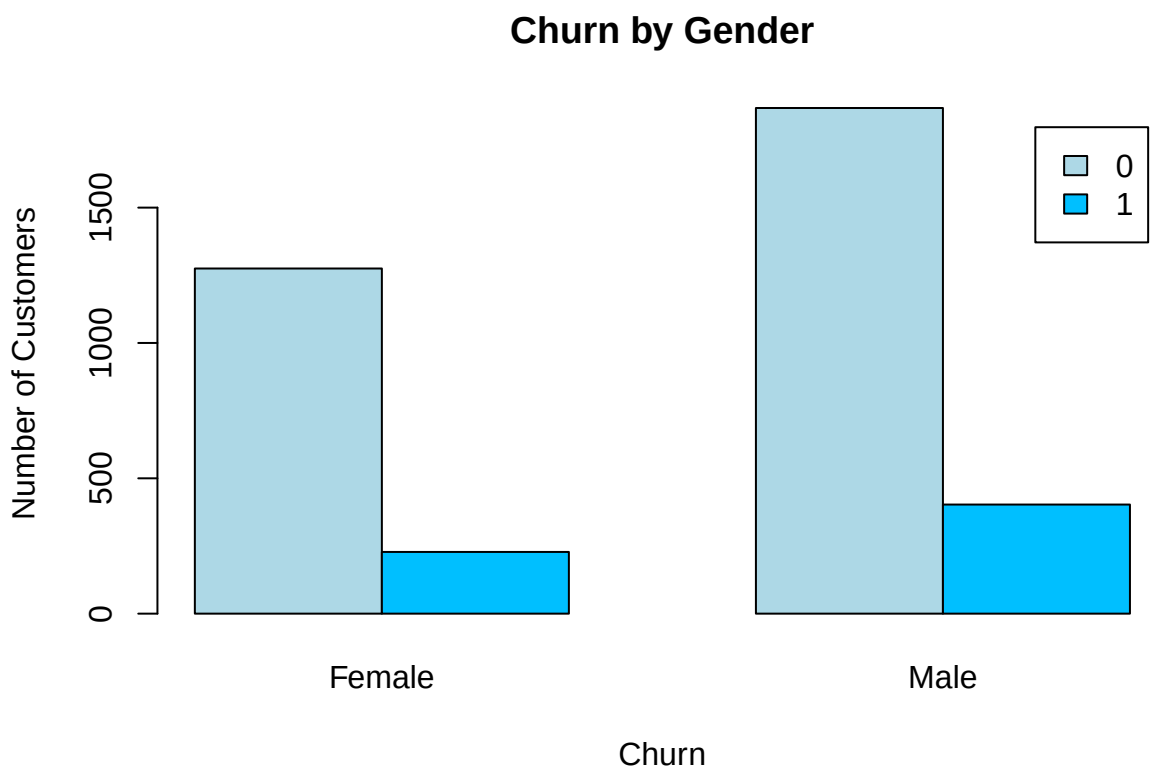
Strong Negative Correlations:

- Tenure and Churn: -0.34. Customers with a longer tenure (longer history with the company) are less likely to churn..

Weak Correlations:

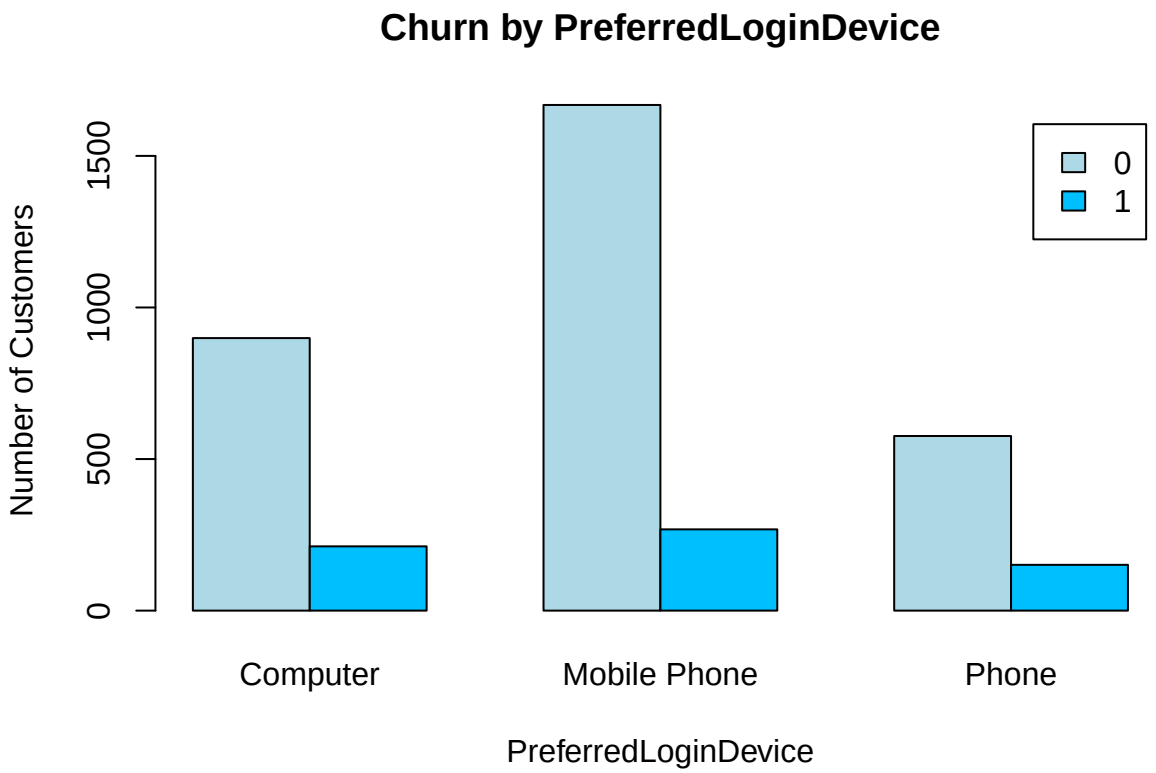
- WarehouseToHome, OrderAmountHikeFromLastYear have very weak correlations with other features (correlation values close to 0).

```
# Churn within categories

churn_gender_table <- table(data$Churn, data$Gender)
barplot(churn_gender_table, beside = TRUE, col = c("lightblue", "deepskyblue"),
        legend = rownames(churn_gender_table),
        xlab = "Churn", ylab = "Number of Customers",
        main = "Churn by Gender")
```

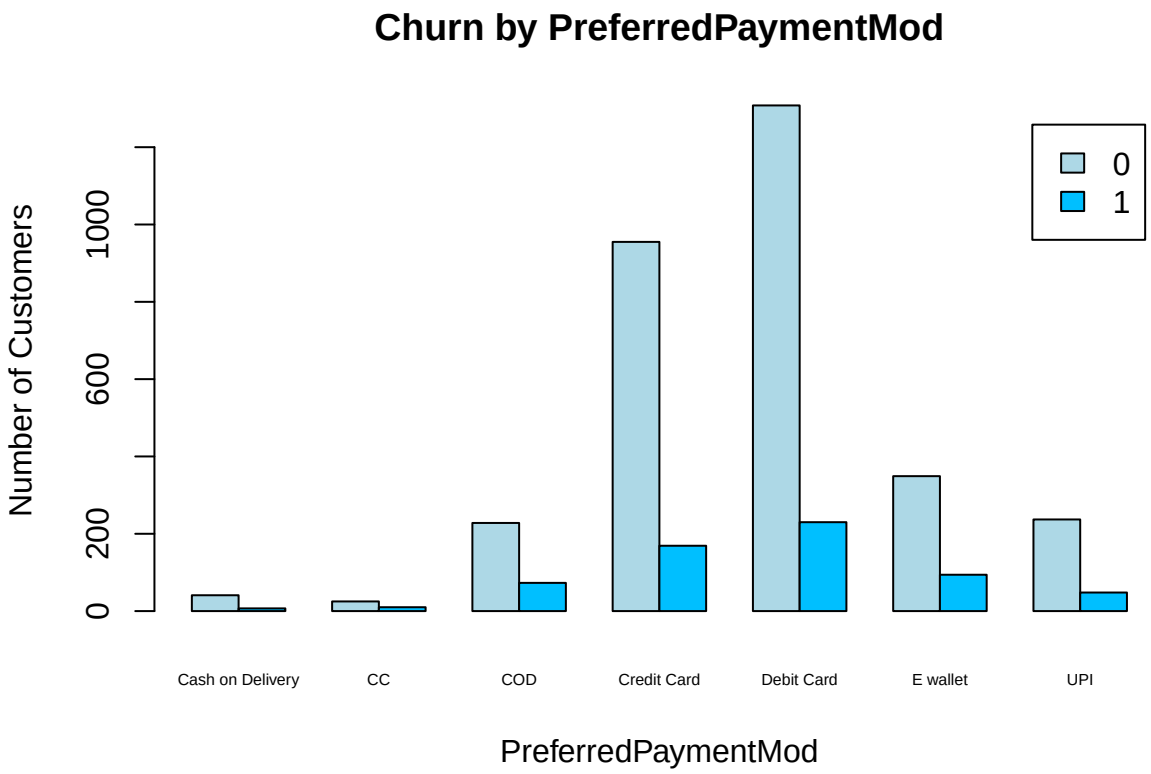



```
churn_gender_table <- table(data$Churn, data$Gender)



churn_PreferredLoginDevice_table <- table(data$Churn, data$PreferredLoginDevice)
barplot(churn_PreferredLoginDevice_table, beside = TRUE, col = c("lightblue", "deepskyblue"),
```

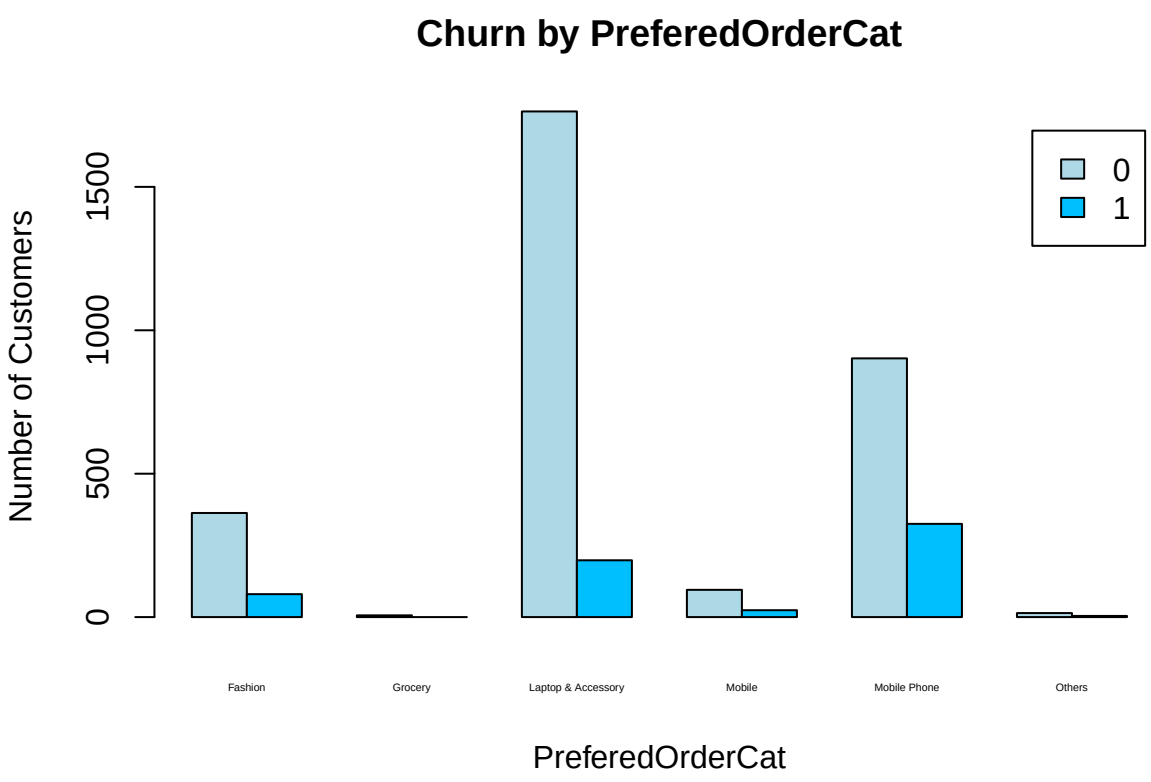
```
        legend = rownames(churn_PreferredLoginDevice_table),
        xlab = "PreferredLoginDevice", ylab = "Number of Customers",
        main = "Churn by PreferredLoginDevice")
```



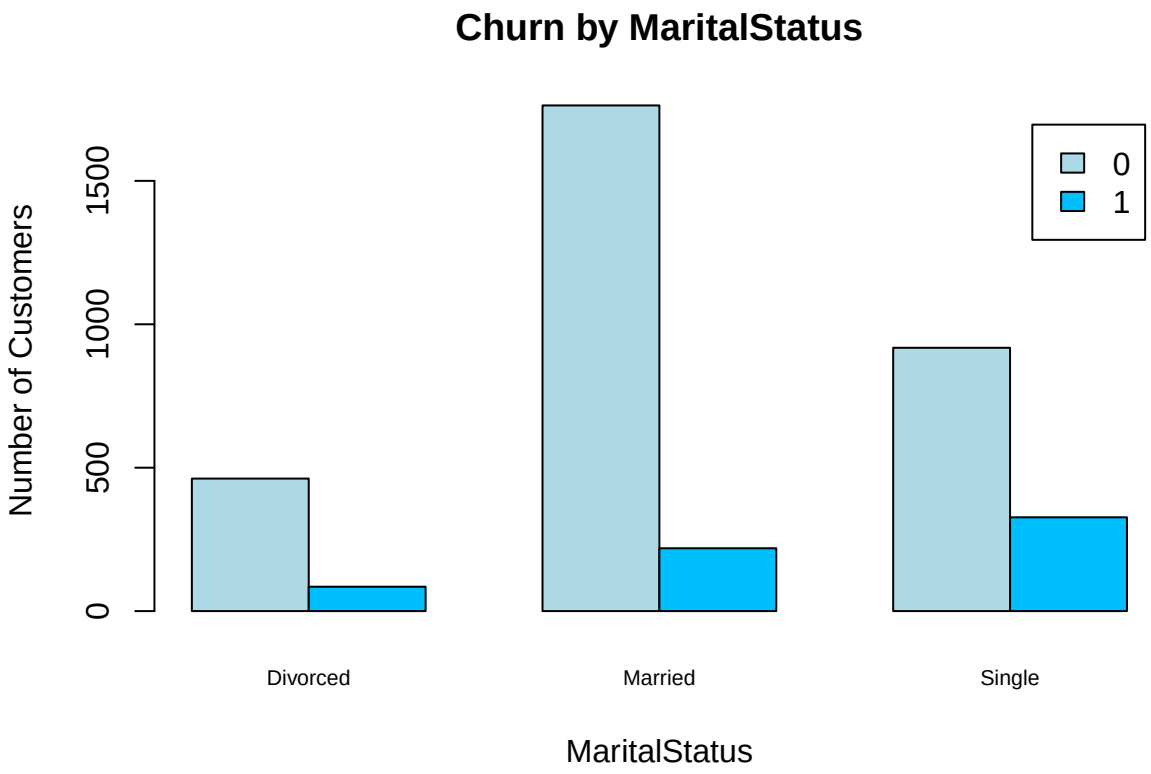

```
churn_PreferredPaymentMod_table <- table(data$Churn, data$PreferredPaymentMod)
barplot(churn_PreferredPaymentMod_table, beside = TRUE, col = c("lightblue", "deepskyblue"),
        legend = rownames(churn_PreferredPaymentMod_table),
        xlab = "PreferredPaymentMod", ylab = "Number of Customers",
        main = "Churn by PreferredPaymentMod",
        cex.names = 0.5)
```




```r
churn_PreferedOrderCat_table <- table(data$Churn, data$PreferedOrderCat)
barplot(churn_PreferedOrderCat_table, beside = TRUE, col = c("lightblue", "deepskyblue"),
        legend = rownames(churn_PreferedOrderCat_table),
        xlab = "PreferedOrderCat", ylab = "Number of Customers",
        main = "Churn by PreferedOrderCat",
        cex.names = 0.3)
```




```r
churn_MaritalStatus_table <- table(data$Churn, data$MaritalStatus)
barplot(churn_MaritalStatus_table, beside = TRUE, col = c("lightblue", "deepskyblue"),
        legend = rownames(churn_MaritalStatus_table),
        xlab = "MaritalStatus", ylab = "Number of Customers",
        main = "Churn by MaritalStatus",
        cex.names = 0.7)
```



5. Method & Model

5.1. Data transformation

To obtain reliable and accurate model results, data preprocessing is essential. As previously mentioned, approximately 5% of the samples contain missing values (NA). These samples were removed from the dataset to ensure uniformity and completeness during subsequent processing stages.

The dataset contains a mix of numerical and categorical variables. To properly handle the categorical variables, all text fields were converted into factors, allowing them to be treated appropriately in statistical models and machine learning algorithms

```
# Convert category variables into factor
data[sapply(data, is.character)] <- lapply(data[sapply(data, is.character)], factor)

# Check result
summary(data)
```

```
##      Churn           Tenure        PreferredLoginDevice    CityTier
##  Min.   :0.0000   Min.   : 0.000   Computer    :1111      Min.   :1.000
##  1st Qu.:0.0000   1st Qu.: 1.000   Mobile Phone:1936      1st Qu.:1.000
##  Median :0.0000   Median : 8.000   Phone       : 727      Median :1.000
##  Mean   :0.1672   Mean   : 8.777                          Mean   :1.708
##  3rd Qu.:0.0000   3rd Qu.:13.000                          3rd Qu.:3.000
```

```
##  Max.   :1.0000   Max.   :51.000                          Max.   :3.000
##
##  WarehouseToHome        PreferredPaymentMode    Gender     HourSpendOnApp
##  Min.   :  5.00   Cash on Delivery:  48      Female:1503   Min.   :0.000
##  1st Qu.:  9.00   CC              :  35      Male  :2271   1st Qu.:2.000
##  Median : 14.00   COD             : 301                    Median :3.000
##  Mean   : 15.74   Credit Card     :1124                    Mean   :2.981
##  3rd Qu.: 21.00   Debit Card      :1538                    3rd Qu.:3.000
##  Max.   :127.00   E wallet        : 443                    Max.   :5.000
##                   UPI             : 285
##  NumberOfDeviceRegistered          PreferedOrderCat SatisfactionScore
##  Min.   :1.000             Fashion           : 443    Min.   :1.000
##  1st Qu.:3.000             Grocery           :   6    1st Qu.:2.000
##  Median :4.000             Laptop & Accessory:1961    Median :3.000
##  Mean   :3.754             Mobile            : 119    Mean   :3.056
##  3rd Qu.:4.000             Mobile Phone      :1227    3rd Qu.:4.000
##  Max.   :6.000             Others            :  18    Max.   :5.000
##
##   MaritalStatus  NumberOfAddress     Complain      OrderAmountHikeFromlastYear
##  Divorced: 547   Min.   : 1.000   Min.   :0.0000   Min.   :11.00
##  Married :1982   1st Qu.: 2.000   1st Qu.:0.0000   1st Qu.:13.00
##  Single  :1245   Median : 3.000   Median :0.0000   Median :15.00
##                  Mean   : 4.216   Mean   :0.2822   Mean   :15.73
##                  3rd Qu.: 6.000   3rd Qu.:1.0000   3rd Qu.:18.00
##                  Max.   :22.000   Max.   :1.0000   Max.   :26.00
##
##    CouponUsed      OrderCount     DaySinceLastOrder CashbackAmount
##  Min.   : 0.00   Min.   : 1.000   Min.   : 0.000    Min.   :  0.0
##  1st Qu.: 1.00   1st Qu.: 1.000   1st Qu.: 2.000    1st Qu.:148.2
##  Median : 1.00   Median : 2.000   Median : 3.000    Median :160.0
##  Mean   : 1.72   Mean   : 2.825   Mean   : 4.526    Mean   :164.2
##  3rd Qu.: 2.00   3rd Qu.: 3.000   3rd Qu.: 7.000    3rd Qu.:178.0
##  Max.   :16.00   Max.   :16.000   Max.   :46.000    Max.   :325.0
##
```

5.2. Model selection

Apply log transformation for continuous right-skew variables

```
# Check "zero-value"
vars_to_check <- c("Tenure", "WarehouseToHome", "NumberOfAddress",
                   "OrderAmountHikeFromlastYear", "CouponUsed",
                   "OrderCount", "DaySinceLastOrder", "CashbackAmount")
sapply(vars_to_check, function(var) {
  sum(data[[var]] == 0)
})
```

```
##                     Tenure             WarehouseToHome
##                        303                           0
##            NumberOfAddress OrderAmountHikeFromlastYear
##                          0                           0
##                 CouponUsed                  OrderCount
##                        610                           0
```

```
##           DaySinceLastOrder              CashbackAmount
##                         207                           4
```

```
#Log - transformation

data$log_Tenure <- log(data$Tenure + 1)
data$log_WarehouseToHome <- log(data$WarehouseToHome)
data$log_NumberOfAddress <- log(data$NumberOfAddress)
data$log_OrderAmountHikeFromlastYear <- log(data$OrderAmountHikeFromlastYear)
data$log_CouponUsed <- log(data$CouponUsed + 1)
data$log_OrderCount <- log(data$OrderCount)
data$log_DaySinceLastOrder <- log(data$DaySinceLastOrder + 1)
data$log_CashbackAmount <- log(data$CashbackAmount + 1)
```

```
# Compare logit and probit model

mylogit <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + PreferredPa
  family = binomial(link = "logit"))

summary(mylogit)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + PreferredPaymentMode + Gender + HourSpendOnApp +
##     NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
##     MaritalStatus + log_NumberOfAddress + Complain + log_OrderAmountHikeFromlastYear +
##     log_CouponUsed + log_OrderCount + log_DaySinceLastOrder +
##     log_CashbackAmount, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                                    Estimate Std. Error z value Pr(>|z|)
## (Intercept)                        -4.44543    2.89485  -1.536 0.124628
## log_Tenure                         -1.71957    0.08123 -21.170  < 2e-16 ***
## PreferredLoginDeviceMobile Phone   -0.52842    0.14229  -3.714 0.000204 ***
## PreferredLoginDevicePhone          -0.31936    0.18092  -1.765 0.077522 .
## CityTier                            0.28865    0.07756   3.721 0.000198 ***
## log_WarehouseToHome                 0.68573    0.12190   5.625 1.85e-08 ***
## PreferredPaymentModeCC             -0.81463    0.89271  -0.913 0.361484
## PreferredPaymentModeCOD            -0.16753    0.62870  -0.266 0.789872
## PreferredPaymentModeCredit Card    -0.70332    0.60164  -1.169 0.242404
## PreferredPaymentModeDebit Card     -0.52641    0.59676  -0.882 0.377716
## PreferredPaymentModeE wallet       -0.04716    0.61785  -0.076 0.939153
## PreferredPaymentModeUPI            -0.77897    0.63084  -1.235 0.216900
## GenderMale                          0.25921    0.12546   2.066 0.038812 *
## HourSpendOnApp                      0.09154    0.09971   0.918 0.358606
## NumberOfDeviceRegistered            0.38531    0.06785   5.679 1.36e-08 ***
## PreferedOrderCatGrocery           -12.47451  313.81841  -0.040 0.968292
## PreferedOrderCatLaptop & Accessory -1.77436    0.22183  -7.999 1.26e-15 ***
## PreferedOrderCatMobile             -1.29480    0.51844  -2.498 0.012507 *
## PreferedOrderCatMobile Phone       -0.90115    0.26043  -3.460 0.000540 ***
## PreferedOrderCatOthers              1.57843    0.71870   2.196 0.028076 *
## SatisfactionScore                   0.25602    0.04568   5.604 2.09e-08 ***
```

```
## MaritalStatusMarried                -0.29964    0.18096  -1.656 0.097759 .
## MaritalStatusSingle                  0.75270    0.18137   4.150 3.32e-05 ***
## log_NumberOfAddress                  1.28533    0.11808  10.885  < 2e-16 ***
## Complain                             1.66321    0.12714  13.082  < 2e-16 ***
## log_OrderAmountHikeFromlastYear     -0.06925    0.27770  -0.249 0.803062
## log_CouponUsed                      -0.26687    0.18436  -1.448 0.147744
## log_OrderCount                       0.85508    0.15964   5.356 8.49e-08 ***
## log_DaySinceLastOrder               -0.66471    0.11163  -5.954 2.61e-09 ***
## log_CashbackAmount                   0.08730    0.53588   0.163 0.870591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1865.8  on 3744  degrees of freedom
## AIC: 1925.8
##
## Number of Fisher Scoring iterations: 13
```

```
myprobit <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + PreferredPa
  family = binomial(link = "probit"))

summary(myprobit)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + PreferredPaymentMode + Gender + HourSpendOnApp +
##     NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
##     MaritalStatus + log_NumberOfAddress + Complain + log_OrderAmountHikeFromlastYear +
##     log_CouponUsed + log_OrderCount + log_DaySinceLastOrder +
##     log_CashbackAmount, family = binomial(link = "probit"), data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -1.93279    1.37017  -1.411 0.158356
## log_Tenure                        -0.89854    0.04026 -22.316  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.28244    0.07562  -3.735 0.000188 ***
## PreferredLoginDevicePhone         -0.15296    0.09802  -1.560 0.118660
## CityTier                           0.15460    0.04119   3.753 0.000175 ***
## log_WarehouseToHome                0.36502    0.06481   5.632 1.78e-08 ***
## PreferredPaymentModeCC            -0.70616    0.46244  -1.527 0.126758
## PreferredPaymentModeCOD           -0.32780    0.29433  -1.114 0.265409
## PreferredPaymentModeCredit Card   -0.64103    0.27724  -2.312 0.020767 *
## PreferredPaymentModeDebit Card    -0.58226    0.27466  -2.120 0.034010 *
## PreferredPaymentModeE wallet      -0.35276    0.28777  -1.226 0.220265
## PreferredPaymentModeUPI           -0.72544    0.29718  -2.441 0.014642 *
## GenderMale                         0.13393    0.06673   2.007 0.044749 *
## HourSpendOnApp                     0.04268    0.05283   0.808 0.419106
## NumberOfDeviceRegistered           0.20128    0.03597   5.596 2.19e-08 ***
## PreferedOrderCatGrocery           -4.40245   80.50886  -0.055 0.956391
## PreferedOrderCatLaptop & Accessory -0.98555    0.11372  -8.667  < 2e-16 ***
## PreferedOrderCatMobile            -0.78459    0.27906  -2.812 0.004931 **
```

```
## PreferedOrderCatMobile Phone       -0.55055    0.13446  -4.094 4.23e-05 ***
## PreferedOrderCatOthers              0.70534    0.38982   1.809 0.070393 .
## SatisfactionScore                   0.12567    0.02407   5.220 1.79e-07 ***
## MaritalStatusMarried               -0.16706    0.09632  -1.734 0.082832 .
## MaritalStatusSingle                 0.39684    0.09715   4.085 4.41e-05 ***
## log_NumberOfAddress                 0.67003    0.06185  10.834  < 2e-16 ***
## Complain                            0.87246    0.06719  12.985  < 2e-16 ***
## log_OrderAmountHikeFromlastYear    -0.04511    0.14829  -0.304 0.760959
## log_CouponUsed                     -0.16686    0.09553  -1.747 0.080692 .
## log_OrderCount                      0.48010    0.08150   5.891 3.84e-09 ***
## log_DaySinceLastOrder              -0.35997    0.05878  -6.124 9.14e-10 ***
## log_CashbackAmount                  0.04711    0.24789   0.190 0.849284
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1908.7  on 3744  degrees of freedom
## AIC: 1968.7
##
## Number of Fisher Scoring iterations: 13
```

To model customer churn effectively, we chose the Binary Logistic Regression Model. Our dataset's target variable, Churn, is a binary outcome (1/0), making logistic regression a natural and appropriate choice. Additionally, our dataset includes a combination of numerical and categorical predictors. After converting all categorical variables into factors, we ensure the logistic model can handle them appropriately without needing extensive additional preprocessing.

We prefer the binary logit model because it provides clear interpretability and better model performance. Furthermore, we compared the performance of the logistic and probit models using the Akaike Information Criterion (AIC). The logit model achieved a lower AIC value (1925) compared to the probit model (1968), indicating that the logit model fits the data better while maintaining simplicity and ease of interpretation.

The **logit model** is based on the **logistic function**, an S-shaped curve that maps any real-valued number into the interval (0, 1). This makes it ideal for modeling probabilities—such as the likelihood of **churn**, which is the focus of our study.

Mathematical Formulation

The **binary logit model** can be expressed as:

$$P(Y = 1 \mid X) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_n X_n)}$$

where:

- $P(Y = 1 \mid X)$ is the **probability of churn** given the predictor variables X.
- $\beta_0, \beta_1, \ldots, \beta_n$ are the **parameters** of the model to be estimated.
- $X_1, \ldots, X_n$ are the **explanatory variables** or **predictors** influencing churn.

Interpretation

The goal of estimating these parameters is to understand how changes in the predictor variables affect the **odds of churn**. Each coefficient β_i represents the **log-odds change** in churn associated with a one-unit increase in X_i, **holding all other variables constant**.

5.3. Methodology

```
# First model (full model)
summary(mylogit)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + PreferredPaymentMode + Gender + HourSpendOnApp +
##     NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
##     MaritalStatus + log_NumberOfAddress + Complain + log_OrderAmountHikeFromlastYear +
##     log_CouponUsed + log_OrderCount + log_DaySinceLastOrder +
##     log_CashbackAmount, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.44543    2.89485  -1.536 0.124628
## log_Tenure                        -1.71957    0.08123 -21.170  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.52842    0.14229  -3.714 0.000204 ***
## PreferredLoginDevicePhone         -0.31936    0.18092  -1.765 0.077522 .
## CityTier                           0.28865    0.07756   3.721 0.000198 ***
## log_WarehouseToHome                0.68573    0.12190   5.625 1.85e-08 ***
## PreferredPaymentModeCC            -0.81463    0.89271  -0.913 0.361484
## PreferredPaymentModeCOD           -0.16753    0.62870  -0.266 0.789872
## PreferredPaymentModeCredit Card   -0.70332    0.60164  -1.169 0.242404
## PreferredPaymentModeDebit Card    -0.52641    0.59676  -0.882 0.377716
## PreferredPaymentModeE wallet      -0.04716    0.61785  -0.076 0.939153
## PreferredPaymentModeUPI           -0.77897    0.63084  -1.235 0.216900
## GenderMale                         0.25921    0.12546   2.066 0.038812 *
## HourSpendOnApp                     0.09154    0.09971   0.918 0.358606
## NumberOfDeviceRegistered           0.38531    0.06785   5.679 1.36e-08 ***
## PreferedOrderCatGrocery          -12.47451  313.81841  -0.040 0.968292
## PreferedOrderCatLaptop & Accessory -1.77436    0.22183  -7.999 1.26e-15 ***
## PreferedOrderCatMobile            -1.29480    0.51844  -2.498 0.012507 *
## PreferedOrderCatMobile Phone      -0.90115    0.26043  -3.460 0.000540 ***
## PreferedOrderCatOthers             1.57843    0.71870   2.196 0.028076 *
## SatisfactionScore                  0.25602    0.04568   5.604 2.09e-08 ***
## MaritalStatusMarried              -0.29964    0.18096  -1.656 0.097759 .
## MaritalStatusSingle                0.75270    0.18137   4.150 3.32e-05 ***
## log_NumberOfAddress                1.28533    0.11808  10.885  < 2e-16 ***
## Complain                           1.66321    0.12714  13.082  < 2e-16 ***
## log_OrderAmountHikeFromlastYear   -0.06925    0.27770  -0.249 0.803062
## log_CouponUsed                    -0.26687    0.18436  -1.448 0.147744
## log_OrderCount                     0.85508    0.15964   5.356 8.49e-08 ***
## log_DaySinceLastOrder             -0.66471    0.11163  -5.954 2.61e-09 ***
## log_CashbackAmount                 0.08730    0.53588   0.163 0.870591
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1865.8  on 3744  degrees of freedom
```

```
## AIC: 1925.8
##
## Number of Fisher Scoring iterations: 13
```

Remove statistically insignificant variables

We sequentially remove statistically insignificant variables based on the regression results from the `summary(mylogit)` function. Specifically, we retain variables with p values less than 0.05 (shown in the `Pr(>|z|)` column), corresponding to a significance level of 5%.

Insignificant variables (p > 0.05)

- `PreferedOrderCatGrocery` (p = 0.968292)
- `PreferredPaymentModeE wallet` (p = 0.939153)
- `CashbackAmount` (p = 0.870591)
- `OrderAmountHikeFromlastYear` (p = 0.803062)
- `PreferredPaymentModeCOD` (p = 0.789872)
- `PreferredPaymentModeDebit Card` (p = 0.377716)
- `PreferredPaymentModeCC` (p = 0.361484)
- `HourSpendOnApp` (p = 0.358606)
- `PreferredPaymentModeCredit Card` (p = 0.242404)
- `PreferredPaymentModeUPI` (p = 0.216900)
- `CouponUsed` (p = 0.147744)
- `MaritalStatusMarried` (p = 0.097759)
- `PreferredLoginDevicePhone` (p = 0.077522)

```
null_logit = glm(Churn~1, data=data, family=binomial(link="logit"))
lrtest(mylogit, null_logit)
```

Compare Full model and Null model

```
## Likelihood ratio test
##
## Model 1: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
## Model 2: Churn ~ 1
##   #Df   LogLik  Df  Chisq Pr(>Chisq)
## 1  30  -932.88
## 2   1 -1703.63 -29 1541.5  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The p-value is extremely small (p < 0.001), indicating strong evidence against the null model. In other words, the full model provides a significantly better fit to the data than the null model.The set of explanatory variables jointly contributes significantly to explaining the likelihood of churn. Therefore, including these predictors in the model is statistically justified.

Variance Inflation Factor (VIF) To check for multicollinearity in your model using the Variance Inflation Factor (VIF), you can use the vif() function from the car package in R. The VIF indicates how much the variance of a regression coefficient is inflated due to collinearity with other predictors. A higher VIF value suggests multicollinearity.

```
vif(mylogit)
```

```
##                                     GVIF Df GVIF^(1/(2*Df))
## log_Tenure                      1.566864  1        1.251744
## PreferredLoginDevice            1.405962  2        1.088913
## CityTier                        1.561601  1        1.249640
## log_WarehouseToHome             1.075885  1        1.037249
## PreferredPaymentMode            2.829258  6        1.090534
## Gender                          1.038172  1        1.018907
## HourSpendOnApp                  1.311566  1        1.145236
## NumberOfDeviceRegistered        1.176532  1        1.084681
## PreferedOrderCat                5.952965  5        1.195290
## SatisfactionScore               1.068287  1        1.033580
## MaritalStatus                   1.100013  2        1.024117
## log_NumberOfAddress             1.307668  1        1.143533
## Complain                        1.096877  1        1.047319
## log_OrderAmountHikeFromlastYear 1.056426  1        1.027826
## log_CouponUsed                  2.334681  1        1.527966
## log_OrderCount                  2.760417  1        1.661450
## log_DaySinceLastOrder           1.460998  1        1.208718
## log_CashbackAmount              2.271725  1        1.507224
```

To check for multicollinearity among explanatory variables, we calculated the **Generalized Variance Inflation Factor (GVIF)** for each variable. We adjusted the GVIF values using the formula:

$$GVIF^{1/(2 \cdot Df)}$$

to account for variables with multiple degrees of freedom.

All adjusted GVIF values were below 2, which suggests **no serious multicollinearity** among the predictors.

The highest adjusted GVIF was for `log_OrderCount` (1.661450), which is still well within acceptable limits.

5.4. General-to-specific method to variables selection

Model 1: Remove "CashbackAmount" variable

```
mylogit1 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + PreferredPa
  family = binomial(link = "logit"))

summary(mylogit1)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + PreferredPaymentMode + Gender + HourSpendOnApp +
##     NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
```

```
##     MaritalStatus + log_NumberOfAddress + Complain + log_OrderAmountHikeFromlastYear +
##     log_CouponUsed + log_OrderCount + log_DaySinceLastOrder,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.00905    1.09946  -3.646 0.000266 ***
## log_Tenure                        -1.71897    0.08112 -21.190  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.52677    0.14193  -3.711 0.000206 ***
## PreferredLoginDevicePhone         -0.32324    0.17929  -1.803 0.071405 .
## CityTier                           0.28864    0.07756   3.721 0.000198 ***
## log_WarehouseToHome                0.68557    0.12190   5.624 1.86e-08 ***
## PreferredPaymentModeCC            -0.81491    0.89321  -0.912 0.361594
## PreferredPaymentModeCOD           -0.16765    0.62941  -0.266 0.789960
## PreferredPaymentModeCredit Card   -0.70370    0.60237  -1.168 0.242723
## PreferredPaymentModeDebit Card    -0.52645    0.59750  -0.881 0.378270
## PreferredPaymentModeE wallet      -0.04661    0.61857  -0.075 0.939936
## PreferredPaymentModeUPI           -0.77957    0.63154  -1.234 0.217054
## GenderMale                         0.26000    0.12537   2.074 0.038093 *
## HourSpendOnApp                     0.09432    0.09831   0.959 0.337365
## NumberOfDeviceRegistered           0.38652    0.06745   5.731 1.00e-08 ***
## PreferedOrderCatGrocery          -12.46431  314.37643  -0.040 0.968374
## PreferedOrderCatLaptop & Accessory -1.78663  0.20890  -8.553  < 2e-16 ***
## PreferedOrderCatMobile            -1.31792    0.49886  -2.642 0.008245 **
## PreferedOrderCatMobile Phone      -0.92219    0.22637  -4.074 4.62e-05 ***
## PreferedOrderCatOthers             1.61775    0.67712   2.389 0.016886 *
## SatisfactionScore                  0.25620    0.04567   5.610 2.03e-08 ***
## MaritalStatusMarried              -0.30005    0.18096  -1.658 0.097293 .
## MaritalStatusSingle                0.75278    0.18139   4.150 3.32e-05 ***
## log_NumberOfAddress                1.28746    0.11736  10.970  < 2e-16 ***
## Complain                           1.66402    0.12707  13.096  < 2e-16 ***
## log_OrderAmountHikeFromlastYear   -0.06897    0.27769  -0.248 0.803865
## log_CouponUsed                    -0.26568    0.18419  -1.442 0.149195
## log_OrderCount                     0.85548    0.15960   5.360 8.32e-08 ***
## log_DaySinceLastOrder             -0.66269    0.11093  -5.974 2.32e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1865.8  on 3745  degrees of freedom
## AIC: 1923.8
##
## Number of Fisher Scoring iterations: 13
```

```
lrtest(mylogit1, mylogit)
```

```
## Likelihood ratio test
##
## Model 1: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
```

```
##     log_OrderCount + log_DaySinceLastOrder
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  29 -932.90
## 2  30 -932.88  1 0.0286     0.8658
```

The p-value is extremely high (0.8658), indicating that the variable `CashbackAmount` does **not** provide any statistically significant improvement in model fit.

Conclusion: There is no evidence to justify keeping `CashbackAmount` in the model. Removing it does not reduce model performance, and its exclusion helps simplify the model without loss of explanatory power.

Model 2: Remove "CashbackAmount", "OrderAmountHikeFromlastYear" variables

```
mylogit2 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + PreferredPa
  family = binomial(link = "logit"))

summary(mylogit2)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + PreferredPaymentMode + Gender + HourSpendOnApp +
##     NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
##     MaritalStatus + log_NumberOfAddress + Complain + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.19165    0.81854  -5.121 3.04e-07 ***
## log_Tenure                        -1.71898    0.08111 -21.192  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.52658    0.14195  -3.710 0.000207 ***
## PreferredLoginDevicePhone         -0.32267    0.17926  -1.800 0.071854 .
## CityTier                           0.28962    0.07745   3.739 0.000184 ***
## log_WarehouseToHome                0.68355    0.12162   5.620 1.91e-08 ***
## PreferredPaymentModeCC            -0.80675    0.89289  -0.904 0.366246
## PreferredPaymentModeCOD           -0.15821    0.62889  -0.252 0.801378
## PreferredPaymentModeCredit Card   -0.69675    0.60235  -1.157 0.247388
## PreferredPaymentModeDebit Card    -0.51989    0.59755  -0.870 0.384278
## PreferredPaymentModeE wallet      -0.03953    0.61852  -0.064 0.949040
## PreferredPaymentModeUPI           -0.77263    0.63156  -1.223 0.221186
## GenderMale                         0.26128    0.12523   2.086 0.036951 *
## HourSpendOnApp                     0.09460    0.09831   0.962 0.335937
## NumberOfDeviceRegistered           0.38558    0.06733   5.727 1.02e-08 ***
## PreferedOrderCatGrocery          -12.47079  314.61006  -0.040 0.968381
## PreferedOrderCatLaptop & Accessory -1.79148    0.20794  -8.615  < 2e-16 ***
## PreferedOrderCatMobile            -1.31546    0.49830  -2.640 0.008294 **
## PreferedOrderCatMobile Phone      -0.92730    0.22534  -4.115 3.87e-05 ***
```

```
## PreferedOrderCatOthers               1.60880    0.67739   2.375 0.017549 *
## SatisfactionScore                    0.25581    0.04565   5.604 2.09e-08 ***
## MaritalStatusMarried                -0.30018    0.18088  -1.660 0.097009 .
## MaritalStatusSingle                  0.75203    0.18132   4.148 3.36e-05 ***
## log_NumberOfAddress                  1.28734    0.11739  10.966  < 2e-16 ***
## Complain                             1.66471    0.12704  13.104  < 2e-16 ***
## log_CouponUsed                      -0.26792    0.18403  -1.456 0.145427
## log_OrderCount                       0.85527    0.15965   5.357 8.45e-08 ***
## log_DaySinceLastOrder               -0.66214    0.11090  -5.970 2.37e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1865.9  on 3746  degrees of freedom
## AIC: 1921.9
##
## Number of Fisher Scoring iterations: 13
```

```
lrtest(mylogit2, mylogit1)
```

```
## Likelihood ratio test
##
## Model 1: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_CouponUsed + log_OrderCount + log_DaySinceLastOrder
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  28 -932.93
## 2  29 -932.90  1 0.0617     0.8038
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0"))
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df Chisq Pr(>Chisq)
```

```
## 1   3746
## 2   3744  2 0.088     0.9569
```

Likelihood ratio test: Since the p-value is far greater than 0.05 (0.8038), we fail to reject the null hypothesis.This means that OrderAmountHikeFromlastYear does not significantly improve the model when added. Therefore, it can be removed from the model without reducing its explanatory power

Wald Test: Joint Significance of CashbackAmount and OrderAmountHikeFromlastYear: The Wald test evaluates whether two coefficients—CashbackAmount and OrderAmountHikeFromlastYear—are jointly equal to zero.

Hypotheses:

- H0 (null): Both coefficients are 0 (i.e., they have no effect).
- H1 (alternative): At least one coefficient is non-zero.

With a p-value of 0.9569, we fail to reject the null hypothesis, indicating that CashbackAmount and OrderAmountHikeFromlastYear are not jointly significant. Their contribution to explaining churn is negligible, and they can be safely excluded from the model to simplify it.

Model 3: Remove "CashbackAmount", "OrderAmountHikeFromlastYear", "PreferredPaymentMode" variables

```
mylogit3 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + Gender + H
  family = binomial(link = "logit"))

summary(mylogit3)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_CouponUsed + log_OrderCount + log_DaySinceLastOrder,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.78648    0.59248  -8.079 6.55e-16 ***
## log_Tenure                        -1.70938    0.08039 -21.263  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.52409    0.14019  -3.738 0.000185 ***
## PreferredLoginDevicePhone         -0.36251    0.17840  -2.032 0.042158 *
## CityTier                           0.37998    0.06968   5.453 4.96e-08 ***
## log_WarehouseToHome                0.66824    0.11991   5.573 2.51e-08 ***
## GenderMale                         0.24653    0.12387   1.990 0.046565 *
## HourSpendOnApp                     0.10130    0.09717   1.043 0.297140
## NumberOfDeviceRegistered           0.38479    0.06654   5.783 7.34e-09 ***
## PreferedOrderCatGrocery          -12.60306  312.49731  -0.040 0.967830
## PreferedOrderCatLaptop & Accessory -1.79826    0.20518  -8.764  < 2e-16 ***
## PreferedOrderCatMobile             -1.32510    0.40728  -3.254 0.001140 **
## PreferedOrderCatMobile Phone       -0.93836    0.22000  -4.265 2.00e-05 ***
## PreferedOrderCatOthers              1.60063    0.69581   2.300 0.021427 *
## SatisfactionScore                   0.25803    0.04533   5.693 1.25e-08 ***
```

```
## MaritalStatusMarried                -0.30670    0.18000  -1.704 0.088390 .
## MaritalStatusSingle                  0.73360    0.18026   4.070 4.71e-05 ***
## log_NumberOfAddress                  1.26377    0.11599  10.896  < 2e-16 ***
## Complain                             1.67576    0.12642  13.256  < 2e-16 ***
## log_CouponUsed                      -0.24996    0.18292  -1.367 0.171781
## log_OrderCount                       0.83933    0.15916   5.274 1.34e-07 ***
## log_DaySinceLastOrder               -0.66623    0.10940  -6.090 1.13e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1879.7  on 3752  degrees of freedom
## AIC: 1923.7
##
## Number of Fisher Scoring iterations: 13
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3752
## 2   3744  8 13.925    0.08375 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wald test

Since p > 0.05, you fail to reject the null hypothesis, meaning: These three variables do not contribute significantly to the model jointly. It is safe to remove PreferredPaymentMod, HourSpendOnApp, OrderAmountHikeFromlastYear,and PreferredPaymentMode from our model.

Model 4: Remove "HourSpendOnApp", "OrderAmountHikeFromlastYear", "CashbackAmount", "PreferredPaymentMode" variables

```r
mylogit4 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + Gender + Nu
  family = binomial(link = "logit"))
summary(mylogit4)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_CouponUsed + log_OrderCount + log_DaySinceLastOrder,
##     family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.64517    0.57566  -8.069 7.07e-16 ***
## log_Tenure                        -1.70153    0.07982 -21.317  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.51646    0.13999  -3.689 0.000225 ***
## PreferredLoginDevicePhone         -0.37523    0.17785  -2.110 0.034869 *
## CityTier                           0.38157    0.06965   5.478 4.29e-08 ***
## log_WarehouseToHome                0.67803    0.11954   5.672 1.41e-08 ***
## GenderMale                         0.24648    0.12385   1.990 0.046575 *
## NumberOfDeviceRegistered           0.39726    0.06555   6.061 1.36e-09 ***
## PreferedOrderCatGrocery          -12.62001  313.17989  -0.040 0.967857
## PreferedOrderCatLaptop & Accessory -1.77272    0.20340  -8.716  < 2e-16 ***
## PreferedOrderCatMobile             -1.29601    0.40626  -3.190 0.001422 **
## PreferedOrderCatMobile Phone       -0.87955    0.21206  -4.148 3.36e-05 ***
## PreferedOrderCatOthers              1.61493    0.69643   2.319 0.020402 *
## SatisfactionScore                   0.26158    0.04521   5.786 7.19e-09 ***
## MaritalStatusMarried               -0.31317    0.17981  -1.742 0.081556 .
## MaritalStatusSingle                 0.72494    0.17990   4.030 5.59e-05 ***
## log_NumberOfAddress                 1.27376    0.11556  11.023  < 2e-16 ***
## Complain                            1.67803    0.12637  13.278  < 2e-16 ***
## log_CouponUsed                     -0.22970    0.18145  -1.266 0.205533
## log_OrderCount                      0.83902    0.15875   5.285 1.26e-07 ***
## log_DaySinceLastOrder              -0.66054    0.10908  -6.056 1.40e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1880.8  on 3753  degrees of freedom
## AIC: 1922.8
##
## Number of Fisher Scoring iterations: 13
```

```r
lrtest(mylogit4, mylogit3)
```

```
## Likelihood ratio test
##
## Model 1: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     Gender + NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
```

```
##     MaritalStatus + log_NumberOfAddress + Complain + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     Gender + HourSpendOnApp + NumberOfDeviceRegistered + PreferedOrderCat +
##     SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_CouponUsed + log_OrderCount + log_DaySinceLastOrder
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  21 -940.41
## 2  22 -939.86  1 1.0882     0.2969
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
## HourSpendOnApp = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3753
## 2   3744  9 14.999    0.09096 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood ratio test

Since p > 0.05, we fail to reject the null hypothesis that the coefficient for HourSpendOnApp is zero.This suggests that HourSpendOnApp is not statistically significant in predicting churn

Wald test

Since p > 0.05, we fail to reject the null hypothesis, meaning: These four variables do not contribute significantly to the model jointly. It is safe to remove PreferredPaymentMod, HourSpendOnApp, OrderAmountHikeFromlastYear,and CashbackAmount from our model.

Model 6: Remove "HourSpendOnApp", "OrderAmountHikeFromlastYear", "CashbackAmount", "PreferredPaymentMode" and "CouponUsed" variables

```r
mylogit5 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + Gender + Nu
  family = binomial(link = "logit"))
summary(mylogit5)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.71281    0.57252  -8.232  < 2e-16 ***
## log_Tenure                        -1.70031    0.07972 -21.329  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.51886    0.13997  -3.707  0.00021 ***
## PreferredLoginDevicePhone         -0.37259    0.17771  -2.097  0.03603 *
## CityTier                           0.38026    0.06961   5.463 4.68e-08 ***
## log_WarehouseToHome                0.68448    0.11937   5.734 9.81e-09 ***
## GenderMale                         0.24871    0.12374   2.010  0.04443 *
## NumberOfDeviceRegistered           0.38911    0.06507   5.980 2.24e-09 ***
## PreferedOrderCatGrocery          -12.54966  316.43693  -0.040  0.96836
## PreferedOrderCatLaptop & Accessory -1.78780    0.20312  -8.802  < 2e-16 ***
## PreferedOrderCatMobile             -1.29576    0.40615  -3.190  0.00142 **
## PreferedOrderCatMobile Phone       -0.90278    0.21114  -4.276 1.90e-05 ***
## PreferedOrderCatOthers              1.59205    0.69922   2.277  0.02279 *
## SatisfactionScore                   0.26153    0.04517   5.790 7.05e-09 ***
## MaritalStatusMarried               -0.29425    0.17923  -1.642  0.10065
## MaritalStatusSingle                 0.73590    0.17983   4.092 4.27e-05 ***
## log_NumberOfAddress                 1.26267    0.11501  10.979  < 2e-16 ***
## Complain                            1.68154    0.12636  13.308  < 2e-16 ***
## log_OrderCount                      0.70290    0.11715   6.000 1.97e-09 ***
## log_DaySinceLastOrder              -0.65750    0.10897  -6.034 1.60e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3407.3  on 3773  degrees of freedom
## Residual deviance: 1882.4  on 3754  degrees of freedom
## AIC: 1922.4
##
## Number of Fisher Scoring iterations: 13
```

```r
lrtest(mylogit5, mylogit4)
```

```
## Likelihood ratio test
##
## Model 1: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     Gender + NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
##     MaritalStatus + log_NumberOfAddress + Complain + log_OrderCount +
```

```
##     log_DaySinceLastOrder
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     Gender + NumberOfDeviceRegistered + PreferedOrderCat + SatisfactionScore +
##     MaritalStatus + log_NumberOfAddress + Complain + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1  20 -941.21
## 2  21 -940.41  1 1.5952     0.2066
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
## HourSpendOnApp = 0
## log_CouponUsed = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df Chisq Pr(>Chisq)
## 1   3754
## 2   3744 10 16.55    0.08495 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Likelihood ratio test

Since p > 0.05, we fail to reject the null hypothesis that the coefficient for CouponUsed is zero.This suggests that CouponUsed is not statistically significant in predicting churn

Wald test

Since p > 0.05, we fail to reject the null hypothesis, meaning: These five variables do not contribute significantly to the model jointly. It is safe to remove HourSpendOnApp, OrderAmountHikeFromlastYear,CashbackAmount,PreferredPaymentMode and CouponUsed from our model.

Model 6: Remove "HourSpendOnApp", "OrderAmountHikeFromlastYear", "CashbackAmount", "PreferredPaymentMode", "CouponUsed" and "PreferedOrderCatGrocery" variables

```r
# Remove "Grocery" and relevel
data$PreferedOrderCat <- factor(data$PreferedOrderCat, levels = setdiff(levels(data$PreferedOrderCat), "

# Fit updated model
mylogit6 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + Gender + Nu
  family = binomial(link = "logit"))

summary(mylogit6)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -4.71281    0.57250  -8.232  < 2e-16 ***
## log_Tenure                       -1.70031    0.07971 -21.331  < 2e-16 ***
## PreferredLoginDeviceMobile Phone -0.51886    0.13996  -3.707  0.00021 ***
## PreferredLoginDevicePhone        -0.37259    0.17771  -2.097  0.03602 *
## CityTier                          0.38026    0.06960   5.463 4.67e-08 ***
## log_WarehouseToHome               0.68448    0.11937   5.734 9.79e-09 ***
## GenderMale                        0.24871    0.12373   2.010  0.04442 *
## NumberOfDeviceRegistered          0.38911    0.06507   5.980 2.23e-09 ***
## PreferedOrderCatLaptop & Accessory -1.78780    0.20311  -8.802  < 2e-16 ***
## PreferedOrderCatMobile           -1.29576    0.40614  -3.190  0.00142 **
## PreferedOrderCatMobile Phone     -0.90278    0.21113  -4.276 1.90e-05 ***
## PreferedOrderCatOthers            1.59205    0.69920   2.277  0.02279 *
## SatisfactionScore                 0.26153    0.04517   5.790 7.04e-09 ***
## MaritalStatusMarried             -0.29425    0.17923  -1.642  0.10064
## MaritalStatusSingle               0.73590    0.17982   4.092 4.27e-05 ***
## log_NumberOfAddress               1.26267    0.11500  10.979  < 2e-16 ***
## Complain                          1.68154    0.12635  13.308  < 2e-16 ***
## log_OrderCount                    0.70290    0.11714   6.000 1.97e-09 ***
## log_DaySinceLastOrder            -0.65750    0.10897  -6.034 1.60e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3405.1  on 3767  degrees of freedom
## Residual deviance: 1882.4  on 3749  degrees of freedom
##   (6 observations deleted due to missingness)
## AIC: 1920.4
##
## Number of Fisher Scoring iterations: 6
```

```r
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
## HourSpendOnApp = 0
## log_CouponUsed = 0
## PreferedOrderCatGrocery = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3755
## 2   3744 11 16.551     0.1219
```

Wald test

Since p > 0.05, we fail to reject the null hypothesis, meaning: These six variables do not contribute significantly to the model jointly. It is safe to remove HourSpendOnApp, OrderAmountHikeFromlastYear,CashbackAmount,PreferredPaymentMode, CouponUsed and PreferedOrderCatGrocery from our model.

Model 7: Remove "HourSpendOnApp", "OrderAmountHikeFromlastYear", "CashbackAmount", "PreferredPaymentMode", "CouponUsed", "PreferedOrderCatGrocery" and "MaritalStatusMarried" variables

```
# Remove "Married" category from MaritalStatus and fit the model again
data$MaritalStatus <- factor(data$MaritalStatus, levels = setdiff(levels(data$MaritalStatus), "Married"

mylogit7 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + Gender + Nu
  family = binomial(link = "logit"))


summary(mylogit7)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
```

```
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -4.89439    0.75192  -6.509 7.56e-11 ***
## log_Tenure                        -1.74908    0.10942 -15.984  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.59996    0.18604  -3.225 0.001260 **
## PreferredLoginDevicePhone         -0.46179    0.24379  -1.894 0.058198 .
## CityTier                           0.32792    0.09119   3.596 0.000323 ***
## log_WarehouseToHome                0.79739    0.16044   4.970 6.69e-07 ***
## GenderMale                         0.21714    0.16538   1.313 0.189182
## NumberOfDeviceRegistered           0.39327    0.08859   4.439 9.04e-06 ***
## PreferedOrderCatLaptop & Accessory -1.67644    0.27755  -6.040 1.54e-09 ***
## PreferedOrderCatMobile             -1.37396    0.52809  -2.602 0.009274 **
## PreferedOrderCatMobile Phone       -0.93292    0.28728  -3.247 0.001164 **
## PreferedOrderCatOthers              2.15199    0.99723   2.158 0.030930 *
## SatisfactionScore                   0.36126    0.06402   5.643 1.67e-08 ***
## MaritalStatusSingle                 0.83072    0.18764   4.427 9.54e-06 ***
## log_NumberOfAddress                 1.05197    0.15390   6.835 8.17e-12 ***
## Complain                            1.85945    0.17426  10.671  < 2e-16 ***
## log_OrderCount                      0.97958    0.16346   5.993 2.06e-09 ***
## log_DaySinceLastOrder              -0.90233    0.15316  -5.892 3.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1931.9  on 1790  degrees of freedom
## Residual deviance: 1037.7  on 1773  degrees of freedom
##   (1983 observations deleted due to missingness)
## AIC: 1073.7
##
## Number of Fisher Scoring iterations: 6
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
## HourSpendOnApp = 0
## log_CouponUsed = 0
## PreferedOrderCatGrocery = 0
## MaritalStatusMarried = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
```

```
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3756
## 2   3744 12 19.167    0.08458 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wald test

Since p > 0.05, we fail to reject the null hypothesis, meaning: These seven variables do not contribute significantly to the model jointly. It is safe to remove HourSpendOnApp, OrderAmountHikeFromlastYear,CashbackAmount,PreferredPaymentMode, CouponUsed, PreferedOrderCatGrocery and MaritalStatusMarried from our model.

Model 8: Remove "HourSpendOnApp", "OrderAmountHikeFromlastYear", "CashbackAmount", "PreferredPaymentMode", "CouponUsed", "PreferedOrderCatGrocery", "MaritalStatusMarried" and "PreferredLoginDevicePhone" variables

```
# Remove "Phone" category from PreferredLoginDevice and fit the model again
data$PreferredLoginDevice <- factor(data$PreferredLoginDevice, levels = setdiff(levels(data$PreferredLog

mylogit8 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + Gender + Nu
  family = binomial(link = "logit"))


summary(mylogit8)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -5.20046    0.84795  -6.133 8.62e-10 ***
## log_Tenure                        -1.72081    0.11994 -14.347  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.57483    0.18750  -3.066 0.002172 **
## CityTier                           0.44867    0.10086   4.448 8.65e-06 ***
## log_WarehouseToHome                0.84275    0.17970   4.690 2.74e-06 ***
## GenderMale                         0.07522    0.18457   0.408 0.683617
## NumberOfDeviceRegistered           0.35347    0.09863   3.584 0.000339 ***
## PreferedOrderCatLaptop & Accessory -1.66792    0.28414  -5.870 4.36e-09 ***
## PreferedOrderCatMobile             0.14697    1.18295   0.124 0.901126
## PreferedOrderCatMobile Phone      -0.82414    0.29743  -2.771 0.005591 **
## PreferedOrderCatOthers             2.19226    0.99832   2.196 0.028095 *
## SatisfactionScore                  0.34524    0.07136   4.838 1.31e-06 ***
## MaritalStatusSingle                0.81816    0.21135   3.871 0.000108 ***
## log_NumberOfAddress                1.11583    0.17885   6.239 4.41e-10 ***
```

```
## Complain                            1.90614    0.19669   9.691  < 2e-16 ***
## log_OrderCount                      0.92139    0.17774   5.184 2.17e-07 ***
## log_DaySinceLastOrder              -0.83346    0.16963  -4.913 8.95e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1516.95  on 1422  degrees of freedom
## Residual deviance:  823.21  on 1406  degrees of freedom
##   (2351 observations deleted due to missingness)
## AIC: 857.21
##
## Number of Fisher Scoring iterations: 6
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
## HourSpendOnApp = 0
## log_CouponUsed = 0
## PreferedOrderCatGrocery = 0
## MaritalStatusMarried = 0
## PreferredLoginDevicePhone = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3757
## 2   3744 13 23.374     0.0374 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wald test

Since $p < 0.05$ (0.0374), we need to reject the null hypothesis, meaning: These eight variables do contribute significantly to the model jointly. It is not safe to remove PreferredLoginDevicePhone from our model.

Model 9: Remove "HourSpendOnApp", "OrderAmountHikeFromlastYear", "CashbackAmount", "PreferredPaymentMode", "CouponUsed", "PreferedOrderCatGrocery", "Mari-

talStatusMarried" and "Gender" variables

```
mylogit9 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + NumberOfDev
  family = binomial(link = "logit"))


summary(mylogit9)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + NumberOfDeviceRegistered + PreferedOrderCat +
##     SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)                       -5.14019    0.83376  -6.165 7.04e-10 ***
## log_Tenure                        -1.71981    0.11976 -14.360  < 2e-16 ***
## PreferredLoginDeviceMobile Phone  -0.57218    0.18748  -3.052 0.002274 **
## CityTier                           0.44956    0.10086   4.457 8.30e-06 ***
## log_WarehouseToHome                0.83694    0.17885   4.680 2.87e-06 ***
## NumberOfDeviceRegistered           0.35636    0.09840   3.622 0.000293 ***
## PreferedOrderCatLaptop & Accessory -1.66625    0.28372  -5.873 4.28e-09 ***
## PreferedOrderCatMobile             0.17419    1.17777   0.148 0.882424
## PreferedOrderCatMobile Phone      -0.82256    0.29709  -2.769 0.005628 **
## PreferedOrderCatOthers             2.18772    0.99336   2.202 0.027641 *
## SatisfactionScore                  0.34594    0.07134   4.849 1.24e-06 ***
## MaritalStatusSingle                0.81462    0.21103   3.860 0.000113 ***
## log_NumberOfAddress                1.11165    0.17840   6.231 4.63e-10 ***
## Complain                           1.90343    0.19657   9.683  < 2e-16 ***
## log_OrderCount                     0.91528    0.17686   5.175 2.28e-07 ***
## log_DaySinceLastOrder             -0.83797    0.16934  -4.948 7.48e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1516.95  on 1422  degrees of freedom
## Residual deviance:  823.37  on 1407  degrees of freedom
##   (2351 observations deleted due to missingness)
## AIC: 855.37
##
## Number of Fisher Scoring iterations: 6
```

```
# Wald test for joint significance
linearHypothesis(mylogit, c("log_CashbackAmount = 0", "log_OrderAmountHikeFromlastYear = 0","PreferredPa
```

```
##
## Linear hypothesis test:
## log_CashbackAmount = 0
## log_OrderAmountHikeFromlastYear = 0
```

```
## PreferredPaymentModeCC = 0
## PreferredPaymentModeCOD = 0
## PreferredPaymentModeCredit Card = 0
## PreferredPaymentModeDebit Card = 0
## PreferredPaymentModeE wallet = 0
## PreferredPaymentModeUPI = 0
## HourSpendOnApp = 0
## log_CouponUsed = 0
## PreferedOrderCatGrocery = 0
## MaritalStatusMarried = 0
## GenderMale = 0
##
## Model 1: restricted model
## Model 2: Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome +
##     PreferredPaymentMode + Gender + HourSpendOnApp + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderAmountHikeFromlastYear + log_CouponUsed +
##     log_OrderCount + log_DaySinceLastOrder + log_CashbackAmount
##
##   Res.Df Df  Chisq Pr(>Chisq)
## 1   3757
## 2   3744 13 23.074    0.04079 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wald test

Since $p < 0.05$ (0.04079), we need to reject the null hypothesis, meaning: These eight variables do contribute significantly to the model jointly. It is not safe to remove Gender from our model.

Model 10: Add interactions between variables

```
mylogit10 <- glm(Churn ~ log_Tenure + PreferredLoginDevice + CityTier + log_WarehouseToHome + NumberOfD
  family = binomial(link = "logit"))
summary(mylogit10)
```

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + NumberOfDeviceRegistered + PreferedOrderCat +
##     SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder + CouponUsed *
##     log_OrderCount + Complain * log_OrderCount + log_Tenure *
##     CityTier, family = binomial(link = "logit"), data = data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -4.73350    0.89857  -5.268 1.38e-07 ***
## log_Tenure                       -2.08504    0.24392  -8.548  < 2e-16 ***
## PreferredLoginDeviceMobile Phone -0.52903    0.18921  -2.796 0.005175 **
## CityTier                          0.22891    0.16900   1.354 0.175577
## log_WarehouseToHome               0.84523    0.18076   4.676 2.92e-06 ***
## NumberOfDeviceRegistered          0.39334    0.10123   3.885 0.000102 ***
```

```
## PreferedOrderCatLaptop & Accessory -1.61552    0.28626  -5.644 1.67e-08 ***
## PreferedOrderCatMobile              0.14779    1.19761   0.123 0.901789
## PreferedOrderCatMobile Phone       -0.79111    0.30544  -2.590 0.009595 **
## PreferedOrderCatOthers              2.29801    1.01756   2.258 0.023923 *
## SatisfactionScore                   0.33523    0.07182   4.668 3.05e-06 ***
## MaritalStatusSingle                 0.78535    0.21224   3.700 0.000215 ***
## log_NumberOfAddress                 1.16911    0.18213   6.419 1.37e-10 ***
## Complain                            2.09657    0.33010   6.351 2.13e-10 ***
## log_OrderCount                      1.08941    0.27275   3.994 6.49e-05 ***
## log_DaySinceLastOrder              -0.84132    0.17069  -4.929 8.26e-07 ***
## CouponUsed                         -0.31460    0.18493  -1.701 0.088920 .
## log_OrderCount:CouponUsed           0.10839    0.07717   1.404 0.160178
## Complain:log_OrderCount            -0.19024    0.31043  -0.613 0.539991
## log_Tenure:CityTier                 0.16490    0.10145   1.625 0.104064
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1516.95  on 1422  degrees of freedom
## Residual deviance:  816.98  on 1403  degrees of freedom
##   (2351 observations deleted due to missingness)
## AIC: 856.98
##
## Number of Fisher Scoring iterations: 6
```

All interactions have p-value > 0.05, they do not contribute statistically and may overcomplicate the model. We should drop interactions.

```
summary(mylogit7)
```

Final model

```
##
## Call:
## glm(formula = Churn ~ log_Tenure + PreferredLoginDevice + CityTier +
##     log_WarehouseToHome + Gender + NumberOfDeviceRegistered +
##     PreferedOrderCat + SatisfactionScore + MaritalStatus + log_NumberOfAddress +
##     Complain + log_OrderCount + log_DaySinceLastOrder, family = binomial(link = "logit"),
##     data = data)
##
## Coefficients:
##                                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)                      -4.89439    0.75192  -6.509 7.56e-11 ***
## log_Tenure                       -1.74908    0.10942 -15.984  < 2e-16 ***
## PreferredLoginDeviceMobile Phone -0.59996    0.18604  -3.225 0.001260 **
## PreferredLoginDevicePhone        -0.46179    0.24379  -1.894 0.058198 .
## CityTier                          0.32792    0.09119   3.596 0.000323 ***
## log_WarehouseToHome               0.79739    0.16044   4.970 6.69e-07 ***
## GenderMale                        0.21714    0.16538   1.313 0.189182
## NumberOfDeviceRegistered          0.39327    0.08859   4.439 9.04e-06 ***
```

```
## PreferedOrderCatLaptop & Accessory -1.67644    0.27755  -6.040 1.54e-09 ***
## PreferedOrderCatMobile             -1.37396    0.52809  -2.602 0.009274 **
## PreferedOrderCatMobile Phone       -0.93292    0.28728  -3.247 0.001164 **
## PreferedOrderCatOthers              2.15199    0.99723   2.158 0.030930 *
## SatisfactionScore                   0.36126    0.06402   5.643 1.67e-08 ***
## MaritalStatusSingle                 0.83072    0.18764   4.427 9.54e-06 ***
## log_NumberOfAddress                 1.05197    0.15390   6.835 8.17e-12 ***
## Complain                            1.85945    0.17426  10.671  < 2e-16 ***
## log_OrderCount                      0.97958    0.16346   5.993 2.06e-09 ***
## log_DaySinceLastOrder              -0.90233    0.15316  -5.892 3.82e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1931.9  on 1790  degrees of freedom
## Residual deviance: 1037.7  on 1773  degrees of freedom
##   (1983 observations deleted due to missingness)
## AIC: 1073.7
##
## Number of Fisher Scoring iterations: 6
```

5.5. Quality Table

```
model_list <- list(
  "Logit (full model)" = mylogit,
  "Probit (full model)" = myprobit,
  "Intermediate model" = mylogit3,
  "Final Model" = mylogit7
)

modelsummary(model_list,
             statistic = "std.error",
             gof_omit = ".*IC|Log.Lik|Deviance",
             stars = TRUE)
```

```
## Warning in n * resp: longer object length is not a multiple of shorter object
## length
## Warning in n * resp: longer object length is not a multiple of shorter object
## length
## Warning in n * resp: longer object length is not a multiple of shorter object
## length
## Warning in n * resp: longer object length is not a multiple of shorter object
## length
```

5.6. Marginal effects

```
marginal_effects <- margins(mylogit7)
summary(marginal_effects)
```

	Logit (full model)	Probit (full model)	Intermediate model	Final Model
(Intercept)	−4.445	−1.933	−4.786***	−4.894***
	(2.895)	(1.370)	(0.592)	(0.752)
log_Tenure	−1.720***	−0.899***	−1.709***	−1.749***
	(0.081)	(0.040)	(0.080)	(0.109)
PreferredLoginDeviceMobile Phone	−0.528***	−0.282***	−0.524***	−0.600**
	(0.142)	(0.076)	(0.140)	(0.186)
PreferredLoginDevicePhone	−0.319+	−0.153	−0.363*	−0.462+
	(0.181)	(0.098)	(0.178)	(0.244)
CityTier	0.289***	0.155***	0.380***	0.328***
	(0.078)	(0.041)	(0.070)	(0.091)
log_WarehouseToHome	0.686***	0.365***	0.668***	0.797***
	(0.122)	(0.065)	(0.120)	(0.160)
PreferredPaymentModeCC	−0.815	−0.706		
	(0.893)	(0.462)		
PreferredPaymentModeCOD	−0.168	−0.328		
	(0.629)	(0.294)		
PreferredPaymentModeCredit Card	−0.703	−0.641*		
	(0.602)	(0.277)		
PreferredPaymentModeDebit Card	−0.526	−0.582*		
	(0.597)	(0.275)		
PreferredPaymentModeE wallet	−0.047	−0.353		
	(0.618)	(0.288)		
PreferredPaymentModeUPI	−0.779	−0.725*		
	(0.631)	(0.297)		
GenderMale	0.259*	0.134*	0.247*	0.217
	(0.125)	(0.067)	(0.124)	(0.165)
HourSpendOnApp	0.092	0.043	0.101	
	(0.100)	(0.053)	(0.097)	
NumberOfDeviceRegistered	0.385***	0.201***	0.385***	0.393***
	(0.068)	(0.036)	(0.067)	(0.089)
PreferedOrderCatGrocery	−12.475	−4.402	−12.603	
	(313.818)	(80.509)	(312.497)	
PreferedOrderCatLaptop & Accessory	−1.774***	−0.986***	−1.798***	−1.676***
	(0.222)	(0.114)	(0.205)	(0.278)
PreferedOrderCatMobile	−1.295*	−0.785**	−1.325**	−1.374**
	(0.518)	(0.279)	(0.407)	(0.528)
PreferedOrderCatMobile Phone	−0.901***	−0.551***	−0.938***	−0.933**
	(0.260)	(0.134)	(0.220)	(0.287)
PreferedOrderCatOthers	1.578*	0.705+	1.601*	2.152*
	(0.719)	(0.390)	(0.696)	(0.997)
SatisfactionScore	0.256***	0.126***	0.258***	0.361***
	(0.046)	(0.024)	(0.045)	(0.064)
MaritalStatusMarried	−0.300+	−0.167+	−0.307+	
	(0.181)	(0.096)	(0.180)	

```
##                               factor     AME     SE        z      p   lower
##                             CityTier  0.0293 0.0080   3.6512 0.0003  0.0136
##                             Complain  0.1663 0.0137  12.1333 0.0000  0.1395
##                           GenderMale  0.0193 0.0147   1.3205 0.1867 -0.0094
##                log_DaySinceLastOrder -0.0807 0.0132  -6.1075 0.0000 -0.1066
##                  log_NumberOfAddress  0.0941 0.0133   7.0950 0.0000  0.0681
##                       log_OrderCount  0.0876 0.0142   6.1809 0.0000  0.0598
##                          log_Tenure -0.1565 0.0066 -23.7781 0.0000 -0.1694
##                  log_WarehouseToHome  0.0713 0.0140   5.1016 0.0000  0.0439
##                  MaritalStatusSingle  0.0724 0.0155   4.6616 0.0000  0.0420
##             NumberOfDeviceRegistered  0.0352 0.0078   4.5077 0.0000  0.0199
##  PreferedOrderCatLaptop & Accessory -0.1602 0.0276  -5.8026 0.0000 -0.2143
##               PreferedOrderCatMobile -0.1358 0.0467  -2.9061 0.0037 -0.2273
##         PreferedOrderCatMobile Phone -0.0966 0.0296  -3.2588 0.0011 -0.1547
##                PreferedOrderCatOthers  0.2728 0.1292   2.1110 0.0348  0.0195
##    PreferredLoginDeviceMobile Phone -0.0552 0.0173  -3.1931 0.0014 -0.0891
##                     SatisfactionScore  0.0323 0.0055   5.8301 0.0000  0.0215
##    upper
##   0.0451
##   0.1932
##   0.0481
##  -0.0548
##   0.1201
##   0.1154
##  -0.1436
##   0.0987
##   0.1029
##   0.0505
##  -0.1061
##  -0.0442
##  -0.0385
##   0.5261
##  -0.0213
##   0.0432
```

Marginal effects in a logistic regression model represent the change in the predicted probability of the outcome (customer churn) for a one-unit change in a predictor variable, holding all other predictors constant. Below is the interpretation of marginal effects for statistically significant variables in the our final model:

Complain: Customers who have lodged a complaint are associated with a 16.63 percentage point increase in the predicted probability of churn, holding all other factors constant.

log_Tenure: A one-unit increase in the logarithm of tenure is associated with a 15.65 percentage point decrease in the probability of churn, all else being equal.

log_DaySinceLastOrder: An increase in the time since the last order (log scale) decreases the likelihood of churn by 8.07 percentage points, ceteris paribus.

log_NumberOfAddress: Customers with more recorded addresses (log scale) are 9.41 percentage points more likely to churn.

log_OrderCount: Each unit increase in the log of order count increases the predicted probability of churn by 8.76 percentage points, holding other variables constant.

log_WarehouseToHome: Greater delivery distance (log-transformed) increases the likelihood of churn by 7.13 percentage points.

MaritalStatusSingle: Being single is associated with a 7.24 percentage point increase in the probability of churn, all else equal.

NumberOfDeviceRegistered: Each additional registered device is associated with a 3.52 percentage point increase in churn probability.

CityTier: Customers living in higher-tier cities have a 2.93 percentage point higher probability of churning.

PreferredLoginDeviceMobile Phone: Using a mobile phone to log in reduces the predicted probability of churn by 5.52 percentage points.

PreferedOrderCatLaptop & Accessory: Preference for laptops and accessories is associated with a 16.02 percentage point decrease in the probability of churn.

PreferedOrderCatMobile: Preference for mobile products reduces churn probability by 13.58 percentage points, all else constant.

PreferedOrderCatMobile Phone: Preference for mobile phones is linked to a 9.66 percentage point decrease in predicted churn.

PreferedOrderCatOthers: Customers preferring "Other" categories are 27.28 percentage points more likely to churn — the largest positive marginal effect observed.

SatisfactionScore: A one-point increase in satisfaction score is surprisingly associated with a 3.23 percentage point increase in churn probability, which may indicate a complex relationship requiring further investigation.

5.7. Odds Ratios

In logistic regression, the **odds ratio (OR)** for a predictor indicates how the **odds of the outcome** (here, customer churn) change with a one-unit increase in that predictor, holding all other variables constant. Odds ratios are calculated by exponentiating the model coefficients:

```
# Calculate odds ratios and 95% CI
odds_ratios <- exp(coef(mylogit7))
conf_int <- exp(confint(mylogit7))
```

```
## Waiting for profiling to be done...
```

```
odds_table <- data.frame(
  Variable = names(odds_ratios),
  OR = odds_ratios,
  CI_lower = conf_int[, 1],
  CI_upper = conf_int[, 2]
)
print(odds_table)
```

```
##                                                                 Variable
## (Intercept)                                                   (Intercept)
## log_Tenure                                                     log_Tenure
## PreferredLoginDeviceMobile Phone     PreferredLoginDeviceMobile Phone
## PreferredLoginDevicePhone                   PreferredLoginDevicePhone
## CityTier                                                         CityTier
## log_WarehouseToHome                                 log_WarehouseToHome
## GenderMale                                                     GenderMale
## NumberOfDeviceRegistered                     NumberOfDeviceRegistered
## PreferedOrderCatLaptop & Accessory PreferedOrderCatLaptop & Accessory
```

```
## PreferedOrderCatMobile                          PreferedOrderCatMobile
## PreferedOrderCatMobile Phone                PreferedOrderCatMobile Phone
## PreferedOrderCatOthers                          PreferedOrderCatOthers
## SatisfactionScore                                    SatisfactionScore
## MaritalStatusSingle                                MaritalStatusSingle
## log_NumberOfAddress                                log_NumberOfAddress
## Complain                                                      Complain
## log_OrderCount                                          log_OrderCount
## log_DaySinceLastOrder                                log_DaySinceLastOrder
##                                              OR    CI_lower    CI_upper
## (Intercept)                          0.007488452 0.001675966  0.03206093
## log_Tenure                           0.173934743 0.139392324  0.21416163
## PreferredLoginDeviceMobile Phone     0.548832370 0.380695912  0.79000051
## PreferredLoginDevicePhone            0.630151651 0.389494589  1.01387984
## CityTier                             1.388077986 1.161840165  1.66172336
## log_WarehouseToHome                  2.219738593 1.624927122  3.04960464
## GenderMale                           1.242522562 0.899570839  1.72131318
## NumberOfDeviceRegistered             1.481821339 1.247657410  1.76634279
## PreferedOrderCatLaptop & Accessory   0.187037924 0.108360235  0.32232693
## PreferedOrderCatMobile               0.253102033 0.089111281  0.70914091
## PreferedOrderCatMobile Phone         0.393402273 0.223936558  0.69173118
## PreferedOrderCatOthers               8.601942976 0.989711447 58.48346236
## SatisfactionScore                    1.435131635 1.267829844  1.62992204
## MaritalStatusSingle                  2.294966458 1.596385949  3.33378834
## log_NumberOfAddress                  2.863279426 2.126373623  3.88946917
## Complain                             6.420232972 4.584244317  9.08348686
## log_OrderCount                       2.663332340 1.938577760  3.68191007
## log_DaySinceLastOrder                0.405624057 0.299490255  0.54622137
```

Interpretation of selected odds ratios:

- **Complain**: Customers who filed a complaint are **6.42 times more likely** to churn compared to those who didn't.
- **log_Tenure**: A one-unit increase in tenure (log-transformed) is associated with an **82.6% reduction** in the odds of churn.
- **log_OrderCount**: More frequent ordering (log-transformed) increases the odds of churn by **166%**.
- **log_NumberOfAddress**: Customers with more delivery addresses are **2.86 times more likely** to churn.
- **MaritalStatusSingle**: Single customers are **2.29 times more likely** to churn than married ones.
- **SatisfactionScore**: Each additional point in satisfaction increases churn odds by **43.5%**, suggesting potential non-linearity or dissatisfaction despite higher scores.
- **PreferedOrderCatLaptop & Accessory**: Preference for this category **reduces churn odds by 81.3%**.
- **log_DaySinceLastOrder**: Longer time since the last order **reduces churn odds by about 59.4%**, indicating re-engagement behavior.

5.8. Diagnostics

The link test is a diagnostic tool used to assess the specification of a logistic regression model. It helps determine whether the model is correctly specified or if key predictors may have been omitted or if non-linearities remain unaddressed.

The logic behind the link test is that if a model is properly specified, adding the predicted value (_hat) should be statistically significant (as it captures the systematic part of the variation), while the square of the predicted value (_hatsq) should not be significant (as it would otherwise suggest a mis-specification).

```
# Update final data
data_clean <- subset(data,
               PreferedOrderCat != "Grocery" &
               MaritalStatus != "Married" )

data$PreferedOrderCat <- droplevels(data$PreferedOrderCat)
data$MaritalStatus <- droplevels(data$MaritalStatus)

dim(data)
```

```
## [1] 3774   27
```

```
#Link test
data_clean$hat <- fitted(mylogit7)
data_clean$hat_sq <- data_clean$hat^2

link_test_model <- glm(Churn ~ hat + hat_sq, family = binomial, data = data_clean)
summary(link_test_model)
```

```
##
## Call:
## glm(formula = Churn ~ hat + hat_sq, family = binomial, data = data_clean)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -3.2585     0.1639 -19.885  < 2e-16 ***
## hat           5.6644     0.9879   5.734  9.8e-09 ***
## hat_sq        1.2489     1.1973   1.043    0.297
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1931.9  on 1790  degrees of freedom
## Residual deviance: 1024.5  on 1788  degrees of freedom
## AIC: 1030.5
##
## Number of Fisher Scoring iterations: 5
```

The coefficient for hat is statistically significant ($p < 0.001$), while the coefficient for hat_sq is not statistically significant ($p = 0.297$). This result indicates that the model is correctly specified and there is no evidence of omitted variables or incorrect functional form. Thus, the model passes the link test.

```
# R-squared statistics
PseudoR2(myprobit)
```

```
##         McFadden     Adj.McFadden        Cox.Snell       Nagelkerke
##        0.4398153        0.4216189        0.3277176        0.5511770
```

```
## McKelvey.Zavoina           Effron            Count        Adj.Count
##        0.6228493        0.4808940        0.9067303        0.4421553
##              AIC    Corrected.AIC
##     1968.6984204     1969.1953480
```

The logistic regression model's performance was evaluated using a variety of pseudo R sq and related fit measures:

- **McKelvey & Zavoina R_sq (0.623)** is regarded as the best approximation of the traditional R^2 in binary outcome models, showing that 62.3% of the variance in the underlying latent variable is explained.
- **Count R_sq (0.907)** indicates that 90.7% of observations were correctly classified. While impressive, this metric can be inflated in imbalanced datasets.
- **Adjusted Count R_sq (0.442)** corrects for potential baseline bias in classification accuracy, and still reflects solid predictive power.

Overall, these fit statistics confirm that the final model provides a strong and robust explanation for customer churn behavior.

```r
# Hosmer-Lemeshow Test
hl_test <- hoslem.test(x = mylogit7$y, y = fitted(mylogit7), g = 10)
print(hl_test)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  mylogit7$y, fitted(mylogit7)
## X-squared = 45.704, df = 8, p-value = 2.706e-07
```

```r
pred_class <- ifelse(fitted(mylogit7) > 0.5, 1, 0)
confusionMatrix(as.factor(pred_class), as.factor(mylogit7$y))
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction    0    1
##          0 1308  128
##          1   71  284
##
##                Accuracy : 0.8889
##                  95% CI : (0.8734, 0.9031)
##     No Information Rate : 0.77
##     P-Value [Acc > NIR] : < 2.2e-16
##
##                   Kappa : 0.6704
##
##  Mcnemar's Test P-Value : 7.195e-05
##
##             Sensitivity : 0.9485
##             Specificity : 0.6893
##          Pos Pred Value : 0.9109
```

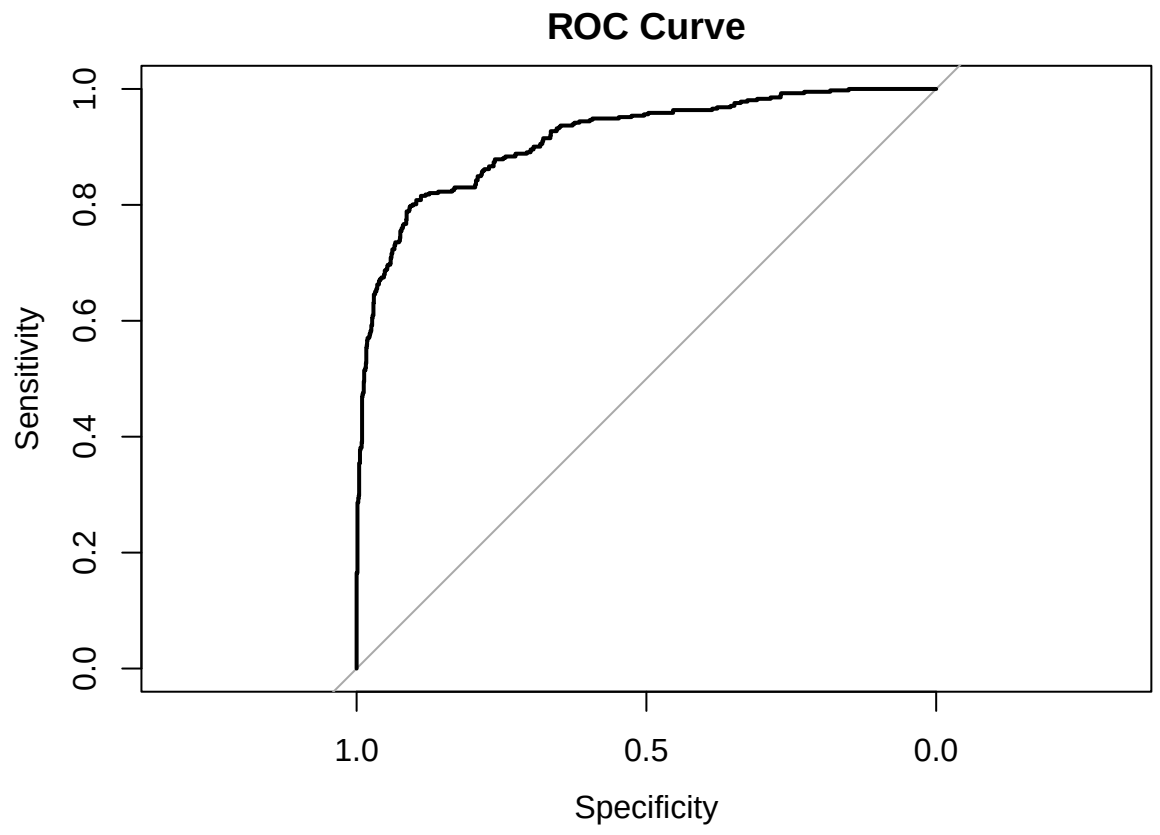
```
##          Neg Pred Value : 0.8000
##              Prevalence : 0.7700
##          Detection Rate : 0.7303
##    Detection Prevalence : 0.8018
##       Balanced Accuracy : 0.8189
##
##        'Positive' Class : 0
##
```

```
#ROC Curve and AUC
roc_curve <- roc(mylogit7$y, fitted(mylogit7))
```

```
## Setting levels: control = 0, case = 1
```

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

```
plot(roc_curve, main = "ROC Curve")
```




```
auc(roc_curve)
```

```
## Area under the curve: 0.9147
```

The logistic regression model demonstrates strong classification performance, as shown by:

- High accuracy (88.9%)
- High sensitivity (94.85%)
- A well-performing ROC curve (AUC 0.9)
- Substantial Kappa statistic (0.67)

Despite a significant Hosmer–Lemeshow test indicating some model misspecification, the model performs well in practical terms, especially in minimizing false negatives for non-churners.However, the relatively lower specificity (68.93%) means that some churners are still being misclassified. Future improvements could include exploring interaction terms, non-linear effects, or machine learning methods to capture more complex relationships in the data.

6. Hypotheses

Based on the insights from prior research and exploratory analysis, we formulate the following primary and secondary hypotheses to test the determinants of customer churn in e-commerce.

Hypothesis 1: Complaints and Churn

H0: Filing a complaint does not affect the likelihood of customer churn.
H1: Customers who lodge complaints are more likely to churn.

The marginal effects indicate that complaints are associated with a **16.63 percentage point increase** in the probability of churn, holding all other variables constant. The p-value associated with the complaint variable is statistically significant at the 1% level. Thus, we **reject the null hypothesis**, suggesting that customer complaints are a strong predictor of churn.

Hypothesis 2: Customer Tenure and Churn

H0: Customer tenure has no impact on the likelihood of churn.
H1: Longer tenure is associated with a lower probability of churn.

The variable *log_Tenure* is statistically significant and is associated with a **15.65 percentage point decrease** in churn probability. Therefore, we **reject the null hypothesis** and conclude that longer-tenured customers are less likely to churn.

Hypothesis 3: Order Activity and Churn

H0: Order frequency does not influence customer churn.
H1: Higher order frequency reduces the probability of churn.

log_OrderCount shows a significant **8.76 percentage point increase** in churn probability for higher order frequency. Interestingly, rather than decreasing churn, more frequent orders are positively associated with churn in our dataset. We **reject the null hypothesis**, but the direction of the relationship suggests complex customer dynamics possibly related to dissatisfaction despite higher purchase activity.

Hypothesis 4: Distance to Delivery and Churn

H0: Delivery distance has no effect on churn likelihood.
H1: Longer delivery distances increase the probability of churn.

The *log_WarehouseToHome* variable is positively associated with churn (7.13 percentage point increase), and is statistically significant. Hence, we **reject the null hypothesis** and conclude that delivery logistics impact customer retention.

Hypothesis 5: Product Category Preference and Churn

H0: Product category preference does not influence churn.
H1: Certain product category preferences are associated with lower churn probability.

Customers who preferred *Laptop & Accessory*, *Mobile*, or *Mobile Phone* categories showed **lower predicted churn probabilities** (16.02, 13.58, and 9.66 percentage point decreases respectively). In contrast, those preferring "Others" showed the **highest increase** (27.28 percentage points). These findings are statistically significant, leading us to **reject the null hypothesis**.

Hypothesis 6: Marital Status and Churn

H0: Marital status does not affect churn.
H1: Single customers are more likely to churn.

Single customers are associated with a **7.24 percentage point increase** in churn probability, and the result is statistically significant. Therefore, we **reject the null hypothesis** and confirm marital status as a relevant demographic factor.

7. Findings and conclusion

This study set out to identify the key drivers of customer churn in the e-commerce sector using a binary logistic regression model. By analyzing a comprehensive set of variables, including behavioral indicators, demographic information, and customer preferences, the research revealed several statistically significant predictors of churn. Customers who submitted complaints were substantially more likely to churn, while those with longer tenure and more recent purchase activity were less likely to do so. Other important predictors included the number of devices registered, marital status, delivery distance, and product category preferences. Some product categories, such as laptops or mobile phones, were associated with reduced churn, whereas others, particularly "Other" categories, had a strong positive relationship with churn risk.

The marginal effects analysis provided further insights into the magnitude of these relationships. For instance, a complaint was associated with a 16.63 percentage point increase in churn probability, while each unit increase in log tenure decreased the probability of churn by 15.65 percentage points. Interestingly, the satisfaction score, which would typically be expected to lower churn, showed a small but significant positive association with churn. This finding may reflect unobserved factors such as inflated satisfaction ratings or unmet expectations despite high scores, indicating the need for deeper qualitative assessment in future research.

From a model performance perspective, the logistic regression demonstrated robust predictive ability. The overall classification accuracy was 88.89%, with high sensitivity (94.85%) and a well-balanced specificity (68.93%). The ROC curve showed a strong area under the curve (AUC), indicating high discriminative power. Pseudo-R sq statistics, including McFadden's R sq (0.44) and McKelvey-Zavoina R sq (0.62), confirmed that the model captured a substantial portion of variance in customer churn behavior. Although the Hosmer-Lemeshow goodness-of-fit test yielded a significant p-value, this is a common occurrence in large samples and does not necessarily undermine the model's overall validity.

In conclusion, this research confirms that customer churn is a multifactorial outcome influenced by a range of behavioral, demographic, and experiential factors. The findings align well with existing literature and highlight the practical importance of monitoring complaints, purchase recency, and order behavior in predicting churn. These insights can support the development of targeted retention strategies and customer relationship management practices. Future work could enhance prediction accuracy through non-linear models, machine learning algorithms, or deeper exploration of the satisfaction-churn paradox.

8. Bibliography

1. Ahmad, A., Jafar, A., & Aljoumaa, K. (2019). Customer churn prediction in telecom using machine learning in big data platform. Journal of Big Data, 6(1), 1–24.

2. Berger, P., & Kompan, M. (2019). Predicting customer churn in e-commerce using behavior-based models. International Journal of Information Management, 47, 150–162.

3. Bhattacharya, S. (2021). Predicting e-commerce customer churn using transactional data. Electronic Commerce Research and Applications, 45, 101024.

4. Dahiya, M., & Bhatia, M. P. S. (2020). Predictive analytics for customer churn using machine learning techniques. Procedia Computer Science, 167, 2319–2328.

5. Jaiswal, A. K., & Niraj, R. (2011). Examining mediating role of attitudinal loyalty and satisfaction on customer behavior. Journal of Services Marketing, 25(3), 165–175.

6. Li, M. (2022). Customer churn prediction on e-commerce platform using Random Forest. International Journal of Business Analytics, 9(4), 45–57.

7. Liu, Q., & Wang, Y. (2010). Predicting customer churn in the telecommunications industry–An application of survival analysis modeling using SAS. SAS Global Forum.