Exploration data analysis

2024-06-05

Contents

Read file

```
Churn_file <- read.csv("~/Carnegie Mellon/24_Software design for data scientist/Final_Project/cmu-95829-
```

General summary of the data

We have 21 columns and attributes with 7,043 records. The median of monthly charges by customer is $70.35 meanwhile the Total charges $1,397.The average tenure is 32 months while 1,869 customers have left the company this quarter.

```
#No col and rows
nrow(Churn_file)
```

```
[1] 7043
```

```
ncol(Churn_file)
```

```
[1] 20
```

```
# Top 5 data rows
head (Churn_file, 5)
```

```
           gender SeniorCitizen Partner Dependents tenure PhoneService
7590-VHVEG Female             0     Yes         No      1           No
5575-GNVDE   Male             0      No         No     34          Yes
3668-QPYBK   Male             0      No         No      2          Yes
7795-CFOCW   Male             0      No         No     45           No
9237-HQITU Female             0      No         No      2          Yes
              MultipleLines InternetService OnlineSecurity OnlineBackup
7590-VHVEG No phone service             DSL             No          Yes
5575-GNVDE               No             DSL            Yes           No
```

```
3668-QPYBK              No             DSL            Yes          Yes
7795-CFOCW No phone service            DSL            Yes           No
9237-HQITU              No     Fiber optic             No           No
           DeviceProtection TechSupport StreamingTV StreamingMovies
7590-VHVEG               No          No          No              No
5575-GNVDE              Yes          No          No              No
3668-QPYBK               No          No          No              No
7795-CFOCW              Yes         Yes          No              No
9237-HQITU               No          No          No              No
                 Contract PaperlessBilling             PaymentMethod
7590-VHVEG Month-to-month              Yes          Electronic check
5575-GNVDE       One year               No              Mailed check
3668-QPYBK Month-to-month              Yes              Mailed check
7795-CFOCW       One year               No Bank transfer (automatic)
9237-HQITU Month-to-month              Yes          Electronic check
           MonthlyCharges TotalCharges Churn
7590-VHVEG          29.85        29.85    No
5575-GNVDE          56.95      1889.50    No
3668-QPYBK          53.85       108.15   Yes
7795-CFOCW          42.30      1840.75    No
9237-HQITU          70.70       151.65   Yes
```

```
# Summary of key attributes
summary(Churn_file[,c(5,18,19,20)])
```

```
     tenure      MonthlyCharges    TotalCharges    Churn
 Min.   : 0.00   Min.   : 18.25   Min.   :  18.8   No :5174
 1st Qu.: 9.00   1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
 Median :29.00   Median : 70.35   Median :1397.5
 Mean   :32.37   Mean   : 64.76   Mean   :2283.3
 3rd Qu.:55.00   3rd Qu.: 89.85   3rd Qu.:3794.7
 Max.   :72.00   Max.   :118.75   Max.   :8684.8
                                  NA’s   :11
```

```
str(Churn_file)
```

```
’data.frame':   7043 obs. of  20 variables:
 $ 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 ...
```

We have monthly charges, total charges and tenure as numerical values. The others are factors or categorical values. ## Data quality and preparation The database contains 11 missing values in "total charges" column.

```
# To validate if there are missing values per column
missing_values <- colSums(is.na(Churn_file))
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               11                0
```

We will remove those 11 rows.

```
# To remove rows
Churn_file <- na.omit(Churn_file)
nrow(Churn_file)
```

```
[1] 7032
```

Now, we have 7,032 records.

We also need to convert senior citizen as factor

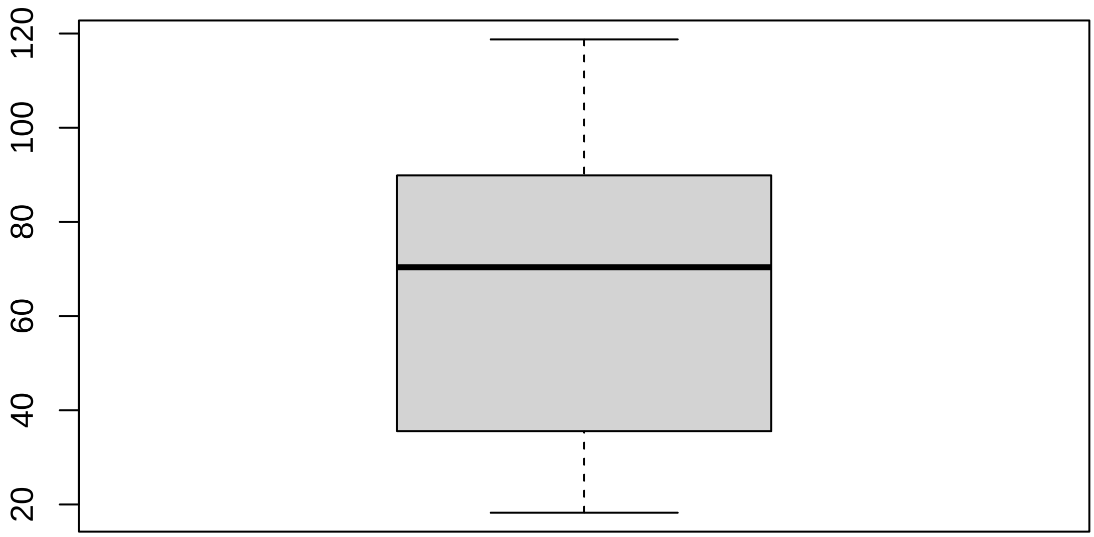
```
Churn_file <- Churn_file %>%
  mutate(SeniorCitizen=as.factor(SeniorCitizen))
```

Now, we change the churn attribute in a numeric value in a new column

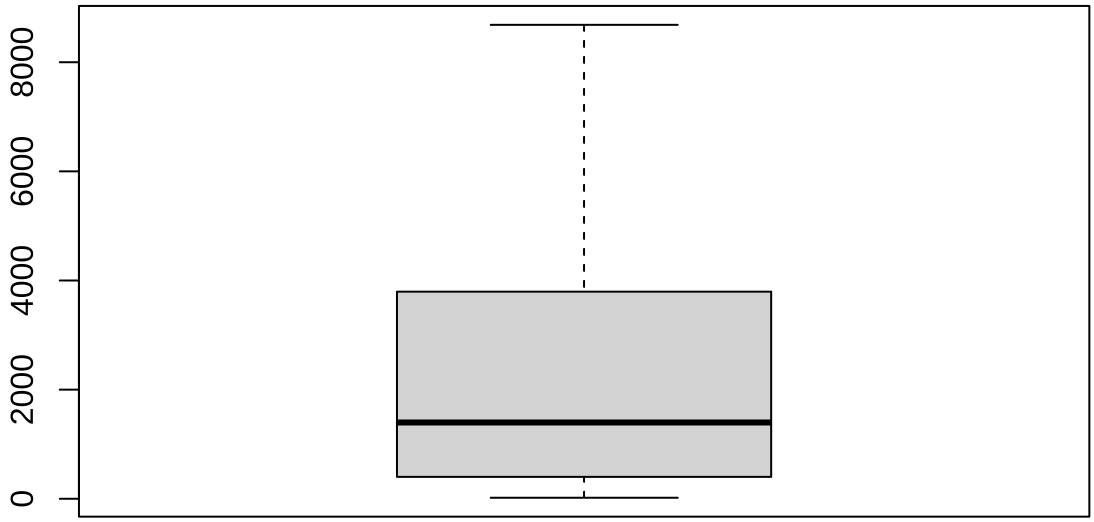
```
# Change the target variable in a numeric value
Churn_file$churn_numeric <- ifelse(Churn_file$Churn == "Yes", 1, 0)
```

Checking outliers for numeric values: tenure, total and monthly charges

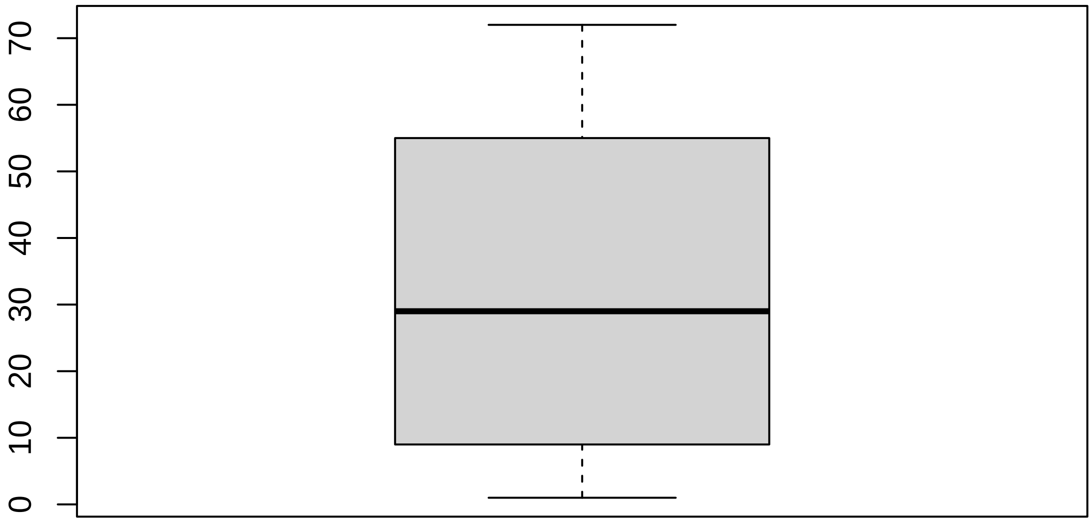
```
# Validate if there are outliers
boxplot(Churn_file$MonthlyCharges)
```



```
boxplot(Churn_file$TotalCharges)
```



```
boxplot(Churn_file$tenure)
```

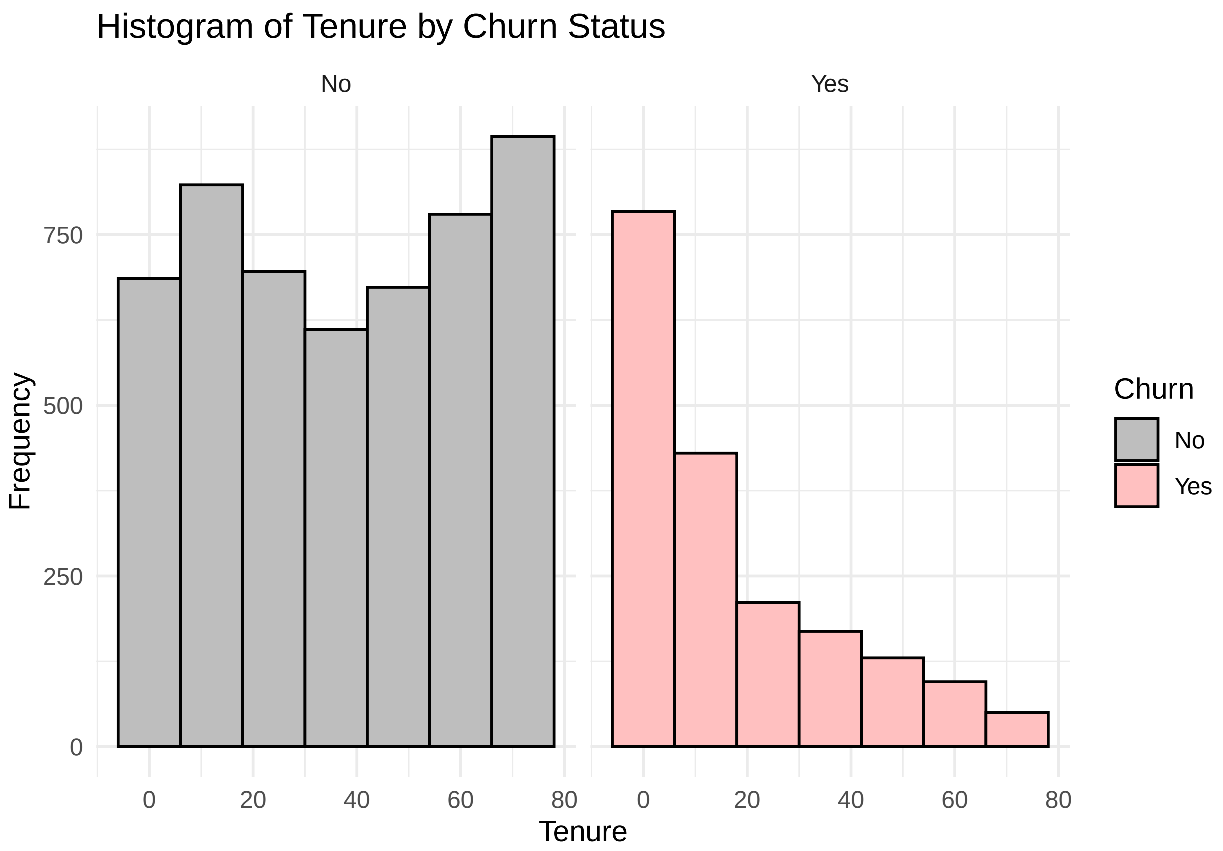


In this dataset we do not have outliers.

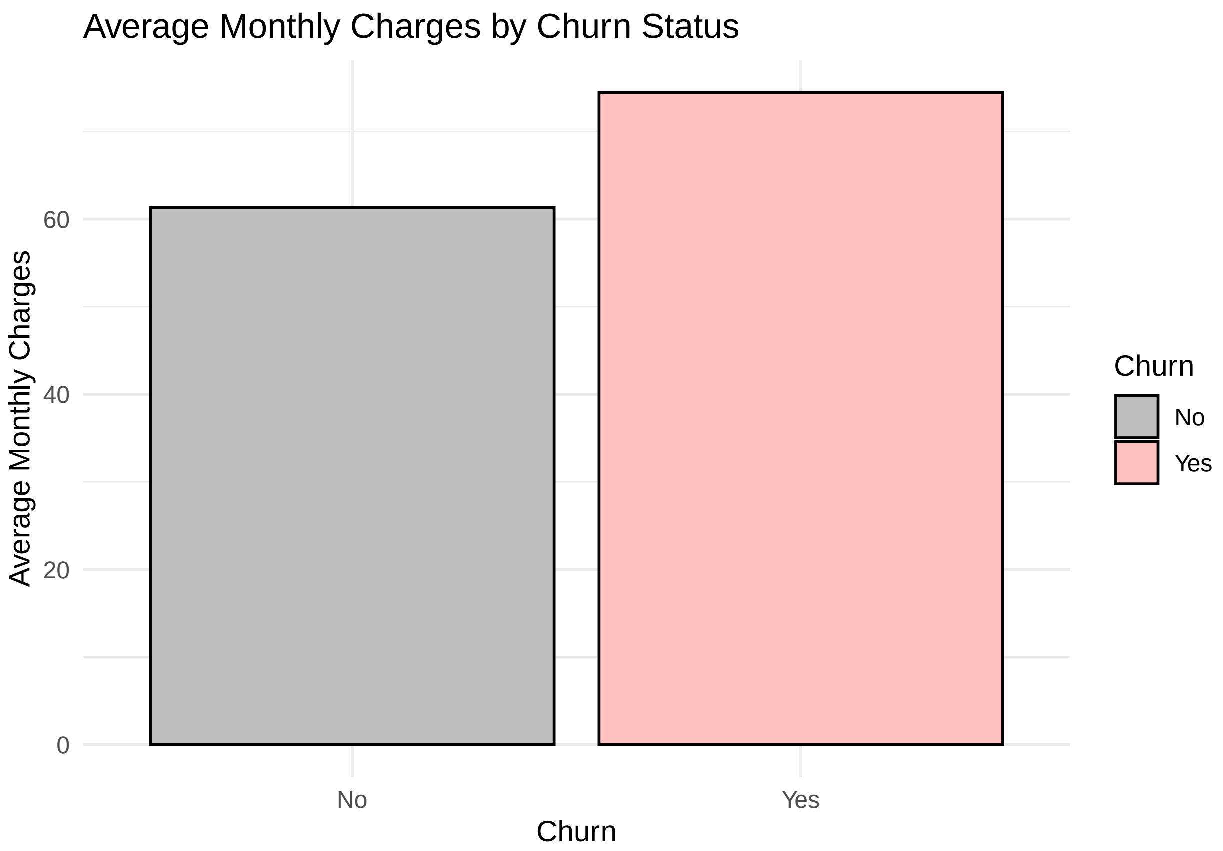
Exploration data analysis

In the histogram, We see tenure is left skewed with churn customers, this means that they tend to leave the company within 12 months. On the other hand, the bar graph shows the average monthly charges in Churn customers was higher than non-churn customers.

```r
light_red <- rgb(255, 192, 192, maxColorValue = 255)

#Histograms by tenure
ggplot(Churn_file, aes(x = tenure, fill = Churn)) +
  geom_histogram(binwidth = 12, color = "black", position = "dodge") +
  facet_wrap(~ Churn) +
  scale_fill_manual(values = c("Yes" = light_red, "No" = "gray")) +
  labs(title = "Histogram of Tenure by Churn Status", x = "Tenure", y = "Frequency") +
  theme_minimal()
```

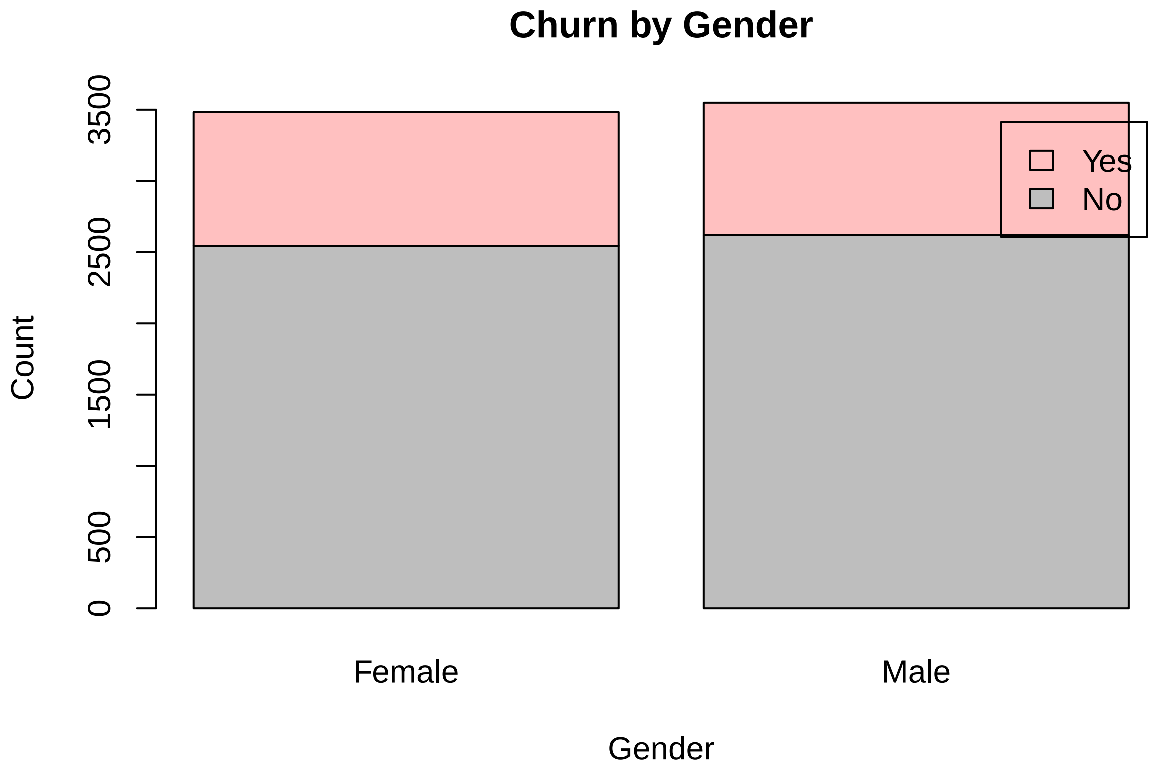


```
#Bar graphs of monthly charges
avg_churn <- Churn_file %>%
  group_by(Churn) %>%
  summarise(AverageMonthlyCharges = mean(MonthlyCharges, na.rm = TRUE))

ggplot(avg_churn, aes(x = Churn, y = AverageMonthlyCharges, fill = Churn)) +
  geom_bar(stat = "identity", color = "black") +
  scale_fill_manual(values = c("Yes" = light_red, "No" = "gray")) +
  labs(title = "Average Monthly Charges by Churn Status", x = "Churn", y = "Average Monthly Charges") +
  theme_minimal()
```



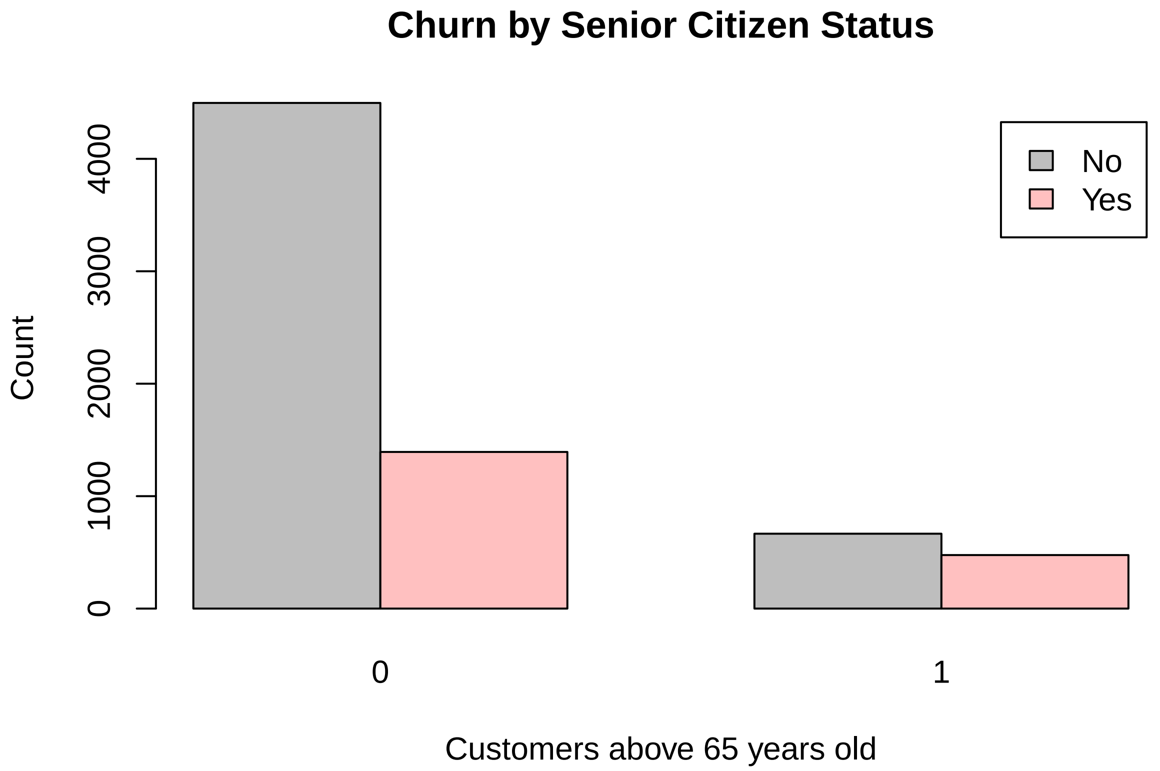
Gender by churn is not a relevant attribute considering the proportion of churn is similar among male and females.

```r
churn_gender_table <- table(Churn_file$Churn, Churn_file$gender)

# gender bar plot
barplot(churn_gender_table, legend = rownames(churn_gender_table), col = c("gray", light_red),
        main = "Churn by Gender", xlab = "Gender", ylab = "Count")
```



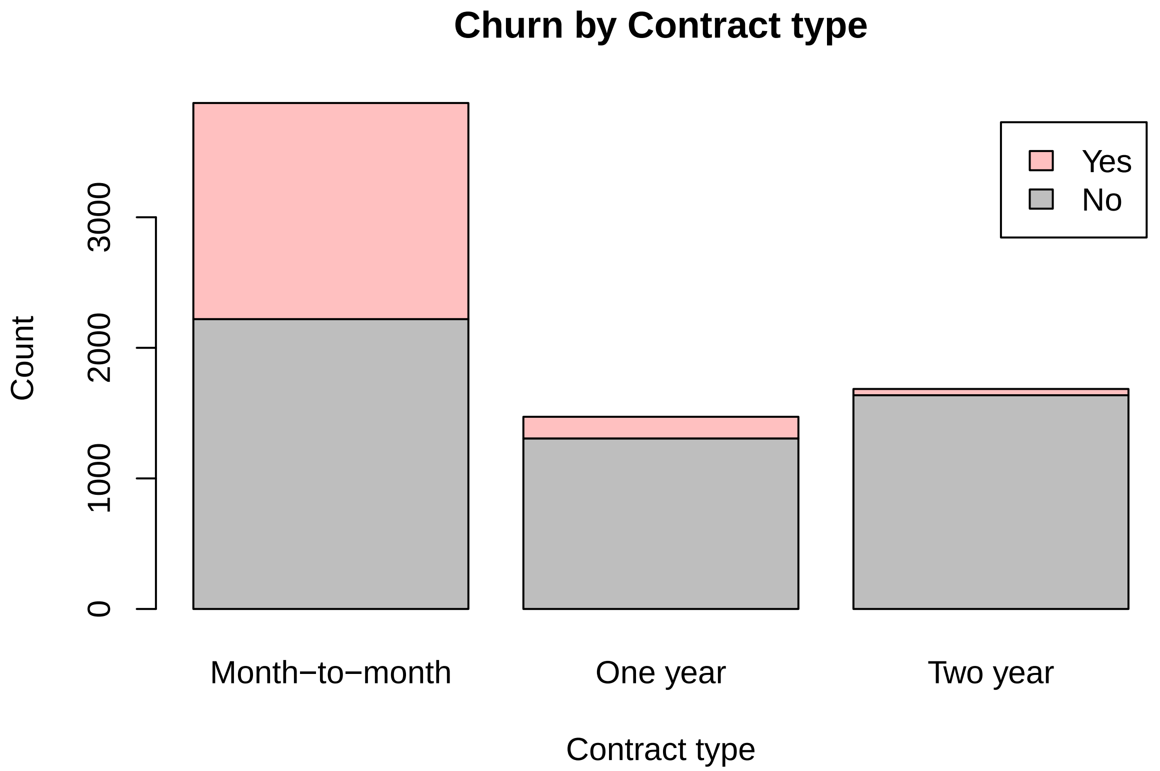
We can see the majority of churn is in people below 65 yeards old. However, in those senior citizen customers (above 65yrs) the churn rate is higher.

```
churn_sc_table <- table(Churn_file$Churn, Churn_file$SeniorCitizen)

# Seniorcitizen graph bar
barplot(churn_sc_table, beside = TRUE, legend = TRUE,
        col = c("gray", light_red),
        main = "Churn by Senior Citizen Status",
        xlab = "Customers above 65 years old", ylab = "Count",)
```



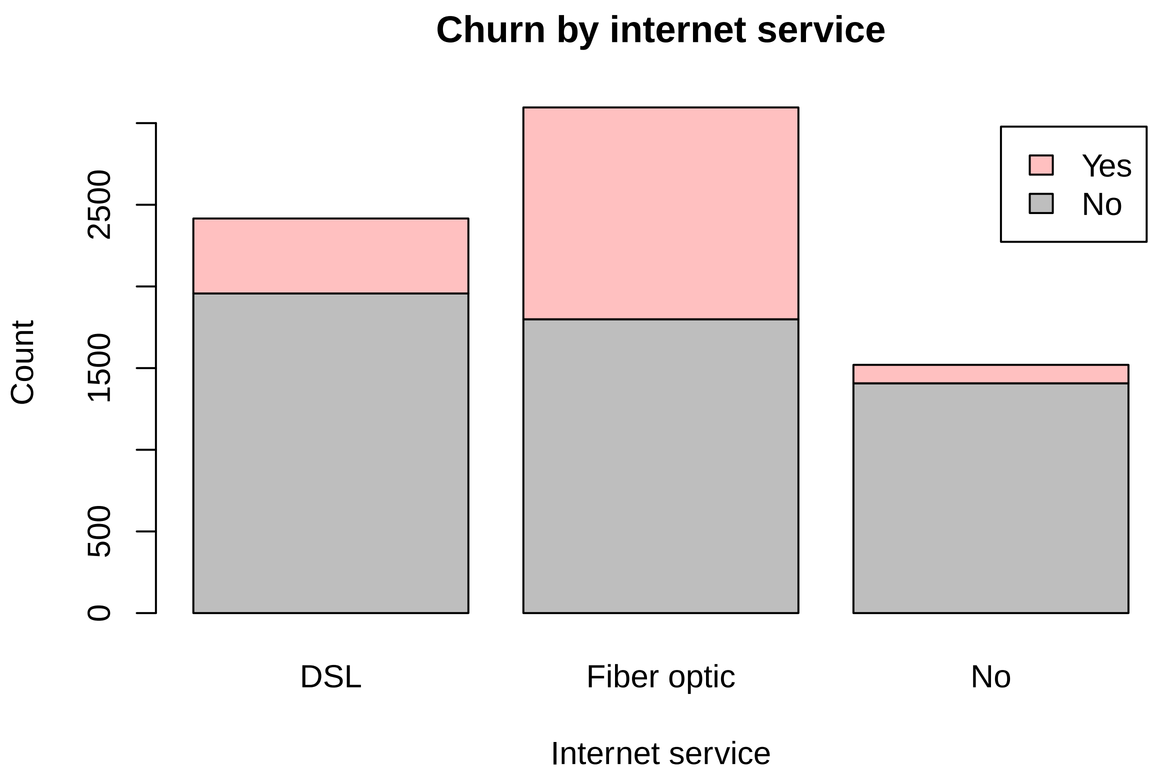
Churn is more common in customers that pay in a monthly basis. Having a long term contract (>1 year) with customers could lead to more retention.

```r
churn_contract_table <- table(Churn_file$Churn, Churn_file$Contract)

# Contract bar plot
barplot(churn_contract_table, legend = rownames(churn_contract_table), col = c("gray", light_red),
        main = "Churn by Contract type", xlab = "Contract type", ylab = "Count")
```



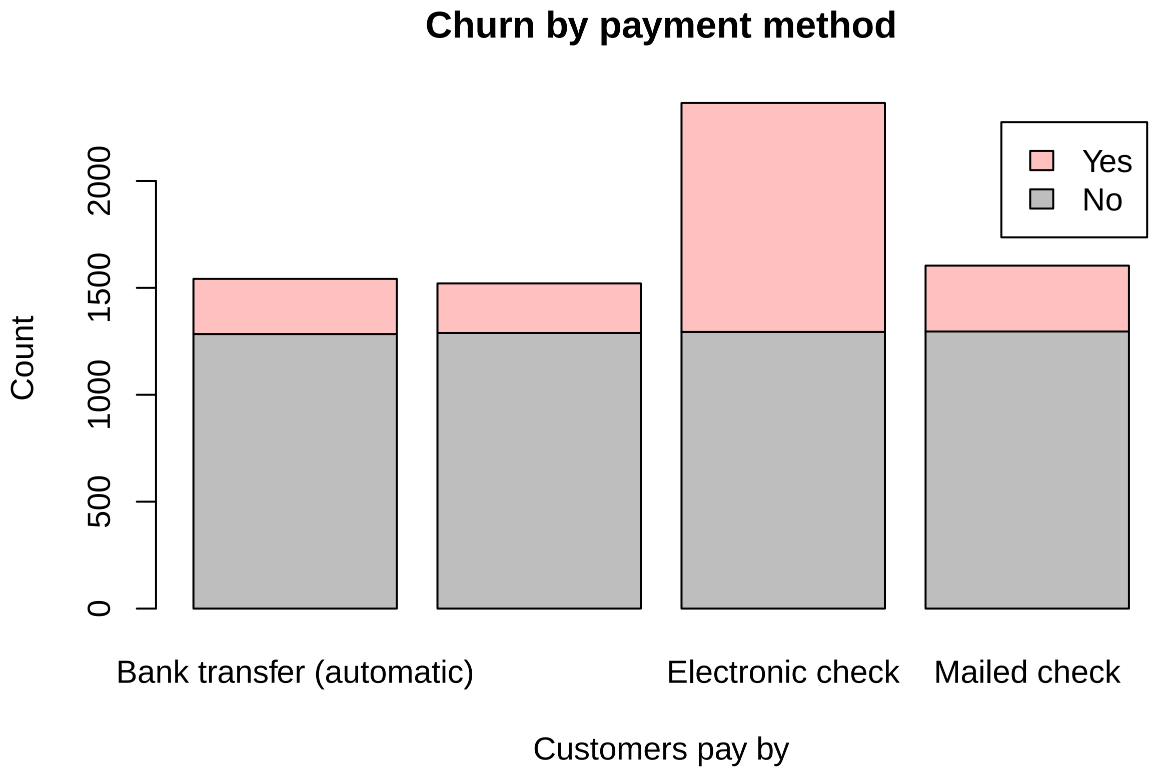
Customers with fiber optic have more chances to churn.

```
churn_internet_table <- table(Churn_file$Churn, Churn_file$InternetService)

# Internet service bar plot
barplot(churn_internet_table, legend = rownames(churn_internet_table), col = c("gray", light_red),
        main = "Churn by internet service", xlab = "Internet service", ylab = "Count")
```



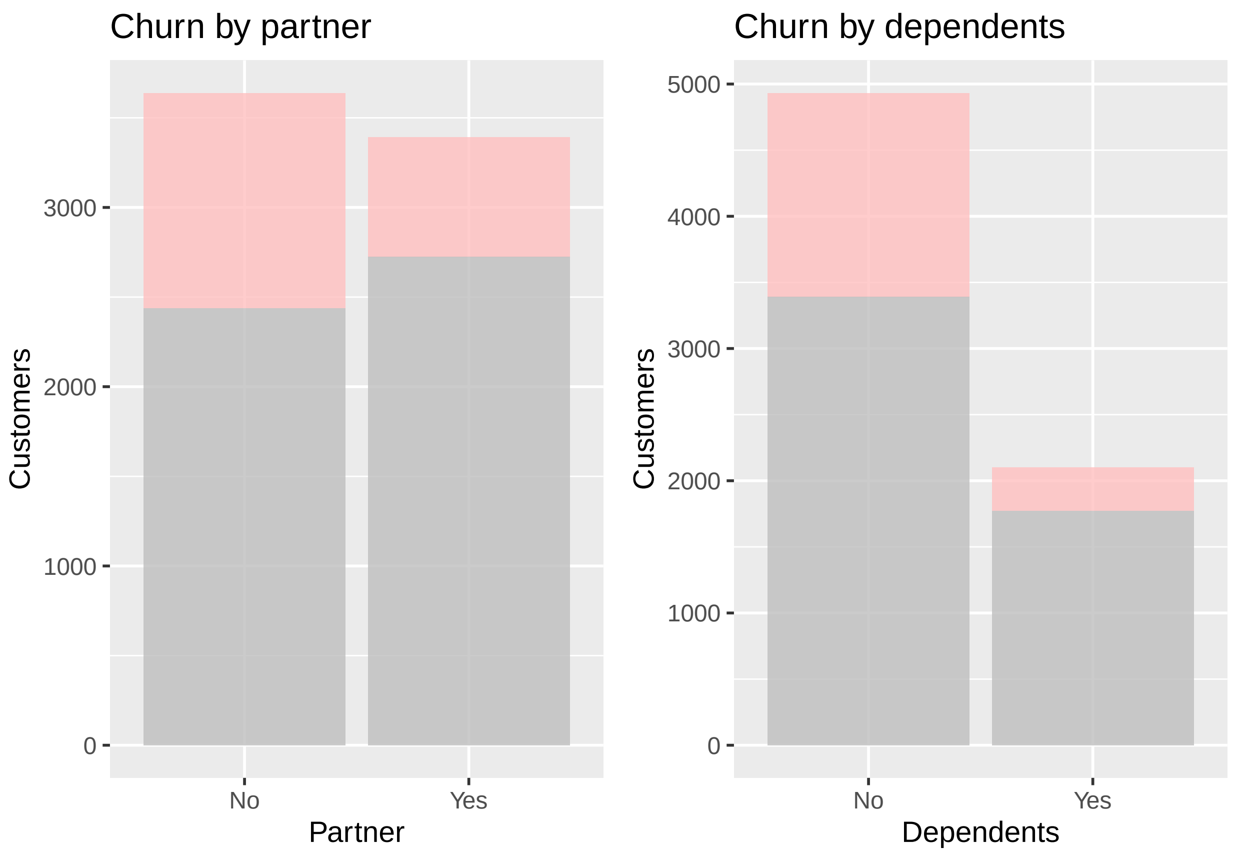
Customers that pay via transfer or credit/debit card have less churn rate.

```
churn_pay_table <- table(Churn_file$Churn, Churn_file$PaymentMethod)

# Payment method bar plot
barplot(churn_pay_table, legend = rownames(churn_pay_table), col = c("gray", light_red),
        main = "Churn by payment method", xlab = "Customers pay by", ylab = "Count")
```



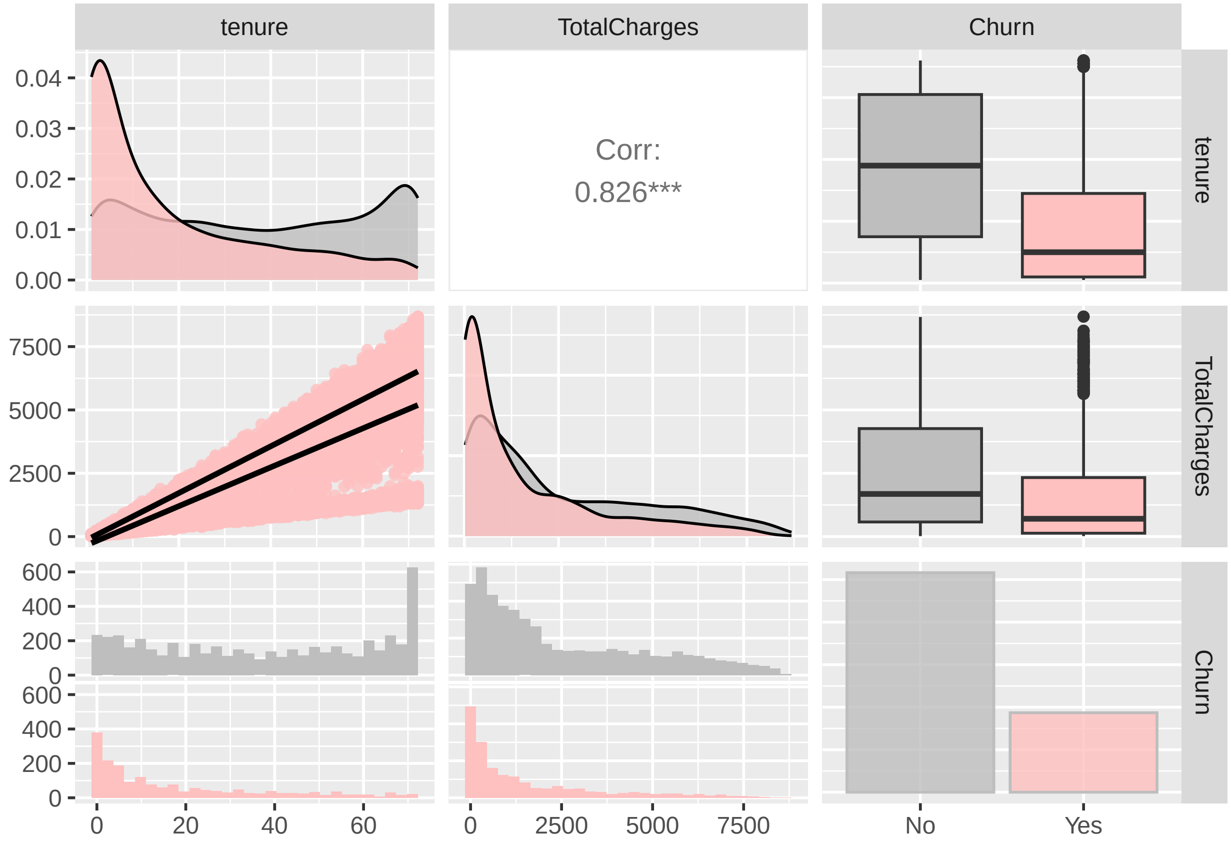
Additionally, we can see customers with dependents and partners are less likely to churn.

```r
partner_plot <- Churn_file %>% ggplot(aes(x=Partner, fill=fct_rev(Churn_file$Churn))) +  geom_bar(alpha
dependents_plot <- Churn_file %>% ggplot(aes(x=Dependents, fill=fct_rev(Churn_file$Churn)))+  geom_bar(

grid.arrange(partner_plot, dependents_plot, ncol=2)
```



At the beginning we see that churn customers have higher monthly charges and less tenure. Now, let's graph the correlation of total charges and tenure.

```
Churn_file %>%
  dplyr::select(tenure, TotalCharges, Churn) %>%
ggpairs(aes(fill = Churn),
        diag = list(continuous = wrap("densityDiag", alpha = 0.8),
                    discrete = wrap("barDiag", alpha = 0.8, color = "gray")),
        lower = list(continuous = wrap("smooth", alpha = 0.8, color = light_red),
                      discrete = wrap("points"))) +
  scale_fill_manual(values = c("gray", light_red))
```



We see a strong positive linear relationship between the tenure and total charges with 0.8. The boxplots show a significant difference in the median of tenure by churn and active customers.