BDA_Week 6_Logistic Regression

110077443

3/24/2022

Please note that all code in this document is presented in a grey box and the output reflected below each box getwd() - The below code allows lengthy lines of code to display neatly within the grey box (wrapping it)

```
knitr::opts_chunk$set(tidy.opts = list(width.cutoff = 60), tidy = TRUE)
```

Import Data

```
tele <- read.csv("Model development.csv", stringsAsFactors = TRUE)
```

Exploratory Data Analysis and Data Cleaning

```
str(tele)
```

```
## 'data.frame':    4140 obs. of  12 variables:
##  $ ï..customerID   : Factor w/ 4140 levels "0003-MKNFE","0004-TLHLJ",..: 3853 1744 3699 817 918 1008
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 2 2 2 2 2 ...
##  $ SeniorCitizen   : int  1 0 0 0 1 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 1 2 1 2 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
##  $ tenure          : int  38 70 39 30 60 50 1 14 52 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 1 2 2 ...
##  $ Contract        : Factor w/ 2 levels "Long term","Short term": 2 1 2 2 2 2 2 2 1 1 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 2 1 1 2 ...
##  $ MonthlyCharges  : num  75 49.9 35.5 51.2 99 ...
##  $ TotalCharges    : num  2870 3370 1309 1562 6018 ...
##  $ Renew           : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 2 1 1 ...
```

```
names(tele)[1] <- "CustomerID"  # Change variable name
tele <- cbind(tele, tele$Renew)
str(tele)
```

```
## 'data.frame':    4140 obs. of  13 variables:
##  $ CustomerID      : Factor w/ 4140 levels "0003-MKNFE","0004-TLHLJ",..: 3853 1744 3699 817 918 1008
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 2 2 2 2 2 ...
##  $ SeniorCitizen   : int  1 0 0 0 1 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 1 2 1 2 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
##  $ tenure          : int  38 70 39 30 60 50 1 14 52 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 1 2 2 ...
##  $ Contract        : Factor w/ 2 levels "Long term","Short term": 2 1 2 2 2 2 2 2 1 1 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 2 1 1 2 ...
##  $ MonthlyCharges  : num  75 49.9 35.5 51.2 99 ...
##  $ TotalCharges    : num  2870 3370 1309 1562 6018 ...
##  $ Renew           : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 2 1 1 ...
##  $ tele$Renew      : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 2 1 1 ...
```

```
names(tele)[13] <- "Churn"
tele$Churn <- ifelse(tele$Churn == "Yes", 1, 2)  # Convert Churn variable to number
tele$SeniorCitizen <- as.factor(ifelse(tele$SeniorCitizen ==
    1, "Yes", "No"))
# tele[12]<-lapply(tele[12], as.factor)
str(tele)
```

```
## ’data.frame':    4140 obs. of  13 variables:
##  $ CustomerID      : Factor w/ 4140 levels "0003-MKNFE","0004-TLHLJ",..: 3853 1744 3699 817 918 1008
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 1 2 1 2 2 2 2 2 2 ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 1 1 1 1 1 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 2 1 2 2 2 1 2 1 2 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 2 1 1 2 1 ...
##  $ tenure          : int  38 70 39 30 60 50 1 14 52 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 2 1 1 1 2 2 2 1 2 2 ...
##  $ Contract        : Factor w/ 2 levels "Long term","Short term": 2 1 2 2 2 2 2 2 1 1 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 2 1 1 2 1 2 1 1 2 ...
##  $ MonthlyCharges  : num  75 49.9 35.5 51.2 99 ...
##  $ TotalCharges    : num  2870 3370 1309 1562 6018 ...
##  $ Renew           : Factor w/ 2 levels "No","Yes": 2 1 1 2 1 1 1 2 1 1 ...
##  $ Churn           : num  1 2 2 1 2 2 2 1 2 2 ...
```

```
tele$Churn <- ifelse(tele$Churn == "1", "No", "Yes")
head(tele)
```

```
##   CustomerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 9286-DOJGF Female           Yes     Yes         No     38          Yes
## 2 4312-GVYNH Female            No     Yes         No     70           No
## 3 8898-KASCD   Male            No      No         No     39           No
## 4 2091-MJTFX Female            No     Yes        Yes     30           No
## 5 2277-DJJDL   Male           Yes     Yes         No     60          Yes
## 6 2511-MORQY   Male            No     Yes        Yes     50          Yes
##     Contract PaperlessBilling MonthlyCharges TotalCharges Renew Churn
## 1 Short term              Yes          74.95      2869.85   Yes    No
## 2  Long term              Yes          49.85      3370.20    No   Yes
## 3 Short term               No          35.55      1309.15    No   Yes
## 4 Short term               No          51.20      1561.50   Yes    No
## 5 Short term              Yes          99.00      6017.90    No   Yes
## 6 Short term               No          54.90      2614.10    No   Yes
```

There are 10 independent variables that can be classified into 3 groups in the data set:

- 1. Demographic
- 2. Customer Account
- 3. Services

Visualize Demographic Distribution

```
# Load Packages
require(ggplot2)
```

```
## Loading required package: ggplot2
```

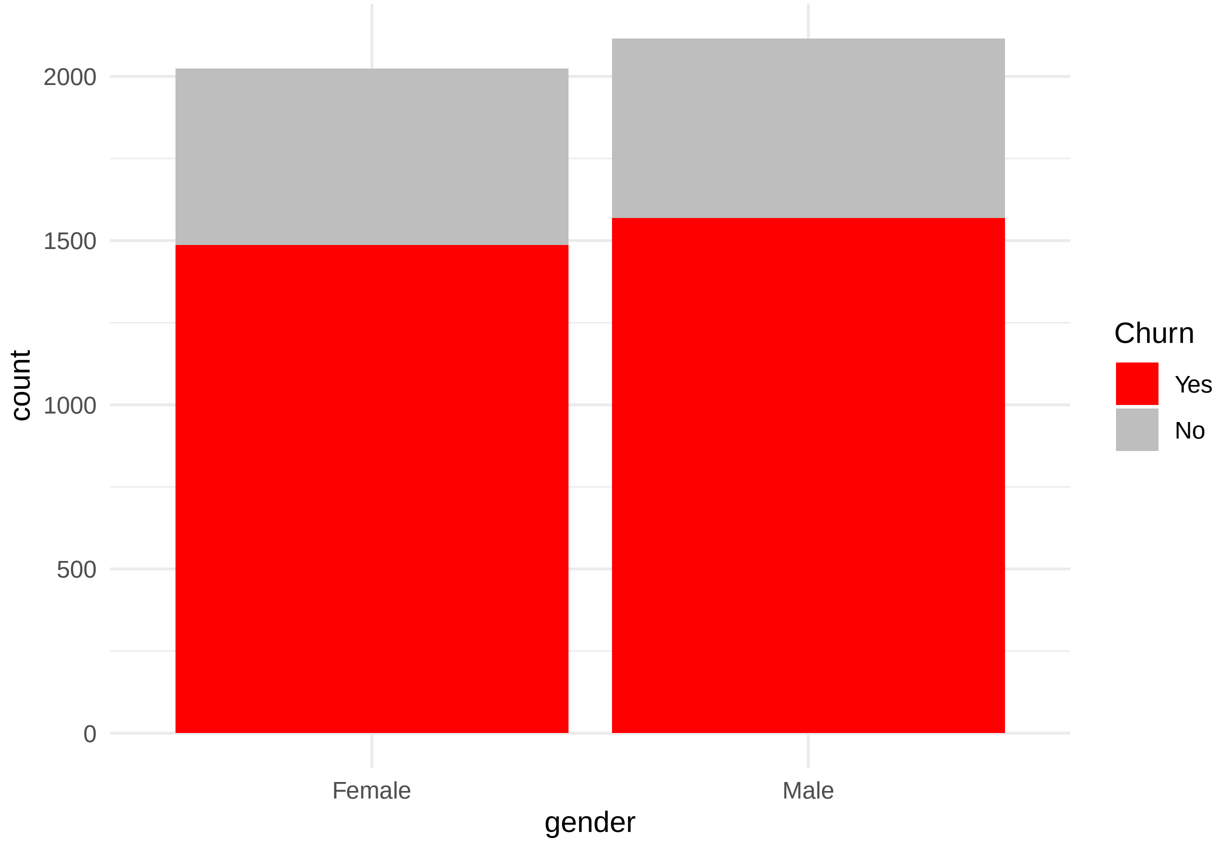
```
require(gridExtra)
```

```
## Loading required package: gridExtra
```

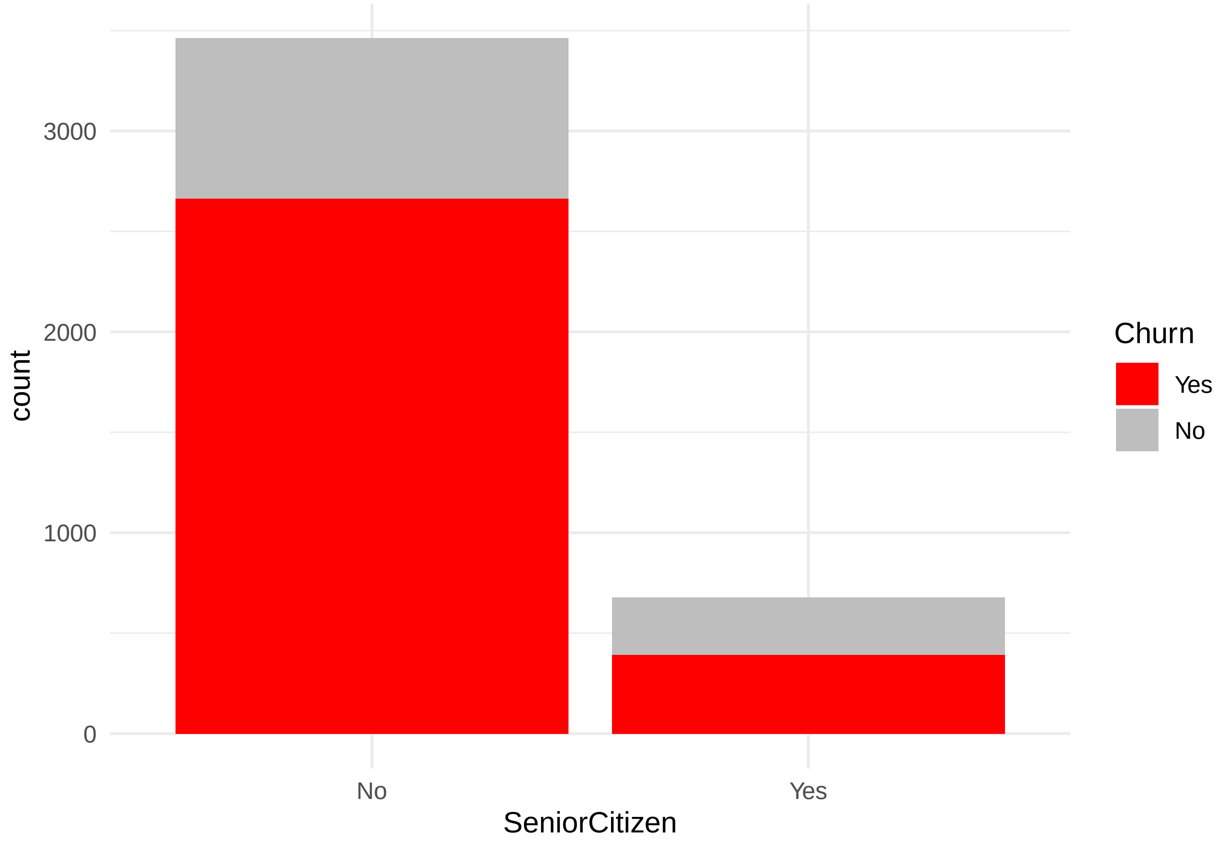
```
require(grid)
```

```
## Loading required package: grid
```

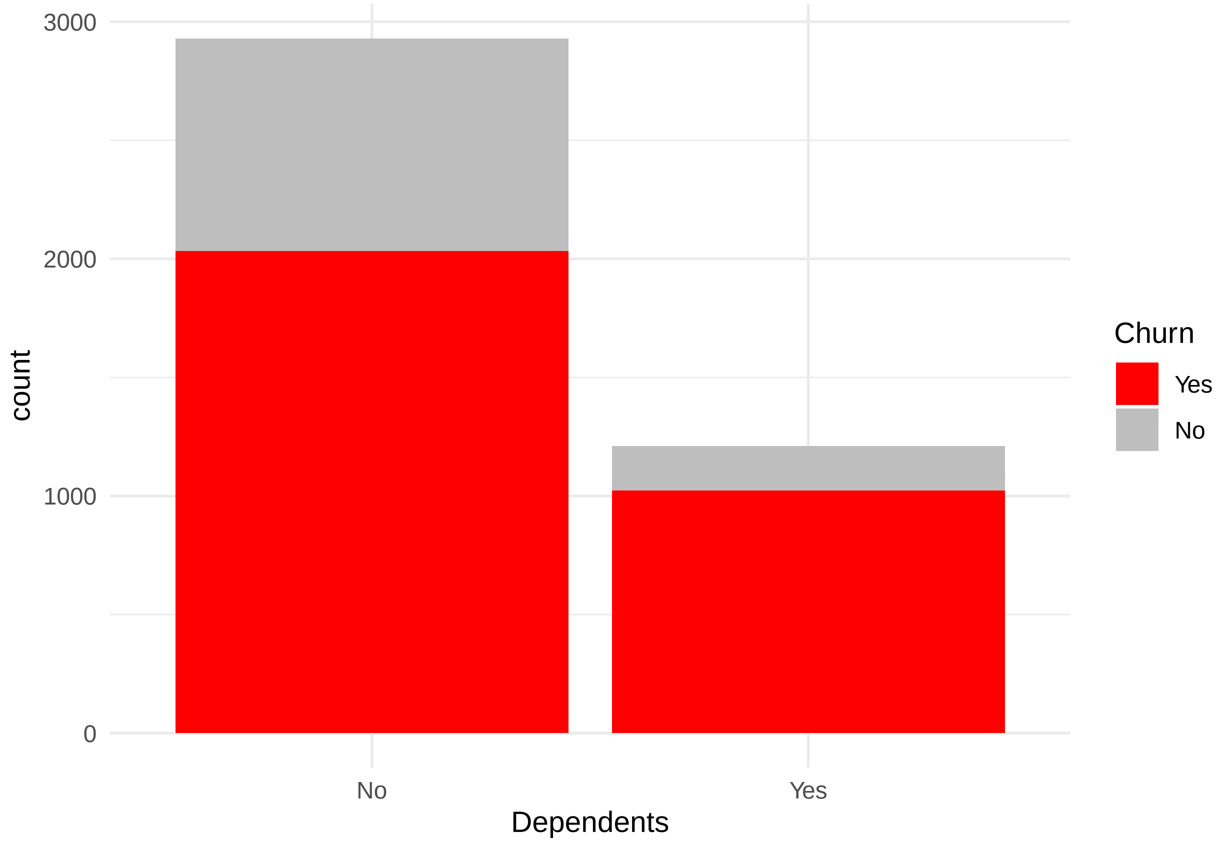
```
gender_plot <- ggplot(tele, aes(x = gender, fill = Churn)) +
    geom_bar(show.legend = TRUE) + scale_fill_manual(values = c(Yes = "Red",
    No = "Gray")) + theme_minimal()
gender_plot
```



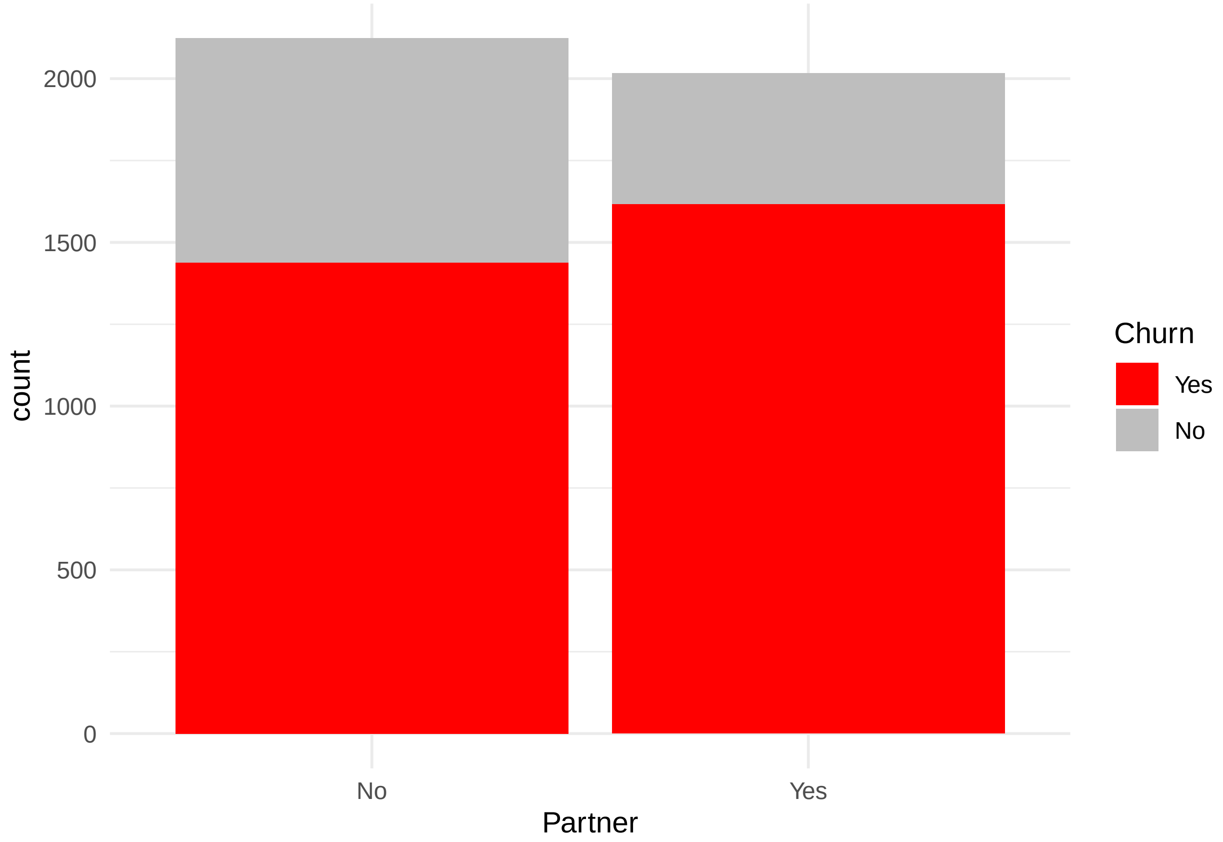
```
SeniorCitizen_plot <- ggplot(tele, aes(x = SeniorCitizen, fill = Churn)) +
    geom_bar(show.legend = TRUE) + scale_fill_manual(values = c(Yes = "Red",
    No = "Gray")) + theme_minimal()
SeniorCitizen_plot
```



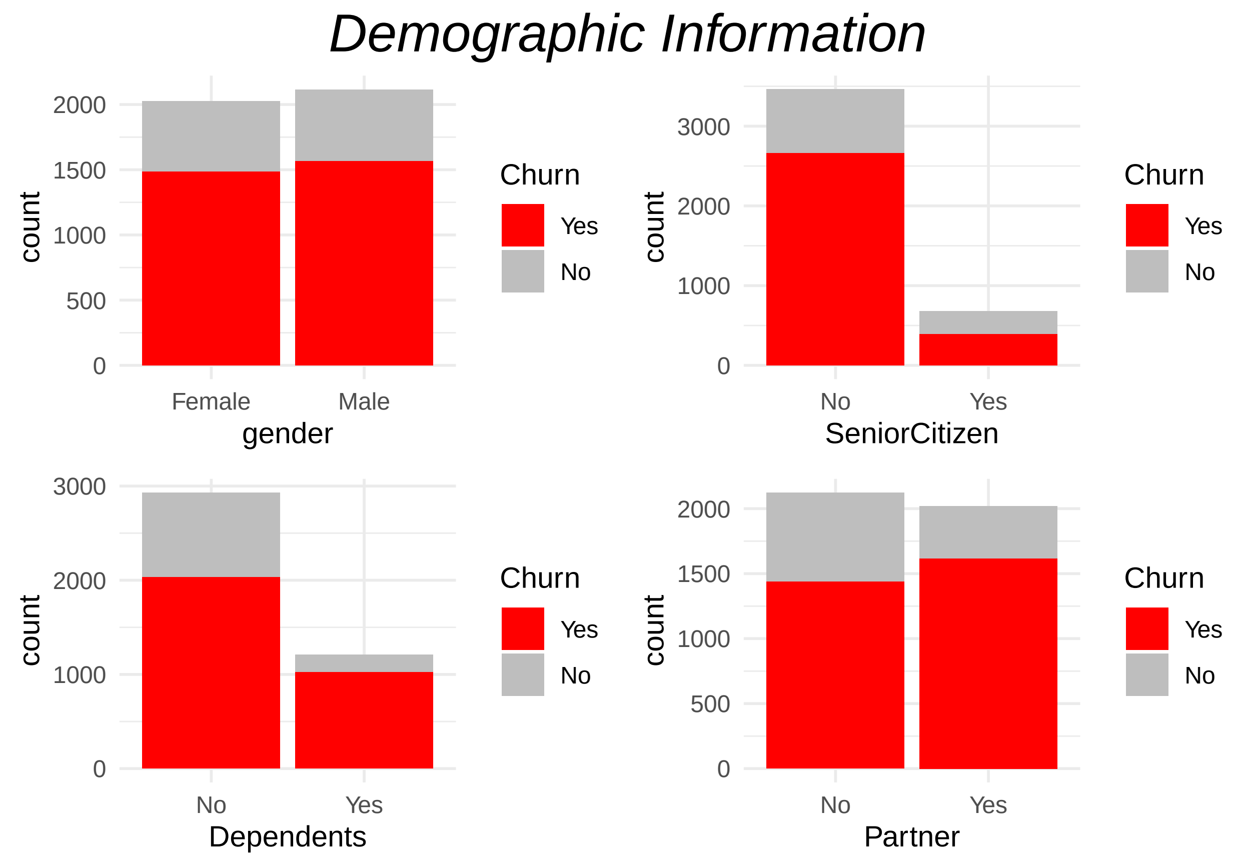
```
dependents_plot <- ggplot(tele, aes(x = Dependents, fill = Churn)) +
    geom_bar(show.legend = TRUE) + scale_fill_manual(values = c(Yes = "Red",
    No = "Gray")) + theme_minimal()
dependents_plot
```



```
partner_plot <- ggplot(tele, aes(x = Partner, fill = Churn)) +
    geom_bar(show.legend = TRUE) + scale_fill_manual(values = c(Yes = "Red",
    No = "Gray")) + theme_minimal()
partner_plot
```



```r
# Plot neatly
grid.arrange(gender_plot, SeniorCitizen_plot, dependents_plot,
    partner_plot, nrow = 2, top = textGrob("Demographic Information",
        gp = gpar(fontsize = 20, font = 3)))
```



Churn by contract and tenure

```
require(dplyr)
```

```
## Loading required package: dplyr

##
## Attaching package: 'dplyr'

## The following object is masked from 'package:gridExtra':
##
##     combine

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

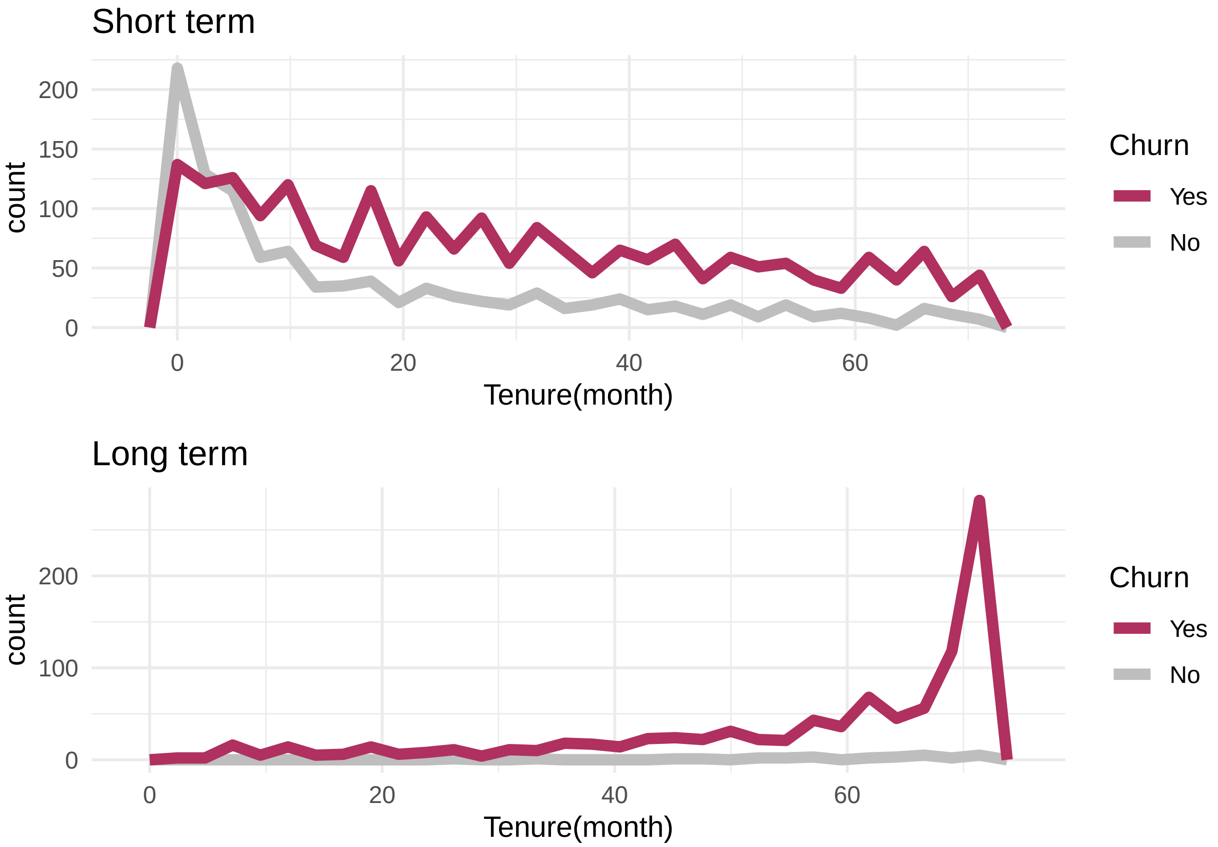
```
short_term <- ggplot(subset(tele, Contract %in% c("Short term")),
    aes(x = tenure, color = Churn)) + geom_freqpoly(size = 2) +
    theme_minimal() + labs(title = "Short term", x = "Tenure(month)") +
    scale_color_manual(values = c(Yes = "Maroon", No = "Gray"))

long_term <- ggplot(subset(tele, Contract %in% c("Long term")),
    aes(x = tenure, color = Churn)) + geom_freqpoly(size = 2) +
    theme_minimal() + labs(title = "Long term", x = "Tenure(month)") +
    scale_color_manual(values = c(Yes = "Maroon", No = "Gray"))

grid.arrange(short_term, long_term)
```

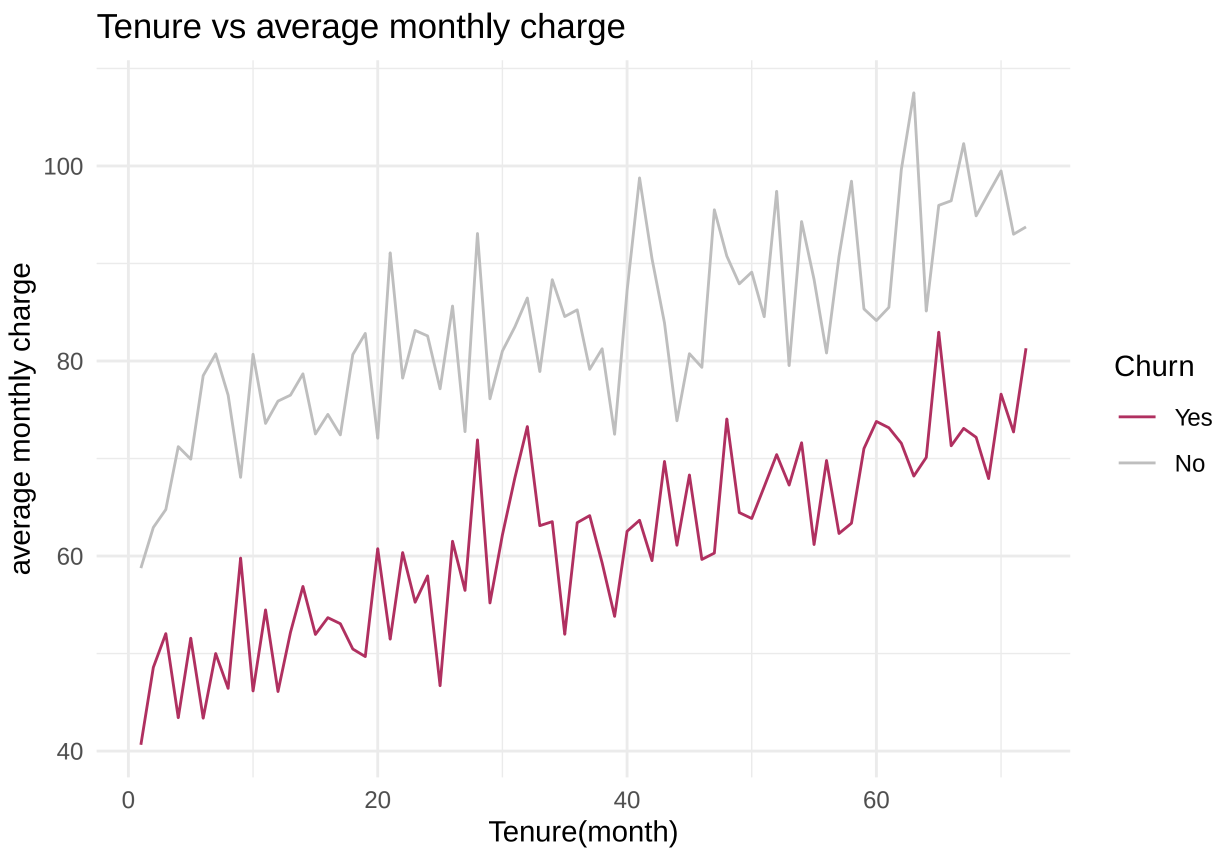
```
## ‘stat_bin()‘ using ‘bins = 30‘. Pick better value with ‘binwidth‘.

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



Average Monthly Charge

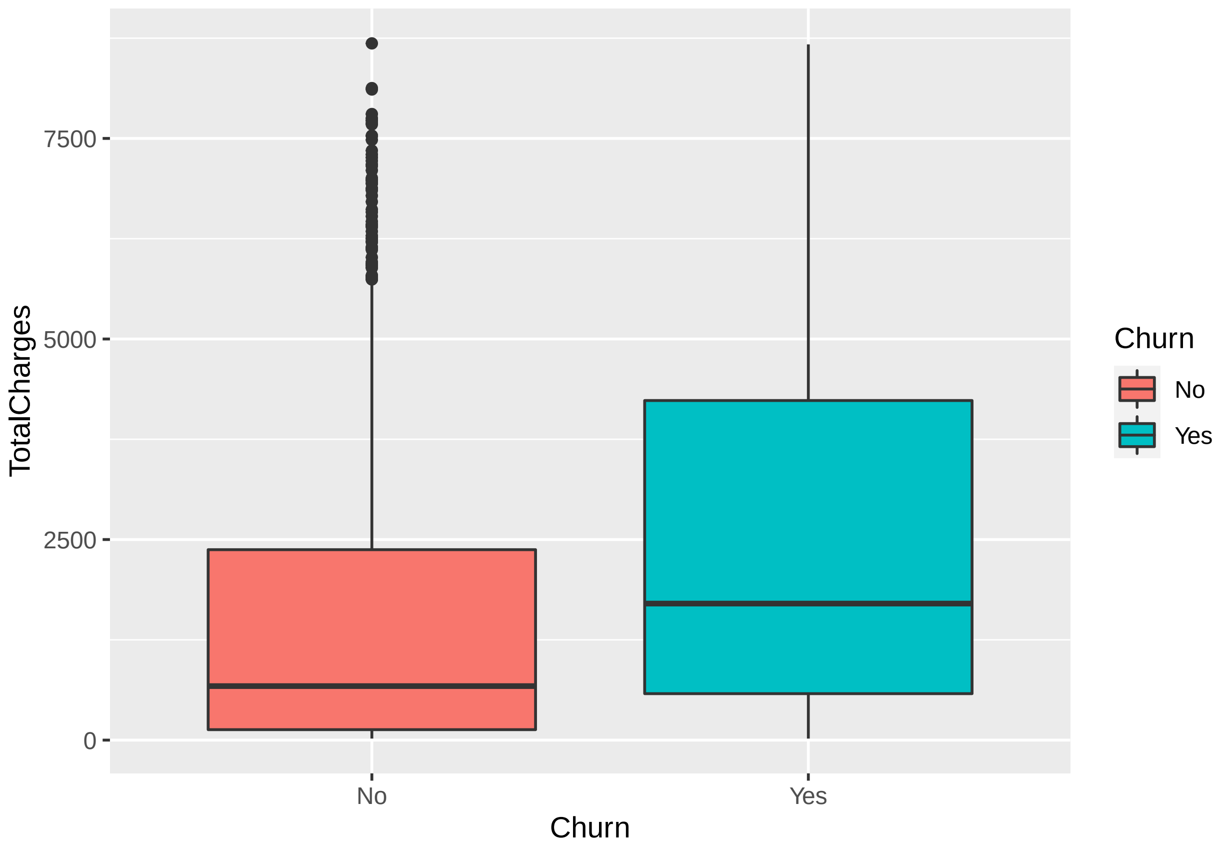
```
ggplot(data = tele) + geom_line(aes(x = tenure, y = MonthlyCharges,
    color = Churn), stat = "summary", fun = "mean") + labs(title = "Tenure vs average monthly charge",
    x = "Tenure(month)", y = "average monthly charge") + scale_color_manual(values = c(Yes = "Maroon",
    No = "Gray")) + theme_minimal()
```



- Customers who churn, are perhaps in the price sensitive category in that their average monthly charge is less than those that do not churn.
- It may also be due to paying for an inferior service that lead them to leave.

Customer Churn vs tenure

```
ggplot(data = tele) + geom_boxplot(aes(x = Churn, y = TotalCharges,
    fill = Churn))
```


TotalCharges
7500
5000
2500
0
No
Yes
Churn
Churn
No
Yes