Assignment 02 Solution

SQB7005 Statistical Laboratory

Jiali Tian (S2126002)

2024-07-06

```r
# Importing packages
library(tidyverse)
library(dplyr)
library(plyr)
library(ggplot2)
library(cowplot)
library(MASS)
library(gridExtra)
library(knitr)
```

1. Introduction

(1) Research Topic

Telco data exploratory analysis and predictive analysis of customer churn.

(2) Background

The Telco customer churn data contains information about a fictional telco company that provided home phone and Internet services to 7043 customers in California in Q3. It indicates which customers have left, stayed, or signed up for their service. Multiple important demographics are included for each customer, as well as a Satisfaction Score, Churn Score, and Customer Lifetime Value (CLTV) index.

(3) Data Sources and Information

Data Sources: The data objects and sources studied in this article are as follows:

Should give different name

Dataset Name	URL
heart failure	https://www.kaggle.com/code/farazrahman/telco-customer-churn-logisticregression

Basic Information:

a. Customers who left within the last month – the column is called Churn

b. Services that each customer has signed up for – phone, multiple lines, internet, online security, online backup, device protection, tech support, and streaming TV and movies

c. Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges

d. Demographic info about customers – gender, age range, and if they have partners and dependents

The meaning of each variable is explained in the following table:

Variable Name	Type	Description
customerID	chr	Customer ID
gender	chr	Whether the customer is a male or a female
SeniorCitizen	int	Whether the customer is a senior citizen or not (1, 0)
Partner	chr	Whether the customer has a partner or not (Yes, No)
Dependents	chr	Whether the customer has dependents or not (Yes, No)
tenure	int	Number of months the customer has stayed with the company
PhoneService	chr	Whether the customer has a phone service or not (Yes, No)
MultipleLines	chr	Whether the customer has multiple lines or not (Yes, No, No phone service)
InternetService	chr	Customer's internet service provider (DSL, Fiber optic, No)
OnlineSecurity	chr	Whether the customer has online security or not (Yes, No, No internet service)
OnlineBackup	chr	Whether the customer has online backup or not (Yes, No, No internet service)
DeviceProtection	chr	Whether the customer has device protection or not (Yes, No, No internet service)
TechSupport	chr	Whether the customer has tech support or not (Yes, No, No internet service)
StreamingTV	chr	Whether the customer has streaming TV or not (Yes, No, No internet service)
StreamingMovies	chr	Whether the customer has streaming movies or not (Yes, No, No internet service)
Contract	chr	The contract term of the customer (Month-to-month, One year, Two year)
PaperlessBilling	chr	Whether the customer has paperless billing or not (Yes, No)
PaymentMethod	chr	The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic))
MonthlyCharges	num	The amount charged to the customer monthly
TotalCharges	num	The total amount charged to the customer
Churn	chr	Whether the customer churned or not (Yes or No)

Content: Each row represents a customer, each column contains customer's attributes described on the column Metadata.

```r
# Load data
file_path = "WA_Fn-UseC_-Telco-Customer-Churn.csv"
Telco = read.csv(file_path, header = TRUE)
# Display first five rows
head(Telco, 5)
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
## 5 9237-HQITU Female             0      No         No      2          Yes
##      MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1 No phone service             DSL             No          Yes               No
## 2               No             DSL            Yes           No              Yes
## 3               No             DSL            Yes          Yes               No
## 4 No phone service             DSL            Yes           No              Yes
## 5               No     Fiber optic             No           No                No
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
## 5          No          No              No Month-to-month              Yes
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No
## 5          Electronic check          70.70       151.65   Yes
```

```
# Check the type of variables
str(Telco)
```

```
## ’data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : chr  "Female" "Male" "Male" "Male" ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : chr  "Yes" "No" "No" "No" ...
##  $ Dependents      : chr  "No" "No" "No" "No" ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : chr  "No" "Yes" "Yes" "No" ...
##  $ MultipleLines   : chr  "No phone service" "No" "No" "No phone service" ...
##  $ InternetService : chr  "DSL" "DSL" "DSL" "DSL" ...
##  $ OnlineSecurity  : chr  "No" "Yes" "Yes" "Yes" ...
##  $ OnlineBackup    : chr  "Yes" "No" "Yes" "No" ...
##  $ DeviceProtection: chr  "No" "Yes" "No" "Yes" ...
##  $ TechSupport     : chr  "No" "No" "No" "Yes" ...
##  $ StreamingTV     : chr  "No" "No" "No" "No" ...
##  $ StreamingMovies : chr  "No" "No" "No" "No" ...
##  $ Contract        : chr  "Month-to-month" "One year" "Month-to-month" "One year" ...
##  $ PaperlessBilling: chr  "Yes" "No" "Yes" "No" ...
##  $ PaymentMethod   : chr  "Electronic check" "Mailed check" "Mailed check" "Bank transfer (automatic
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : chr  "No" "No" "Yes" "No" ...
```

(4) The Purpose of Analysis

The main purpose of this analysis is to extract valuable insights from the Telco dataset. It aims to achieve the following key objectives:

Exploratory Analysis: Use Exploratory Data Analysis (EDA) techniques to reveal those variables that have a large impact. Analyze the potential patterns in the data, find any outliers, identify missing values, and reveal potential trends in the dataset.

Correlation and Impact Assessment: Find the correlation between user churn and various behaviors. Relationships and potential correlations between behavioral characteristics to reveal interdependencies. Understand the reasons for customer churn by considering different service methods and fees charged separately.

Prediction Model: Based on the information provided, develop a logistic regression model to assess and predict whether a customer will churn.

2. Problem Statement

Problem Statement

In the Internet age, telecommunication network services are an essential part of life. Telecommunications business customers are very picky about the type of telecommunications services they receive and judge the entire company based on one experience. As telecommunications services continue to evolve, there are generally several different telecommunications companies for customers to choose from. Once a customer is lost, it will take more time to get him back. So customer churn analysis becomes very critical! The higher the customer churn rate, the more customers stop buying from your business, which directly affects revenue! Therefore, based on the insights gained from customer churn analysis, companies can develop strategies, target market segments, and improve the quality of services provided to improve customer experience, thereby cultivating trust with customers.

Questions

In order to arouse the reader's interest and make the framework of the tedious research content below clearer. Based on the content of this data, several research questions were designed and raised as follows:

Question 1: Analyze the proportion of people who choose different types of telecommunications services. Is there any connection between these services?

Question 2: As the customers subscribe service period extends, will the number of customers churn increase?

Question 3: Will the increase in value-added telecommunications services lead to customer churn?

Question 4: Is customer churn related to the per months fees they pay?

Question 5: Fit a logistic regression model to find and explain the effects of significant variables.

Note: These questions correspond to the Exploratory Analysis and Visualization section later.

3. Results and Discussion

(1) Data Preparation and Cleaning

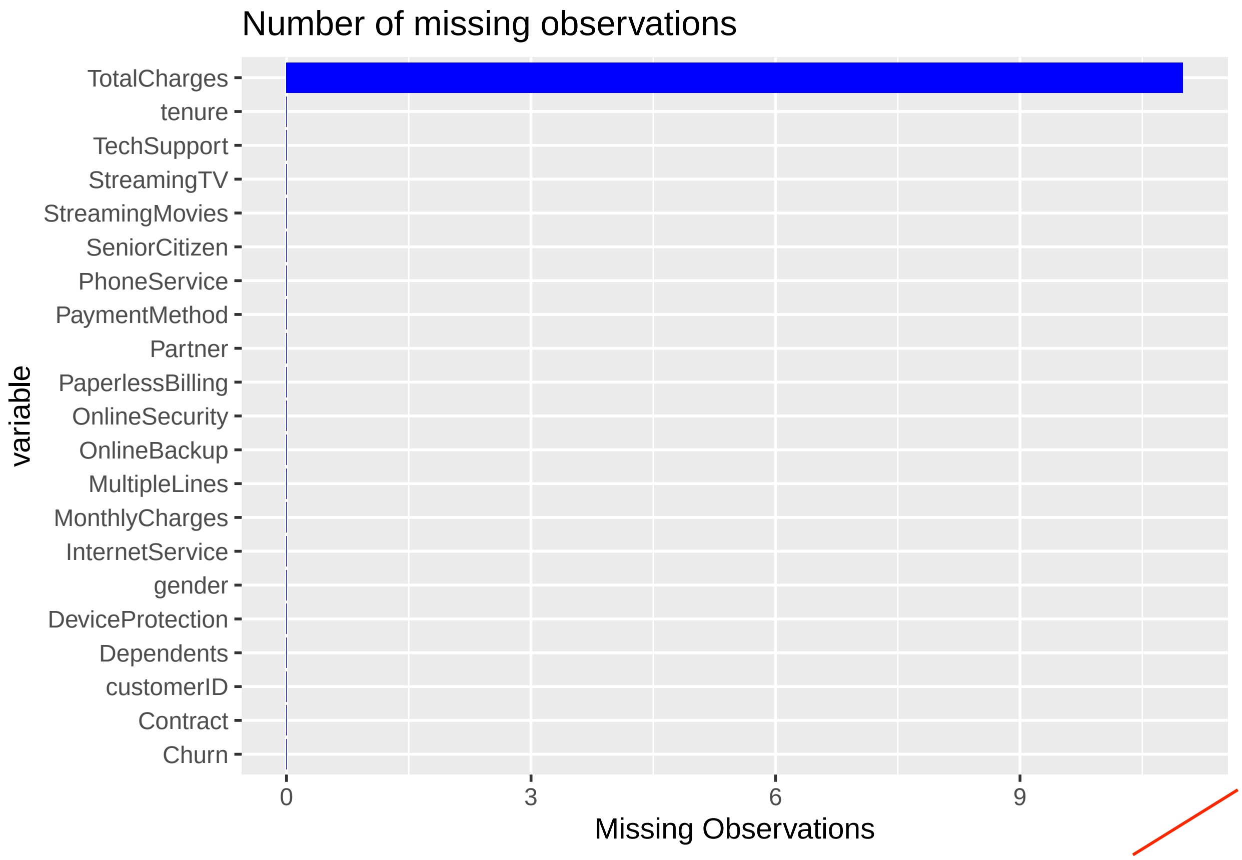
The purpose of checking and cleaning data is to allow the research to proceed correctly and smoothly. By inspecting the data set, it was confirmed that the data content and sample size were correct.

i) Check missing values

```r
# Check for missing values in the entire dataset
any(is.na(Telco))
```

```
## [1] TRUE
```

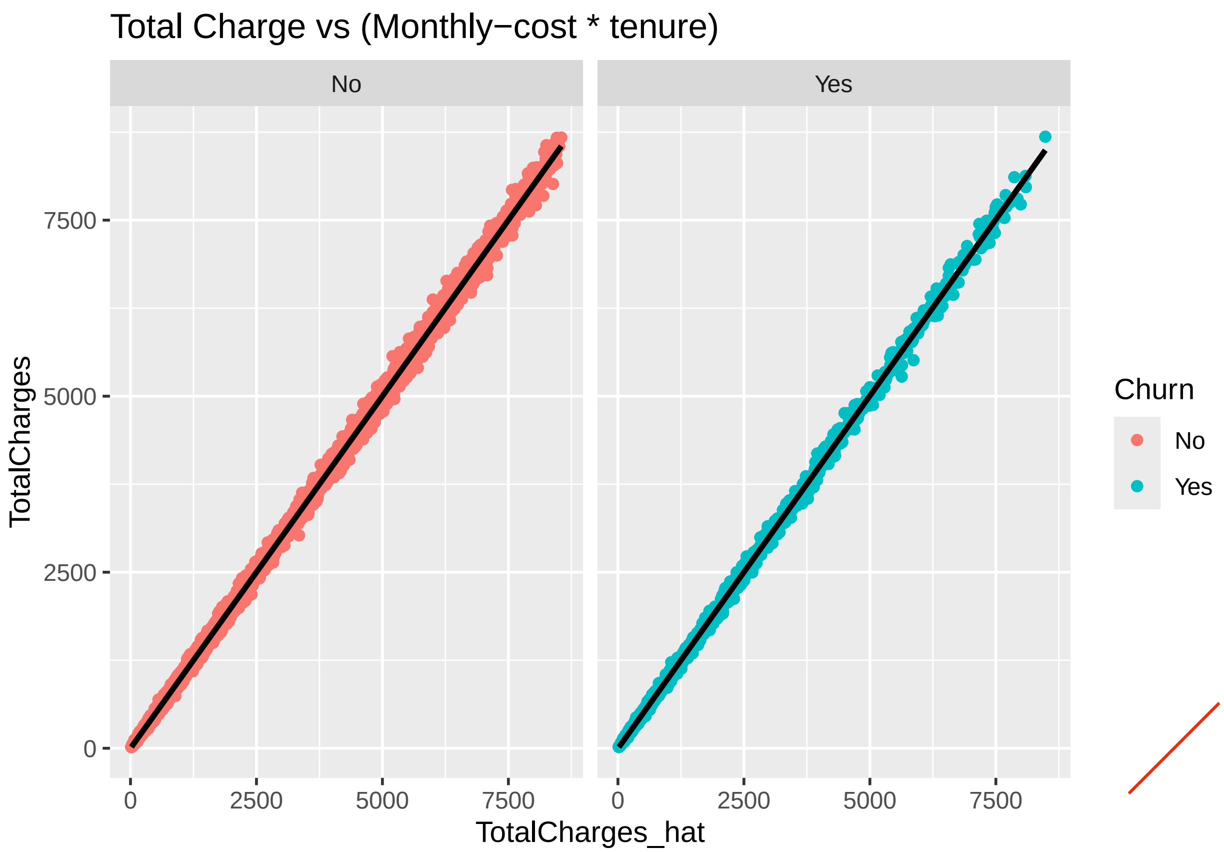
```r
# Missing variables
Telco %>%
  map_df(function(x) sum(is.na(x))) %>%
  gather(variable, value, customerID:Churn) %>%
  arrange(desc(value)) %>%
  ggplot(aes(x = variable, y = value)) +
  geom_col(fill = "blue") +
  coord_flip() +
  labs(y = "Missing Observations") +
  ggtitle("Number of missing observations")
```



There are some missing values (NA) in the TotalCharges variable.

```
# Correlation between Total Charges and Total Charges hat
Telco %>%
  dplyr::select(tenure, MonthlyCharges, TotalCharges, Churn) %>%
  mutate(TotalCharges_hat = MonthlyCharges * tenure) %>%
  filter(!is.na(TotalCharges)) %>%  # Correct filtering of missing values
  ggplot(aes(x = TotalCharges_hat, y = TotalCharges, color = Churn)) +
  geom_point() +
  facet_grid(. ~ Churn) +
  geom_smooth(method = "lm", colour = "black") +
  ggtitle("Total Charge vs (Monthly-cost * tenure)")
```

```
## ‘geom_smooth()‘ using formula = ’y ~ x’
```



```
# Replace missing values with product of MonthlyCharges times tenure
Telco <- Telco %>%
  mutate(TotalCharges = ifelse(is.na(TotalCharges), MonthlyCharges * tenure, TotalCharges))
# Check for missing values in the entire dataset
any(is.na(Telco))
```

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

It can be seen that there is a strong linear relationship between "TotalCharges" and "MonthlyCharges * tenure", so interpolation is used to fill in the missing values.

ii) Check duplicates data

```
# Check for duplicate rows by using duplicated() function
# Extract not unique elements
duplicate_rows <- Telco[duplicated(Telco), ]
# Display duplicate rows (if any)
if (nrow(duplicate_rows) > 0) {
  print("Duplicate Rows:")
  print(duplicate_rows)
} else {
  print("No duplicate rows found.")
}
```

```
## [1] "No duplicate rows found."
```

iii) Changed type of data

There are three continuous variables and they are Tenure, MonthlyCharges and TotalCharges. SeniorCitizen is in 'int' form, that can be changed to categorical.

```
Telco <- Telco[complete.cases(Telco),]
Telco$SeniorCitizen <- as.factor(ifelse(Telco$SeniorCitizen==1, 'YES', 'NO'))
```

(2) Exploratory Data Analysis and Visualization

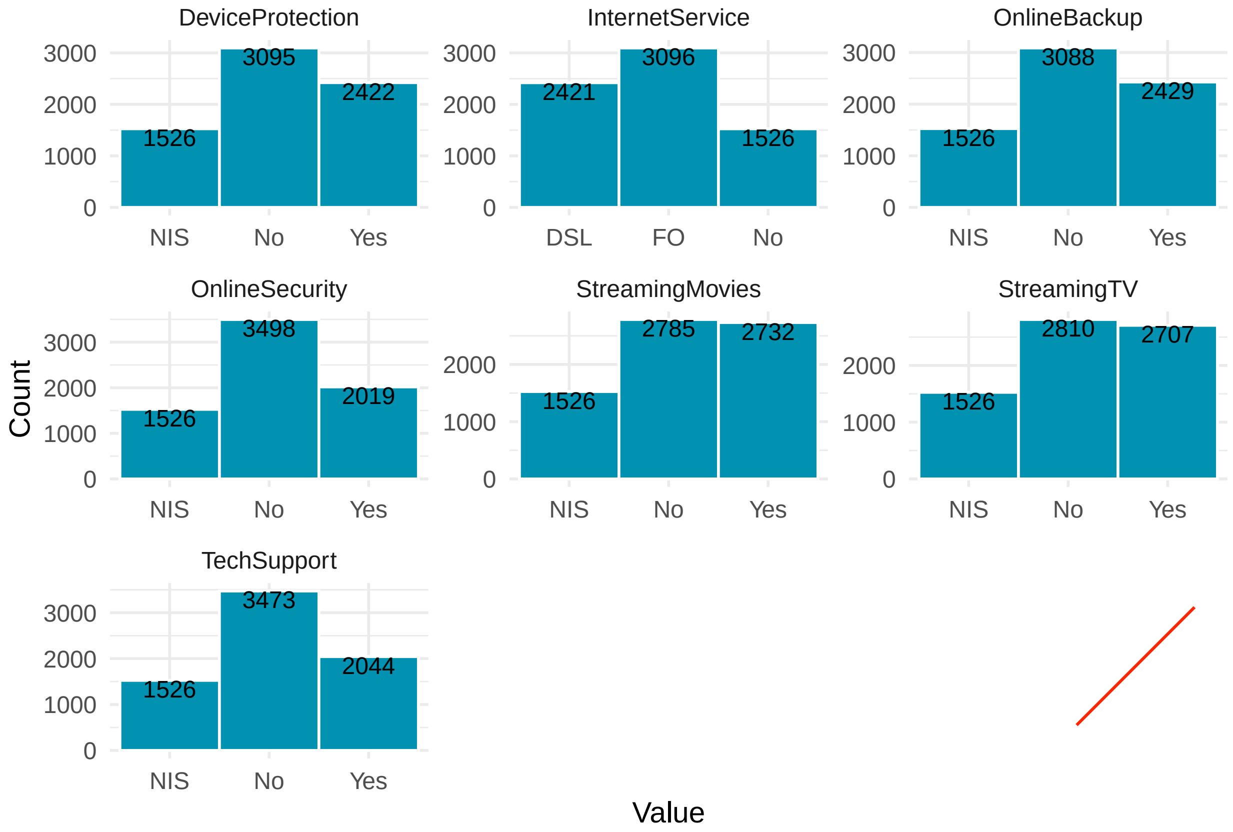
i) Data distribution

For the presentation of categorical data, a part of the histogram and part of the pie chart are used, where the histogram gives the specific quantity and the pie chart uses the proportion.

```
# Prepare data for plotting histograms of categorical variables
category = c("OnlineSecurity", "OnlineBackup", "DeviceProtection",
             "StreamingMovies", "TechSupport", "StreamingTV" , "InternetService")
# Select only the categorical columns
Telco_cat <- Telco[, category]
# Convert wide format to long format
Telco_long <- Telco_cat %>% gather(key = "variable", value = "value")
```

```
## Warning: attributes are not identical across measure variables; they will be
## dropped
```

```
# Plot histograms for each categorical variable
hist_plot <- ggplot(Telco_long, aes(x = value)) +
  geom_bar(color = "white", fill = "#0192B1", width = 1) +
  geom_text(
    stat = "count", size = 3, vjust = 1, col = "black",
    aes(label = after_stat(count)),
    position = position_dodge(1)
  ) +
  labs(x = "Value", y = "Count") +
  theme_minimal() +
  facet_wrap(~ variable, scales = "free", ncol = 3)
# Display the plot
print(hist_plot)
```



For the histogram above:

a. The number of people who chose "No internet service" was all 1526.

b. The ratios of "StreamingMovies" and "StreamingTV" are almost the same.

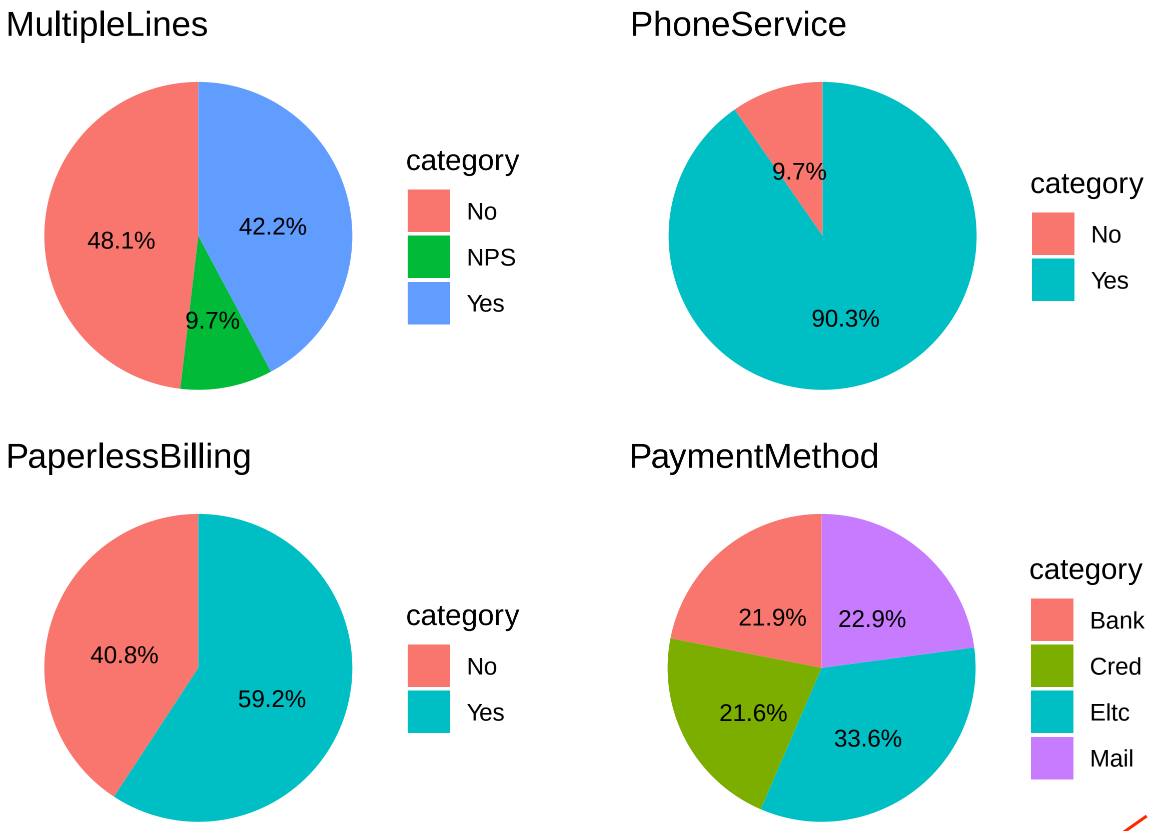
c. The ratios of "InternetService", "DeviceProtection" and "OnlineBackup" are almost the same.

d. The ratios of "TechSupport" and "OnlineSecurity" are almost the same.

For b,c,d. These two services can be classified into the same category. It shows that there is a connection between the two.

What is two services? Need to be clearer

```
# Function to create a pie chart with proportions
create_pie_chart <- function(data, column) {
  df <- as.data.frame(table(data[[column]]))
  colnames(df) <- c("category", "count")
  df$proportion <- df$count / sum(df$count) * 100
  ggplot(df, aes(x = "", y = count, fill = category)) +
    geom_bar(width = 1, stat = "identity") +
    coord_polar(theta = "y", start = 0) +
    theme_void() +
    geom_text(
      size = 3, vjust = 1, col = "black",
      aes(label = paste0(round(proportion, 1), "%")),
      position = position_stack(vjust = 0.5)
      ) +
    ggtitle(paste(column)) +
    theme(legend.title = element_blank()) +
    theme_void()
}
# Create pie charts
p2 <- create_pie_chart(Telco, "PaperlessBilling")
p3 <- create_pie_chart(Telco, "MultipleLines")
p4 <- create_pie_chart(Telco, "PhoneService")
p5 <- create_pie_chart(Telco, "PaymentMethod")
# Combine all pie charts into a single plot
grid.arrange(p3, p4, p2, p5, nrow = 2)
```



For the pie chart above:

a: For pie chart “MultipleLines” and “PhoneService”:

The proportion of “No phone service” in ‘MultipleLines’ is equal to “No” in ‘PhoneService’, that is 9.7%. And 48.1% + 42.2% = 90.3%. This pie chart indicates that the variable ‘PhoneService’ is fully subsumed within the ‘MultipleLines’ variables.

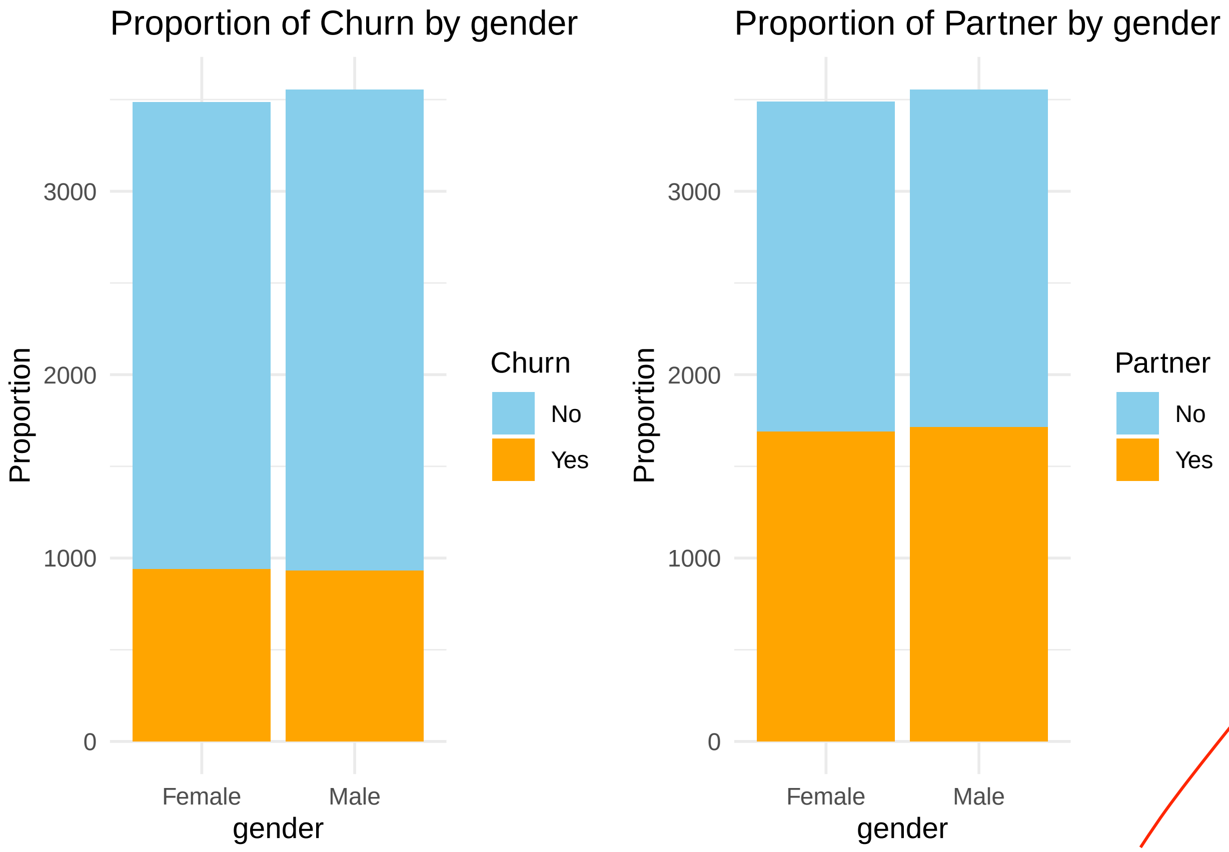
‘No’ can be interpreted as a single phone line,

‘Yes’ can be interpreted as a multiple phone lines,

‘No phone service’ can be interpreted as no phone line.

b: For pie chart “PaperlessBilling” and “PaymentMethod”: we can know the proportion from “PaperlessBilling” and “PaymentMethod”.

```r
# create a stacked bar plot to show the proportion of Churn within each gender
s1 = ggplot(Telco, aes(gender, fill = Churn)) +
  geom_bar() +
  labs(title = "Proportion of Churn by gender", y = "Proportion") +
  theme_minimal()  +
  scale_fill_manual(values = c("No" = "skyblue", "Yes" = "orange"))
s2 = ggplot(Telco, aes(gender, fill = Partner)) +
  geom_bar() +
  labs(title = "Proportion of Partner by gender", y = "Proportion") +
  theme_minimal()  +
  scale_fill_manual(values = c("No" = "skyblue", "Yes" = "orange"))
# Combine pie charts into a single plot
grid.arrange(s1, s2, nrow = 1)
```



For the stacked bar chart above:

a: For the first graph. The proportion of customer churn among males is very close to that among females, which shows that the cause of customer churn is not strongly related to gender.

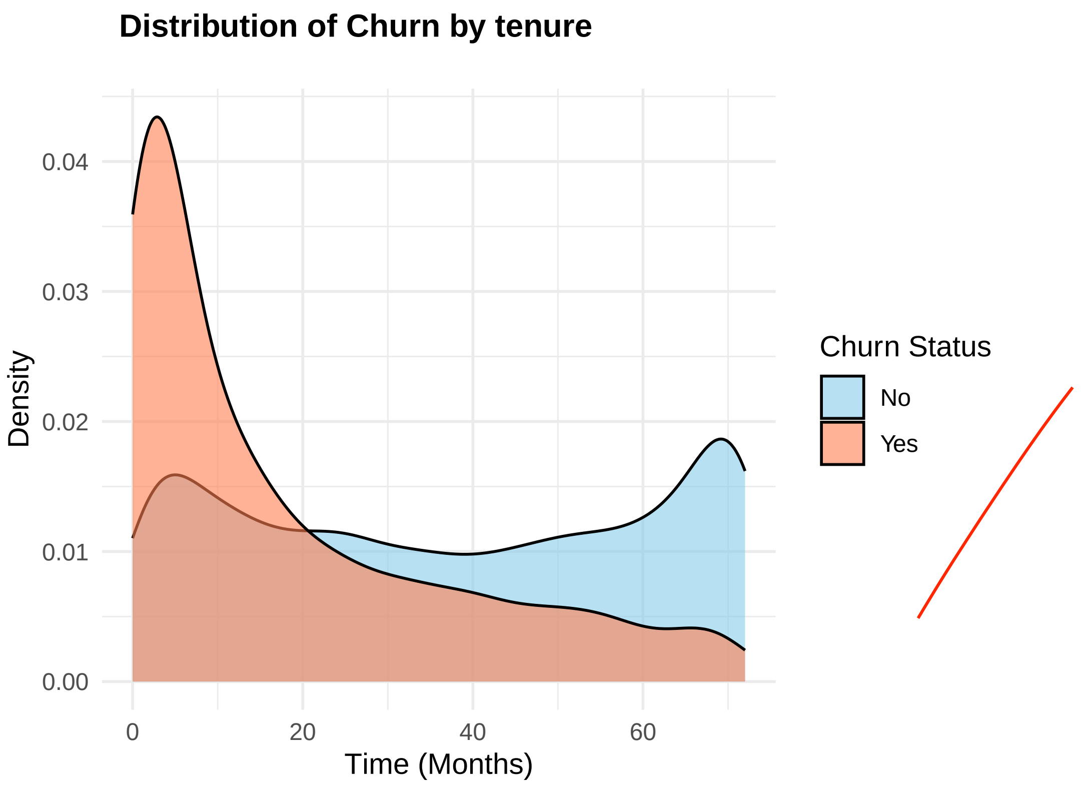
b: For the second graph. The proportion of men and women who have a partner is almost the same.

Answer **Question 1**:

The Telco dataset contains 21 variables with relationships between them. There is a strong linear relationship between “TotalCharges” and “MonthlyCharges * tenure”. Different service types can be grouped together, for example, services can be grouped into TV services (StreamingTV and StreamingMovies), support services (DeviceProtection and TechSupport), and online protection services (OnlineSecurity and OnlineBackup).

ii) The relationship between tenure and customer churn

```
# Create a density to show the distribution of churn under tenure
Churn_tenure <- ggplot(Telco, aes(x = tenure, fill = factor(Churn))) +
  geom_density(position = "identity", alpha = 0.6) +              # for two group plot the density
  labs(x = "Time (Months)", y = "Density") +
  theme_minimal() +
  scale_fill_manual(
    name = "Churn Status",                                   # set the legend title
    values = c("skyblue", "#FF7F50")
  ) +
  theme(
    legend.position = "right",                                    # set the position of legend
    plot.title = element_text(size = 12, hjust = 0.5),            # Adjust plot title size and position
    plot.margin = margin(0, 0, 0, 0, "cm")                        # Add margins around the plot
  )
# Resize the graph using the plot_grid() of the cowplot package
plot_grid(Churn_tenure, label_size = 12,         # add label title, and adjust the size
          labels = c("Distribution of Churn by tenure"),
          ncol = 1, nrow = 1, scale = c(0.8, 0.8) )
```



Answer **Question 2**:

The churned curve (Status Yes) is gradually decreasing. The fewer months a customer stays with the company, the more customers churn.

iii) The impact of total services and average costs on churn

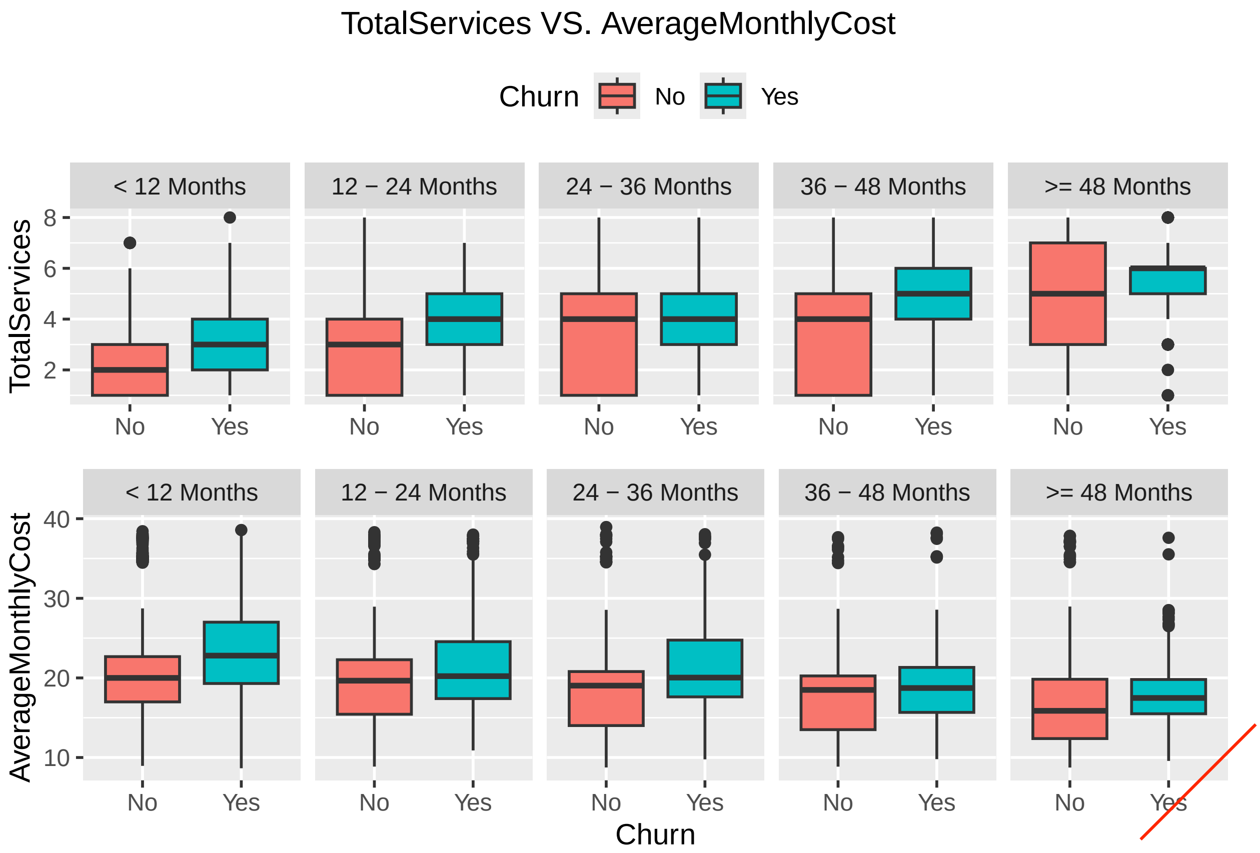
For the analysis of part i), perform the following operations on the Telco data:

I. Group services into TV services (StreamingTV and StreamingMovies), Support services (DeviceProtection and TechSupport), Online protection service (OnlineSecurity and OnlineBackup).

II. Divided the tenure months into annual bands.

III. Generates additional variables (6 new variables) with a focus of increasing predictive power of the model in upcoming sections. That contains "TotalServices", "AverageMonthlyCost", "Tenure_bands", "TV_Services", "Support_Services", "OnlineProtection_Services."

```r
# Capture all services which contained, using box-plot to draw the interrelationship
Telco <-
  Telco %>%
  # Group services into TV services, Support services & Online protection service
  mutate(TotalServices =
           ifelse(PhoneService == "Yes", 1, 0) +
           ifelse(InternetService != "No", 1, 0) +
           ifelse(OnlineSecurity == "Yes", 1, 0) +
           ifelse(OnlineBackup == "Yes", 1, 0) +
           ifelse(DeviceProtection == "Yes", 1, 0) +
           ifelse(TechSupport == "Yes", 1, 0) +
           ifelse(StreamingTV == "Yes", 1, 0) +
           ifelse(StreamingMovies == "Yes", 1, 0)
         ) %>%
  # Divided the tenure months into annual bands.
  mutate(
    AverageMonthlyCost = MonthlyCharges / TotalServices,
    Tenure_bands = factor(
      case_when(tenure < 12 ~ "< 12 Months",
                tenure >= 12 & tenure < 24 ~ "12 - 24 Months",
                tenure >= 24 & tenure < 36 ~ "24 - 36 Months",
                tenure >= 36 & tenure < 48 ~ "36 - 48 Months",
                tenure >= 48 ~ ">= 48 Months"),
      # capture customers that do subscribe to any internet service
      levels = c("< 12 Months", "12 - 24 Months", "24 - 36 Months","36 - 48 Months", ">= 48 Months")),
        TV_Services = factor(
          ifelse(StreamingTV == "Yes" | StreamingMovies == "Yes", "Yes", "No"), levels = c("Yes", "No"
        Support_Services = factor(
          ifelse(DeviceProtection == "Yes" | TechSupport == "Yes", "Yes", "No"), levels = c("Yes", "No
        OnlineProtection_Services = factor(
          ifelse(OnlineSecurity == "Yes" | OnlineBackup == "Yes", "Yes", "No"), levels = c("Yes", "No"
# Box-plot for Total services and AverageMonthlyCost
grid.arrange(
  top = "TotalServices VS. AverageMonthlyCost",
  Telco %>%
    dplyr::select(TotalServices, Churn, Tenure_bands) %>%
    ggplot(aes(x = Churn, y = TotalServices, fill = Churn)) +
    geom_boxplot() +
    facet_grid(. ~ Tenure_bands) +
    theme(plot.title = element_text(hjust = 0.5, size = 20, face = "bold"), legend.position="top") +
    labs(x = NULL),
  Telco %>%
    dplyr::select(AverageMonthlyCost, Churn, Tenure_bands) %>%
    ggplot(aes(x = Churn, y = AverageMonthlyCost, fill = Churn)) +
    geom_boxplot() +
    facet_grid(. ~ Tenure_bands) +
    theme(legend.position="none"), nrow = 2)
```



Answer **Question 3**:

From the TotalServices chart we can see that, in general, customers who subscribe to more services are more likely to churn, especially as the service term increases.

Answer **Question 4**:

The higher the monthly payment for a customer who has subscribed to a telco's service for a shorter period of time, the higher the risk of churn.

(3) Modelling

After performing a comprehensive exploratory analysis on a Telco data set, we may be very interested in knowing the model results for this data. Next, a logistic regression model were fitted.

i) Logistic regression model

Since the target variable "Churn is a categorical variable, a logistic regression model is easy to obtain, but the logistic regression model may be affected by its dependence on feature scales. The result is as follows:

```r
# remove some variables
remove_col <- c("customerID","StreamingMovies","DeviceProtection","OnlineSecurity","gender","tenure")
Telco_update <- Telco[, !(names(Telco) %in% remove_col)]
#Build the first model using all variables
model1 <- glm(Churn ~., family = "binomial", data = Telco_update)
summary(model1)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = "binomial", data = Telco_update)
##
## Coefficients: (4 not defined because of singularities)
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -1.819e+00  5.637e-01  -3.227 0.001250 **
## SeniorCitizenYES             2.215e-01  8.463e-02   2.617 0.008871 **
## PartnerYes                  -1.574e-02  7.775e-02  -0.202 0.839528
## DependentsYes               -1.398e-01  8.958e-02  -1.561 0.118606
## PhoneServiceYes             -1.323e+00  3.154e-01  -4.196 2.72e-05 ***
## MultipleLinesNPS                    NA         NA      NA       NA
## MultipleLinesYes             6.151e-02  1.363e-01   0.451 0.651717
## InternetServiceFO           -3.336e-01  5.589e-01  -0.597 0.550624
## InternetServiceNo            2.747e-01  4.212e-01   0.652 0.514293
## OnlineBackupNIS                     NA         NA      NA       NA
## OnlineBackupYes              1.458e-01  1.157e-01   1.260 0.207779
## TechSupportNIS                      NA         NA      NA       NA
## TechSupportYes              -3.078e-01  1.109e-01  -2.775 0.005527 **
## StreamingTVNIS                      NA         NA      NA       NA
## StreamingTVYes               1.171e-01  1.232e-01   0.950 0.342025
## Contract1Y                  -7.646e-01  1.077e-01  -7.100 1.25e-12 ***
## Contract2Y                  -1.733e+00  1.788e-01  -9.694  < 2e-16 ***
## PaperlessBillingYes          3.477e-01  7.466e-02   4.658 3.20e-06 ***
## PaymentMethodCred           -8.557e-02  1.134e-01  -0.754 0.450622
## PaymentMethodEltc            3.196e-01  9.448e-02   3.382 0.000719 ***
## PaymentMethodMail           -8.821e-03  1.146e-01  -0.077 0.938622
## MonthlyCharges               5.970e-02  2.186e-02   2.731 0.006321 **
## TotalCharges                -1.654e-04  5.454e-05  -3.033 0.002422 **
## TotalServices               -2.971e-01  1.744e-01  -1.703 0.088504 .
## AverageMonthlyCost          -1.871e-02  1.537e-02  -1.217 0.223684
## Tenure_bands12 - 24 Months  -7.815e-01  1.015e-01  -7.696 1.40e-14 ***
## Tenure_bands24 - 36 Months  -1.057e+00  1.369e-01  -7.722 1.15e-14 ***
## Tenure_bands36 - 48 Months  -8.952e-01  1.814e-01  -4.936 7.98e-07 ***
## Tenure_bands>= 48 Months    -9.711e-01  2.500e-01  -3.884 0.000103 ***
## TV_ServicesNo                2.383e-01  1.495e-01   1.594 0.110961
## Support_ServicesNo           8.649e-02  1.287e-01   0.672 0.501527
## OnlineProtection_ServicesNo  4.735e-01  1.437e-01   3.294 0.000986 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8150.1  on 7042  degrees of freedom
## Residual deviance: 5850.3  on 7015  degrees of freedom
## AIC: 5906.3
```

not significant for Internet Service; in comparison to DSL as reference category

```
##
## Number of Fisher Scoring iterations: 6
```

```
# reduce variables by AIC
model2 <- stepAIC(model1, direction="both", trace = 0)
summary(model2)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + MultipleLines +
##     TechSupport + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + TotalServices + Tenure_bands +
##     OnlineProtection_Services, family = "binomial", data = Telco_update)
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                 -2.664e+00  2.286e-01 -11.654  < 2e-16 ***
## SeniorCitizenYES             2.185e-01  8.403e-02   2.600 0.009319 **
## DependentsYes               -1.489e-01  8.137e-02  -1.830 0.067260 .
## MultipleLinesNPS             1.139e+00  1.497e-01   7.609 2.76e-14 ***
## MultipleLinesYes             1.217e-01  8.410e-02   1.447 0.148009
## TechSupportNIS               7.542e-02  1.759e-01   0.429 0.668129
## TechSupportYes              -3.572e-01  9.914e-02  -3.603 0.000315 ***
## Contract1Y                  -7.768e-01  1.070e-01  -7.262 3.80e-13 ***
## Contract2Y                  -1.767e+00  1.753e-01 -10.077  < 2e-16 ***
## PaperlessBillingYes          3.490e-01  7.441e-02   4.691 2.72e-06 ***
## PaymentMethodCred           -8.367e-02  1.132e-01  -0.739 0.459993
## PaymentMethodEltc            3.303e-01  9.421e-02   3.506 0.000455 ***
## PaymentMethodMail            2.992e-03  1.142e-01   0.026 0.979108
## MonthlyCharges               4.127e-02  3.713e-03  11.115  < 2e-16 ***
## TotalCharges                -1.510e-04  5.378e-05  -2.808 0.004979 **
## TotalServices               -1.432e-01  4.882e-02  -2.933 0.003354 **
## Tenure_bands12 - 24 Months  -7.914e-01  1.009e-01  -7.840 4.50e-15 ***
## Tenure_bands24 - 36 Months  -1.075e+00  1.361e-01  -7.898 2.84e-15 ***
## Tenure_bands36 - 48 Months  -9.181e-01  1.808e-01  -5.079 3.79e-07 ***
## Tenure_bands>= 48 Months    -1.016e+00  2.486e-01  -4.088 4.35e-05 ***
## OnlineProtection_ServicesNo  3.483e-01  8.771e-02   3.971 7.15e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8150.1  on 7042  degrees of freedom
## Residual deviance: 5856.5  on 7022  degrees of freedom
## AIC: 5898.5
##
## Number of Fisher Scoring iterations: 6
```

wrong; this is not significant

wrong; this is not significant

Answer **Question 5**:

The logistic model gives the significant variables "MultipleLines", "InternetService", "TechSupport", "Contract", "PaperlessBilling", "PaymentMethod", and "TotalCharges".

For "MultipleLines", "InternetService", "TechSupport", this shows that users value these three types of services. Good service options will indeed reduce user churn.

For “Contract”, this shows that users value whether they can get contract guarantees. The longer the contract, the less likely users are to churn in the short term.

For “PaperlessBilling”, “PaymentMethod”. Customers prefer the security of paper bills and are accustomed to choosing electronic payment methods.

For “TotalCharges”. Most non-premium customers may care about whether the total cost is reasonable. The total cost is also one of the reasons for customer churn.

Conclusion

(1) Telcos need to create an easy and affordable entry point for their services. Extensive focus on support services, online services, and TV services is needed during the first 6 months period, as this period is the most critical and uncertain for customers. Increase the volume of services to reduce the number of subscribers that churn in the early stages.

(2) Users also attach great importance to Internet services. Therefore, telecommunications companies need to promote the use of multiple lines and fiber-optic cables for telephone services and Internet services respectively. However, this will increase users’ monthly expenses. Therefore, the main obstacle is that the starting point of the monthly fee needs to be adjusted.

(3) In addition, users feel reassured by the guarantee of paper bills. Providing customers with guarantees and allowing users to obtain convenient and fast payment methods is also one of the measures to prevent user churn.

References

1. TANMAY DESHPANDE. (kaggle). *Telco Churn: EDA|CV Score (85%+)| F1 Score (80%+)*. https://www.kaggle.com/code/tanmay111999/telco-churn-eda-cv-score-85-f1-score-80

2. Shakarchi, Ali and Mostafa, Salama and Saringat, Mohd and Mohammed, Dheyaa and Al-Dulaimi, Shihab and Jaber, Mustafa, *A Data Mining Approach for Analysis of Telco Customer Churn*, 07/2023 doi: 10.1109/AICCIT57614.2023.10218161.

Session Information

```
## R version 4.4.0 (2024-04-24)
## Platform: aarch64-apple-darwin20
## Running under: macOS Monterey 12.4
## 
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRblas.0.dylib 
## LAPACK: /Library/Frameworks/R.framework/Versions/4.4-arm64/Resources/lib/libRlapack.dylib;  LAPACK ve
## 
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
## 
## time zone: Asia/Kuala_Lumpur
## tzcode source: internal
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
```

```
##
## other attached packages:
##  [1] knitr_1.47      gridExtra_2.3   MASS_7.3-60.2   cowplot_1.1.3
##  [5] plyr_1.8.9      lubridate_1.9.3 forcats_1.0.0   stringr_1.5.1
##  [9] dplyr_1.1.4     purrr_1.0.2     readr_2.1.5     tidyr_1.3.1
## [13] tibble_3.2.1    ggplot2_3.5.1   tidyverse_2.0.0
##
## loaded via a namespace (and not attached):
##  [1] utf8_1.2.4        generics_0.1.3    lattice_0.22-6    stringi_1.8.4
##  [5] hms_1.1.3         digest_0.6.35     magrittr_2.0.3    evaluate_0.24.0
##  [9] grid_4.4.0        timechange_0.3.0  fastmap_1.2.0     Matrix_1.7-0
## [13] mgcv_1.9-1        fansi_1.0.6       scales_1.3.0      cli_3.6.2
## [17] rlang_1.1.4       munsell_0.5.1     splines_4.4.0     withr_3.0.0
## [21] yaml_2.3.8        tools_4.4.0       tzdb_0.4.0        colorspace_2.1-0
## [25] vctrs_0.6.5       R6_2.5.1          lifecycle_1.0.4   pkgconfig_2.0.3
## [29] pillar_1.9.0      gtable_0.3.5      glue_1.7.0        Rcpp_1.0.12
## [33] xfun_0.44         tidyselect_1.2.1  highr_0.11        rstudioapi_0.16.0
## [37] farver_2.1.2      nlme_3.1-164      htmltools_0.5.8.1 rmarkdown_2.27
## [41] labeling_0.4.3    compiler_4.4.0
```

Comments:
Probably it would be so much better if you divide the training and testing data for logistic regression - then you can evaluate the performance of your logistic regression model