Analysis of Teleco Customer Churn

Valencia Dias

20/02/2020

```r
library(readr)
```

```
## Warning: package 'readr' was built under R version 3.6.2
```

```r
custc <- read.csv("C:/Users/admin/Desktop/MVA/PROJECT/TelEco_Customer_Churn.csv")
summary(custc)
```

```
##       customerID      gender     SeniorCitizen    Partner    Dependents
##  0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933  
##  0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110  
##  0004-TLHLJ:   1                 Median :0.0000                        
##  0011-IGKFF:   1                 Mean   :0.1621                        
##  0013-EXCHZ:   1                 3rd Qu.:0.0000                        
##  0013-MHZWF:   1                 Max.   :1.0000                        
##  (Other)   :7037                                                       
##      tenure       PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421   
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096   
##  Median :29.00                Yes             :2971   No         :1526   
##  Mean   :32.37                                                           
##  3rd Qu.:55.00                                                           
##  Max.   :72.00                                                           
##                                                                          
##              OnlineSecurity              OnlineBackup 
##  No                 :3498   No                 :3088  
##  No internet service:1526   No internet service:1526  
##  Yes                :2019   Yes                :2429  
##                                                       
##                                                       
##                                                       
##                                                       
##             DeviceProtection              TechSupport  
##  No                 :3095    No                 :3473  
##  No internet service:1526    No internet service:1526  
##  Yes                :2422    Yes                :2044  
##                                                        
##                                                        
##                                                        
##                                                        
##               StreamingTV              StreamingMovies           Contract   
##  No                 :2810   No                 :2785   Month-to-month:3875  
##  No internet service:1526   No internet service:1526   One year      :1473  
##  Yes                :2707   Yes                :2732   Two year      :1695  
##                                                                             
##                                                                             
##                                                                             
##                                                                             
##  PaperlessBilling                   PaymentMethod  MonthlyCharges  
##  No :2872         Bank transfer (automatic):1544   Min.   : 18.25  
##  Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50  
##                   Electronic check         :2365   Median : 70.35  
##                   Mailed check             :1612   Mean   : 64.76  
##                                                    3rd Qu.: 89.85  
##                                                    Max.   :118.75  
##                                                                    
##   TotalCharges    Churn     
##  Min.   :  18.8   No :5174  
##  1st Qu.: 401.4   Yes:1869  
##  Median :1397.5             
##  Mean   :2283.3             
##  3rd Qu.:3794.7             
##  Max.   :8684.8             
##  NA's   :11
```

```
dim(custc)
```

```
## [1] 7043   21
```

```
str(custc)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 10
03 4771 5605 4535 ...
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
##  $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
##  $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
##  $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
##  $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
##  $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
##  $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

```
#Finding the missing values in each column
sapply(custc, function(x) sum(is.na(x)))
```

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

```
#Dropping all the rows with the missing values
custc <- custc[complete.cases(custc),]
sapply(custc, function(x) sum(is.na(x)))
```

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

```
dim(custc)
```

```
## [1] 7032   21
```

Comments:We have used “sapply” to check if there are missing values in each columns. We found that there are 11 missing values in “TotalCharges” columns.We have further removed all those rows with missing values.

```
#Factoring the Churn Variable and changing the value of No to 0 and Yes to 1
library(magrittr)
```

```
## Warning: package 'magrittr' was built under R version 3.6.2
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.2
```

```
##
## Attaching package: 'dplyr'
```

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

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

```
library(tidyr)
```

```
## Warning: package 'tidyr' was built under R version 3.6.2
```

```
##
## Attaching package: 'tidyr'
```

```
## The following object is masked from 'package:magrittr':
##
##     extract
```

```
custc$Churn <- factor(ifelse(custc$Churn == 'No', 0, 1))
cus <- custc %>% group_by(Churn)%>%
  summarise(Count = length(Churn)) %>%
  mutate(Rate = Count / sum(Count)*100.0)
cus
```

Churn <fctr>
0
1

2 rows | 1-1 of 3 columns

```
#Churners Rate
library(plyr)
```

```
## Warning: package 'plyr' was built under R version 3.6.2
```

```
## ------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
## ------------------------------------------------------------------------------
```

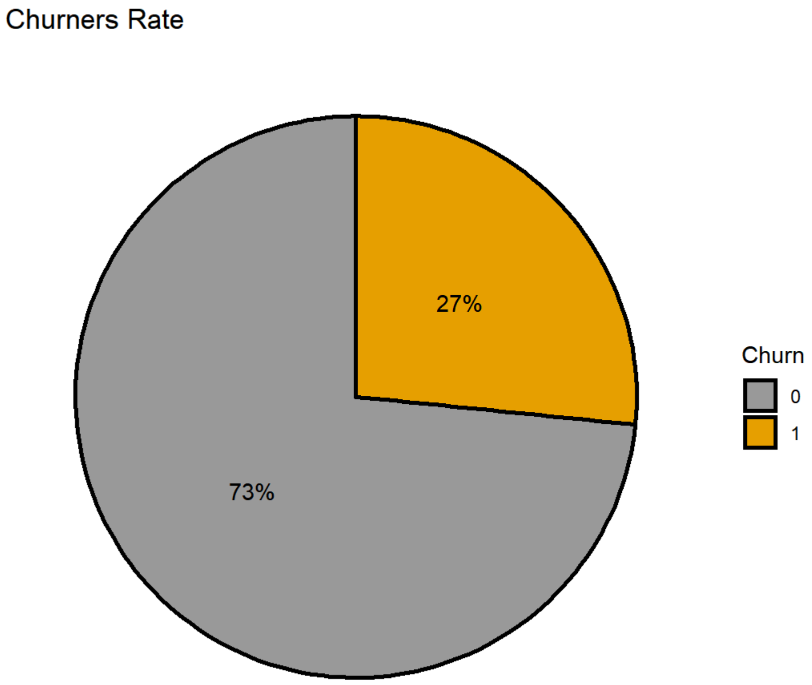
```
##
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.2
```

```
ggplot(cus, aes(x = '', y = Rate, fill = Churn)) +
  geom_bar(width = 1, size = 1, color = 'black', stat = 'identity') +
  coord_polar('y') +
  geom_text(aes(label = paste0(round(Rate), '%')),
            position = position_stack(vjust = 0.5)) +
  scale_fill_manual(values=c("#999999", "#E69F00"))+
  labs(title = 'Churners Rate') +
  theme_classic() +
  theme(axis.line = element_blank(),axis.title.x = element_blank(),axis.title.y = element_blank(),
        axis.ticks = element_blank(),
        axis.text = element_blank())
```



Comments:In our dataset, close to 30% customers churned, while the rest remained with the service provider.

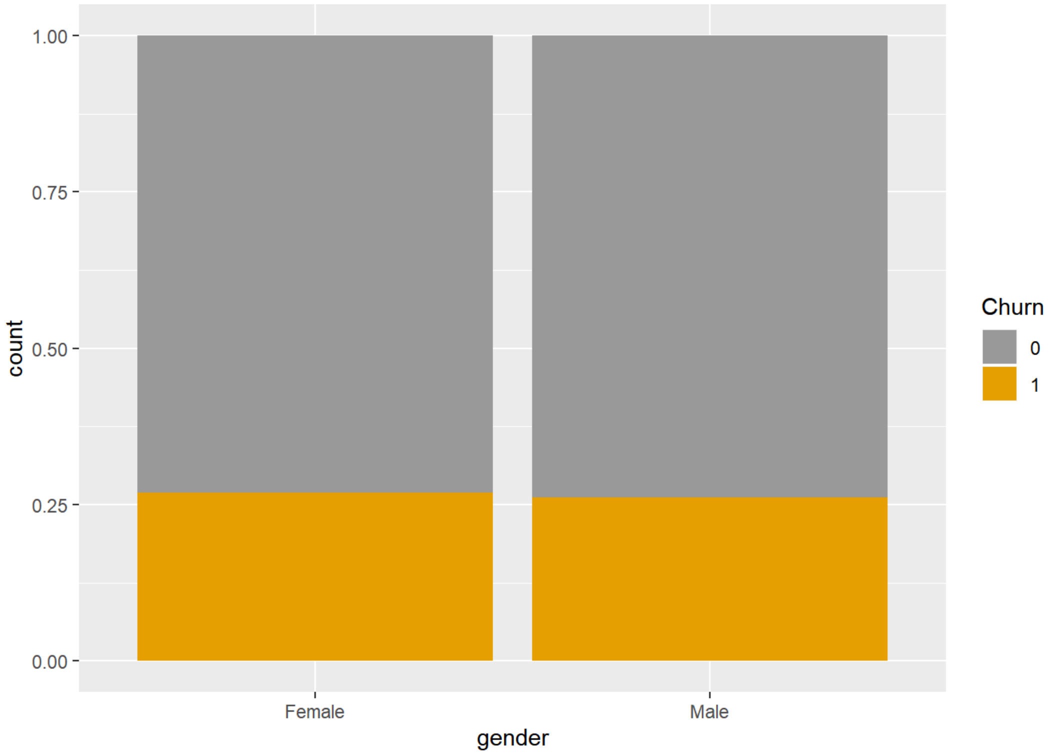
```
custc$MonthlyChargesBin <- NA
custc$MonthlyChargesBin[custc$MonthlyCharges > 0 & custc$MonthlyCharges <= 10] <- '10'
custc$MonthlyChargesBin[custc$MonthlyCharges > 10 & custc$MonthlyCharges <= 20] <- '20'
custc$MonthlyChargesBin[custc$MonthlyCharges > 20 & custc$MonthlyCharges <= 30] <- '30'
custc$MonthlyChargesBin[custc$MonthlyCharges > 30 & custc$MonthlyCharges <= 40] <- '40'
custc$MonthlyChargesBin[custc$MonthlyCharges > 40 & custc$MonthlyCharges <= 50] <- '50'
custc$MonthlyChargesBin[custc$MonthlyCharges > 50 & custc$MonthlyCharges <= 60] <- '60'
custc$MonthlyChargesBin[custc$MonthlyCharges > 60 & custc$MonthlyCharges <= 70] <- '70'
custc$MonthlyChargesBin[custc$MonthlyCharges > 70 & custc$MonthlyCharges <= 80] <- '80'
custc$MonthlyChargesBin[custc$MonthlyCharges > 80 & custc$MonthlyCharges <= 90] <- '90'
custc$MonthlyChargesBin[custc$MonthlyCharges > 90 & custc$MonthlyCharges <= 100] <- '100'
custc$MonthlyChargesBin[custc$MonthlyCharges > 100 & custc$MonthlyCharges <= 110] <- '110'
custc$MonthlyChargesBin[custc$MonthlyCharges > 110 & custc$MonthlyCharges <= 120] <- '120'

custc$MonthlyChargesBin <- factor(custc$MonthlyChargesBin,
                                      levels = c('10', '20', '30', '40', '50', '60', '70', '80', '90','100'
,'110','120'))
```

```
cols_recode1 <- c(10:15)
for (i in 1:ncol(custc[, cols_recode1])) {
  custc[, cols_recode1][, i] <- as.factor(mapvalues(custc[, cols_recode1][, i], from = c("No internet servic
e"), to = c("No")))
}

custc$MultipleLines <- as.factor(mapvalues(custc$MultipleLines, from = c("No phone service"), to = c("No")))

str(custc)
```

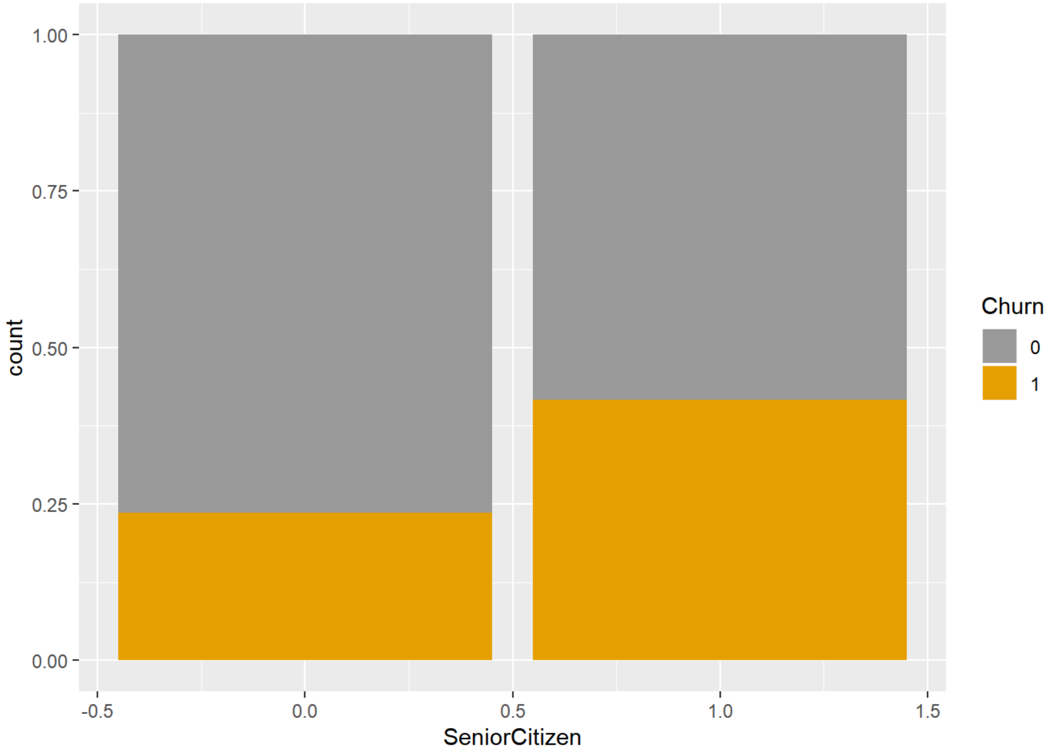
```
## 'data.frame':    7032 obs. of  22 variables:
##  $ customerID       : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 6552 1
003 4771 5605 4535 ...
##  $ gender           : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner          : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents       : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ tenure           : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService     : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines    : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
##  $ InternetService  : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ OnlineSecurity   : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
##  $ OnlineBackup     : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
##  $ DeviceProtection : Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
##  $ TechSupport      : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
##  $ StreamingTV      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
##  $ StreamingMovies  : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
##  $ Contract         : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
##  $ PaperlessBilling : Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ PaymentMethod    : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges   : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges     : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn            : Factor w/ 2 levels "0","1": 1 1 2 1 2 2 1 1 2 1 ...
##  $ MonthlyChargesBin: Factor w/ 12 levels "10","20","30",..: 3 6 6 5 8 10 9 3 11 6 ...
```

Comments:We have changed ‘No internet service’ to ‘No’ for six columns, they are: ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’, ‘TechSupport’, ‘streamingTV’, ’streamingMovies’.

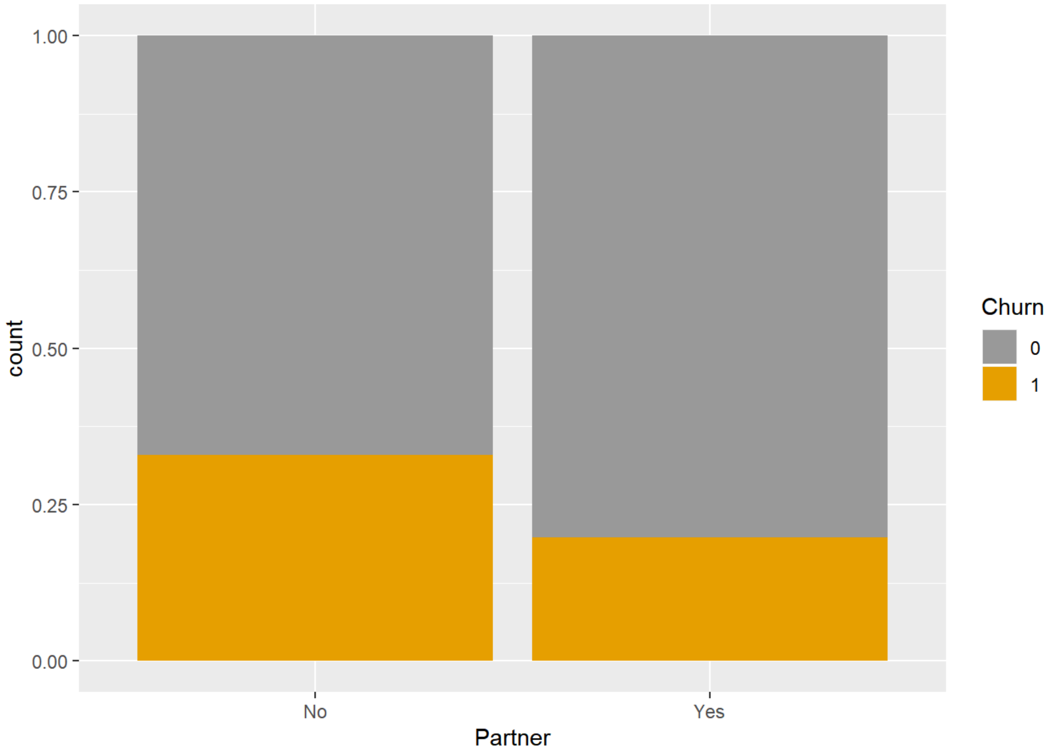
```
#Bar plots of categorical variables
b1 <- ggplot(custc, aes(gender,fill=Churn)) + geom_bar(position='fill') +scale_fill_manual(values=c("#999999
", "#E69F00"))
#+theme(legend.position="none")
b1
```



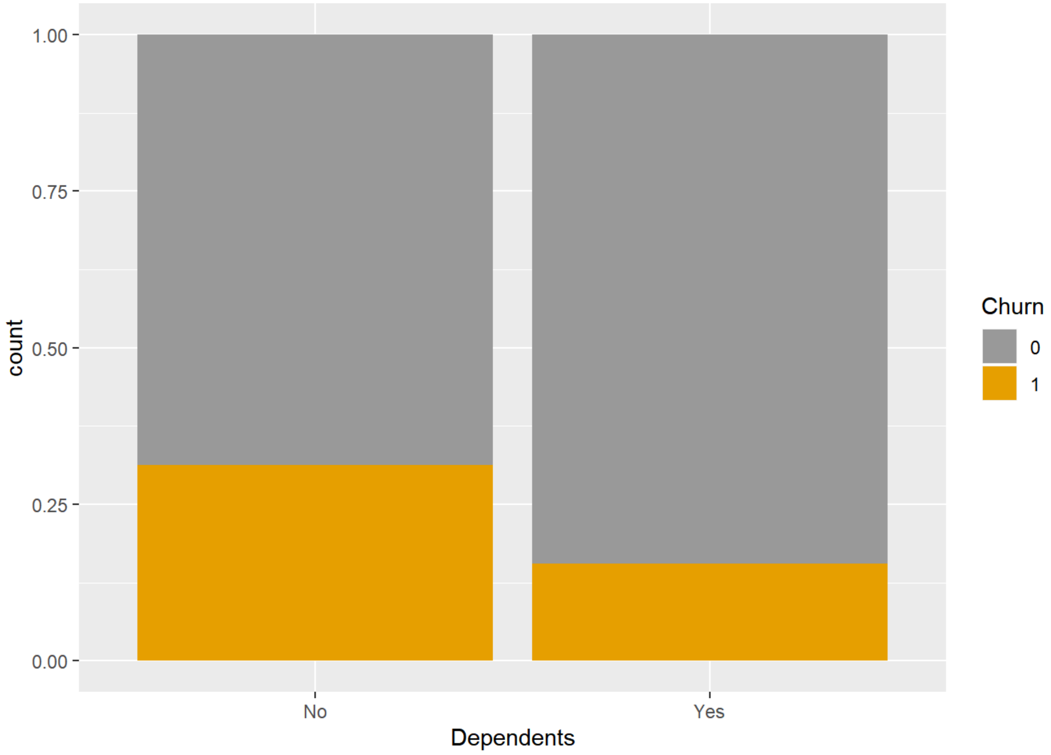
```
b2 <- ggplot(custc, aes(SeniorCitizen, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c("#999999", "#E69F00"))
b2
```



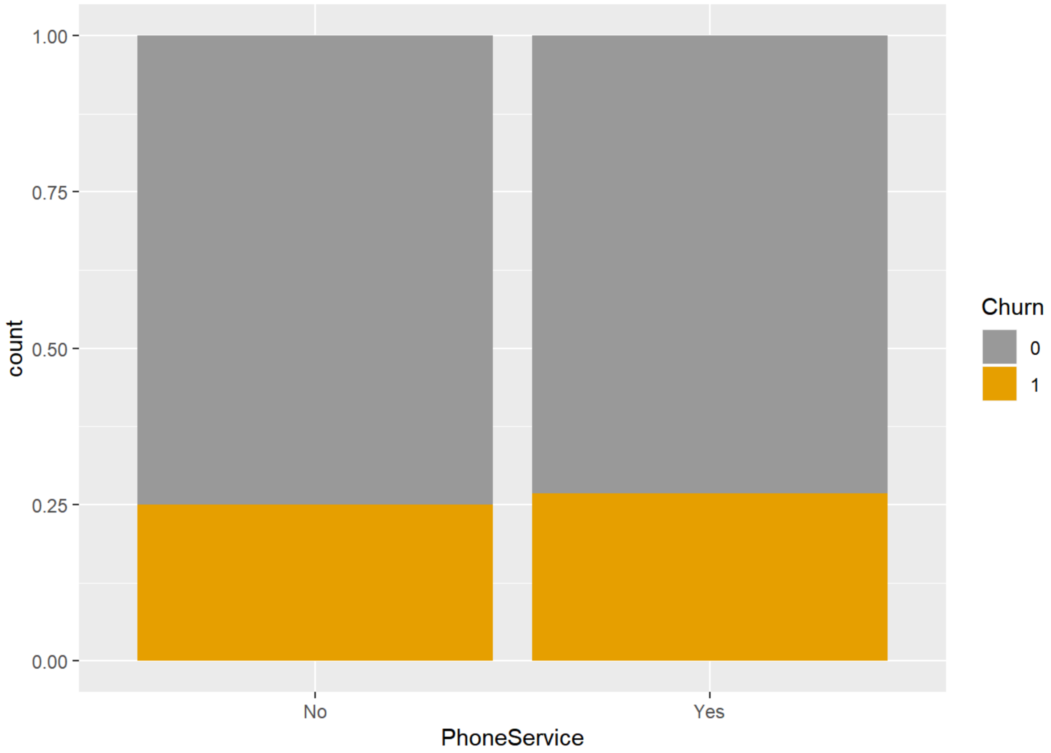
```
b3 <- ggplot(custc, aes(Partner, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c("#999999", "#E69F00"))
b3
```



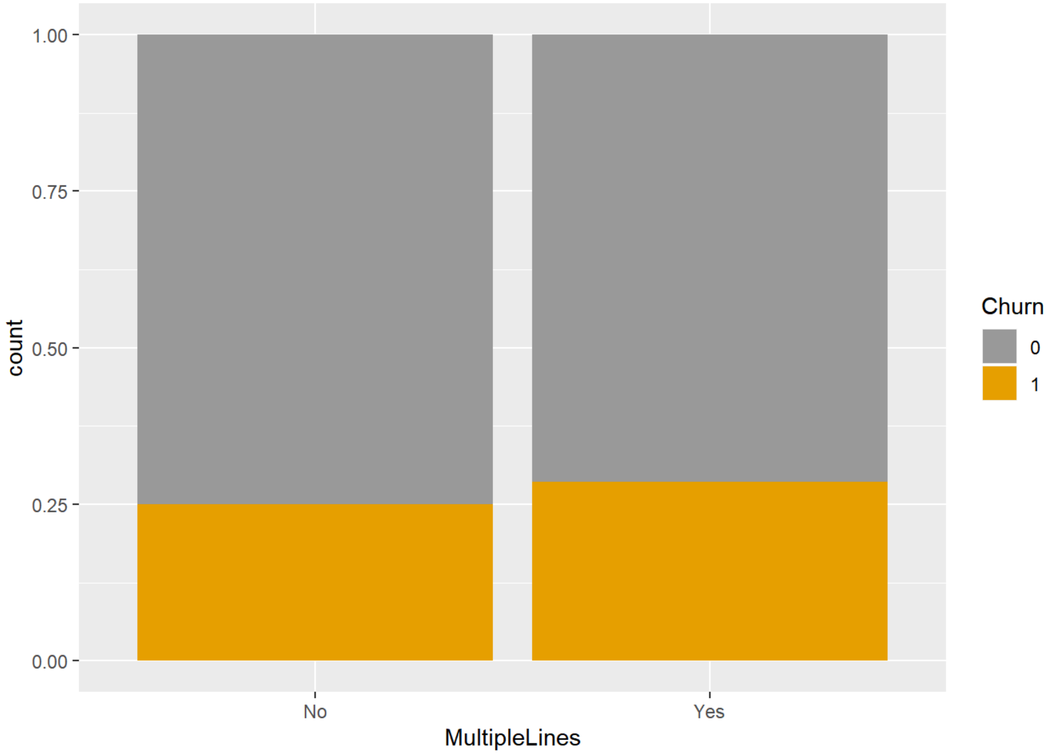
```
b4 <- ggplot(custc, aes(Dependents, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c("#
999999", "#E69F00"))
b4
```



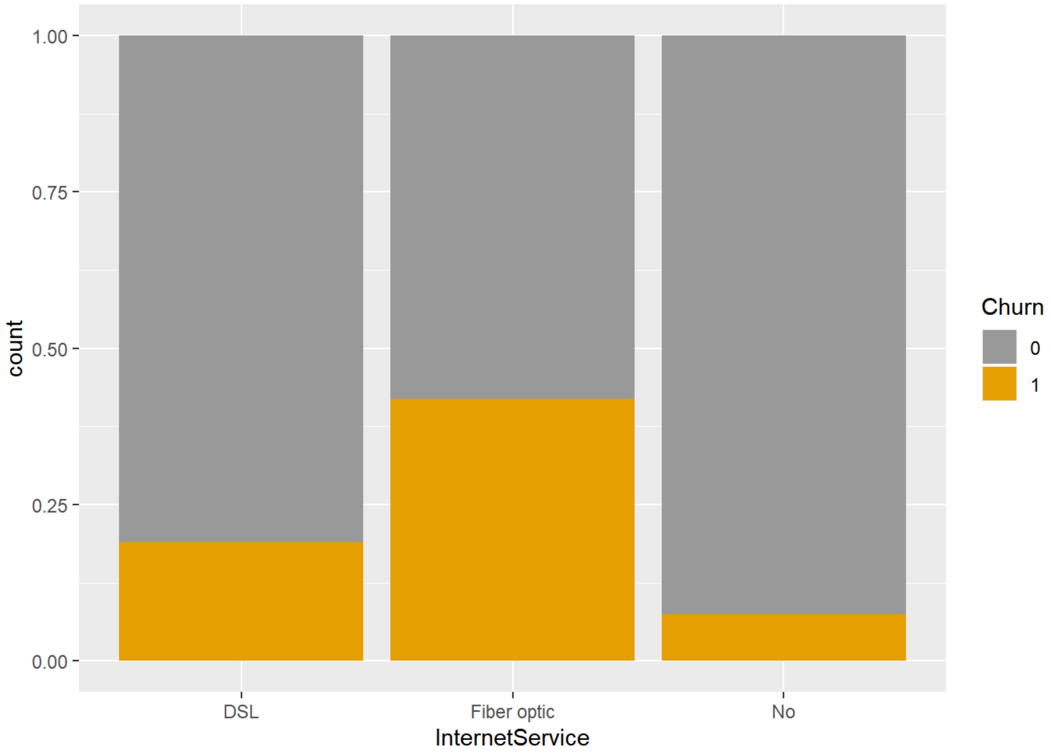
```
b5 <- ggplot(custc, aes(PhoneService, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c(
"#999999", "#E69F00"))
b5
```



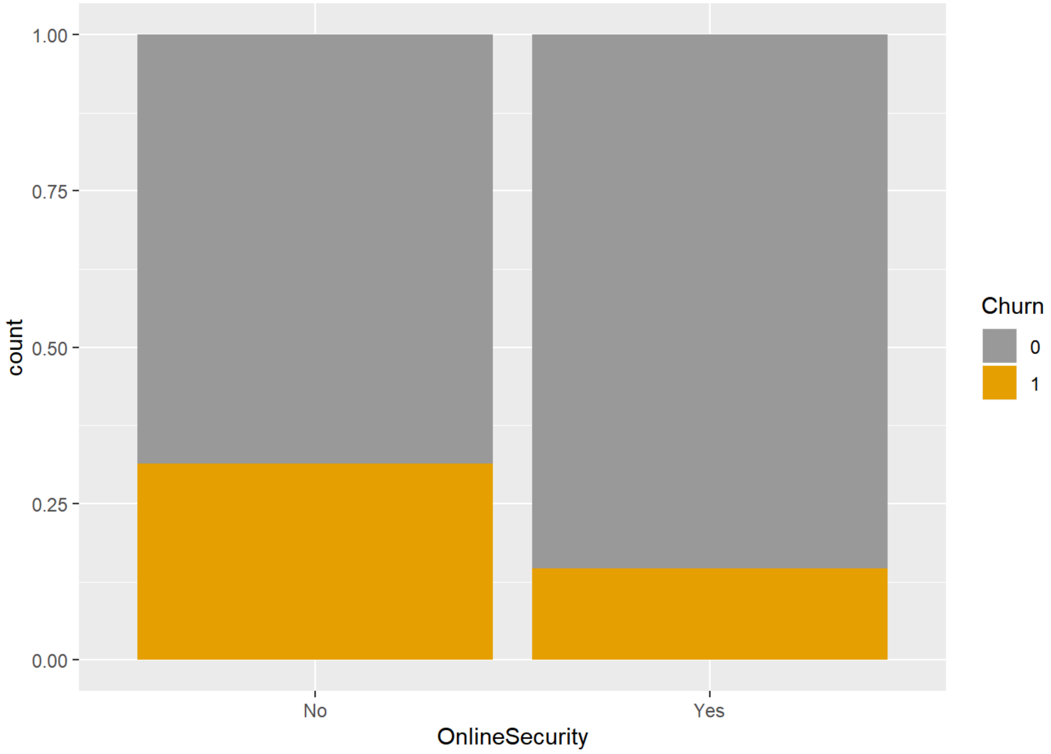
```
b6 <- ggplot(custc, aes(MultipleLines, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c
("#999999", "#E69F00"))
b6
```



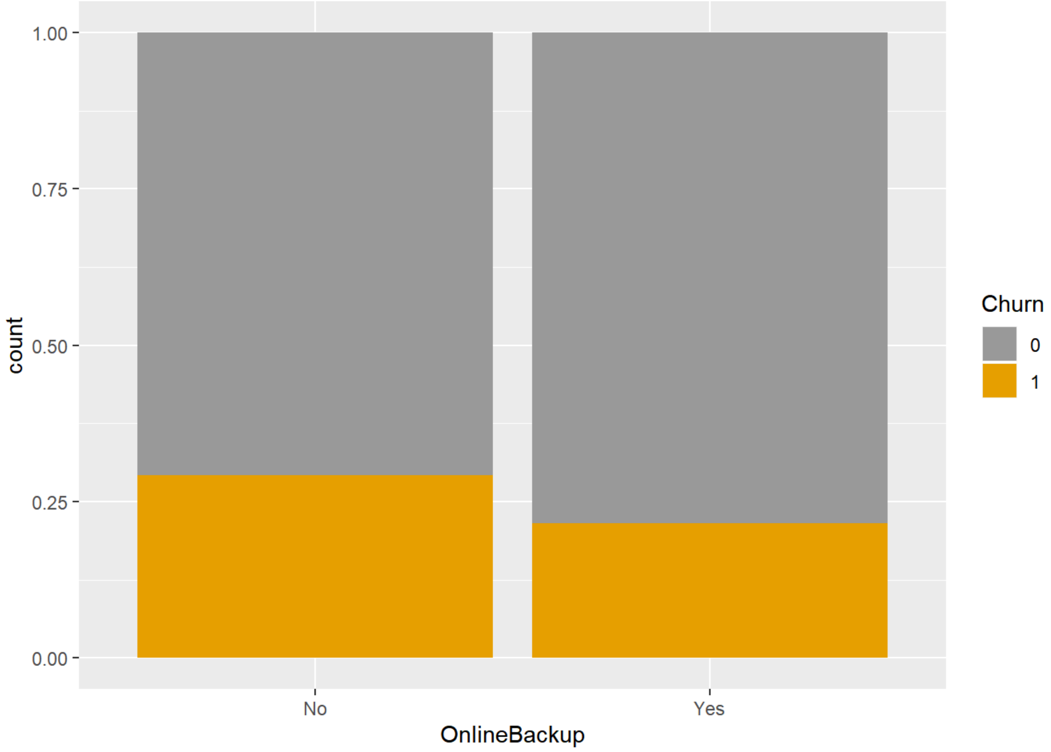
```
b7 <- ggplot(custc, aes(InternetService, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values
=c("#999999", "#E69F00"))
b7
```



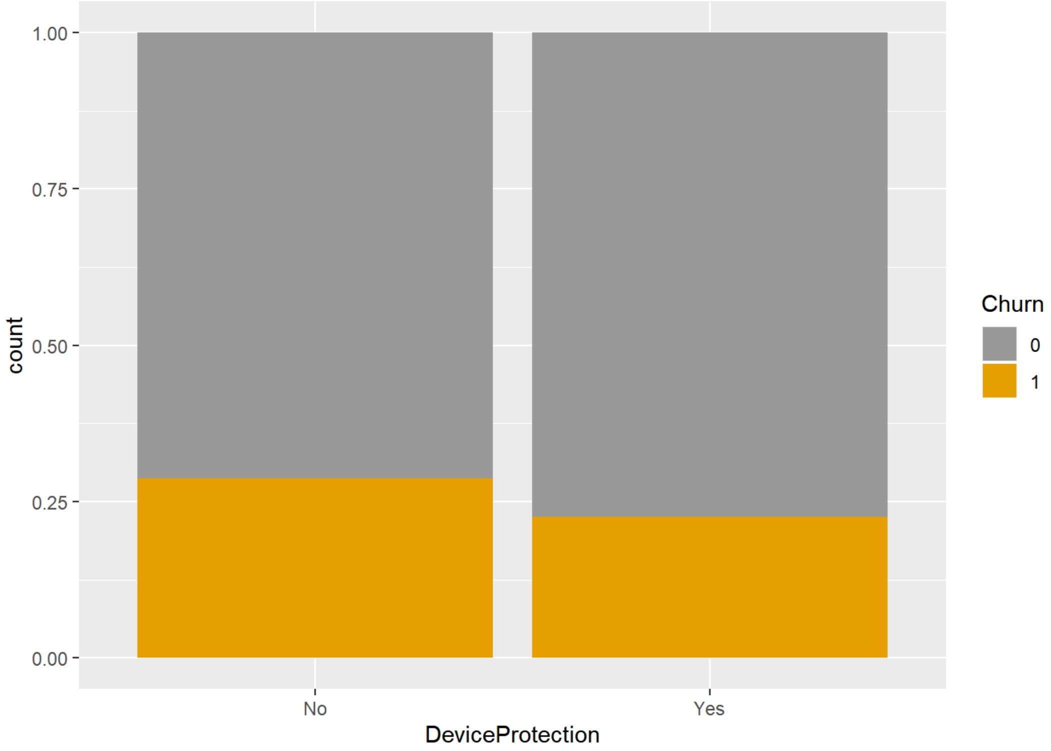
```
b8 <- ggplot(custc, aes(OnlineSecurity, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=
c("#999999", "#E69F00"))
b8
```



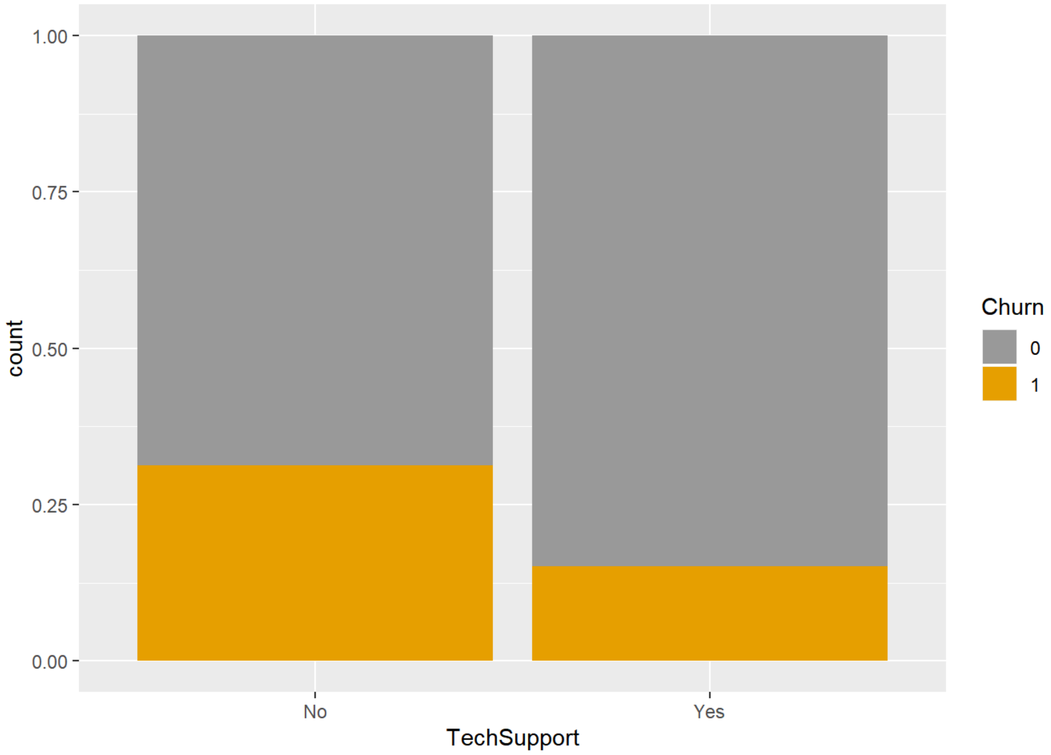
```
b9 <- ggplot(custc, aes(OnlineBackup, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c(
"#999999", "#E69F00"))
b9
```



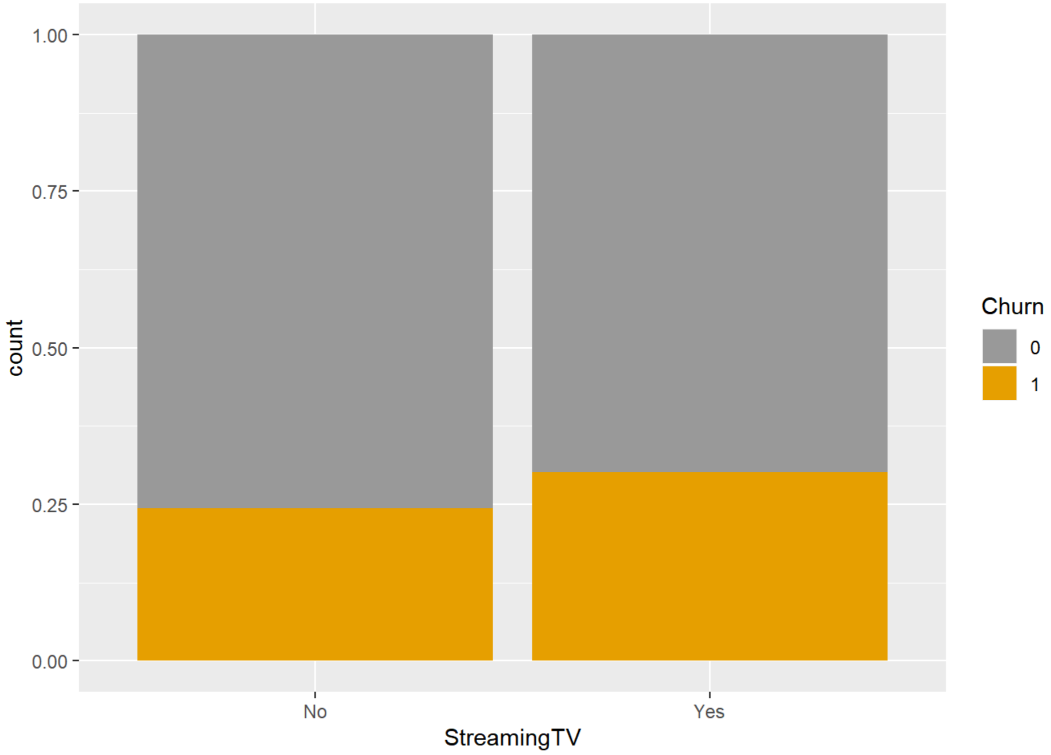
```
b10 <- ggplot(custc, aes(DeviceProtection, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(valu
es=c("#999999", "#E69F00"))
b10
```



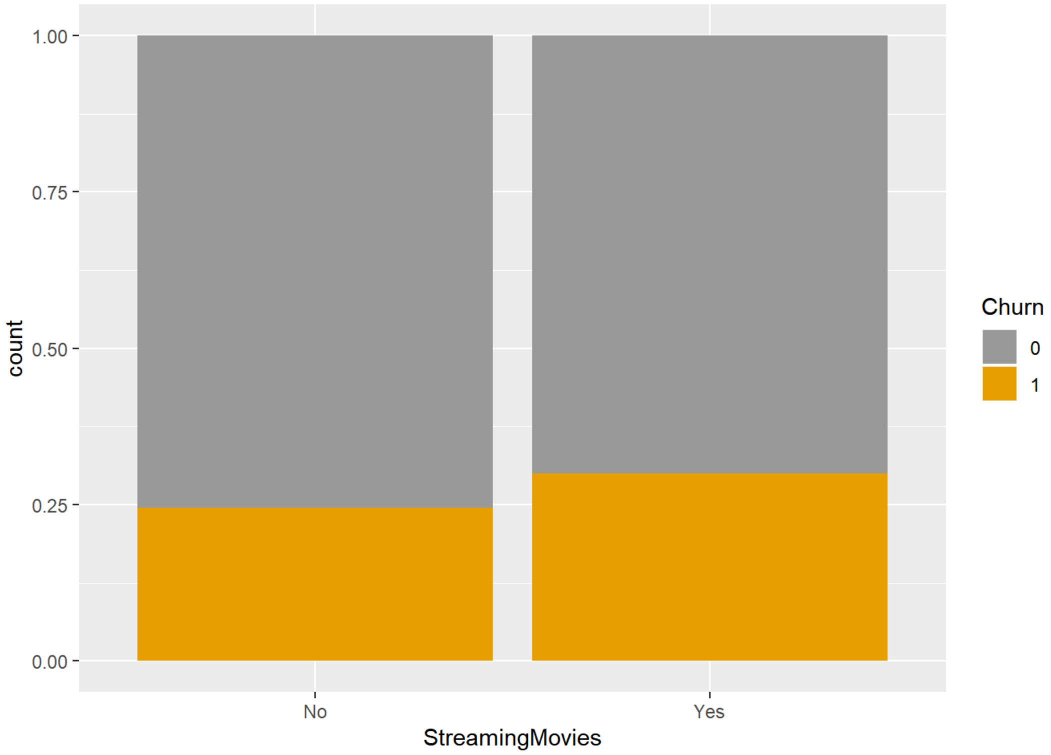
```
b11 <- ggplot(custc, aes(TechSupport, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c(
"#999999", "#E69F00"))
b11
```



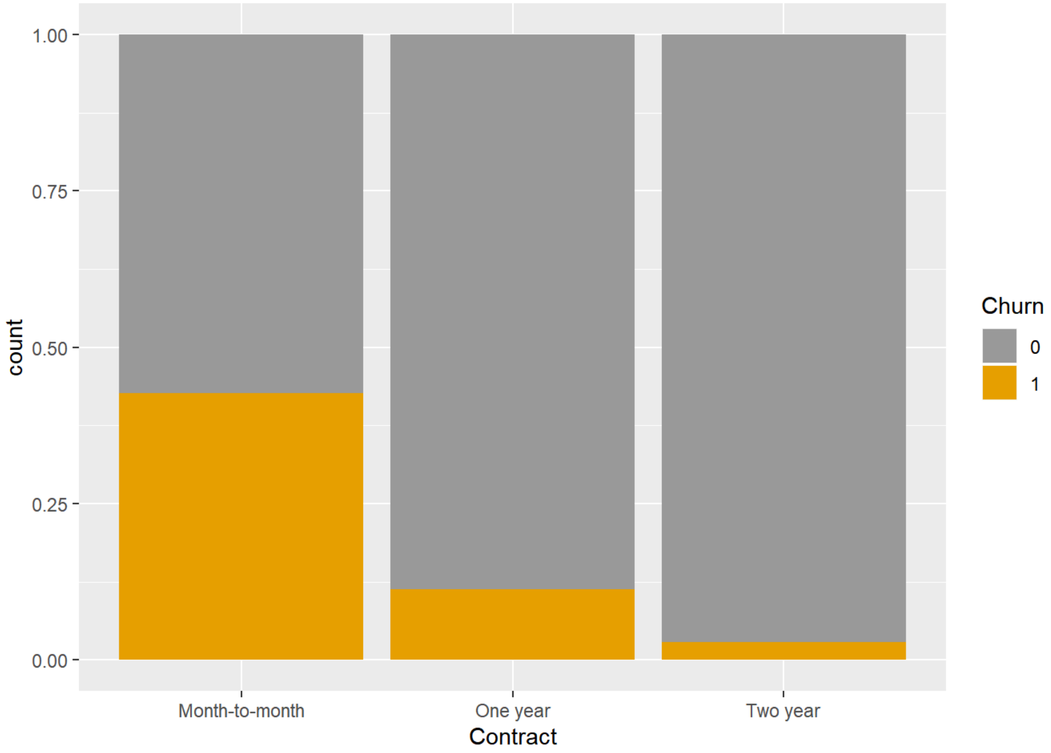
```
b12 <- ggplot(custc, aes(StreamingTV, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c(
"#999999", "#E69F00"))
b12
```



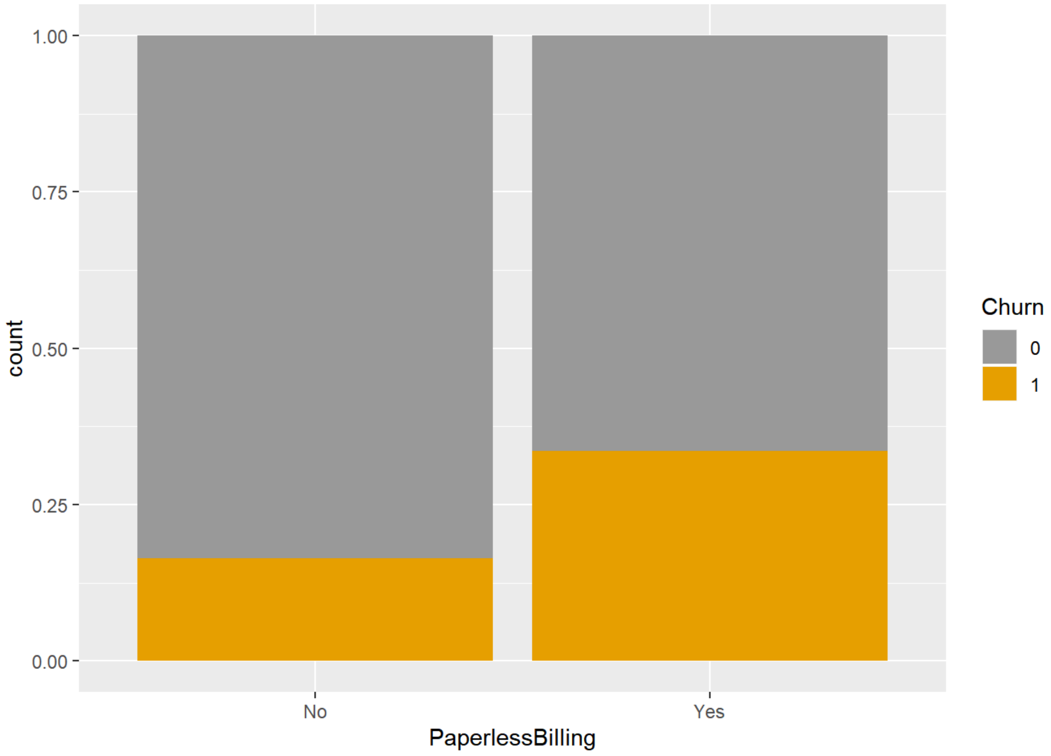
```
b13 <- ggplot(custc, aes(StreamingMovies, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(value
s=c("#999999", "#E69F00"))
b13
```



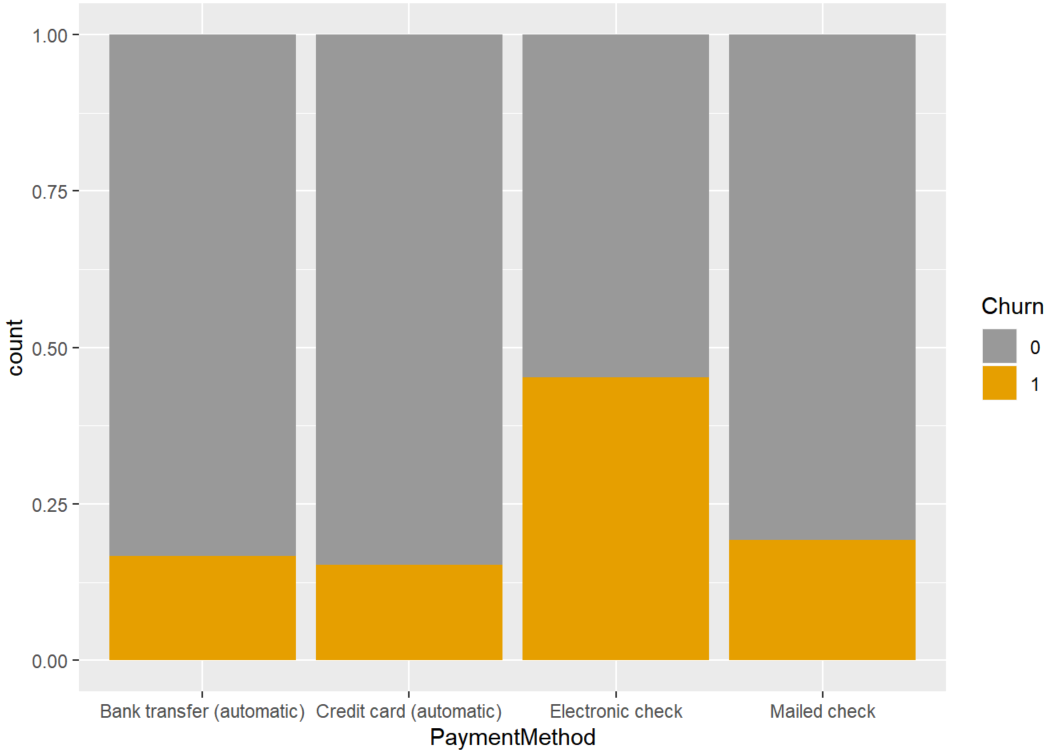
```
b14 <- ggplot(custc, aes(Contract, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c("#999999", "#E69F00"))
b14
```



```
b15 <- ggplot(custc, aes(PaperlessBilling, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=c("#999999", "#E69F00"))
b15
```



```
b16 <- ggplot(custc, aes(PaymentMethod, fill = Churn)) + geom_bar(position='fill')+scale_fill_manual(values=
c("#999999", "#E69F00"))
b16
```



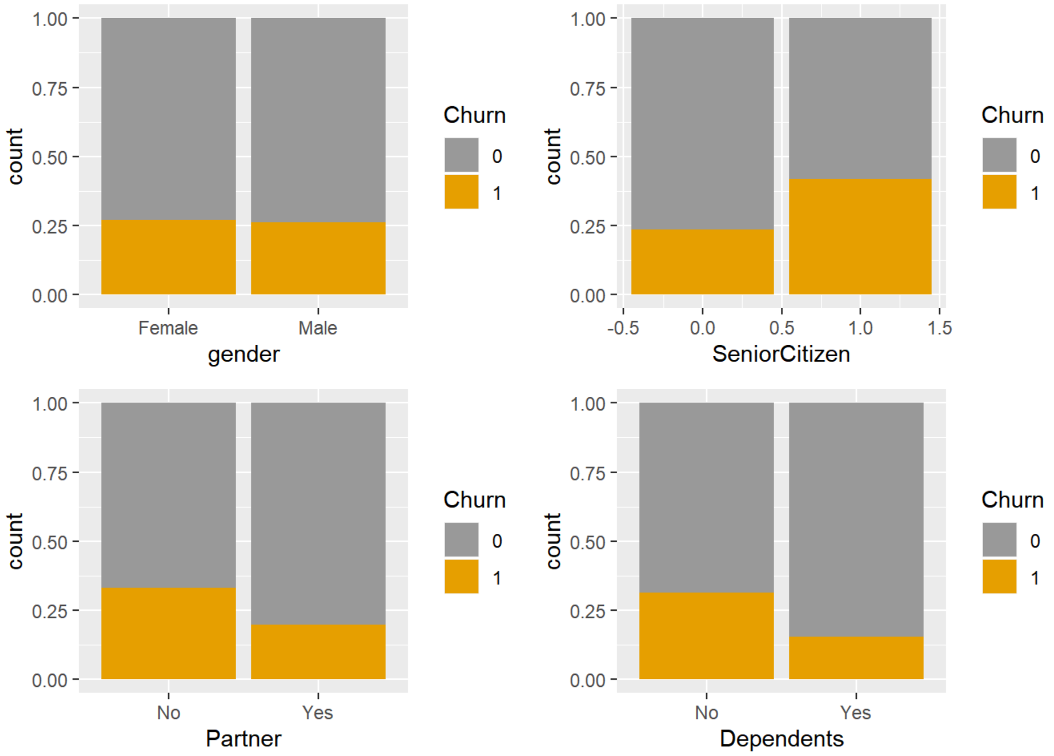
```
library(gridExtra)
```

```
## Warning: package 'gridExtra' was built under R version 3.6.2
```

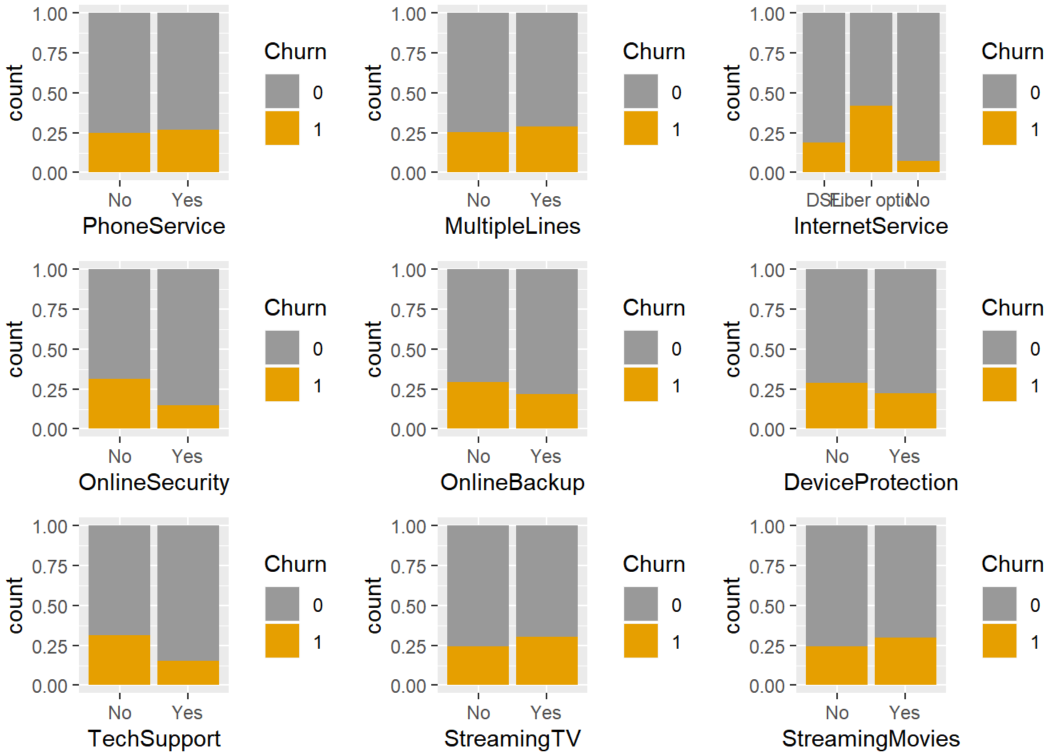
```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

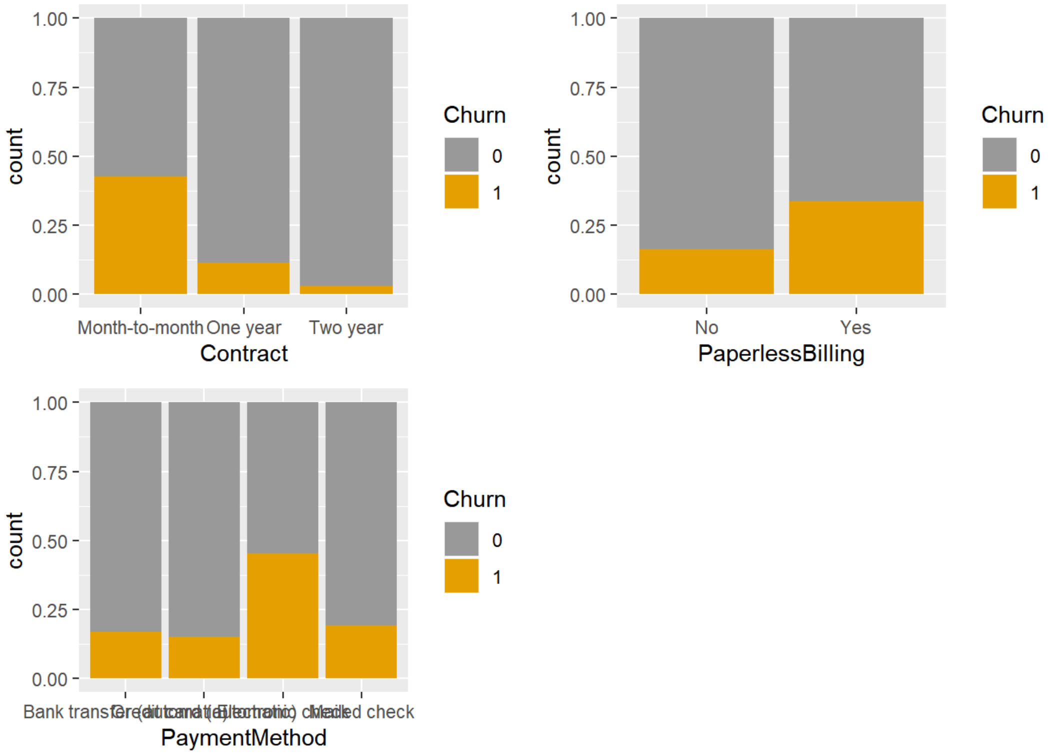
```
grid.arrange(b1,b2,b3,b4, ncol = 2)
```



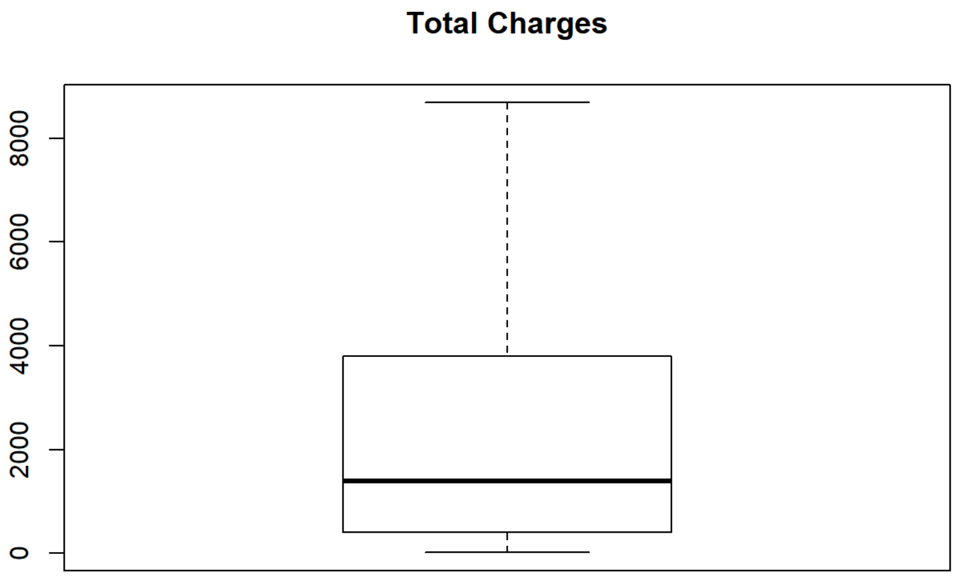
```
grid.arrange(b5,b6,b7,b8,b9,b10,b11,b12,b13, ncol = 3)
```



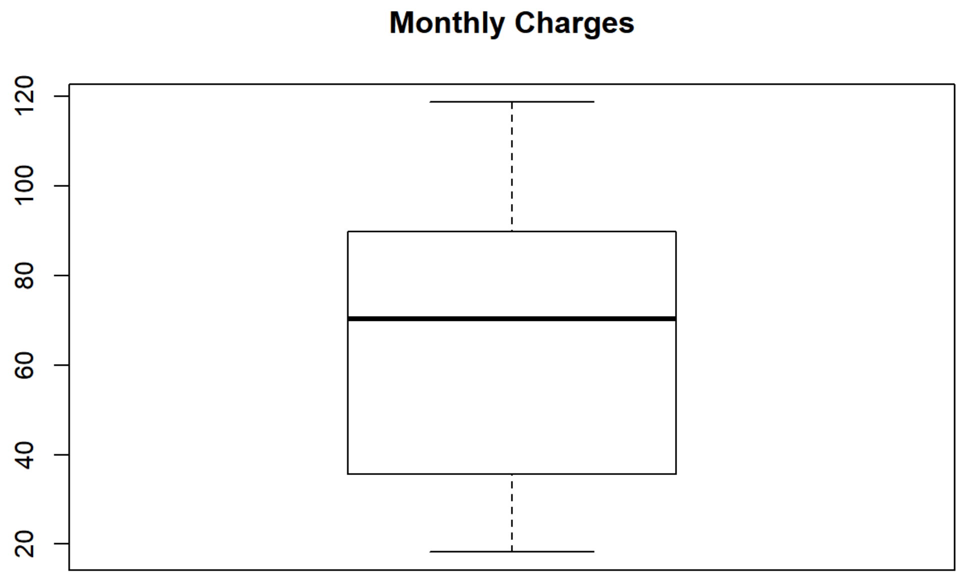
```
grid.arrange(b14,b15,b16, ncol = 2)
```



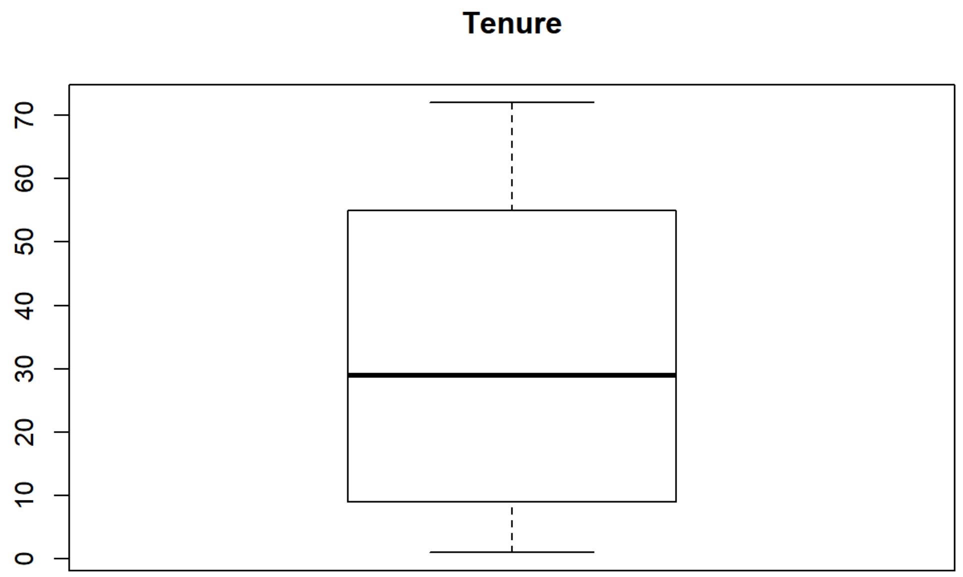
```
boxplot(custc$TotalCharges,data=custc, main="Total Charges")
```



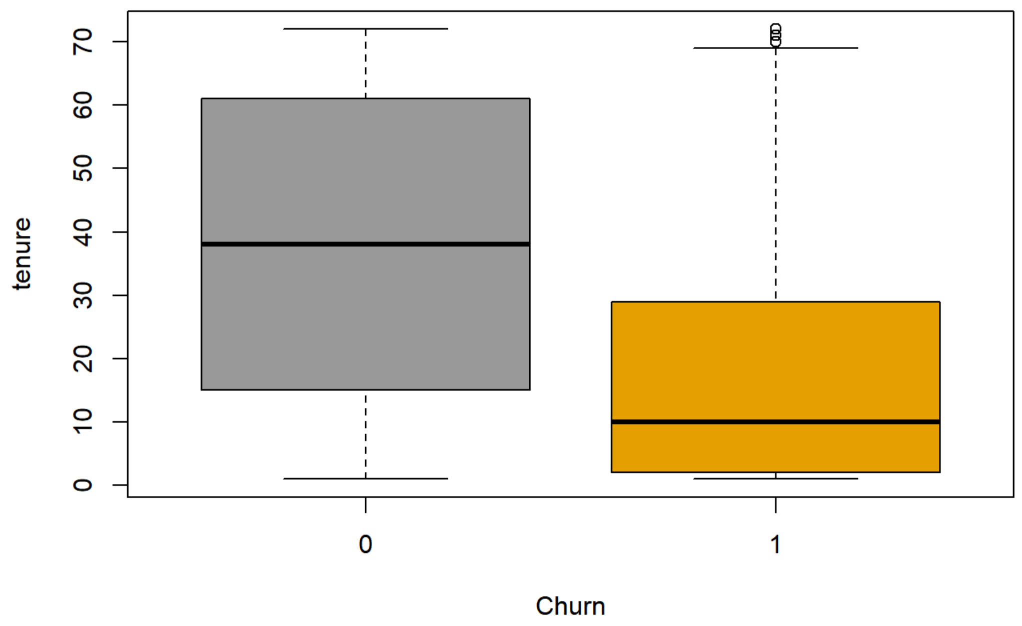
```
boxplot(custc$MonthlyCharges,data=custc, main="Monthly Charges")
```



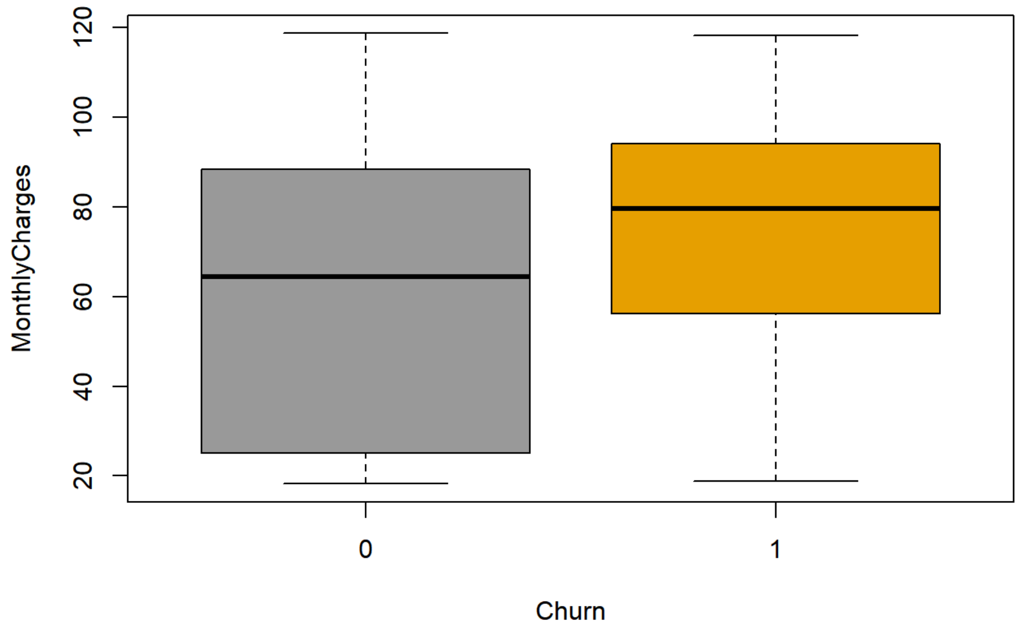
```
boxplot(custc$tenure,data=custc, main="Tenure")
```



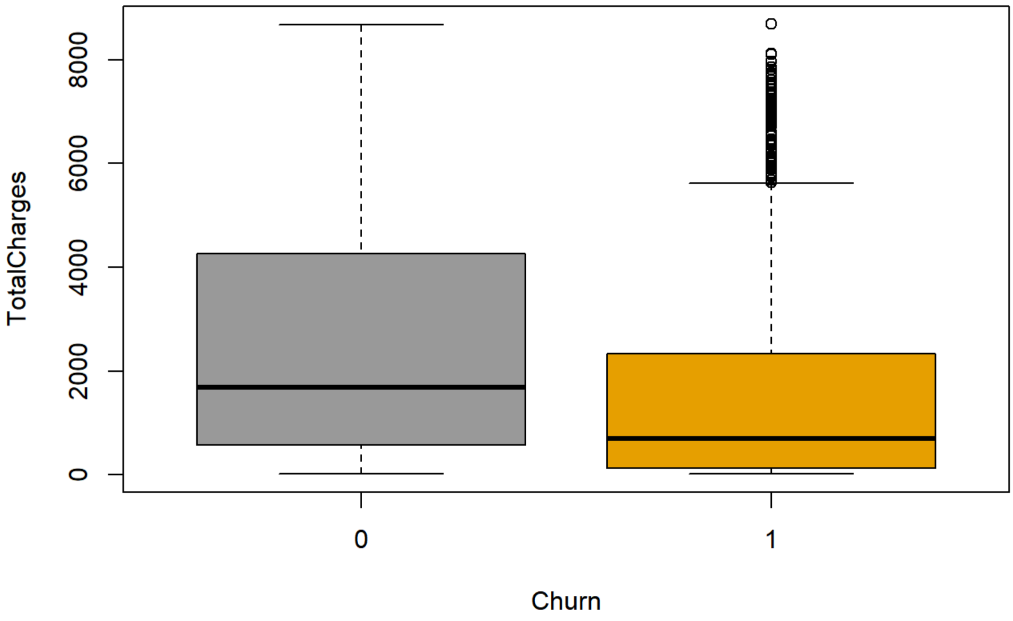
```
#Plotting Box Plots for Tenure ,Monthly Charges and Total Charges
b1 <- boxplot(tenure~Churn,data = custc,col = c("#999999","#E69F00"), xlab ="Churn" , ylab = "tenure")
```



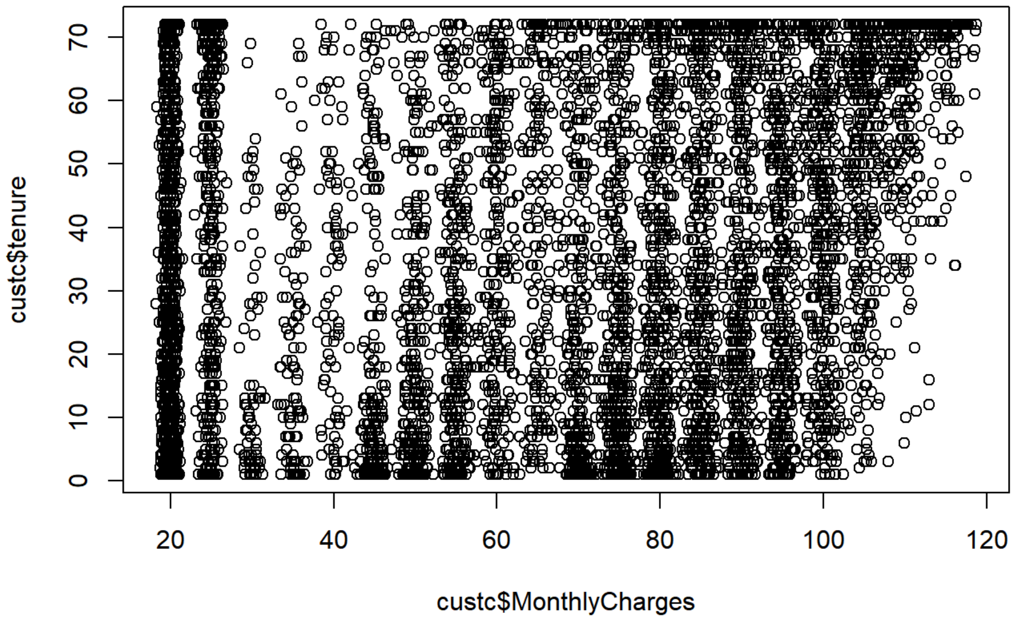
```
b2 <- boxplot(MonthlyCharges~Churn,data = custc,col = c("#999999","#E69F00"), xlab ="Churn" , ylab = "Monthl
yCharges")
```



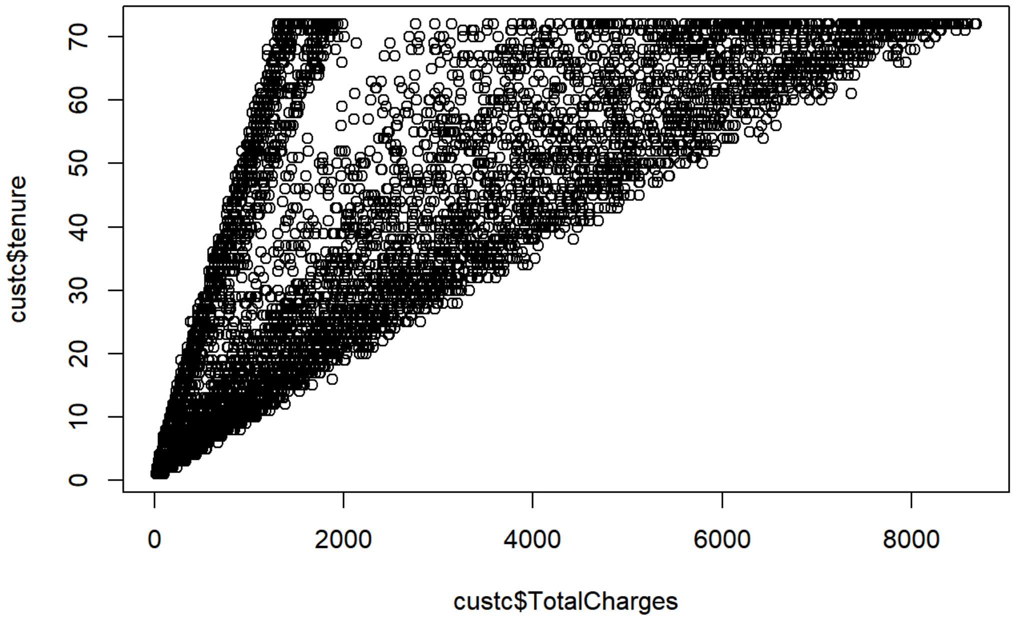
```
b3 <- boxplot(TotalCharges~Churn,data = custc,col = c("#999999","#E69F00"), xlab ="Churn" , ylab = "TotalCha
rges")
```



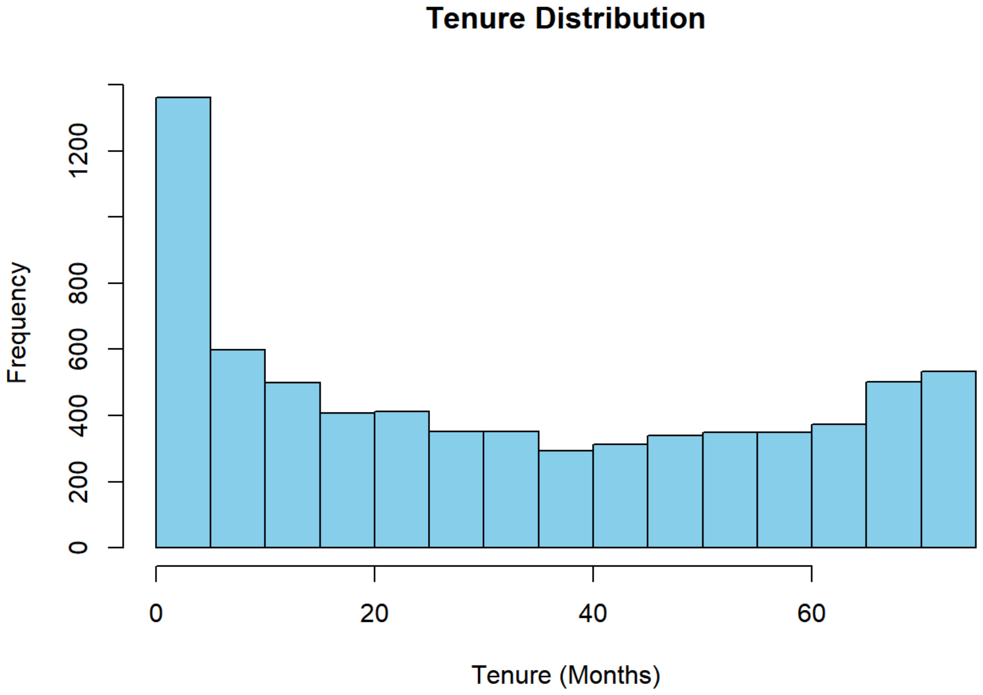
```
plot(custc$MonthlyCharges, custc$tenure)
```



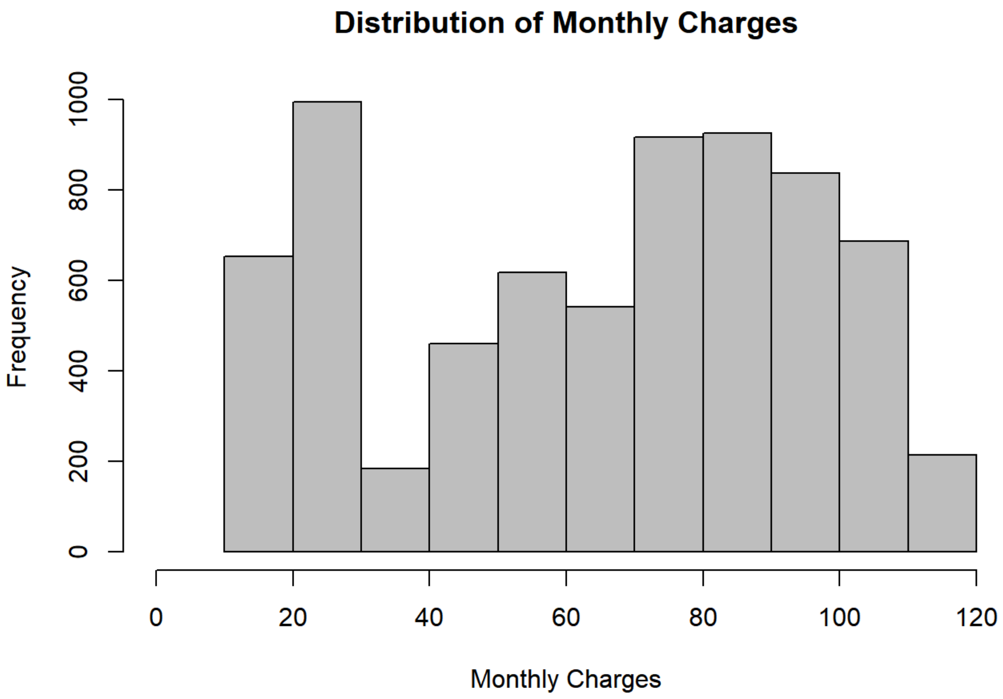
```
plot(custc$TotalCharges, custc$tenure)
```



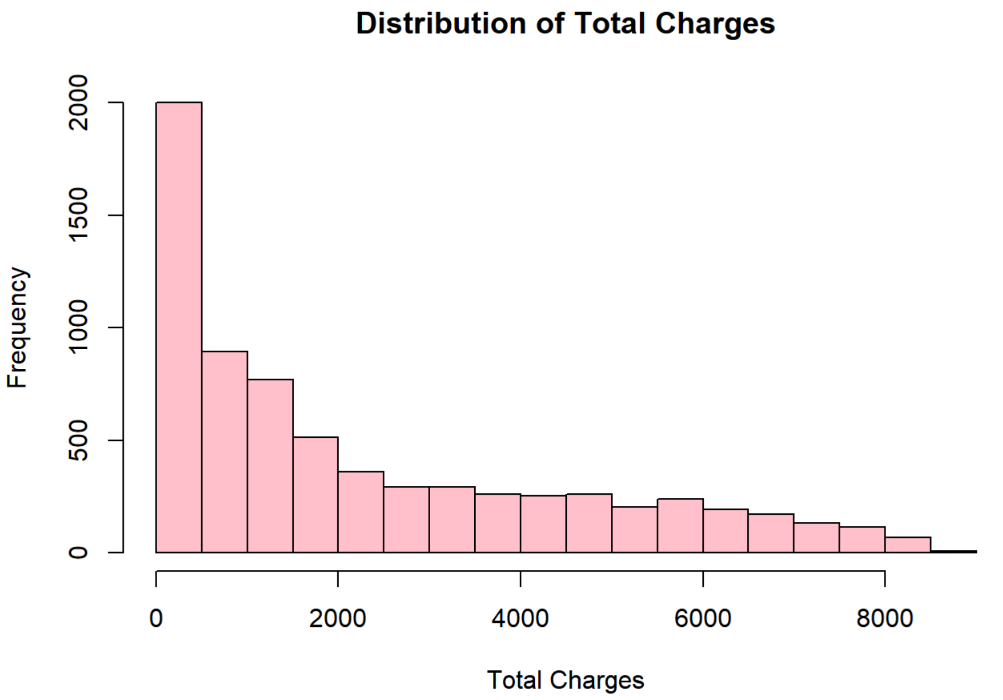
```
hist(custc$tenure, main="Tenure Distribution",col="sky blue",xlab="Tenure (Months)")
```



```
hist(custc$MonthlyCharges, main="Distribution of Monthly Charges",col="grey", xlab="Monthly Charges",xlim=c(
0,120),breaks=12)
```



```
hist(custc$TotalCharges, main="Distribution of Total Charges",col="pink", xlab="Total Charges")
```

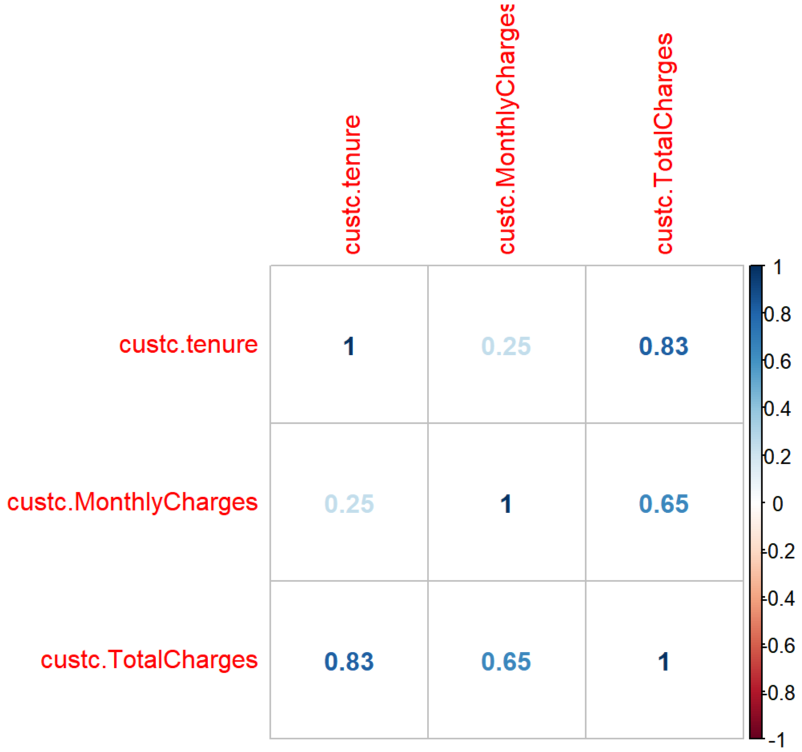


```
#Correlation between numeric variables

library(corrplot)
```

```
## Warning: package 'corrplot' was built under R version 3.6.2
```

```
## corrplot 0.84 loaded
```

```
cor_data <-data.frame(custc$tenure,custc$MonthlyCharges,custc$TotalCharges)
corr <- cor(cor_data)
corrplot(corr, method = "number")
```



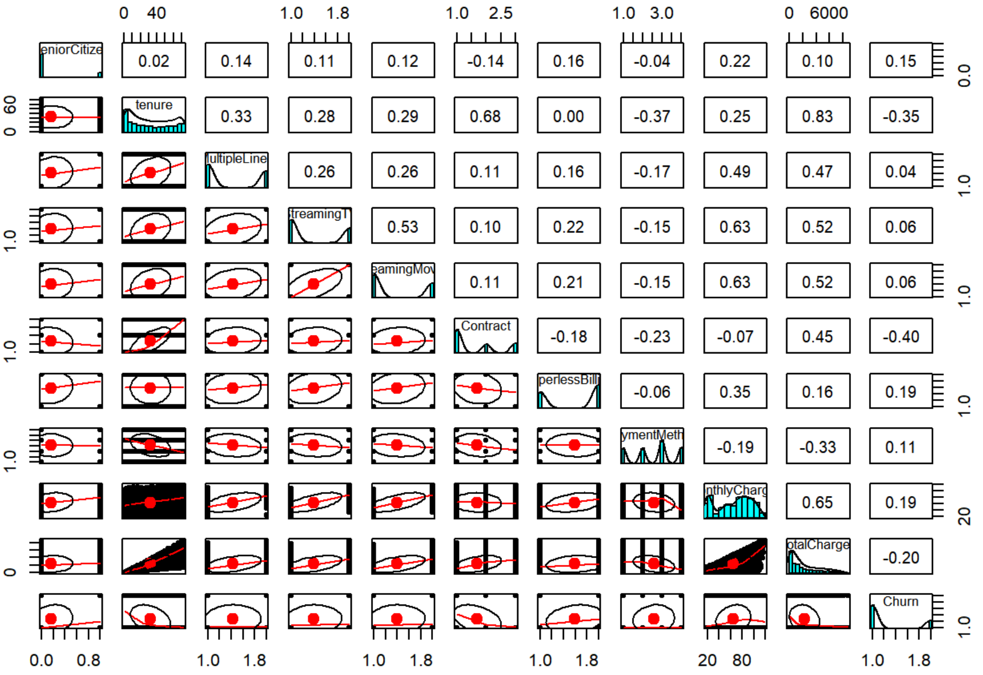
```
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.6.2
```

```
##
## Attaching package: 'psych'
```

```
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
pairs.panels(custc[c(3,6,8,14,15,16,17,18,19,20,21)])
```


eniorCitize
tenure
ultipleLine
treamingT
eamingMov
Contract
perlessBill
ymentMeth
nthlyCharg
otalCharge
Churn