Churn Analysis

Nanda kishore

2023-12-03

AIM:

Business Goals

Churn refers to customers who have exited within the last month, significantly impacting both sales figures and overall business revenue. The root cause of this phenomenon often lies in customer satisfaction levels. Therefore, the business objectives should focus on minimizing customer churn to boost revenue and enhancing overall customer satisfaction.

To effectively address customer churn, it becomes imperative to predict which customers are at a higher risk of churning by identifying early indicators across various sales channels.

The key challenge is to predict if an individual customer will churn or not. The extra challenge is to identify the key components of churning.

questions for exploration include:

- Are there discernible patterns among churned customers?
- What is the percentage distribution between churned and existing customers?
- How are contract types distributed among customers?
- Which variables exhibit correlation with customer churn?
- Conversely, which variables do not display correlation with customer churn?
- What model proves most accurate in predicting churn?
- How are contract types distributed within the customer base?
- What impact does contract length have on customer behavior?
- Is there a relationship between customer usage patterns, additional products, and churn?
- How does the variety of services a customer utilizes influence churn?

About Dataset

Context "Predict behavior to retain customers. You can analyze all relevant customer data and develop focused customer retention programs." [IBM Sample Data Sets]

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

The data set includes information about:

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

- Customer account information – how long they've been a customer, contract, payment method, paperless billing, monthly charges, and total charges
- Demographic info about customers – gender, age range, and if they have partners and dependents

Dataset Attributes

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

reference kaggle

Loading Dataset

```
file_path <- "C:/Users/manda/OneDrive - Western Michigan University/Desktop/wmu/fall 2023/applied data m

# Read the CSV file into a data frame
setwd("C:/Users/manda/OneDrive - Western Michigan University/Desktop/wmu/fall 2023/applied data minnig/p
df <- read.csv("WA_Fn-UseC_-Telco-Customer-Churn.csv")

# View the structure of the data frame
str(df)
```

```
## '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" ...
```

```r
#Data Exploration

summary(df)
```

```
##   customerID           gender          SeniorCitizen      Partner
##  Length:7043        Length:7043        Min.   :0.0000   Length:7043
##  Class :character   Class :character   1st Qu.:0.0000   Class :character
##  Mode  :character   Mode  :character   Median :0.0000   Mode  :character
##                                        Mean   :0.1621
##                                        3rd Qu.:0.0000
##                                        Max.   :1.0000
##
##   Dependents            tenure      PhoneService       MultipleLines
##  Length:7043        Min.   : 0.00   Length:7043        Length:7043
##  Class :character   1st Qu.: 9.00   Class :character   Class :character
##  Mode  :character   Median :29.00   Mode  :character   Mode  :character
##                     Mean   :32.37
##                     3rd Qu.:55.00
##                     Max.   :72.00
##
##  InternetService    OnlineSecurity     OnlineBackup       DeviceProtection
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  TechSupport        StreamingTV        StreamingMovies      Contract
##  Length:7043        Length:7043        Length:7043        Length:7043
##  Class :character   Class :character   Class :character   Class :character
##  Mode  :character   Mode  :character   Mode  :character   Mode  :character
##
##
##
##
##  PaperlessBilling   PaymentMethod      MonthlyCharges    TotalCharges
##  Length:7043        Length:7043        Min.   : 18.25   Min.   :  18.8
##  Class :character   Class :character   1st Qu.: 35.50   1st Qu.: 401.4
##  Mode  :character   Mode  :character   Median : 70.35   Median :1397.5
```

```
##                                         Mean   : 64.76   Mean   :2283.3
##                                         3rd Qu.: 89.85   3rd Qu.:3794.7
##                                         Max.   :118.75   Max.   :8684.8
##                                                          NA's   :11
##     Churn
##  Length:7043
##  Class :character
##  Mode  :character
##
##
##
##
```

pre processing the data

- before everything else we remove the customer id in order for better analysis

```
df <- df[, !(names(df) %in% c("customerID"))]
```

```
# Check for missing values in each column
missing_values <- colSums(is.na(df))

# Display the results
print(missing_values)
```

missing values

```
##           gender    SeniorCitizen          Partner       Dependents
##                0                0                0                0
##           tenure     PhoneService    MultipleLines  InternetService
##                0                0                0                0
##   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport
##                0                0                0                0
##      StreamingTV  StreamingMovies         Contract PaperlessBilling
##                0                0                0                0
##    PaymentMethod   MonthlyCharges     TotalCharges            Churn
##                0                0               11                0
```

```
# 11 missing values in Total charges

# Impute missing values in 'TotalCharges' with the mean
df$TotalCharges[is.na(df$TotalCharges)] <- mean(df$TotalCharges, na.rm = TRUE)

#or

df$TotalCharges[is.na(df$TotalCharges)] <- 0

# Check for missing values in each column
missing_values <- colSums(is.na(df))
```

```r
# Display the results
print(missing_values)
```

```
##           gender    SeniorCitizen          Partner       Dependents
##                0                0                0                0
##           tenure     PhoneService    MultipleLines  InternetService
##                0                0                0                0
##   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport
##                0                0                0                0
##      StreamingTV  StreamingMovies         Contract PaperlessBilling
##                0                0                0                0
##    PaymentMethod   MonthlyCharges     TotalCharges            Churn
##                0                0                0                0
```

- no missing values

```r
# Check for duplicated rows in the data frame
num_duplicated <- sum(duplicated(df))

# Print the number of duplicated values
print(paste("Number of duplicated values in the dataset: ", num_duplicated))
```

duplicate values

```
## [1] "Number of duplicated values in the dataset:  22"
```

```r
df <- unique(df)

# Check for duplicated rows in the data frame
num_duplicated <- sum(duplicated(df))

# Print the number of duplicated values
print(paste("Number of duplicated values in the dataset: ", num_duplicated))
```

```
## [1] "Number of duplicated values in the dataset:  0"
```

```r
columns <- names(df)

numeric_columns <- character(0)
binary_columns <- character(0)
categorical_columns <- character(0)

for (col in columns) {
  unique_vals <- unique(df[[col]])

  if (length(unique_vals) > 10) {
    numeric_columns <- c(numeric_columns, col)
  } else if (length(unique_vals) == 2) {
```

```
    binary_columns <- c(binary_columns, col)
  } else {
    categorical_columns <- c(categorical_columns, col)
  }
}

# Print the categorization
cat("Numeric Columns:", numeric_columns, "\n")
```

```
## Numeric Columns: tenure MonthlyCharges TotalCharges
```

```
cat("Binary Columns:", binary_columns, "\n")
```

```
## Binary Columns: gender SeniorCitizen Partner Dependents PhoneService PaperlessBilling Churn
```

```
cat("Categorical Columns:", categorical_columns, "\n")
```

```
## Categorical Columns: MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection TechS
```

```
# Copy the data frame
df1 <- df

# Identify categorical columns
categoric_columns <- c("gender", "Partner", "Dependents", "PhoneService",
                       "MultipleLines", "InternetService", "OnlineSecurity",
                       "OnlineBackup", "DeviceProtection", "TechSupport",
                       "StreamingTV", "StreamingMovies", "Contract",
                       "PaperlessBilling", "PaymentMethod")

# Convert categorical columns to factors
df1[categoric_columns] <- lapply(df1[categoric_columns], factor)

# Convert 'Churn' column to factor
df1$Churn <- factor(df1$Churn, levels = c("No", "Yes"))

# Print the first few rows of the transformed data frame
head(df1)
```

```
##   gender SeniorCitizen Partner Dependents tenure PhoneService    MultipleLines
## 1 Female             0     Yes         No      1           No No phone service
## 2   Male             0      No         No     34          Yes               No
## 3   Male             0      No         No      2          Yes               No
## 4   Male             0      No         No     45           No No phone service
## 5 Female             0      No         No      2          Yes               No
## 6 Female             0      No         No      8          Yes              Yes
##   InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport
## 1             DSL             No          Yes               No          No
## 2             DSL            Yes           No              Yes          No
## 3             DSL            Yes          Yes               No          No
## 4             DSL            Yes           No              Yes         Yes
## 5     Fiber optic             No           No               No          No
```

```
## 6     Fiber optic             No           No              Yes          No
##   StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No              No Month-to-month              Yes
## 2          No              No       One year               No
## 3          No              No Month-to-month              Yes
## 4          No              No       One year               No
## 5          No              No Month-to-month              Yes
## 6         Yes             Yes 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
## 6          Electronic check          99.65       820.50   Yes
```

```
summary(df[, numeric_columns])
```

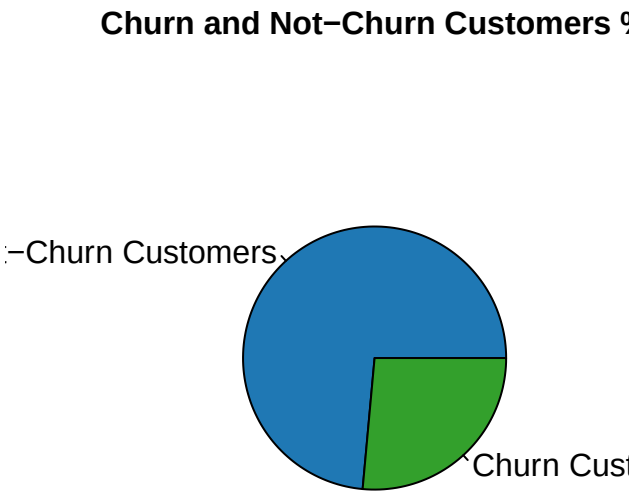
```
##      tenure      MonthlyCharges    TotalCharges
##  Min.   : 0.00   Min.   : 18.25   Min.   :  18.8
##  1st Qu.: 9.00   1st Qu.: 35.75   1st Qu.: 411.1
##  Median :29.00   Median : 70.40   Median :1410.2
##  Mean   :32.47   Mean   : 64.85   Mean   :2290.3
##  3rd Qu.:55.00   3rd Qu.: 89.90   3rd Qu.:3801.7
##  Max.   :72.00   Max.   :118.75   Max.   :8684.8
```

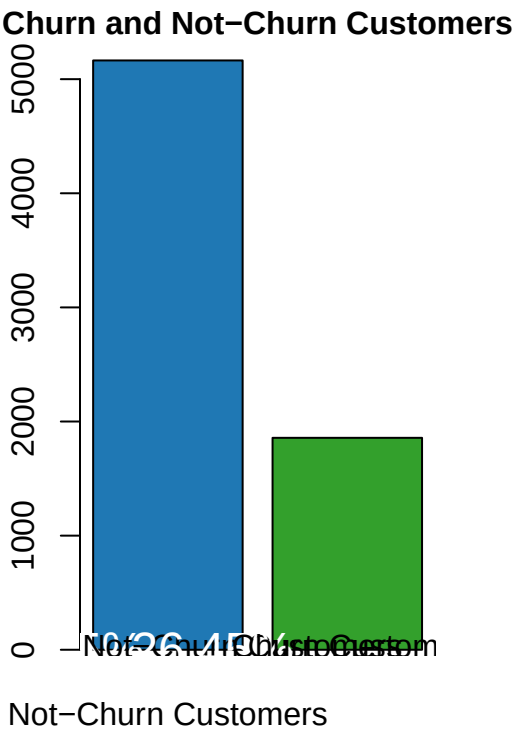
```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 4.3.2
```

```
# Data imbalance check
churn_counts <- table(df$Churn)
pie_values <- prop.table(churn_counts) * 100

# Plotting
par(mfrow = c(1, 2), mar = c(5, 5, 2, 2))

# Pie chart
pie(pie_values, labels = c('Not-Churn Customers', 'Churn Customers'),
    col = c('#1F78B4', '#33A02C'), main = 'Churn and Not-Churn Customers %',
    cex.main = 1, cex.lab = 1)

# Bar chart
bar_colors <- c('#1F78B4', '#33A02C')
barplot(churn_counts, names.arg = c('Not-Churn Customers', 'Churn Customers'),
        col = bar_colors, main = 'Churn and Not-Churn Customers',
        cex.main = 1, cex.lab = 1, border = 'black')

# Adding percentage labels to the pie chart
text(0, 0, paste0(sprintf("%.2f", pie_values[1]), "%"), col = 'white', cex = 1.5)
text(0, 0, 'Not-Churn Customers', col = 'black', cex = 1, pos = 4)
```

```
text(1, 0, paste0(sprintf("%.2f", pie_values[2]), "%"), col = 'white', cex = 1.5)
text(1, 0, 'Churn Customers', col = 'black', cex = 1, pos = 4)
```





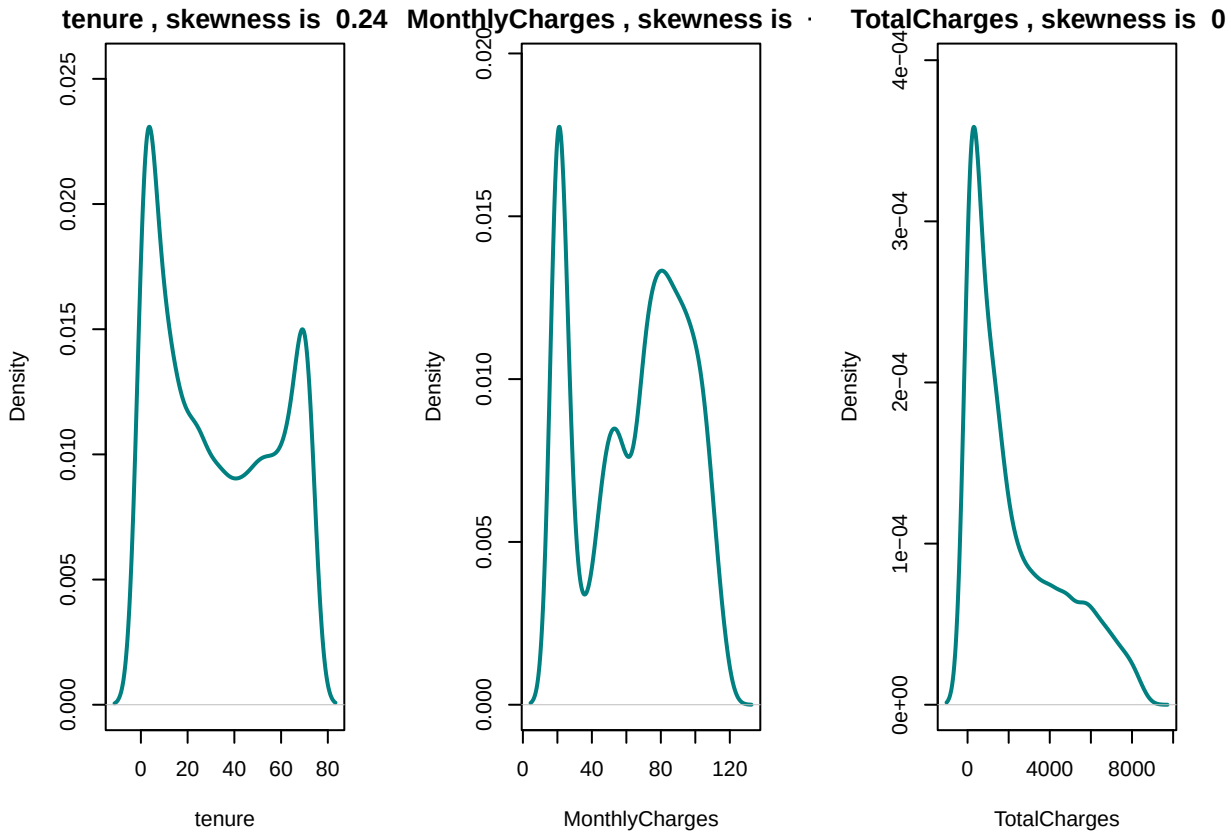

```
# Install the e1071 package if not already installed
if (!requireNamespace("e1071", quietly = TRUE)) {
  install.packages("e1071")
}

# Load the e1071 package
library(e1071)
```

```
## Warning: package 'e1071' was built under R version 4.3.2
```

```
# Assuming df1 is your data frame

dist_custom <- function(dataset, columns_list, rows, cols, suptitle) {
  par(mfrow = c(rows, cols), mar = c(5, 5, 2, 2))
  cat("\n")
  cat(suptitle, "\n", sep = "")

  for (i in seq_along(columns_list)) {
    col <- columns_list[i]

    # Density plot
```

```
    plot(density(dataset[[col]]), main = paste(col, ", skewness is ", round(e1071::skewness(dataset[[col
         col = '#008080', lwd = 2, ylim = c(0, max(density(dataset[[col]])$y) * 1.1),
         xlab = col, ylab = 'Density')
  }
}

# Usage
dist_custom(dataset = df, columns_list = numeric_columns, rows = 1, cols = 3, suptitle = 'Distribution
```

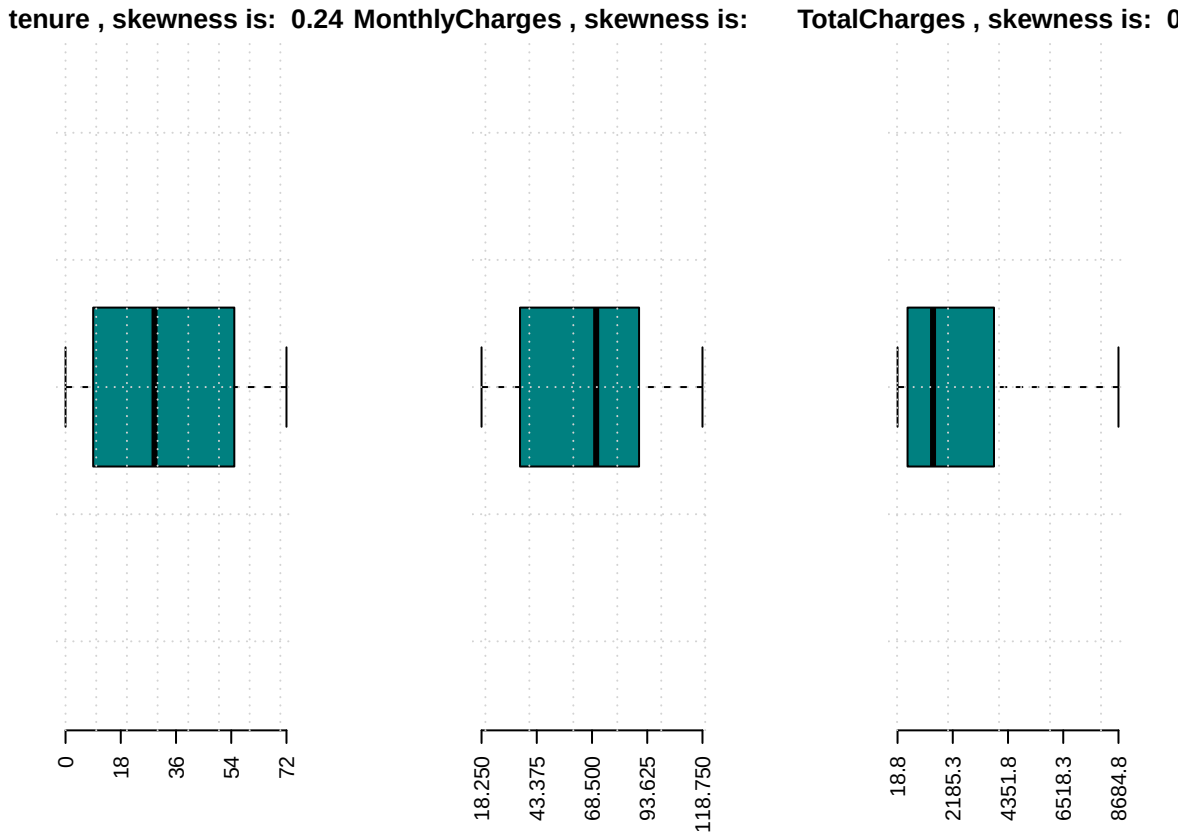
```
##
## Distribution for each numerical feature
```




```
# Assuming df1 is your data frame

boxplots_custom <- function(dataset, columns_list, rows, cols, suptitle) {
  par(mfrow = c(rows, cols), mar = c(5, 5, 2, 2))
  cat("\n")
  cat(suptitle, "\n", sep = "")

  for (i in seq_along(columns_list)) {
    col <- columns_list[i]

    # Boxplot
    boxplot(dataset[[col]], horizontal = TRUE, main = paste(col, ", skewness is: ", round(e1071::skewnes
            col = '#008080', border = 'black', axes = FALSE, boxwex = 0.5)
```

```
    # Add labels and grid
    axis(1, at = seq(min(dataset[[col]]), max(dataset[[col]]), length = 5), las = 2)
    grid()
  }
}


boxplots_custom(dataset = df1, columns_list = numeric_columns, rows = 1, cols = 3, suptitle = 'Boxplots
```

```
##
## Boxplots for numerical features
```



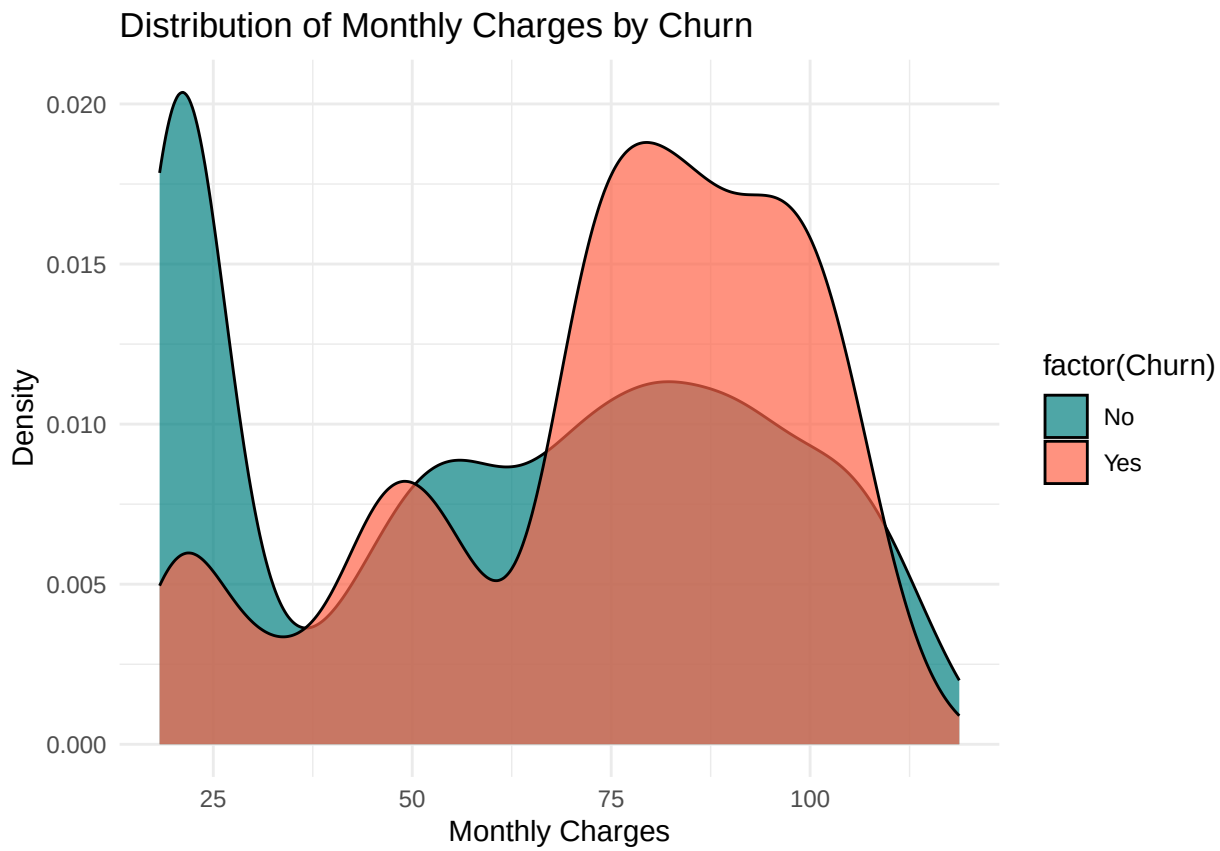
- TotalCharges is rightly skewed.

```
detect_outliers_iqr <- function(data) {
  # Calculate the first quartile (Q1) and third quartile (Q3)
  Q1 <- quantile(data, 0.25, na.rm = TRUE)
  Q3 <- quantile(data, 0.75, na.rm = TRUE)

  # Calculate the IQR
  IQR <- Q3 - Q1

  # Define the lower and upper bounds to identify outliers
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
```

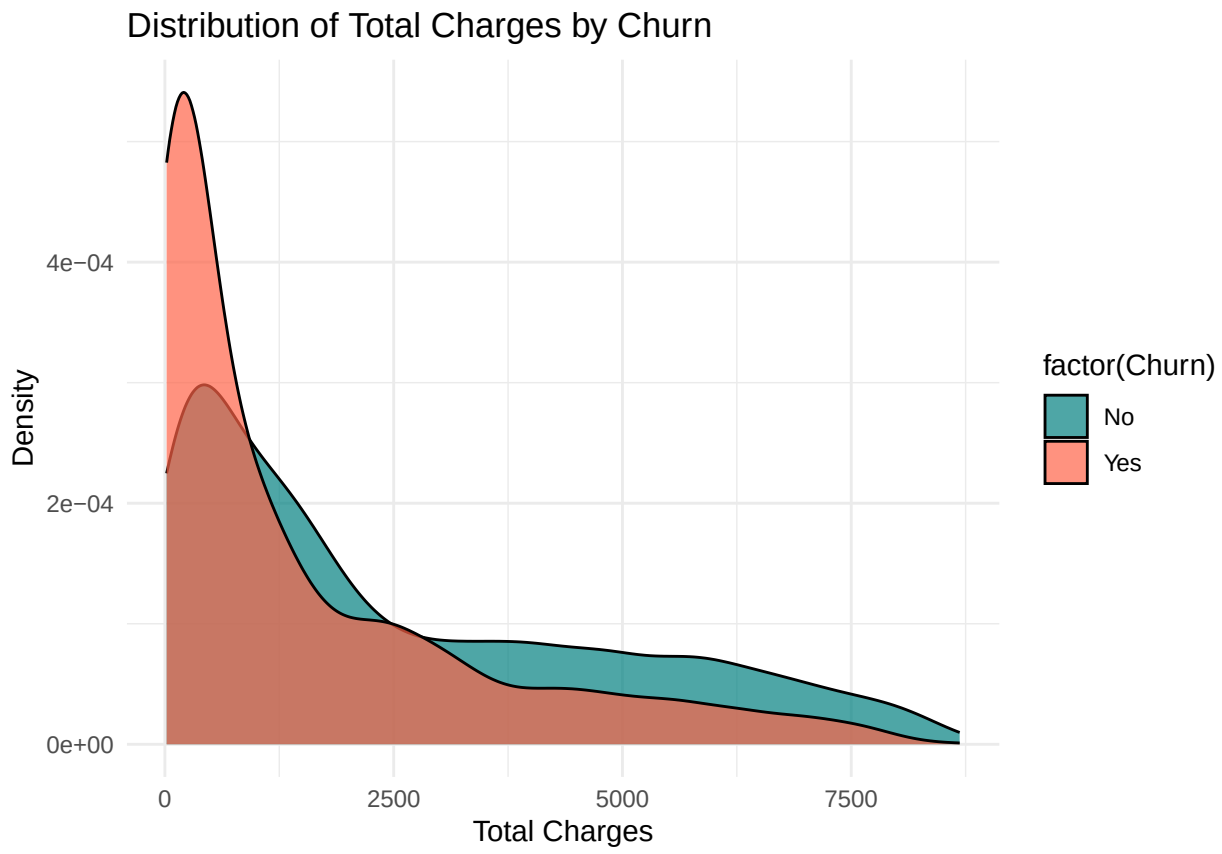
```r
  # Find the outliers
  outliers <- data[data < lower_bound | data > upper_bound]

  return(outliers)
}

# Apply outlier detection to all numerical columns
outliers_by_column <- list()

for (column in numeric_columns) {
  data_column <- df1[[column]]
  outliers <- detect_outliers_iqr(data_column)
  outliers_by_column[[column]] <- outliers
}

# Print the outliers for each column
for (column in names(outliers_by_column)) {
  cat("Outliers in", column, ":", outliers_by_column[[column]], "\n")
}
```

```
## Outliers in tenure :
## Outliers in MonthlyCharges :
## Outliers in TotalCharges :
```

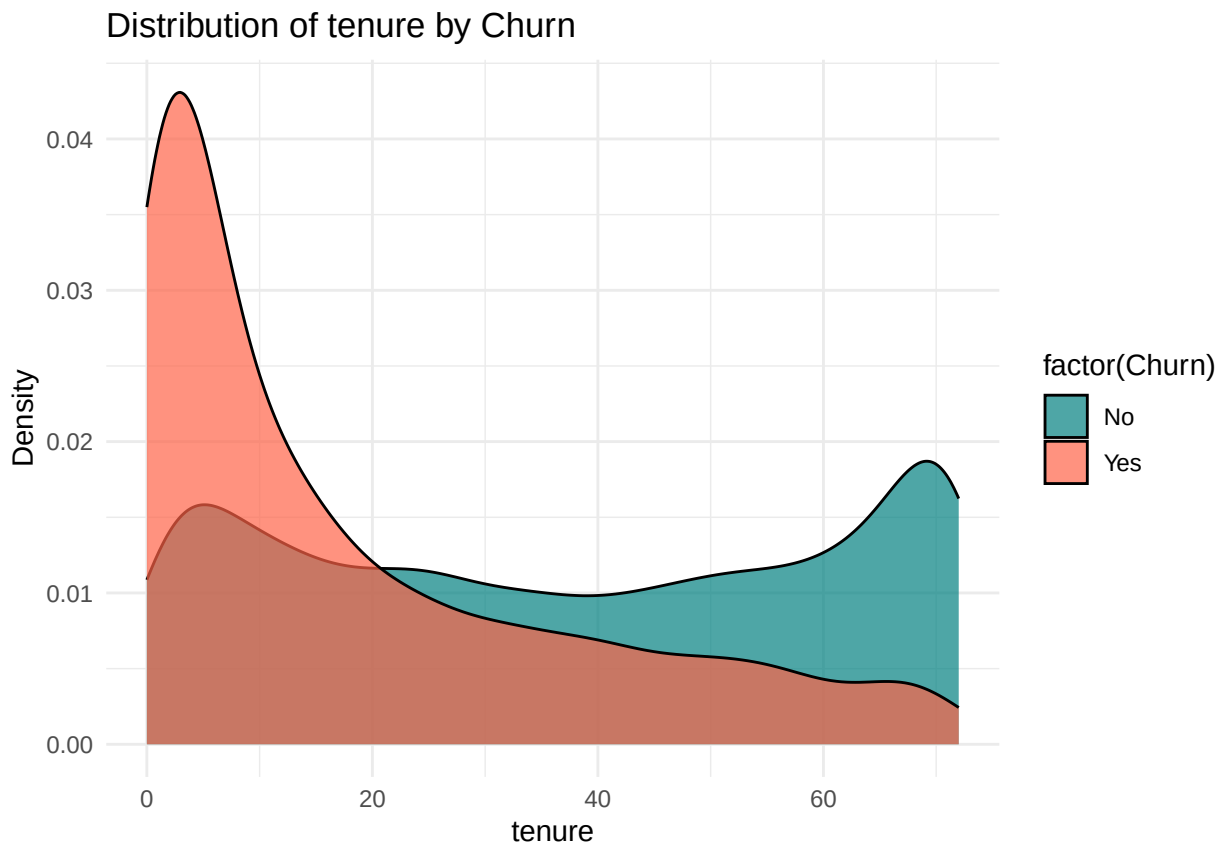
```r
#install.packages("ggplot2")

library(ggplot2)

# Monthly Charges vs Churn
ggplot(df1, aes(x = MonthlyCharges, fill = factor(Churn))) +
  geom_density(alpha = 0.7, color = 'black') +
  scale_fill_manual(values = c('#008080', '#FF6347')) +
  labs(title = 'Distribution of Monthly Charges by Churn', x = 'Monthly Charges', y = 'Density') +
  theme_minimal()
```




```r
# Total Charges vs Churn
ggplot(df1, aes(x = TotalCharges, fill = factor(Churn))) +
  geom_density(alpha = 0.7, color = 'black') +
  scale_fill_manual(values = c('#008080', '#FF6347')) +
  labs(title = 'Distribution of Total Charges by Churn', x = 'Total Charges', y = 'Density') +
  theme_minimal()
```



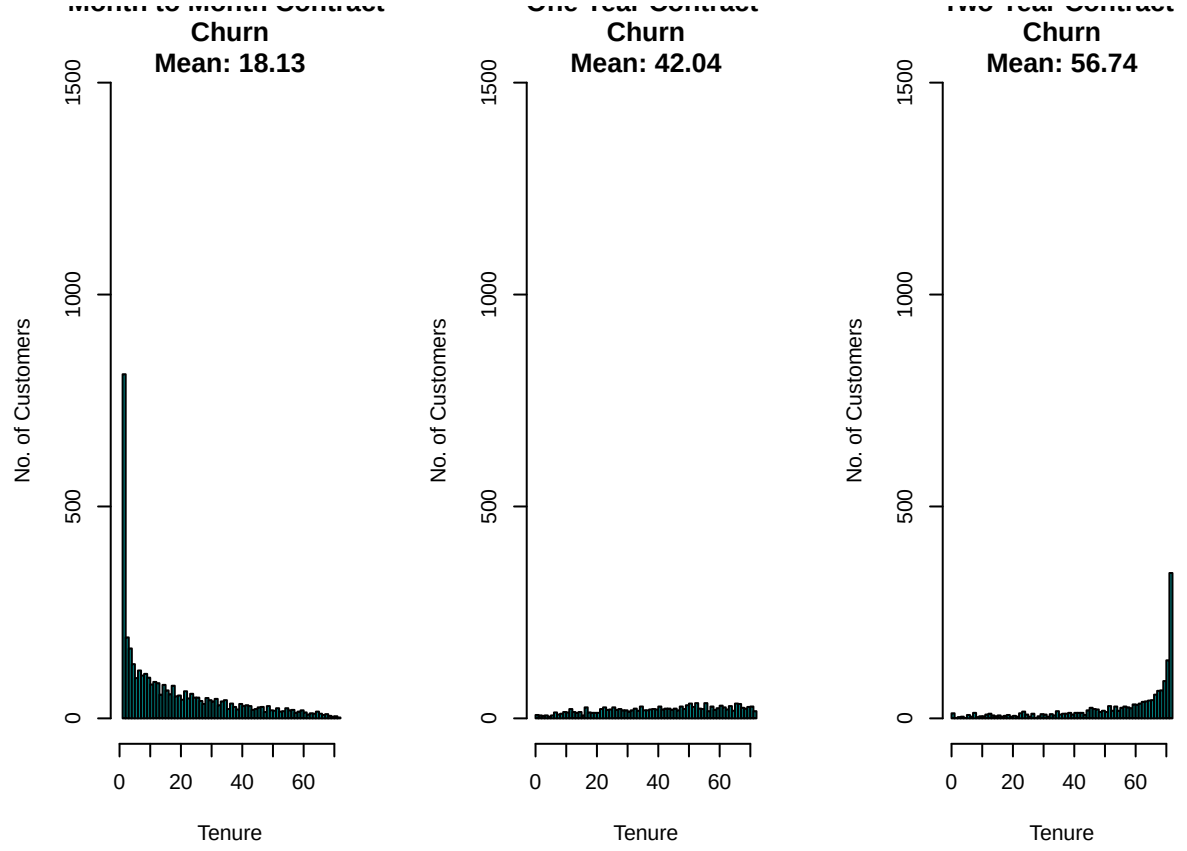

```r
# tenure vs Churn
ggplot(df1, aes(x = tenure, fill = factor(Churn))) +
  geom_density(alpha = 0.7, color = 'black') +
  scale_fill_manual(values = c('#008080', '#FF6347')) +
  labs(title = 'Distribution of tenure by Churn', x = 'tenure', y = 'Density') +
  theme_minimal()
```



- The longer the customer has been with the provider the more likely he will not churn.

```r
# Assuming df is your data frame

# Convert 'Contract' to a factor with appropriate levels
df$Contract <- factor(df$Contract, levels = c("Month-to-month", "One year", "Two year"))

# Calculate mean tenures for different contract types
two_year_mean <- round(mean(df$tenure[df$Contract == "Two year"], na.rm = TRUE), 2)
month_mean <- round(mean(df$tenure[df$Contract == "Month-to-month"], na.rm = TRUE), 2)
year_mean <- round(mean(df$tenure[df$Contract == "One year"], na.rm = TRUE), 2)

# Create histograms for different contract types
par(mfrow = c(1, 3), mar = c(5, 5, 2, 2), xpd = TRUE)

# Function to handle potential issues in plotting histogram
plot_hist <- function(data, contract_type, mean_value) {
  if (length(data) > 0) {
    hist(data, main = paste(contract_type, "\nChurn\nMean:", mean_value),
         xlab = "Tenure", ylab = "No. of Customers", col = "#008080", breaks = 72,
         xlim = c(0, max(data, na.rm = TRUE)), ylim = c(0, 1500))
  } else {
    cat("No data for", contract_type, "\n")
  }
}

# Plot histograms
```

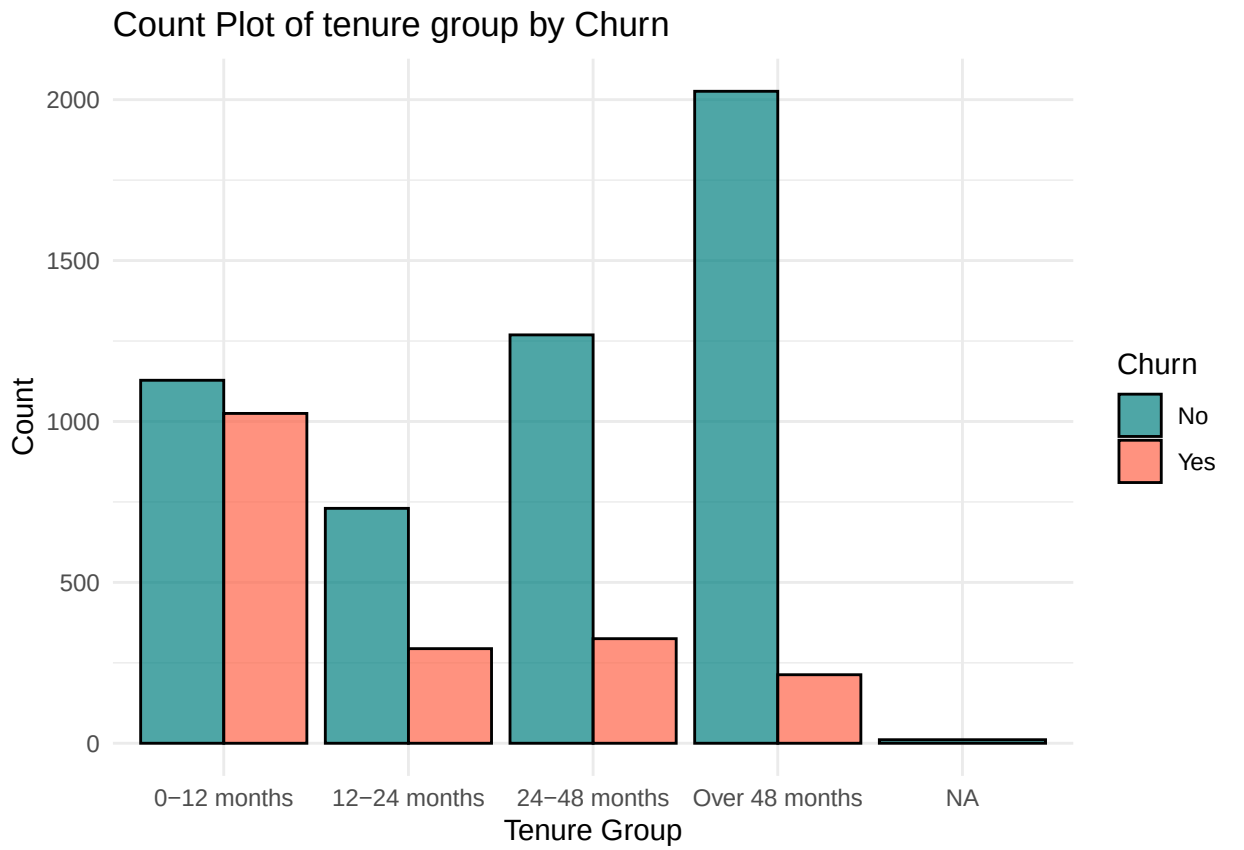
```r
plot_hist(df$tenure[df$Contract == "Month-to-month"], "Month to Month Contract", month_mean)
plot_hist(df$tenure[df$Contract == "One year"], "One Year Contract", year_mean)
plot_hist(df$tenure[df$Contract == "Two year"], "Two Year Contract", two_year_mean)
```



- A significant number of customers churned within the first month of their subscription.
- A considerable portion of customers has remained with the provider for a prolonged period of 72 weeks.
- There is a positive correlation between the length of the contract and the customer's tenure, suggesting that customers tend to stay longer with the provider when they have longer-term contracts. This is indicated by a higher mean tenure for customers with longer contracts.

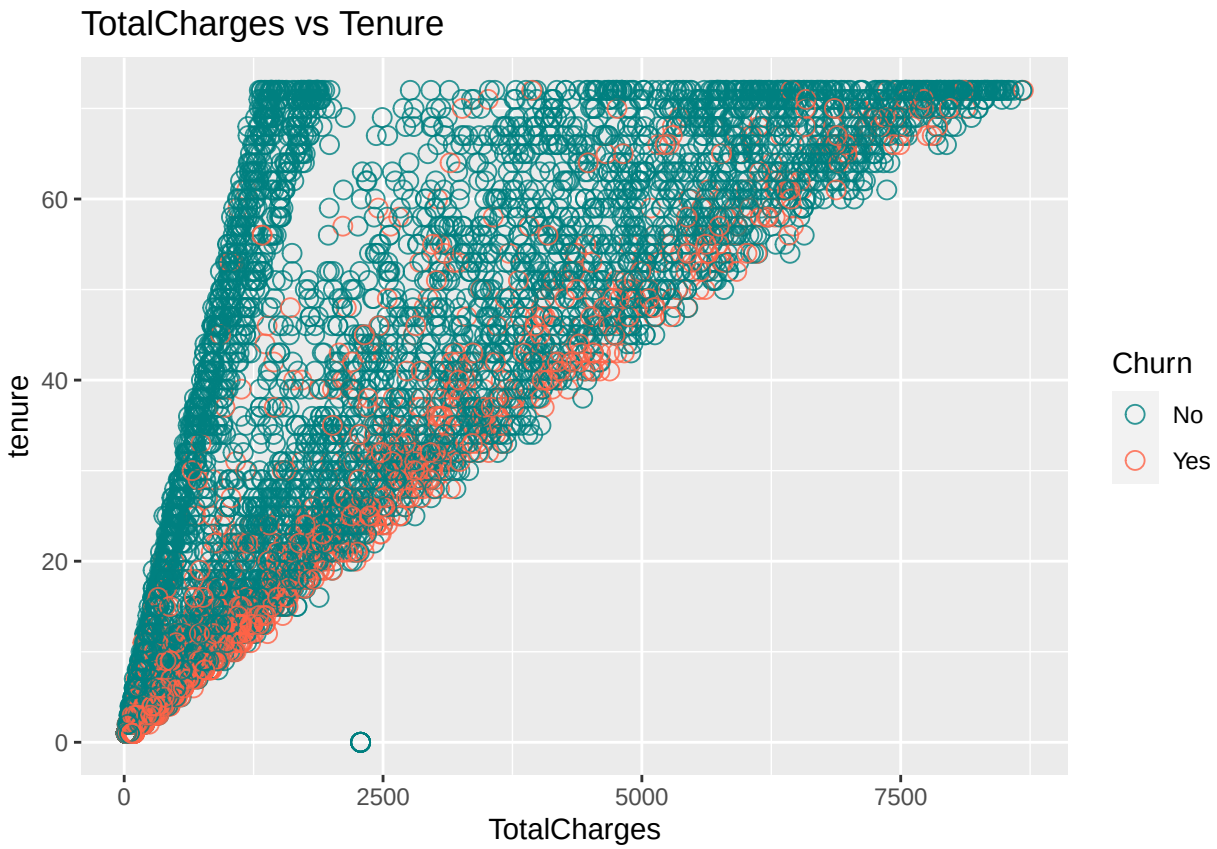
```r
df$tenure <- as.numeric(df$tenure)
# Create a categorical variable for tenure groups
df$tenure.group <- cut(df$tenure, breaks = c(0, 12, 24, 48, Inf),
                       labels = c("0-12 months", "12-24 months", "24-48 months", "Over 48 months"))

# Load required libraries
library(ggplot2)

# Create a count plot
ggplot(df, aes(x = tenure.group, fill = Churn)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  labs(title = "Count Plot of tenure group by Churn") +
  xlab("Tenure Group") +
  ylab("Count") +
```

```
  scale_fill_manual(values = c("#008080", "#FF6347")) +
  theme_minimal()
```

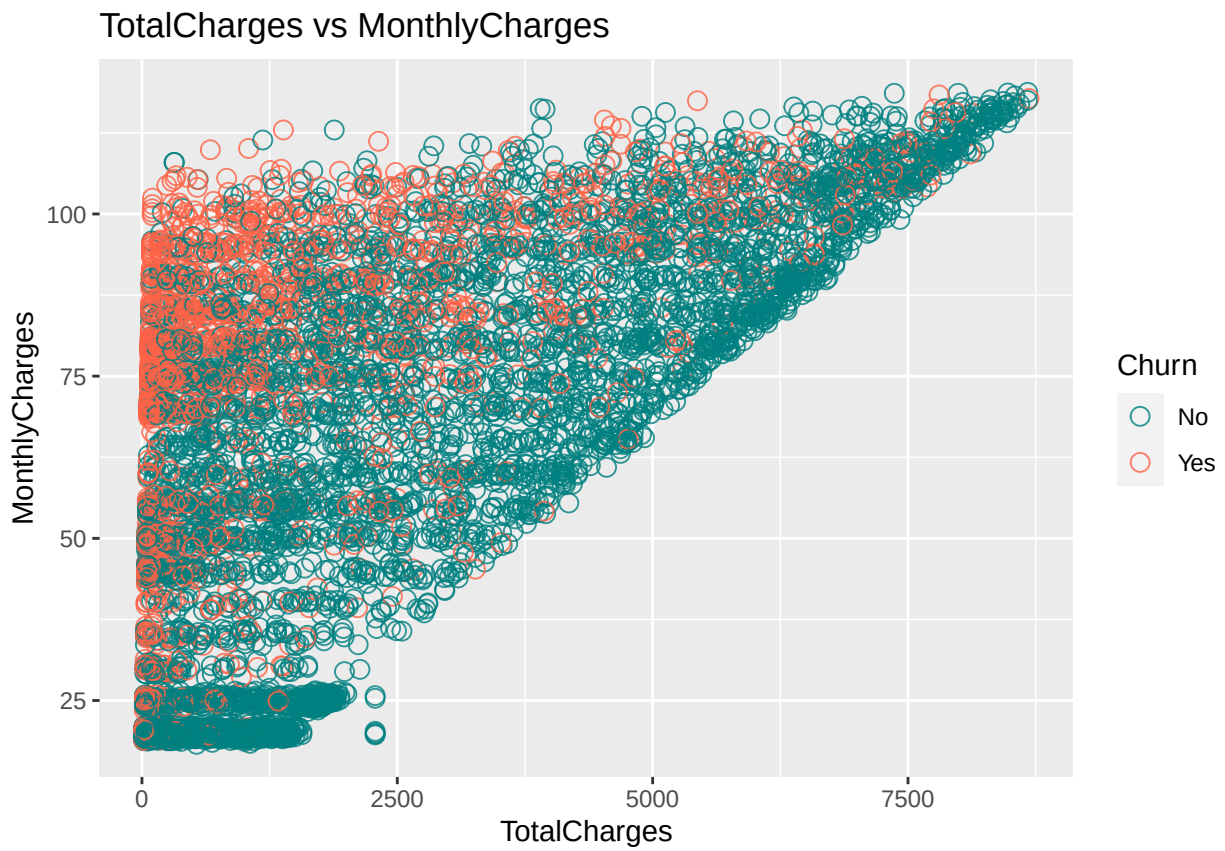


- The analysis gives a bigger picture. HIgher the tenure group, higher chances of retention and lesser chances of churn.

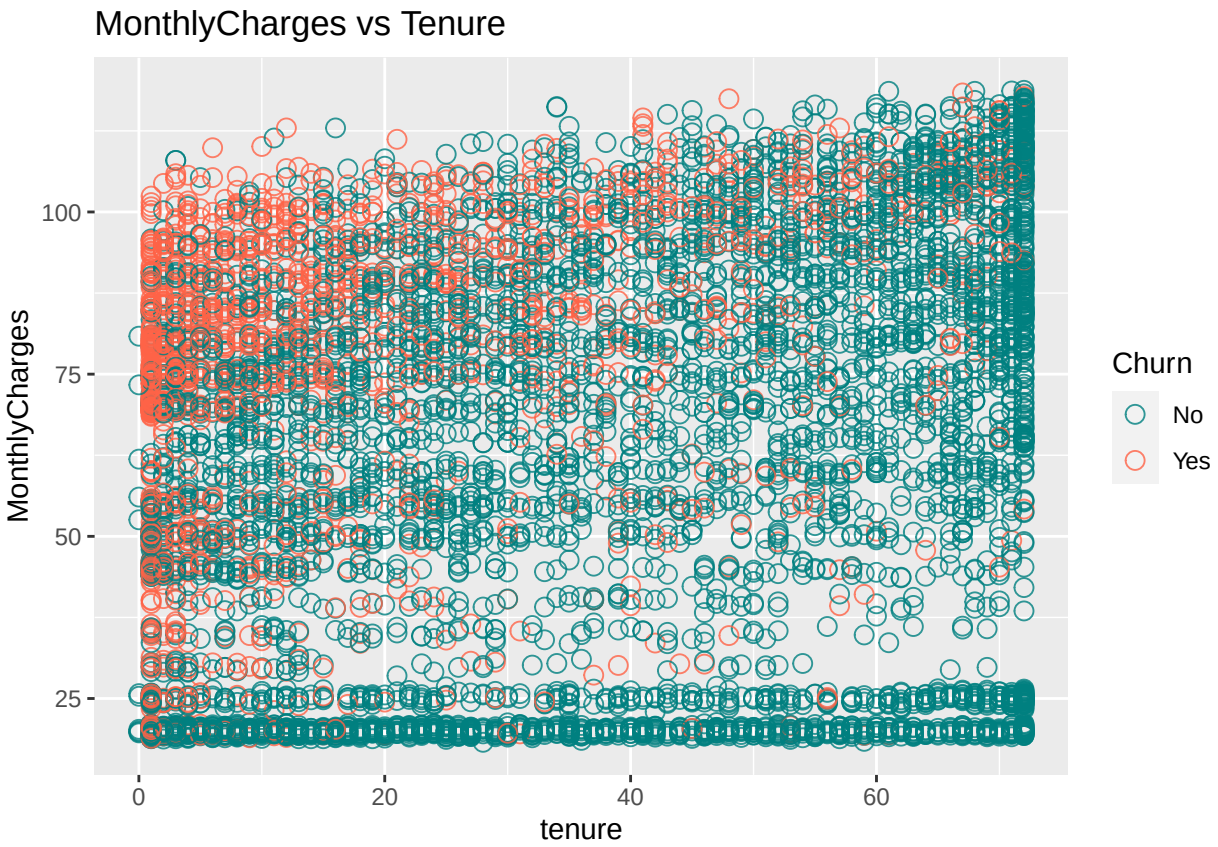
```
# Assuming df is your data frame

library(ggplot2)

# Define the color palette
palette2 <- c('#008080', '#FF6347')

# Scatterplot for TotalCharges vs tenure
ggplot(df, aes(x = TotalCharges, y = tenure, color = Churn)) +
  geom_point(size = 3, alpha = 0.8, shape = 1, stroke = 0.5) +
  scale_color_manual(values = palette2) +
  labs(title = 'TotalCharges vs Tenure')
```




```
# Scatterplot for TotalCharges vs MonthlyCharges
ggplot(df, aes(x = TotalCharges, y = MonthlyCharges, color = Churn)) +
  geom_point(size = 3, alpha = 0.8, shape = 1, stroke = 0.5) +
  scale_color_manual(values = palette2) +
  labs(title = 'TotalCharges vs MonthlyCharges')
```




```
# Scatterplot for MonthlyCharges vs tenure
ggplot(df, aes(x = tenure, y = MonthlyCharges, color = Churn)) +
  geom_point(size = 3, alpha = 0.8, shape = 1, stroke = 0.5) +
  scale_color_manual(values = palette2) +
  labs(title = 'MonthlyCharges vs Tenure')
```



- Many customers leave after just one month.
- A significant number of customers stay with the provider for 72 weeks.
- Customers tend to stay longer if they have a contract, especially those with higher mean scores.
- Customers with higher Monthly Charges are more likely to leave the service.
- The longer a customer has been with the provider, the less likely they are to leave
- Major customers who moved out were having Electronic Check as Payment Method.
- Customers who opted for Credit-Card automatic transfer or Bank Automatic Transfer and Mailed Check as Payment Method were less likely to move out.

```
# Plotting Senior Citizen Distribution

library(tidyverse)
```

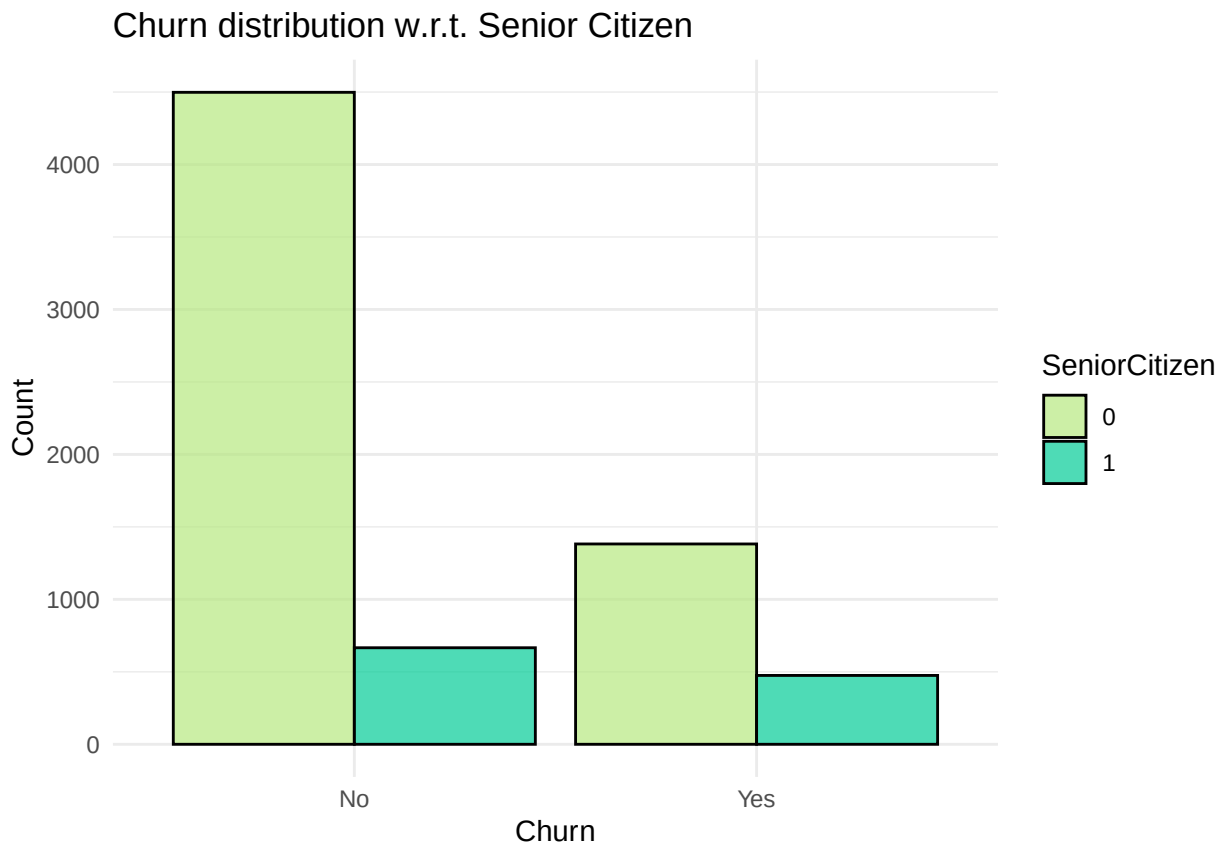
```
## Warning: package 'tidyverse' was built under R version 4.3.2

## Warning: package 'stringr' was built under R version 4.3.2

## Warning: package 'forcats' was built under R version 4.3.2
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.3     v readr     2.1.4
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.0.2     v tidyr     1.3.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
## i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

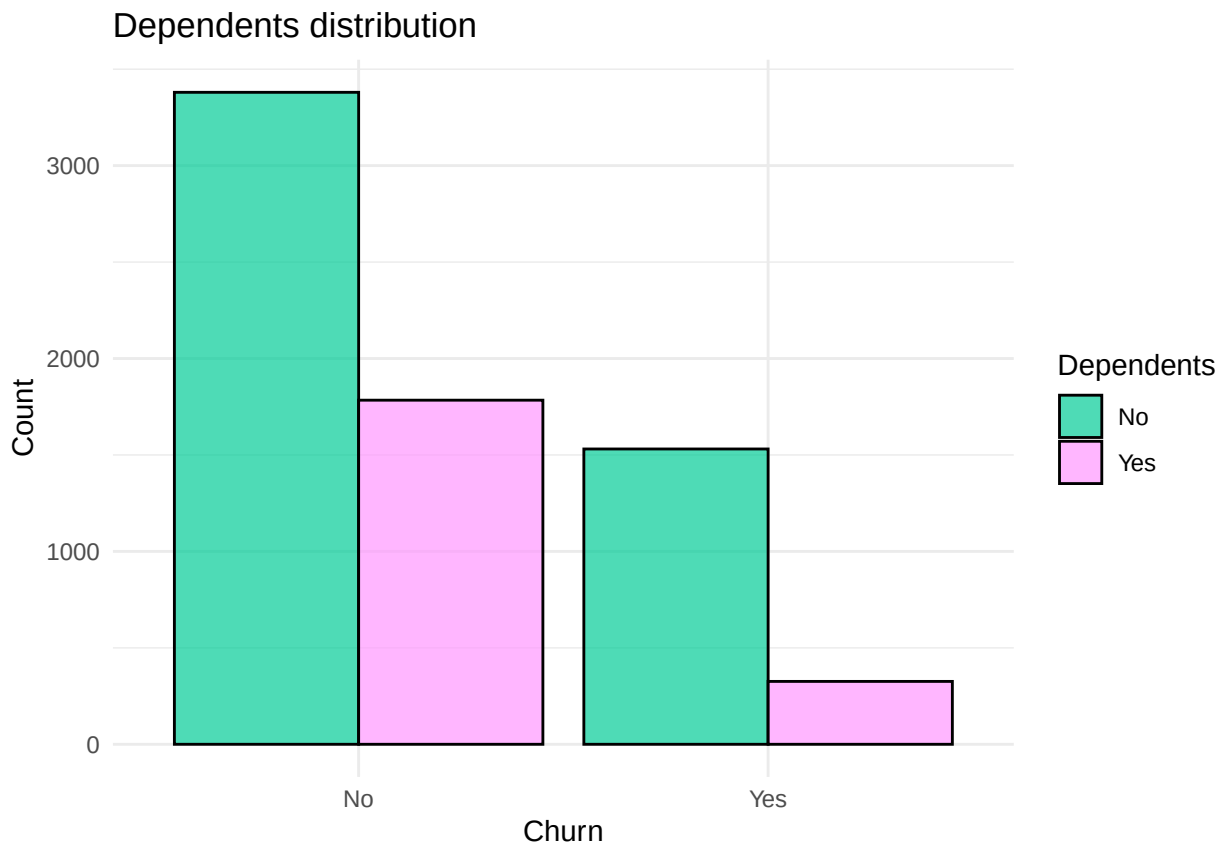
```
library(magrittr)
```

```
##
## Attaching package: 'magrittr'
##
## The following object is masked from 'package:purrr':
##
##     set_names
##
## The following object is masked from 'package:tidyr':
##
##     extract
```

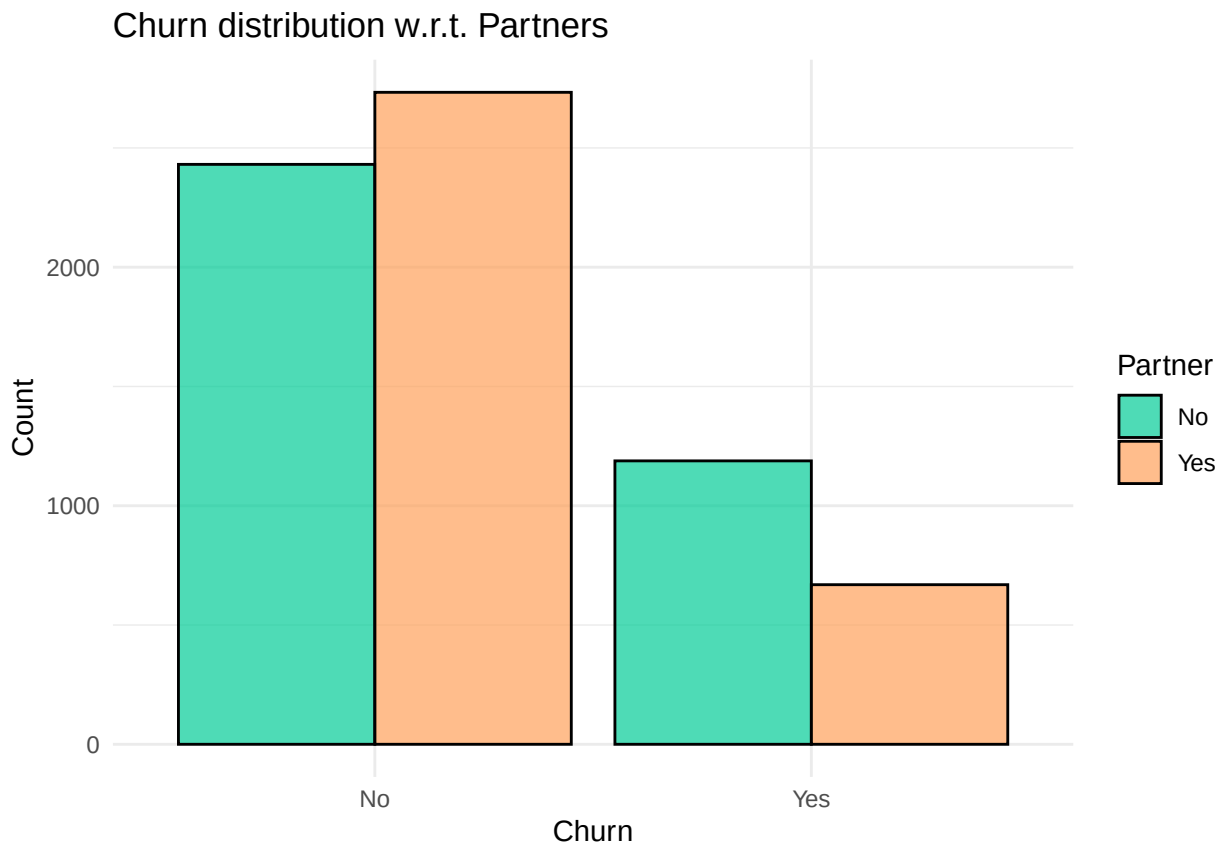
```
# Convert SeniorCitizen to factor
df$SeniorCitizen <- factor(df$SeniorCitizen)
color_map <- c("1" = '#00CC96', "0" = '#B6E880')
df %>%
  ggplot(aes(x = Churn, fill = SeniorCitizen)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  labs(title = "Churn distribution w.r.t. Senior Citizen") +
  xlab("Churn") +
  ylab("Count") +
  scale_fill_manual(values = color_map) +
  theme_minimal()
```



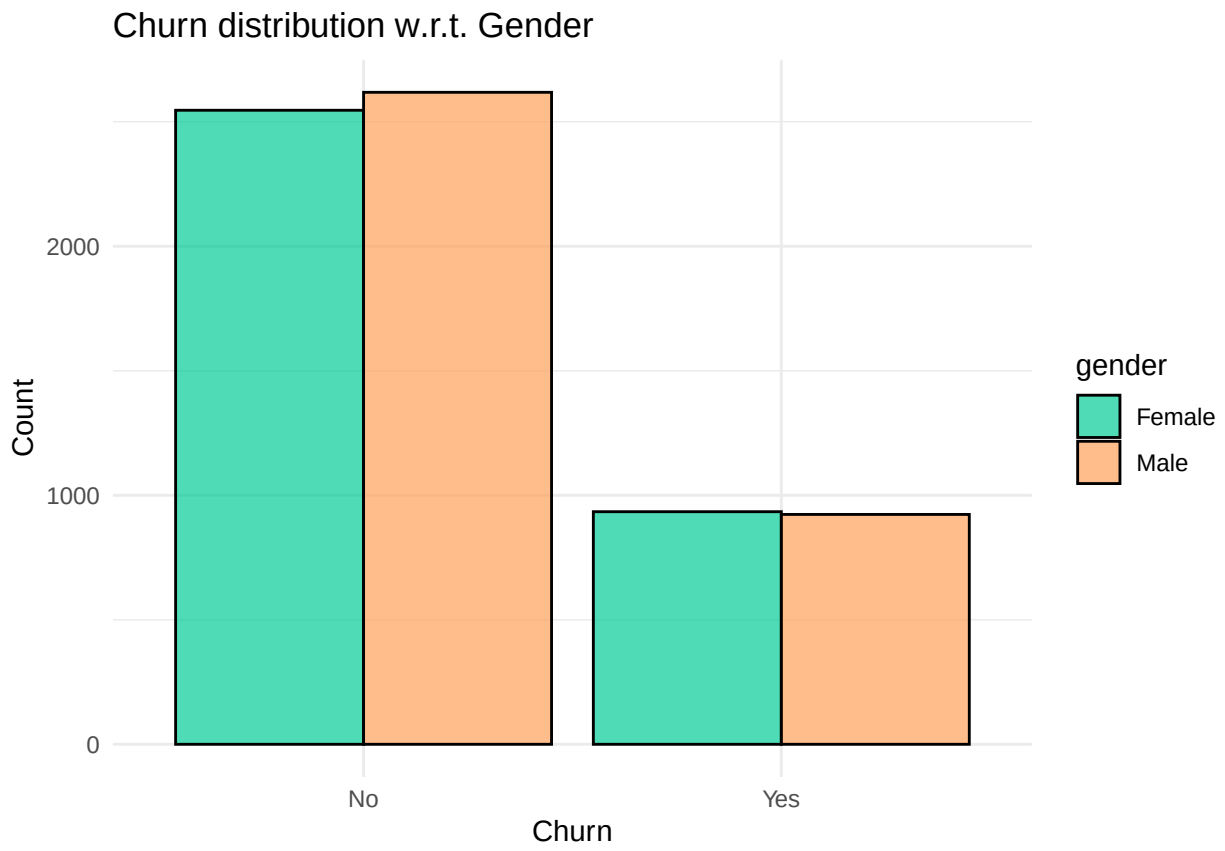

```r
# Plotting Dependents Distribution
color_map <- c("Yes" = "#FF97FF", "No" = "#00CC96")
df %>%
  ggplot(aes(x = Churn, fill = Dependents)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  labs(title = "Dependents distribution") +
  xlab("Churn") +
  ylab("Count") +
  scale_fill_manual(values = color_map) +
  theme_minimal()
```




```r
# Plotting Partner Distribution
color_map <- c("Yes" = "#FFA15A", "No" = "#00CC96")
df %>%
  ggplot(aes(x = Churn, fill = Partner)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  labs(title = "Churn distribution w.r.t. Partners") +
  xlab("Churn") +
  ylab("Count") +
  scale_fill_manual(values = color_map) +
  theme_minimal()
```

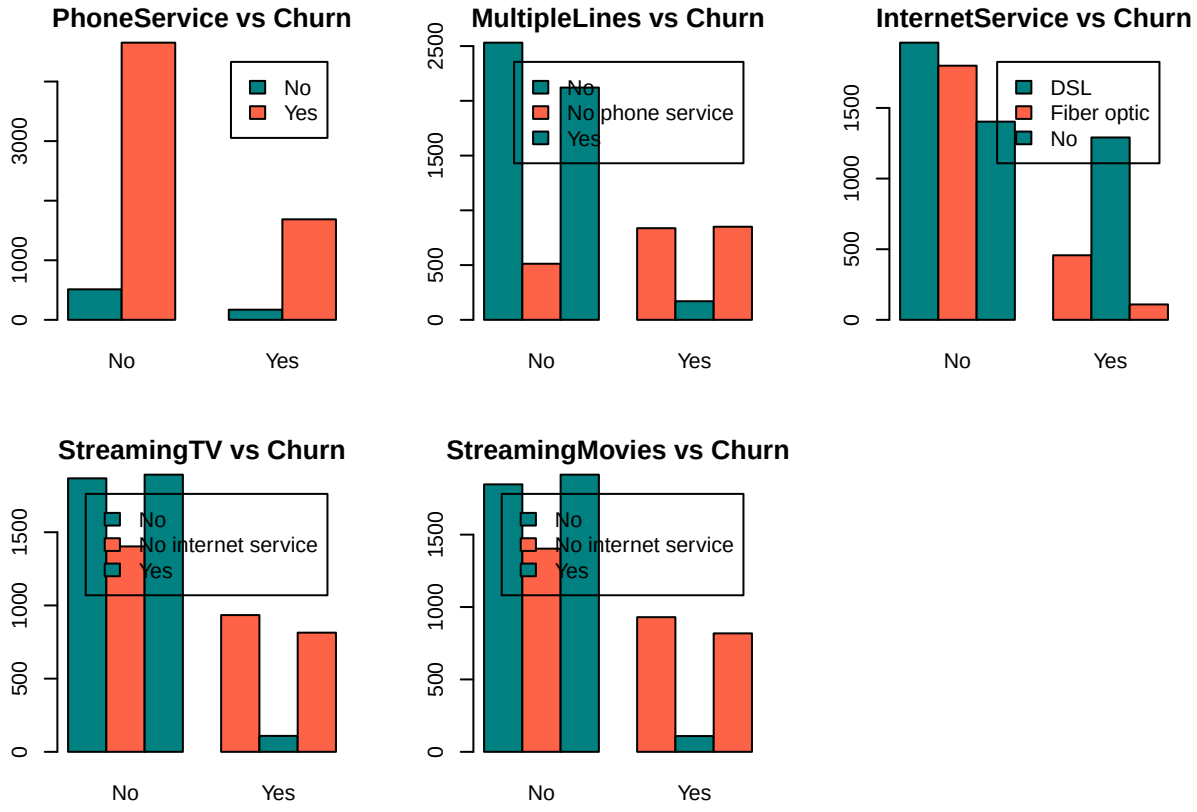



```r
# Plotting Gender Distribution
color_map <- c("Male" = "#FFA15A", "Female" = "#00CC96")
df %>%
  ggplot(aes(x = Churn, fill = gender)) +
  geom_bar(position = "dodge", color = "black", alpha = 0.7) +
  labs(title = "Churn distribution w.r.t. Gender") +
  xlab("Churn") +
  ylab("Count") +
  scale_fill_manual(values = color_map) +
  theme_minimal()
```



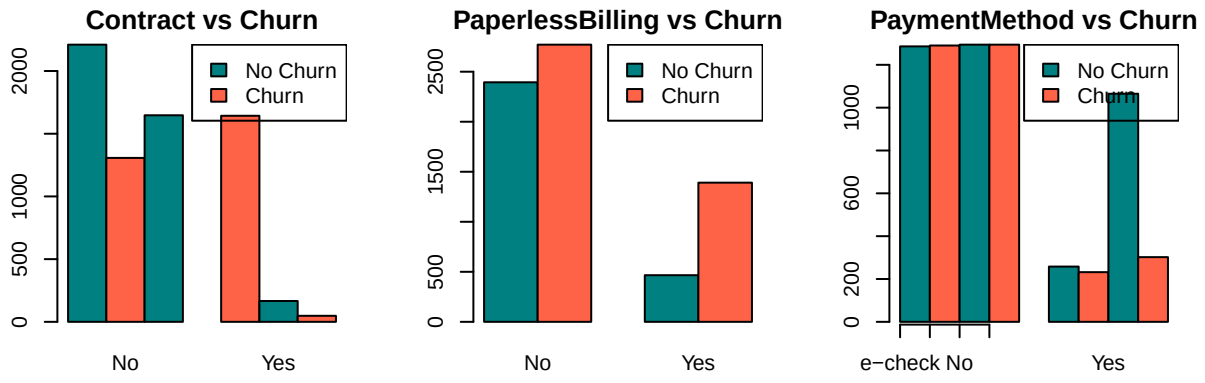
Provided services PhoneService, MultipleLines, InternetService, StreamingTV, StreamingMovies

```
# List of features
list2 <- c('PhoneService', 'MultipleLines', 'InternetService', 'StreamingTV', 'StreamingMovies')

# Create a multi-plot layout
par(mfrow=c(2,3), mar=c(4,4,2,1))

# Plot each feature against Churn
for (i in 1:length(list2)) {
  # Create a count plot
  counts <- table(df[, list2[i]], df$Churn)
  barplot(counts, beside=TRUE, col=palette2, main=paste(list2[i], 'vs Churn'), legend.text=TRUE)
}
```




```r
# Support services
list3 <- c('OnlineSecurity', 'OnlineBackup', 'DeviceProtection', 'TechSupport')
palette <- c('#008080', '#FF6347', '#E50000', '#D2691E')

# Create a multi-plot layout
par(mfrow=c(2,2), mar=c(4,4,2,1))

# Plot each support service against Churn using the specified palette
for (i in 1:length(list3)) {
  # Create a count plot
  counts <- table(df[, list3[i]], df$Churn)
  barplot(counts, beside=TRUE, col=palette2, main=paste(list3[i], 'vs Churn'), legend.text=TRUE)
}
```

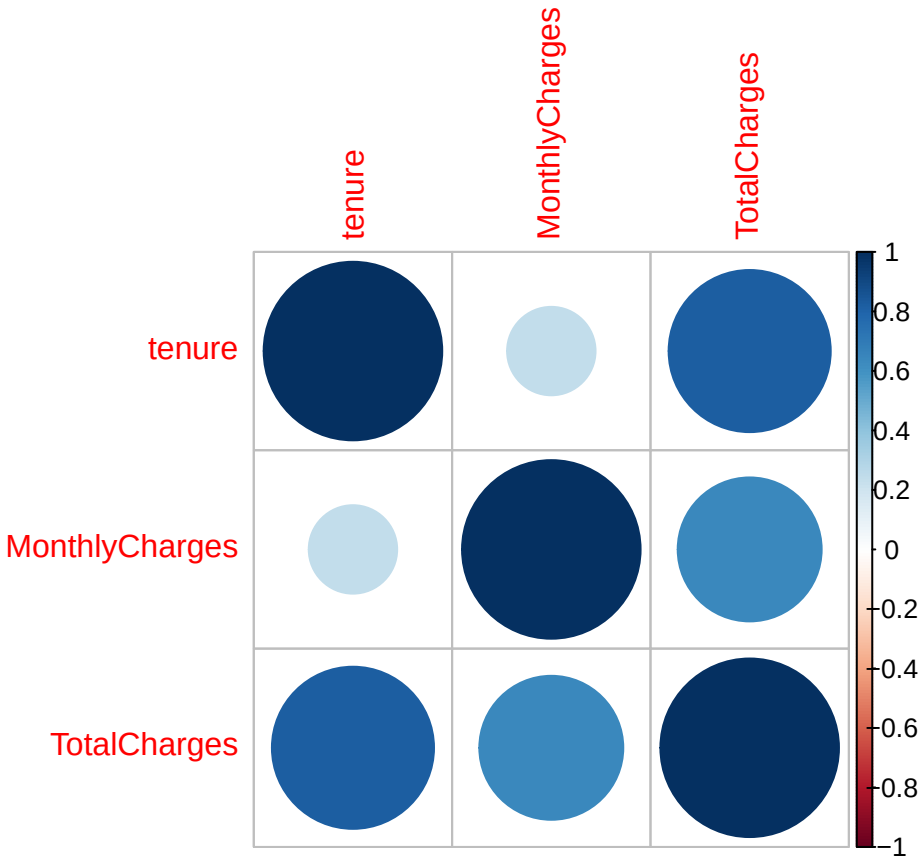

```r
# Specify the columns of interest
list3 <- c('Contract', 'PaperlessBilling', 'PaymentMethod')

# Create a multi-plot layout
par(mfrow=c(2,3), mar=c(4,4,2,1))

# Plot each column against Churn using the specified palette
for (i in 1:length(list3)) {
  # Create a count plot
  counts <- table(df[[list3[i]]], df$Churn)

  # Barplot with custom colors
  barplot(counts, beside=TRUE, col=palette2, main=paste(list3[i], 'vs Churn'))

  # Customize legend and x-axis labels for PaymentMethod
  if (list3[i] == 'PaymentMethod') {
    legend('topright', legend=c('No Churn', 'Churn'), fill=palette2)
    axis(1, at=1:4, labels=c('e-check', 'm-check', 'Bank transfer', 'Credit card'))
  } else {
    legend('topright', legend=c('No Churn', 'Churn'), fill=palette2)
  }
}
```



In simpler words:

1. Many customers opt for Fiber optic internet, but they tend to leave the service more often. This suggests they might not be happy with this type of internet.
2. Most customers use DSL service, and they tend to stay with the provider without leaving as much as Fiber optic users.
3. Customers without dependents are more likely to leave the service.
4. Customers without partners are more likely to leave the service.
5. There are very few senior citizens among the customers, but most of them leave the service.
6. Most customers leave if they don't have online security.
7. Customers with Paperless Billing are more likely to leave.
8. Customers without TechSupport are more likely to switch to another provider.
9. A very small number of customers don't have phone service, and among them, one-third are likely to leave the service.
10. The relationship between gender and the likelihood of churn is not significant. However, more interesting insights can be gained by exploring the features related to having a partner, dependents, and being a senior citizen.

```r
# Select Numeric Variables
numeric_df <- df %>%
  select_if(is.numeric)

# Calculate Correlation Matrix
corr_df <- cor(numeric_df)

# Correlation Plot
corrplot::corrplot(corr_df, method = "circle")
```




```r
# Extract Correlation with the Target Variable (Assuming it's the first variable)
corr_target <- corr_df[1, -1]

# Sort Correlation Values
corr_sorted <- sort(corr_target, decreasing = TRUE)

# Select Top Correlated Variables (Adjust the threshold as needed)
top_corr_vars <- names(corr_sorted[abs(corr_sorted) > 0.1])

top_corr_vars
```

```
## [1] "TotalCharges"   "MonthlyCharges"
```

```r
# Select the columns to be used in the regression
columns <- c(
  "gender", "SeniorCitizen", "Partner", "Dependents",
  "PhoneService", "MultipleLines", "InternetService", "OnlineSecurity",
  "OnlineBackup", "DeviceProtection", "TechSupport", "StreamingTV",
  "StreamingMovies", "Contract", "PaperlessBilling", "PaymentMethod",
   "Churn", "tenure.group"
)

# Create a new dataframe with only the selected columns
df_selected <- df[, columns]

df1 <- df[, columns]

# Convert specified columns to factors
df_selected[columns] <- lapply(df_selected[columns], as.factor)


df_selected$Churn <- as.factor(df_selected$Churn)
df$Churn <- as.factor(df$Churn)
df1$Churn <- as.factor(df_selected$Churn)
# Check levels of all categorical variables
sapply(df_selected, function(x) if (is.factor(x)) length(levels(x)) else NA)
```

```
##           gender    SeniorCitizen          Partner       Dependents
##                2                2                2                2
##     PhoneService    MultipleLines  InternetService   OnlineSecurity
##                2                3                3                3
##     OnlineBackup DeviceProtection      TechSupport      StreamingTV
##                3                3                3                3
##  StreamingMovies         Contract PaperlessBilling    PaymentMethod
##                3                3                2                4
##            Churn     tenure.group
##                2                4
```

```r
# Perform multiple logistic regression to determine impactful factors
model <- glm(Churn ~ ., data = df_selected, family = "binomial")
# summary(model) # view summary of the model

# Obtain p-values for each variable in the model
p_values <- summary(model)$coefficients[, 4]

# Determine significant variables based on p-values less than 0.05
sig_vars <- names(p_values[p_values < 0.05])

sig_vars
```

```
##  [1] "SeniorCitizen1"               "PhoneServiceYes"
##  [3] "MultipleLinesYes"             "InternetServiceFiber optic"
##  [5] "InternetServiceNo"            "OnlineSecurityYes"
##  [7] "OnlineBackupYes"              "TechSupportYes"
##  [9] "StreamingTVYes"               "StreamingMoviesYes"
## [11] "ContractOne year"             "ContractTwo year"
```

```
## [13] "PaperlessBillingYes"           "PaymentMethodElectronic check"
## [15] "tenure.group12-24 months"      "tenure.group24-48 months"
## [17] "tenure.groupOver 48 months"
```

```
# Forward selection
full_model <- glm(Churn ~ ., family = binomial(link = 'logit'), data = df)
forward_model <- step(full_model, direction = 'forward')
```

```
## Start:  AIC=5826.13
## Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
##     PhoneService + MultipleLines + InternetService + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group
```

```
summary(forward_model)
```

```
##
## Call:
## glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents +
##     tenure + PhoneService + MultipleLines + InternetService +
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges + TotalCharges + tenure.group,
##     family = binomial(link = "logit"), data = df)
##
## Coefficients: (7 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          1.064e+00  8.185e-01   1.300  0.19359
## genderMale                          -2.215e-02  6.523e-02  -0.340  0.73412
## SeniorCitizen1                       2.263e-01  8.479e-02   2.668  0.00762 **
## PartnerYes                           9.365e-03  7.817e-02   0.120  0.90464
## DependentsYes                       -1.332e-01  9.008e-02  -1.478  0.13934
## tenure                              -6.206e-02  9.212e-03  -6.737 1.62e-11 ***
## PhoneServiceYes                      1.130e-01  6.507e-01   0.174  0.86211
## MultipleLinesNo phone service               NA         NA      NA       NA
## MultipleLinesYes                     4.775e-01  1.779e-01   2.683  0.00729 **
## InternetServiceFiber optic           1.702e+00  8.004e-01   2.126  0.03350 *
## InternetServiceNo                   -1.672e+00  8.105e-01  -2.063  0.03914 *
## OnlineSecurityNo internet service           NA         NA      NA       NA
## OnlineSecurityYes                   -1.900e-01  1.791e-01  -1.061  0.28879
## OnlineBackupNo internet service             NA         NA      NA       NA
## OnlineBackupYes                      2.242e-02  1.760e-01   0.127  0.89864
## DeviceProtectionNo internet service         NA         NA      NA       NA
## DeviceProtectionYes                  1.649e-01  1.768e-01   0.933  0.35076
## TechSupportNo internet service              NA         NA      NA       NA
## TechSupportYes                      -1.693e-01  1.811e-01  -0.935  0.34993
## StreamingTVNo internet service              NA         NA      NA       NA
## StreamingTVYes                       5.917e-01  3.273e-01   1.808  0.07060 .
## StreamingMoviesNo internet service          NA         NA      NA       NA
## StreamingMoviesYes                   5.943e-01  3.276e-01   1.814  0.06966 .
## ContractOne year                    -6.914e-01  1.084e-01  -6.375 1.83e-10 ***
## ContractTwo year                    -1.432e+00  1.801e-01  -7.949 1.88e-15 ***
```

```
## PaperlessBillingYes                     3.413e-01  7.510e-02   4.545 5.48e-06 ***
## PaymentMethodCredit card (automatic) -8.117e-02  1.140e-01  -0.712  0.47653
## PaymentMethodElectronic check         2.853e-01  9.469e-02   3.013  0.00258 **
## PaymentMethodMailed check            -8.050e-02  1.158e-01  -0.695  0.48703
## MonthlyCharges                       -3.557e-02  3.187e-02  -1.116  0.26431
## TotalCharges                          1.931e-04  7.336e-05   2.633  0.00847 **
## tenure.group12-24 months             -2.596e-01  1.305e-01  -1.989  0.04670 *
## tenure.group24-48 months              1.345e-01  2.169e-01   0.620  0.53534
## tenure.groupOver 48 months            7.608e-01  3.659e-01   2.079  0.03758 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8105.3  on 7009  degrees of freedom
## Residual deviance: 5772.1  on 6983  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 5826.1
##
## Number of Fisher Scoring iterations: 6
```

Certainly! Here's a brief explanation of the logistic regression model summary:

1. **Model Overview:**
   - **AIC:** The AIC value is a measure of model fit, and a lower AIC suggests a better-fitted model. The AIC for this model is 5826.1.
   - **Deviance:** The deviance is a measure of how well the model predicts the data. The residual deviance is 5772.1, indicating an improvement over the null deviance (8105.3).
2. **Coefficients:**
   - The coefficients represent the log-odds of the response variable (Churn).
   - Positive coefficients increase the log-odds of the event, while negative coefficients decrease it.
3. **Significant Predictors:**
   - **SeniorCitizen:** Senior citizens are more likely to churn (positive coefficient).
   - **Tenure:** Longer tenure reduces the likelihood of churn (negative coefficient).
   - **InternetService(Fiber optic):** Fiber optic service increases the likelihood of churn (positive coefficient).
   - **Contract (One year, Two years):** Longer contract terms reduce the likelihood of churn (negative coefficients).
   - **PaperlessBilling:** Paperless billing increases the likelihood of churn (positive coefficient).
   - **PaymentMethod(Electronic check):** Electronic check payment method increases the likelihood of churn (positive coefficient).
   - **TotalCharges:** Higher total charges increase the likelihood of churn (positive coefficient).
4. **Variables with No Internet Service:**
   - Several variables related to internet services have a category "No internet service," and their coefficients are marked as `NA`.
5. **Iterations:**
   - The model fitting process involved six iterations.

In summary, the model identifies key predictors influencing customer churn, considering factors like contract length, billing methods, and internet service types.

```r
# Backward selection
full_model <- glm(Churn ~ ., family = binomial(link = 'logit'), data = df)
backward_model <- step(full_model, direction = 'backward')
```

```
## Start:  AIC=5826.13
## Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
##     PhoneService + MultipleLines + InternetService + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group
##
##
## Step:  AIC=5826.13
## Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
##     MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
##     DeviceProtection + TechSupport + StreamingTV + StreamingMovies +
##     Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +
##     TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## - Partner           1   5772.1 5824.1
## - OnlineBackup      1   5772.1 5824.1
## - gender            1   5772.2 5824.2
## - DeviceProtection  1   5773.0 5825.0
## - TechSupport       1   5773.0 5825.0
## - OnlineSecurity    1   5773.3 5825.3
## - MonthlyCharges    1   5773.4 5825.4
## <none>                  5772.1 5826.1
## - Dependents        1   5774.3 5826.3
## - StreamingTV       1   5775.4 5827.4
## - StreamingMovies   1   5775.4 5827.4
## - InternetService   1   5776.7 5828.7
## - SeniorCitizen     1   5779.2 5831.2
## - TotalCharges      1   5779.3 5831.3
## - PaymentMethod     3   5796.1 5844.1
## - PaperlessBilling  1   5792.9 5844.9
## - MultipleLines     2   5796.6 5846.6
## - tenure.group      3   5802.0 5850.0
## - tenure            1   5820.5 5872.5
## - Contract          2   5859.9 5909.9
##
## Step:  AIC=5824.15
## Churn ~ gender + SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + OnlineBackup + DeviceProtection +
##     TechSupport + StreamingTV + StreamingMovies + Contract +
##     PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges +
##     tenure.group
##
##                    Df Deviance    AIC
## - OnlineBackup      1   5772.2 5822.2
## - gender            1   5772.3 5822.3
## - TechSupport       1   5773.0 5823.0
## - DeviceProtection  1   5773.0 5823.0
```

```
## - OnlineSecurity    1   5773.3 5823.3
## - MonthlyCharges    1   5773.4 5823.4
## <none>                  5772.1 5824.1
## - Dependents        1   5774.6 5824.6
## - StreamingTV       1   5775.4 5825.4
## - StreamingMovies   1   5775.5 5825.5
## - InternetService   1   5776.7 5826.7
## - TotalCharges      1   5779.4 5829.4
## - SeniorCitizen     1   5779.4 5829.4
## - PaymentMethod     3   5796.2 5842.2
## - PaperlessBilling  1   5792.9 5842.9
## - MultipleLines     2   5796.6 5844.6
## - tenure.group      3   5802.0 5848.0
## - tenure            1   5820.6 5870.6
## - Contract          2   5860.0 5908.0
##
## Step:  AIC=5822.16
## Churn ~ gender + SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges + TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## - gender            1   5772.3 5820.3
## - DeviceProtection  1   5774.1 5822.1
## <none>                  5772.2 5822.2
## - Dependents        1   5774.6 5822.6
## - TechSupport       1   5775.0 5823.0
## - OnlineSecurity    1   5775.7 5823.7
## - MonthlyCharges    1   5777.3 5825.3
## - TotalCharges      1   5779.4 5827.4
## - SeniorCitizen     1   5779.4 5827.4
## - StreamingTV       1   5784.2 5832.2
## - StreamingMovies   1   5784.4 5832.4
## - PaymentMethod     3   5796.2 5840.2
## - InternetService   1   5792.3 5840.3
## - PaperlessBilling  1   5793.0 5841.0
## - MultipleLines     2   5797.6 5843.6
## - tenure.group      3   5802.1 5846.1
## - tenure            1   5820.6 5868.6
## - Contract          2   5860.0 5906.0
##
## Step:  AIC=5820.28
## Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges + TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## - DeviceProtection  1   5774.2 5820.2
## <none>                  5772.3 5820.3
## - Dependents        1   5774.8 5820.8
## - TechSupport       1   5775.1 5821.1
## - OnlineSecurity    1   5775.8 5821.8
```

```
## - MonthlyCharges    1   5777.4 5823.4
## - TotalCharges      1   5779.5 5825.5
## - SeniorCitizen     1   5779.5 5825.5
## - StreamingTV       1   5784.3 5830.3
## - StreamingMovies   1   5784.5 5830.5
## - PaymentMethod     3   5796.3 5838.3
## - InternetService   1   5792.4 5838.4
## - PaperlessBilling  1   5793.1 5839.1
## - MultipleLines     2   5797.7 5841.7
## - tenure.group      3   5802.2 5844.2
## - tenure            1   5820.7 5866.7
## - Contract          2   5860.1 5904.1
##
## Step:  AIC=5820.21
## Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## <none>                  5774.2 5820.2
## - Dependents        1   5776.8 5820.8
## - MonthlyCharges    1   5777.4 5821.4
## - TechSupport       1   5780.1 5824.1
## - TotalCharges      1   5781.5 5825.5
## - SeniorCitizen     1   5781.5 5825.5
## - OnlineSecurity    1   5781.8 5825.8
## - StreamingTV       1   5784.6 5828.6
## - StreamingMovies   1   5784.9 5828.9
## - PaymentMethod     3   5798.0 5838.0
## - PaperlessBilling  1   5794.8 5838.8
## - InternetService   1   5796.1 5840.1
## - MultipleLines     2   5798.3 5840.3
## - tenure.group      3   5803.6 5843.6
## - tenure            1   5822.5 5866.5
## - Contract          2   5860.7 5902.7
```

```
summary(backward_model)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients: (4 not defined because of singularities)
##                                      Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          4.248e-01  5.188e-01   0.819  0.41290
## SeniorCitizen1                       2.273e-01  8.426e-02   2.698  0.00698 **
## DependentsYes                       -1.300e-01  8.181e-02  -1.589  0.11198
## tenure                              -6.187e-02  9.191e-03  -6.731 1.68e-11 ***
## MultipleLinesNo phone service        2.186e-01  2.434e-01   0.898  0.36922
```

```
## MultipleLinesYes                       3.937e-01  9.534e-02   4.130 3.63e-05 ***
## InternetServiceFiber optic             1.293e+00  2.772e-01   4.664 3.10e-06 ***
## InternetServiceNo                     -1.259e+00  3.176e-01  -3.964 7.37e-05 ***
## OnlineSecurityNo internet service             NA         NA      NA       NA
## OnlineSecurityYes                     -2.734e-01  9.921e-02  -2.756  0.00585 **
## TechSupportNo internet service                NA         NA      NA       NA
## TechSupportYes                        -2.487e-01  1.024e-01  -2.430  0.01512 *
## StreamingTVNo internet service                NA         NA      NA       NA
## StreamingTVYes                         4.359e-01  1.355e-01   3.217  0.00130 **
## StreamingMoviesNo internet service            NA         NA      NA       NA
## StreamingMoviesYes                     4.392e-01  1.348e-01   3.259  0.00112 **
## ContractOne year                      -6.833e-01  1.082e-01  -6.314 2.72e-10 ***
## ContractTwo year                      -1.418e+00  1.798e-01  -7.888 3.07e-15 ***
## PaperlessBillingYes                    3.392e-01  7.503e-02   4.522 6.14e-06 ***
## PaymentMethodCredit card (automatic) -8.156e-02  1.139e-01  -0.716  0.47398
## PaymentMethodElectronic check          2.829e-01  9.463e-02   2.990  0.00279 **
## PaymentMethodMailed check             -8.232e-02  1.156e-01  -0.712  0.47656
## MonthlyCharges                        -1.908e-02  1.065e-02  -1.791  0.07324 .
## TotalCharges                           1.934e-04  7.331e-05   2.638  0.00833 **
## tenure.group12-24 months              -2.611e-01  1.304e-01  -2.002  0.04525 *
## tenure.group24-48 months               1.297e-01  2.167e-01   0.599  0.54937
## tenure.groupOver 48 months             7.451e-01  3.655e-01   2.039  0.04148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8105.3  on 7009  degrees of freedom
## Residual deviance: 5774.2  on 6987  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 5820.2
##
## Number of Fisher Scoring iterations: 6
```

In the backward logistic regression:

- **Step 1:**
  - Removed "Partner" variable.
  - The AIC reduced from 5826.13 to 5824.15.
  - The model without "Partner" is preferred.
- **Step 2:**
  - Removed "OnlineBackup" variable.
  - The AIC further reduced to 5822.16.
  - The model without "OnlineBackup" is preferred.
- **Step 3:**
  - Removed "gender" variable.
  - The AIC reduced to 5820.28.
  - The model without "gender" is preferred.
- **Step 4:**
  - Removed "DeviceProtection" variable.

  - The AIC reached 5820.21.
  - The model without "DeviceProtection" is preferred.

- **Final Model:**
  - The selected variables are:
    * SeniorCitizen
    * Dependents
    * Tenure
    * MultipleLines
    * InternetService
    * OnlineSecurity
    * TechSupport
    * StreamingTV
    * StreamingMovies
    * Contract
    * PaperlessBilling
    * PaymentMethod
    * MonthlyCharges
    * TotalCharges
    * Tenure group

- **Model Summary:**
  - The model's AIC is 5820.2, indicating a good fit.
  - Deviance dropped from 8105.3 (null model) to 5774.2.
  - Significant predictors include SeniorCitizen, Tenure, MultipleLines, InternetService, OnlineSecurity, TechSupport, StreamingTV, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, and Tenure group.

- **Interpretation:**
  - Senior citizens are more likely to churn.
  - Longer tenure reduces the likelihood of churn.
  - Fiber optic internet service increases churn.
  - Having OnlineSecurity, TechSupport, and certain payment methods decreases churn.
  - PaperlessBilling and certain contract types increase churn.
  - MonthlyCharges and TotalCharges also impact churn.

This final model captures the essential predictors for customer churn in a concise form.

```
# Stepwise selection
full_model <- glm(Churn ~ ., family = binomial(link = 'logit'), data = df)
stepwise_model <- step(full_model)
```

```
## Start:  AIC=5826.13
## Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
##     PhoneService + MultipleLines + InternetService + OnlineSecurity +
##     OnlineBackup + DeviceProtection + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group
##
##
## Step:  AIC=5826.13
## Churn ~ gender + SeniorCitizen + Partner + Dependents + tenure +
##     MultipleLines + InternetService + OnlineSecurity + OnlineBackup +
```

```
##     DeviceProtection + TechSupport + StreamingTV + StreamingMovies +
##     Contract + PaperlessBilling + PaymentMethod + MonthlyCharges +
##     TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## - Partner           1   5772.1 5824.1
## - OnlineBackup      1   5772.1 5824.1
## - gender            1   5772.2 5824.2
## - DeviceProtection  1   5773.0 5825.0
## - TechSupport       1   5773.0 5825.0
## - OnlineSecurity    1   5773.3 5825.3
## - MonthlyCharges    1   5773.4 5825.4
## <none>                  5772.1 5826.1
## - Dependents        1   5774.3 5826.3
## - StreamingTV       1   5775.4 5827.4
## - StreamingMovies   1   5775.4 5827.4
## - InternetService   1   5776.7 5828.7
## - SeniorCitizen     1   5779.2 5831.2
## - TotalCharges      1   5779.3 5831.3
## - PaymentMethod     3   5796.1 5844.1
## - PaperlessBilling  1   5792.9 5844.9
## - MultipleLines     2   5796.6 5846.6
## - tenure.group      3   5802.0 5850.0
## - tenure            1   5820.5 5872.5
## - Contract          2   5859.9 5909.9
##
## Step:  AIC=5824.15
## Churn ~ gender + SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + OnlineBackup + DeviceProtection +
##     TechSupport + StreamingTV + StreamingMovies + Contract +
##     PaperlessBilling + PaymentMethod + MonthlyCharges + TotalCharges +
##     tenure.group
##
##                    Df Deviance    AIC
## - OnlineBackup      1   5772.2 5822.2
## - gender            1   5772.3 5822.3
## - TechSupport       1   5773.0 5823.0
## - DeviceProtection  1   5773.0 5823.0
## - OnlineSecurity    1   5773.3 5823.3
## - MonthlyCharges    1   5773.4 5823.4
## <none>                  5772.1 5824.1
## - Dependents        1   5774.6 5824.6
## - StreamingTV       1   5775.4 5825.4
## - StreamingMovies   1   5775.5 5825.5
## - InternetService   1   5776.7 5826.7
## - TotalCharges      1   5779.4 5829.4
## - SeniorCitizen     1   5779.4 5829.4
## - PaymentMethod     3   5796.2 5842.2
## - PaperlessBilling  1   5792.9 5842.9
## - MultipleLines     2   5796.6 5844.6
## - tenure.group      3   5802.0 5848.0
## - tenure            1   5820.6 5870.6
## - Contract          2   5860.0 5908.0
##
```

```
## Step:  AIC=5822.16
## Churn ~ gender + SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges + TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## - gender            1   5772.3 5820.3
## - DeviceProtection  1   5774.1 5822.1
## <none>                  5772.2 5822.2
## - Dependents        1   5774.6 5822.6
## - TechSupport       1   5775.0 5823.0
## - OnlineSecurity    1   5775.7 5823.7
## - MonthlyCharges    1   5777.3 5825.3
## - TotalCharges      1   5779.4 5827.4
## - SeniorCitizen     1   5779.4 5827.4
## - StreamingTV       1   5784.2 5832.2
## - StreamingMovies   1   5784.4 5832.4
## - PaymentMethod     3   5796.2 5840.2
## - InternetService   1   5792.3 5840.3
## - PaperlessBilling  1   5793.0 5841.0
## - MultipleLines     2   5797.6 5843.6
## - tenure.group      3   5802.1 5846.1
## - tenure            1   5820.6 5868.6
## - Contract          2   5860.0 5906.0
##
## Step:  AIC=5820.28
## Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + DeviceProtection + TechSupport +
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling +
##     PaymentMethod + MonthlyCharges + TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## - DeviceProtection  1   5774.2 5820.2
## <none>                  5772.3 5820.3
## - Dependents        1   5774.8 5820.8
## - TechSupport       1   5775.1 5821.1
## - OnlineSecurity    1   5775.8 5821.8
## - MonthlyCharges    1   5777.4 5823.4
## - TotalCharges      1   5779.5 5825.5
## - SeniorCitizen     1   5779.5 5825.5
## - StreamingTV       1   5784.3 5830.3
## - StreamingMovies   1   5784.5 5830.5
## - PaymentMethod     3   5796.3 5838.3
## - InternetService   1   5792.4 5838.4
## - PaperlessBilling  1   5793.1 5839.1
## - MultipleLines     2   5797.7 5841.7
## - tenure.group      3   5802.2 5844.2
## - tenure            1   5820.7 5866.7
## - Contract          2   5860.1 5904.1
##
## Step:  AIC=5820.21
## Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
```

```
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group
##
##                    Df Deviance    AIC
## <none>                  5774.2 5820.2
## - Dependents        1   5776.8 5820.8
## - MonthlyCharges    1   5777.4 5821.4
## - TechSupport       1   5780.1 5824.1
## - TotalCharges      1   5781.5 5825.5
## - SeniorCitizen     1   5781.5 5825.5
## - OnlineSecurity    1   5781.8 5825.8
## - StreamingTV       1   5784.6 5828.6
## - StreamingMovies   1   5784.9 5828.9
## - PaymentMethod     3   5798.0 5838.0
## - PaperlessBilling  1   5794.8 5838.8
## - InternetService   1   5796.1 5840.1
## - MultipleLines     2   5798.3 5840.3
## - tenure.group      3   5803.6 5843.6
## - tenure            1   5822.5 5866.5
## - Contract          2   5860.7 5902.7
```

```
summary(stepwise_model)
```

```
##
## Call:
## glm(formula = Churn ~ SeniorCitizen + Dependents + tenure + MultipleLines +
##     InternetService + OnlineSecurity + TechSupport + StreamingTV +
##     StreamingMovies + Contract + PaperlessBilling + PaymentMethod +
##     MonthlyCharges + TotalCharges + tenure.group, family = binomial(link = "logit"),
##     data = df)
##
## Coefficients: (4 not defined because of singularities)
##                                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          4.248e-01  5.188e-01   0.819  0.41290
## SeniorCitizen1                       2.273e-01  8.426e-02   2.698  0.00698 **
## DependentsYes                       -1.300e-01  8.181e-02  -1.589  0.11198
## tenure                              -6.187e-02  9.191e-03  -6.731 1.68e-11 ***
## MultipleLinesNo phone service        2.186e-01  2.434e-01   0.898  0.36922
## MultipleLinesYes                     3.937e-01  9.534e-02   4.130 3.63e-05 ***
## InternetServiceFiber optic           1.293e+00  2.772e-01   4.664 3.10e-06 ***
## InternetServiceNo                   -1.259e+00  3.176e-01  -3.964 7.37e-05 ***
## OnlineSecurityNo internet service           NA         NA      NA       NA
## OnlineSecurityYes                   -2.734e-01  9.921e-02  -2.756  0.00585 **
## TechSupportNo internet service              NA         NA      NA       NA
## TechSupportYes                      -2.487e-01  1.024e-01  -2.430  0.01512 *
## StreamingTVNo internet service              NA         NA      NA       NA
## StreamingTVYes                       4.359e-01  1.355e-01   3.217  0.00130 **
## StreamingMoviesNo internet service          NA         NA      NA       NA
## StreamingMoviesYes                   4.392e-01  1.348e-01   3.259  0.00112 **
## ContractOne year                    -6.833e-01  1.082e-01  -6.314 2.72e-10 ***
## ContractTwo year                    -1.418e+00  1.798e-01  -7.888 3.07e-15 ***
## PaperlessBillingYes                  3.392e-01  7.503e-02   4.522 6.14e-06 ***
## PaymentMethodCredit card (automatic) -8.156e-02  1.139e-01  -0.716  0.47398
## PaymentMethodElectronic check        2.829e-01  9.463e-02   2.990  0.00279 **
```

```
## PaymentMethodMailed check            -8.232e-02  1.156e-01  -0.712  0.47656
## MonthlyCharges                       -1.908e-02  1.065e-02  -1.791  0.07324 .
## TotalCharges                          1.934e-04  7.331e-05   2.638  0.00833 **
## tenure.group12-24 months             -2.611e-01  1.304e-01  -2.002  0.04525 *
## tenure.group24-48 months              1.297e-01  2.167e-01   0.599  0.54937
## tenure.groupOver 48 months            7.451e-01  3.655e-01   2.039  0.04148 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8105.3  on 7009  degrees of freedom
## Residual deviance: 5774.2  on 6987  degrees of freedom
##   (11 observations deleted due to missingness)
## AIC: 5820.2
##
## Number of Fisher Scoring iterations: 6
```

The stepwise logistic regression results indicate the model building process. Here's a summary of the steps:

1. **Start:** The initial model includes all the variables listed.
   - AIC = 5826.13
2. **Step 1:**
   - Removed the "Partner" variable.
   - AIC reduced to 5824.15.
   - Variables removed: Partner
   - Remaining variables: gender, SeniorCitizen, Dependents, tenure, MultipleLines, InternetService, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, tenure.group.
3. **Step 2:**
   - Removed the "OnlineBackup" variable.
   - AIC further reduced to 5822.16.
   - Variables removed: OnlineBackup
   - Remaining variables: gender, SeniorCitizen, Dependents, tenure, MultipleLines, InternetService, OnlineSecurity, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, tenure.group.
4. **Step 3:**
   - Removed the "gender" variable.
   - AIC reduced to 5820.28.
   - Variables removed: gender
   - Remaining variables: SeniorCitizen, Dependents, tenure, MultipleLines, InternetService, OnlineSecurity, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, tenure.group.
5. **Step 4:**
   - Removed the "DeviceProtection" variable.
   - AIC reached 5820.21.
   - Variables removed: DeviceProtection
   - Remaining variables: SeniorCitizen, Dependents, tenure, MultipleLines, InternetService, OnlineSecurity, TechSupport, StreamingTV, StreamingMovies, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, tenure.group.

6. **Final Model:**
    - The final selected model includes the remaining variables after step 4.
    - AIC = 5820.2

Interpretation of the Final Model: - The final model includes significant predictors such as SeniorCitizen, tenure, MultipleLines, InternetService, OnlineSecurity, TechSupport, StreamingTV, Contract, PaperlessBilling, PaymentMethod, MonthlyCharges, TotalCharges, and tenure.group.

- Each coefficient's estimate, standard error, z-value, and p-value are provided for interpretation.
- The model provides insights into the factors influencing customer churn based on the selected variables.

Final Model Summary:

Predictors in the Model:

SeniorCitizen: Being a senior citizen increases the odds of churn. Dependents: Having dependents decreases the odds of churn. Tenure: Longer tenure decreases the odds of churn. MultipleLines: Having multiple lines increases the odds of churn. InternetService: Having fiber optic internet service increases the odds of churn. OnlineSecurity: Lack of online security increases the odds of churn. TechSupport: Lack of tech support increases the odds of churn. StreamingTV: Having streaming TV increases the odds of churn. Contract: Shorter contract terms (especially month-to-month) increase the odds of churn. PaperlessBilling: Having paperless billing increases the odds of churn. PaymentMethod: Using electronic check for payment increases the odds of churn. MonthlyCharges: Higher monthly charges increase the odds of churn. TotalCharges: Higher total charges decrease the odds of churn. Tenure.Group: Specific groups (12-24 months, Over 48 months) impact churn odds.

Model Evaluation:

The model is built based on the stepwise selection process, optimizing the AIC. AIC (Akaike Information Criterion) is a measure of the model's goodness of fit, penalizing for complexity. Lower AIC indicates a better-fitting model. Practical Insights: Customer Profiles:

Older customers are more likely to churn. Customers with dependents tend to be more loyal. Long-tenured customers are less likely to churn. Services Impact:

Having additional services like multiple lines, streaming TV, and fiber optic internet may increase churn. Lack of online security and tech support contributes to higher churn. Contract and Billing:

Shorter contract terms (month-to-month) and paperless billing are associated with higher churn. Payment method, especially electronic checks, influences customer retention. Financial Factors:

Monthly charges impact churn, with higher charges leading to increased churn. Total charges, however, have a stabilizing effect on customer retention. Recommendations: Retention Strategies:

Target promotions or discounts to senior citizens to encourage loyalty. Enhance services like online security and tech support. Incentivize longer-term contracts and non-electronic payment methods. Address concerns related to multiple lines and streaming services. Communication Strategies:

Tailor communication and promotions based on customer tenure. Educate customers about the benefits of additional services to improve retention. Financial Management:

Monitor and potentially revise pricing strategies, especially for high-churn services. Consider promotions or loyalty programs for long-tenured customers. Remember, these recommendations are based on statistical associations, and the actual business strategies may need to consider additional factors and nuances specific to your industry and customer base.