Analysis of Categorical Data - Assignment - Phase 1

MATH 1298 Analysis of Categorical Data Project Phase I

Amal Joy (s3644794) & Nupura Sanjay Sawle (s3639703)

08 September 2018

Contents

1 Introduction

The aim of this project is to build a customer churn model for a telecommunication company to predict the customers who are about to get churned so that they can implement different business strategies to retain those customers before they actually get churned. The tool that I am using for this analysis is R-studio. This project will be conducted in 2 different phase where I will be exploring different data analysis techniques to accurately predict the churned customers. Phase 1 of this project will include the detailed descriptive statistical analysis of the data by making use of various R packages to build relevant charts, graphs, and interactions etc. Data preprocessing will be done to clean and transform the data to suit the prediction model. This section will address the issues with the dataset like missing values, outliers etc. The second phase of this project is solely for the model building and will include appropriate statistical procedures, the test of independence etc.

In this phase 1 of the project, the business rules of the telecommunication company are Identified to have a deep knowledge in customer churn in the telecommunication industry. Studies show that 5% the customer retention result in the 25% to 95% the overall profit of the companies (Forbes, Jerry Jao, 2015). The major initiators of churn may include the quality of service, tariffs, unsatisfactory post-sales service etc. Early identification of the customers who have the more chance of falling into the category of churned customers may help the marketing team to device cost-effective ways to build retention offers to target only those customers. It is also important to keep in mind that the incorrect predictions may lead to marketing team spending more time, resources and effort in trying to retain those customers who are not going to leave the company even if they were not provided with those offers. This means a company losing discounts offered to them.

2 Dataset source and description

The following packages are used in this report for data preparation and data modeling.

```r
library(dplyr)
library(knitr)
library(kableExtra)
library(mlr)
library(cowplot)
library(ggplot2)
library(corrplot)
```

The data was read into a data file named 'telcom_churn'. Null values are replaced with 'NA' while reading the file.

```r
telcom_churn <- read.csv(
  "/Users/amaljoy/Study/Categorcal Data/Assignment 1/Telco-Customer-Churn.csv",
  header=T,na.strings=c("","NA")) # Reading the data

dim(telcom_churn) # dimensions of the dataset
```

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

The dataset consists of 21 variables and 7043 observations. Each row in the dataset is the attributes associated to a customer, each column contains customer's attributes. The customer attributes are provided below: * customerID (Unique customer identification)
* gender (female, male)
* SeniorCitizen (Whether the customer is a senior citizen or not (1, 0))
* Partner (Whether the customer has a partner or not (Yes, No))
* Dependents (Whether the customer has dependents or not (Yes, No))

* tenure (Number of months the customer has stayed with the company)
* PhoneService (Whether the customer has a phone service or not (Yes, No))
* MultipleLines (Whether the customer has multiple lines r not (Yes, No, No phone service)
* InternetService (Customer's internet service provider (DSL, Fiber optic, No)
* OnlineSecurity (Whether the customer has online security or not (Yes, No, No internet service)
* OnlineBackup (Whether the customer has an online backup or not (Yes, No, No internet service)
* DeviceProtection (Whether the customer has device protection or not (Yes, No, No internet service)
* TechSupport (Whether the customer has tech support or not (Yes, No, No internet service)
* streamingTV (Whether the customer has streaming TV or not (Yes, No, No internet service)
* streamingMovies (Whether the customer has streaming movies or not (Yes, No, No internet service)
* Contract (The contract term of the customer (Month-to-month, One year, Two year)
* Paperless billing (Whether the customer has paperless billing or not (Yes, No))
* PaymentMethod (The customer's payment method (Electronic check, Mailed check, Bank transfer (automatic), Credit card (automatic)))
* MonthlyCharges (The amount charged to the customer monthly - numeric)
* TotalCharges (The total amount charged to the customer - numeric)
* Churn (Whether the customer churned or not (Yes or No))

The attribute churn will be the target variable. Given below is the first 5 observations in the dataset.

Table 1: Head of the data

customerID	gender	SeniorCitizen	Partner	Dependents	tenure	PhoneService	MultipleLines
7590-VHVEG	Female	0	Yes	No	1	No	No phone service
5575-GNVDE	Male	0	No	No	34	Yes	No
3668-QPYBK	Male	0	No	No	2	Yes	No
7795-CFOCW	Male	0	No	No	45	No	No phone service
9237-HQITU	Female	0	No	No	2	Yes	No
9305-CDSKC	Female	0	No	No	8	Yes	Yes

InternetService	OnlineSecurity	OnlineBackup	DeviceProtection	TechSupport	StreamingTV
DSL	No	Yes	No	No	No
DSL	Yes	No	Yes	No	No
DSL	Yes	Yes	No	No	No
DSL	Yes	No	Yes	Yes	No
Fiber optic	No	No	No	No	No
Fiber optic	No	No	Yes	No	Yes

StreamingMovi	Contract	PaperlessBilling	PaymentMetho	MonthlyCharge	TotalCharges	Churn
No	Month-to-month	Yes	Electronic check	29.85	29.85	No
No	One year	No	Mailed check	56.95	1889.50	No
No	Month-to-month	Yes	Mailed check	53.85	108.15	Yes
No	One year	No	Bank transfer (automatic)	42.30	1840.75	No
No	Month-to-month	Yes	Electronic check	70.70	151.65	Yes
Yes	Month-to-month	Yes	Electronic check	99.65	820.50	Yes

2.1 Summary Statistics

Summary of the dataset is given below. It shows the summary of each variable including the levels in case of factors; range and central tendency values in case of numerical values.

```r
# Summary of the data
summary(telcom_churn)
```

```
##       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 
##  No                  :2810   No                  :2785  
##  No internet service:1526   No internet service:1526  
##  Yes                 :2707   Yes                 :2732  
##                                                         
##                                                         
##                                                         
##                                                         
##            Contract    PaperlessBilling                   PaymentMethod
##  Month-to-month:3875   No :2872         Bank transfer (automatic):1544  
##  One year      :1473   Yes:4171         Credit card (automatic)  :1522  
##  Two year      :1695                    Electronic check         :2365  
##                                         Mailed check             :1612  
##                                                                         
##                                                                         
```

```
##
##  MonthlyCharges    TotalCharges    Churn
##  Min.   : 18.25   Min.   :  18.8   No :5174
##  1st Qu.: 35.50   1st Qu.: 401.4   Yes:1869
##  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
```

From the summary statistics, it is clear that ‘customerID’ is an unusable attribute as is it a factor and is different for each observation. The male and female proportions in the dataset are almost similar. The attribute showing whether the customer is a senior citizen or not is a binary variable. But it is represented as a numerical variable. Data is distributed almost equally between people who have partners and those who do not. People with dependents are under-represented in this dataset. There is data for 72 months as represented by the variable ‘tenure’. The minimum value of ‘tenure’ is 0 showing those customers who have joined within a month of collecting the data. Customers who don’t have a phone service are very less in the sample and are almost 1/10th of those who have a phone service. The variable ‘MultipleLines’ has 3 variables; customers who has/doesn’t have multiple lines, and those who don’t have a phone service. There are customers who don’t subscribe for an internet service and those people are grouped separately as ‘No internet service’ in attributes ‘OnlineSecurity’, ‘OnlineBackup’, ‘DeviceProtection’ ‘TechSupport’, ‘StreamingTV’, and ‘StreamingMovies’. There are three types of contracts 1 year, 2 year and month to month contracts. There are more number of month to month contracts than both 1 year and 2 year contracts combined. Only a few people have signed up for paperless billing. The monthly charge for the customers varies from $18.25 to $118.75. The minimum value any customer has paid in total to the company is $18.8 and the maximum amount any customer has paid in total to the company is $8684.8. There are 11 observations with ‘NA’ values in ‘TotalCharges’. The dataset is highly imbalanced in case of the target variable. There are nearly three times data representing the active customers than the churned customers. This may lead to a biased model creation in the next phase of the project. We may have to deal with this data imbalance according to the type of model we are creating in the next phase.

3 Data Preprocessing

The structure of the data is represented using the below line of code.

```
str(telcom_churn) # Structure of the data
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 655
##  $ 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 ...
```

'SeniorCitizen' has to be converted to a factor. It is labeled as 'Yes', and 'No'.

```
telcom_churn$SeniorCitizen= factor(telcom_churn$SeniorCitizen, c(0,1), labels = c('No','Yes'),
                                   ordered = is.ordered(telcom_churn))
```

There are null values in the dataset. It can be checked using the following code chunk.

```
table(is.na(telcom_churn)) # Check for Null Values.
```

```
## 
##  FALSE   TRUE 
## 147892     11 
```

```
colSums(is.na(telcom_churn)) #totalcharges column have 11 missing values
```

```
##       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 
```

‘TotalCharges’ has 11 ‘NA’ values. For getting a closer picture of this null values, we can create a table of those customers with missing values as below:

```
kable(
  select(telcom_churn,gender, SeniorCitizen, tenure, Contract,
         PaymentMethod, MonthlyCharges, Churn ) %>%
        filter(is.na(telcom_churn$TotalCharges)), caption = "Observations with missing values" ) %>%

  kable_styling("striped", full_width = T,
                bootstrap_options = c("striped", "hover","condensed"),
                latex_options = "HOLD_position",position = "center",
                font_size = 9) %>%

  row_spec(0, angle = 0,font_size = 7,bold = T, background = "#FAE5D3") %>%
  column_spec(1:7, border_left = T)
```

Table 2: Observations with missing values

gender	SeniorCitizen	tenure	Contract	PaymentMetho	MonthlyCharge	Churn
Female	No	0	Two year	Bank transfer (automatic)	52.55	No
Male	No	0	Two year	Mailed check	20.25	No
Female	No	0	Two year	Mailed check	80.85	No
Male	No	0	Two year	Mailed check	25.75	No
Female	No	0	Two year	Credit card (automatic)	56.05	No
Male	No	0	Two year	Mailed check	19.85	No
Male	No	0	Two year	Mailed check	25.35	No
Female	No	0	Two year	Mailed check	20.00	No
Male	No	0	One year	Mailed check	19.70	No
Female	No	0	Two year	Mailed check	73.35	No
Male	No	0	Two year	Bank transfer (automatic)	61.90	No

The table shows that the customers with missing values are all active customers. Their tenure is shown as zero which means that they all have joined within one month of the data collection. All of them are new customers. Customers who have joined recently will give less insight into the churn tendency of a customer. We cannot consider the attributes of a recent customer to train the model. For the stability of the model and better accuracy, we are considering only those customers who are customers of telco for at least 1 year, i.e 12 months.

```
len1 <- nrow(telcom_churn)
cat("Number of observations before:", len1)
```

```
## Number of observations before: 7043
```

```
telcom_churn <- subset(telcom_churn,telcom_churn$tenure>12) # creating new subset

len2 <- nrow(telcom_churn)
cat("\nNumber of observations after:", len2)
```

```
##
## Number of observations after: 4857
```

There are 2186 Observations being removed according to this criteria. Now we have 4857 churned/active customers who are with telco for more than one year. Checking if this removed our null values as well.

```
table(is.na(telcom_churn)) # Check for Null Values.
```

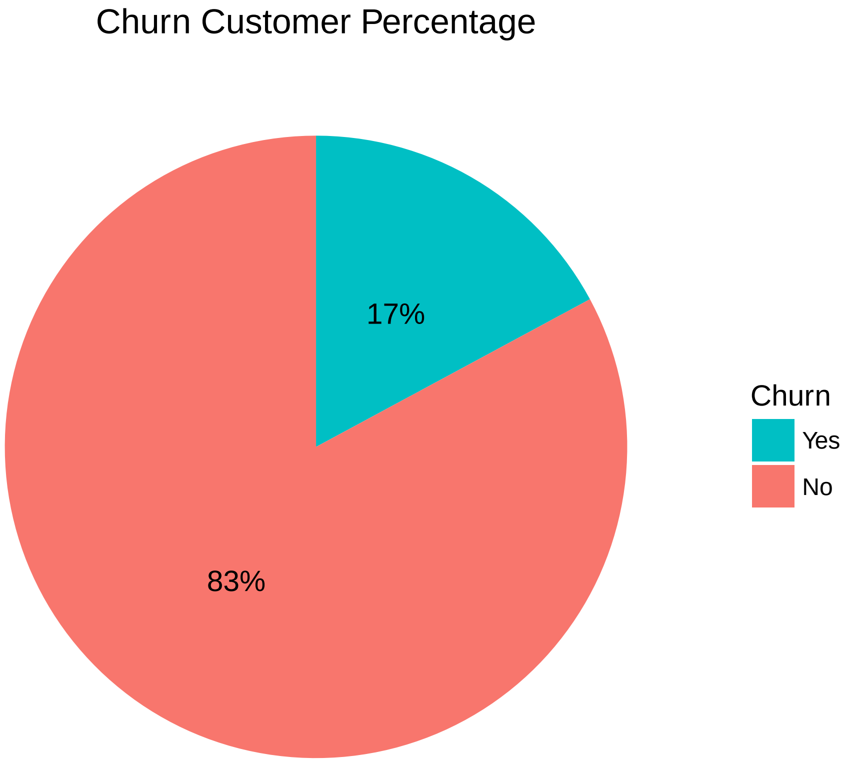
```
##
##  FALSE
## 101997
```

Since the minimum tenure is 12 months and also to reduce the Curse of Dimensionality and the time required to train the algorithm, we choose to factorise the variable 'tenure' in to 5 different groups; 13-24 months, 25-36 months, 37-48 months, 49-60 months, and 61-72 months.

```
telcom_churn$tenure <- cut(telcom_churn$tenure,breaks=5,dig.lab=2,labels=2:6)
```

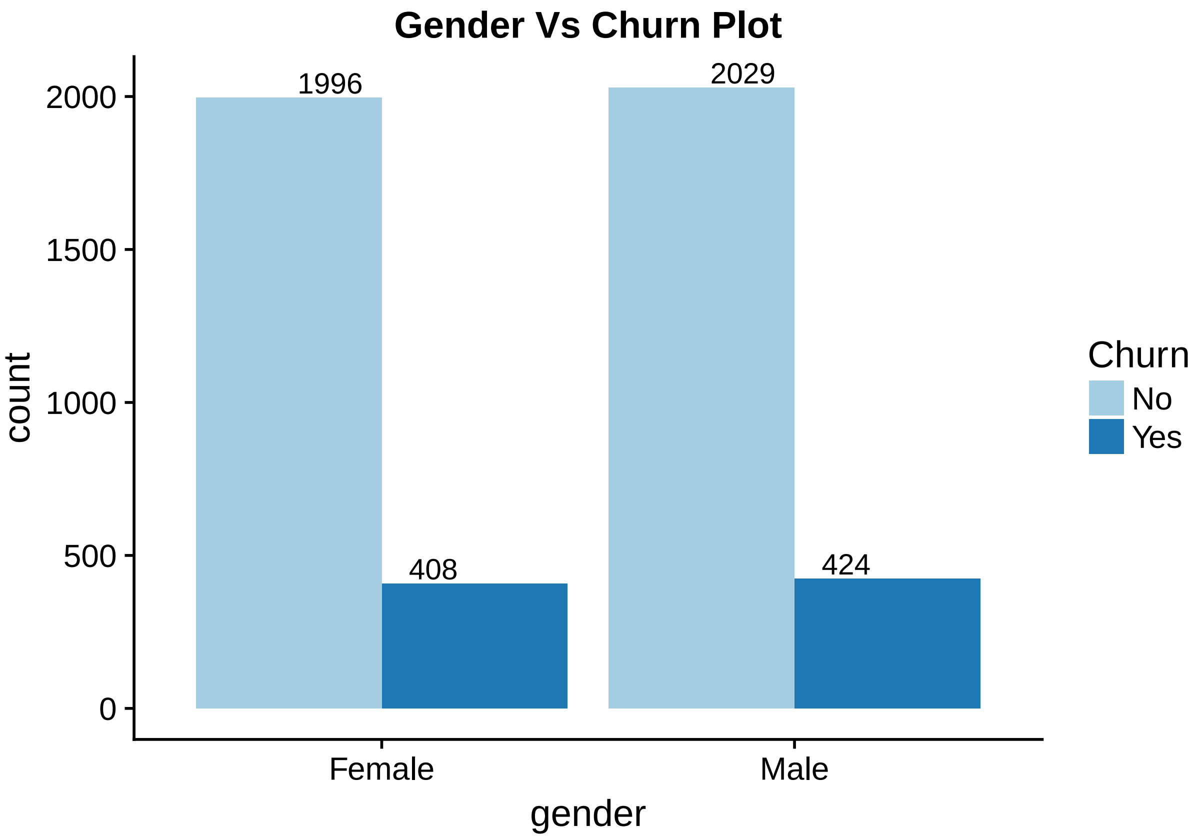
4 Data Exploration

```
data_summary <- telcom_churn %>% group_by(Churn) %>%
summarise(Percent = n() / nrow(.) * 100) # Creating percentage values

ggplot(data = data_summary,
mapping = aes(x = "", y = Percent, fill = Churn)) +
  geom_bar(width = 1, stat = "identity") +
  scale_y_continuous(breaks = round(cumsum(rev(data_summary$Percent)), 1)) +
  coord_polar("y", start = 0) +
  geom_text(aes(label = paste0(round(Percent), "%")),
            position = position_stack(vjust = 0.5)) +
  labs(x = NULL, y = NULL, fill = "Churn",
       title = "Churn Customer Percentage") +
  guides(fill = guide_legend(reverse = TRUE)) +
  theme_classic() +
  theme(axis.line = element_blank(),
        axis.text = element_blank(),
        axis.ticks = element_blank(),
        plot.title = element_text(hjust =0.5))
```



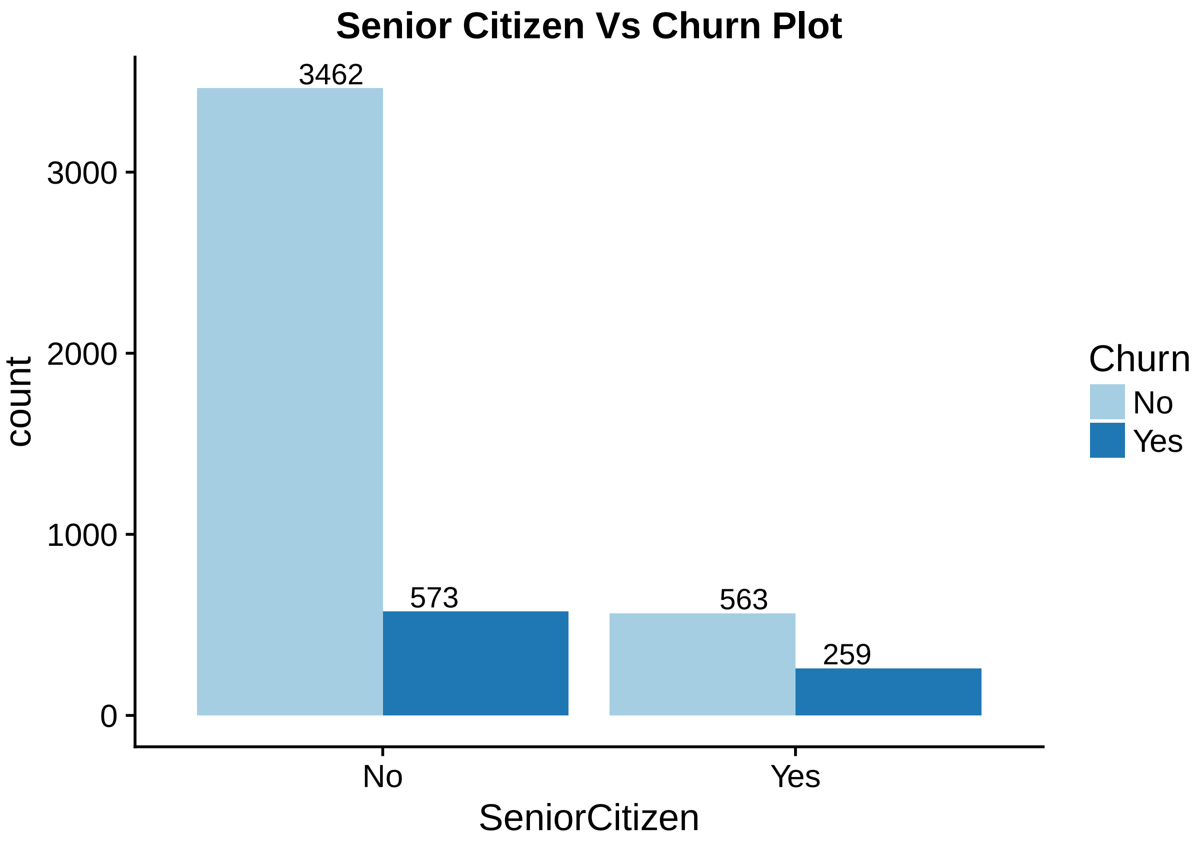
Above graph clearly shows that out of 4857 customers 17% customers churned.

```
# Analysing Gender variable
ggplot(telcom_churn, aes(gender, ..count..,fill = Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Gender Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



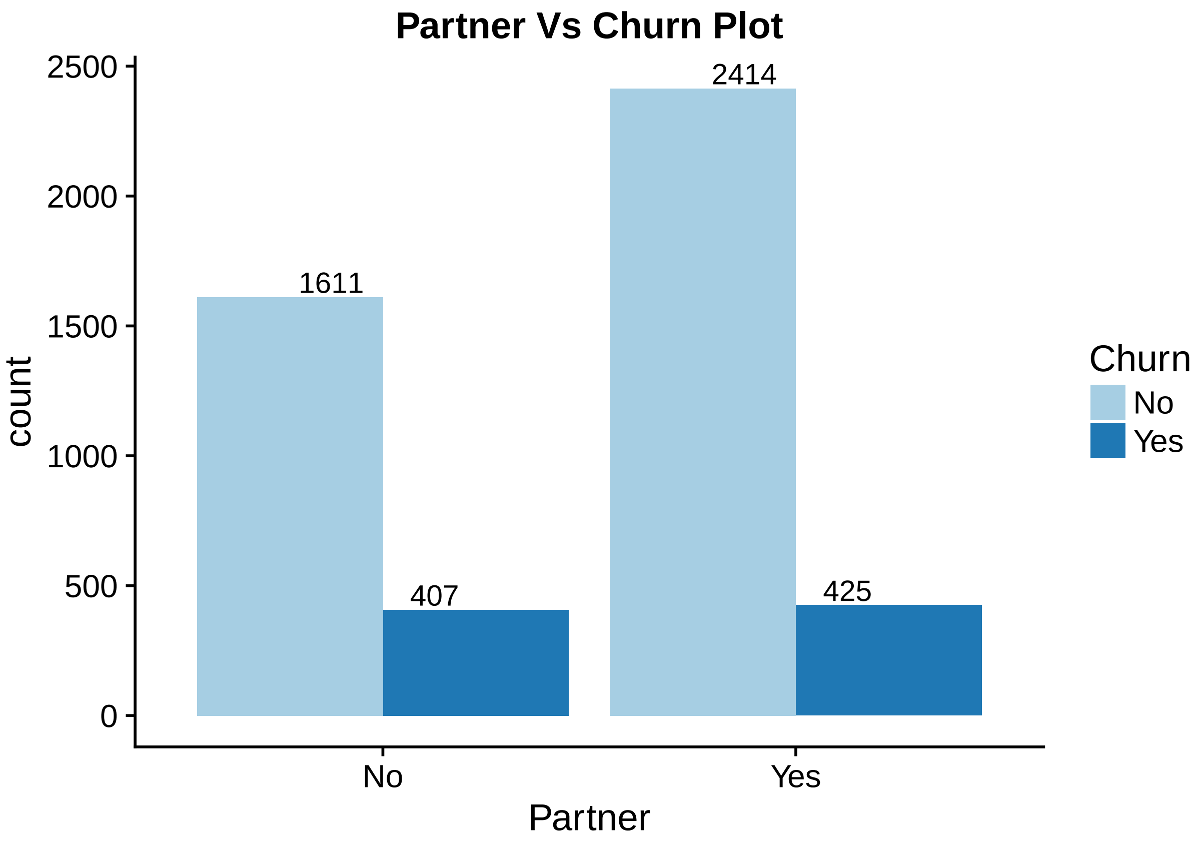
This visualization shows that male and female both have similar chances to churn. Which suggest that gender variable have less impact on predicting which customers are likely to leave Telco telecom service.

```
# Analyse senior citizen status variable
ggplot(telcom_churn, aes(SeniorCitizen,fill = Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Senior Citizen Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



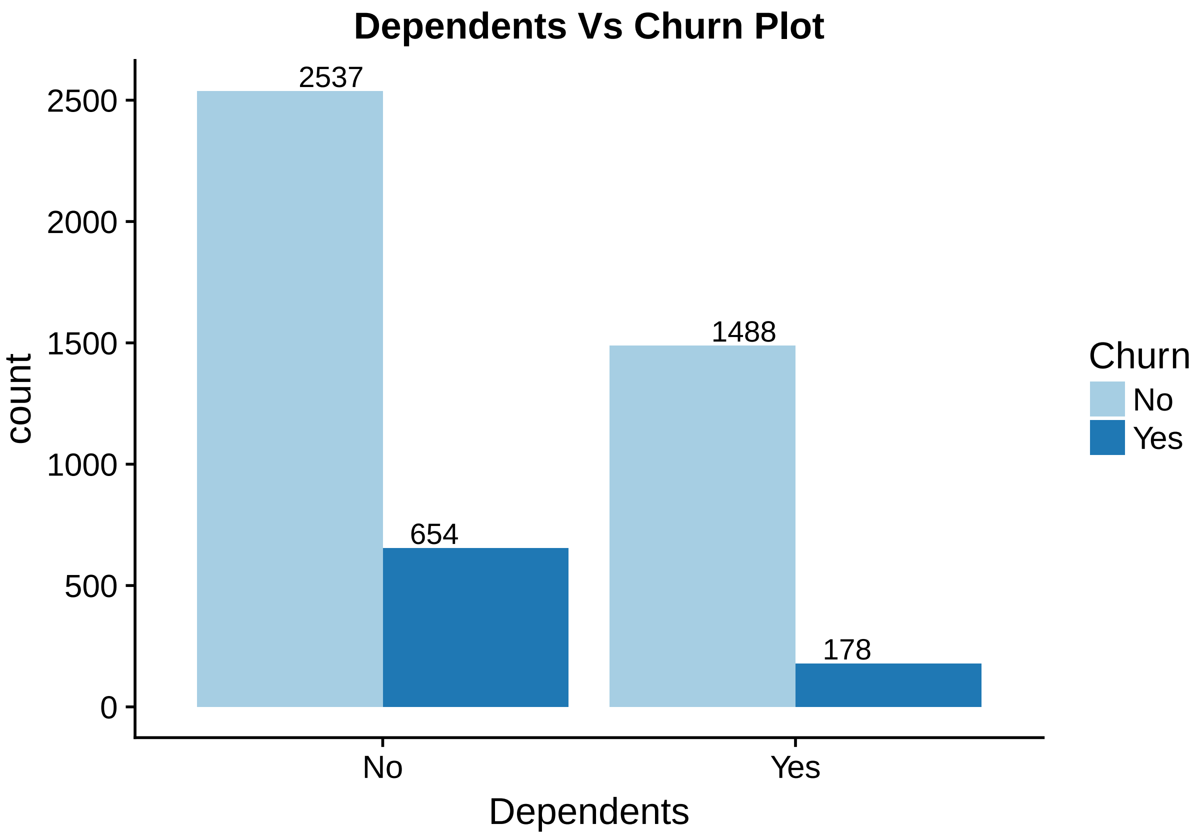
Above visualization shows that the number of senior citizens is less so a number of senior citizens churned are also less.

```
# Analyse Partner variable

ggplot(telcom_churn, aes(Partner,fill = Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Partner Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



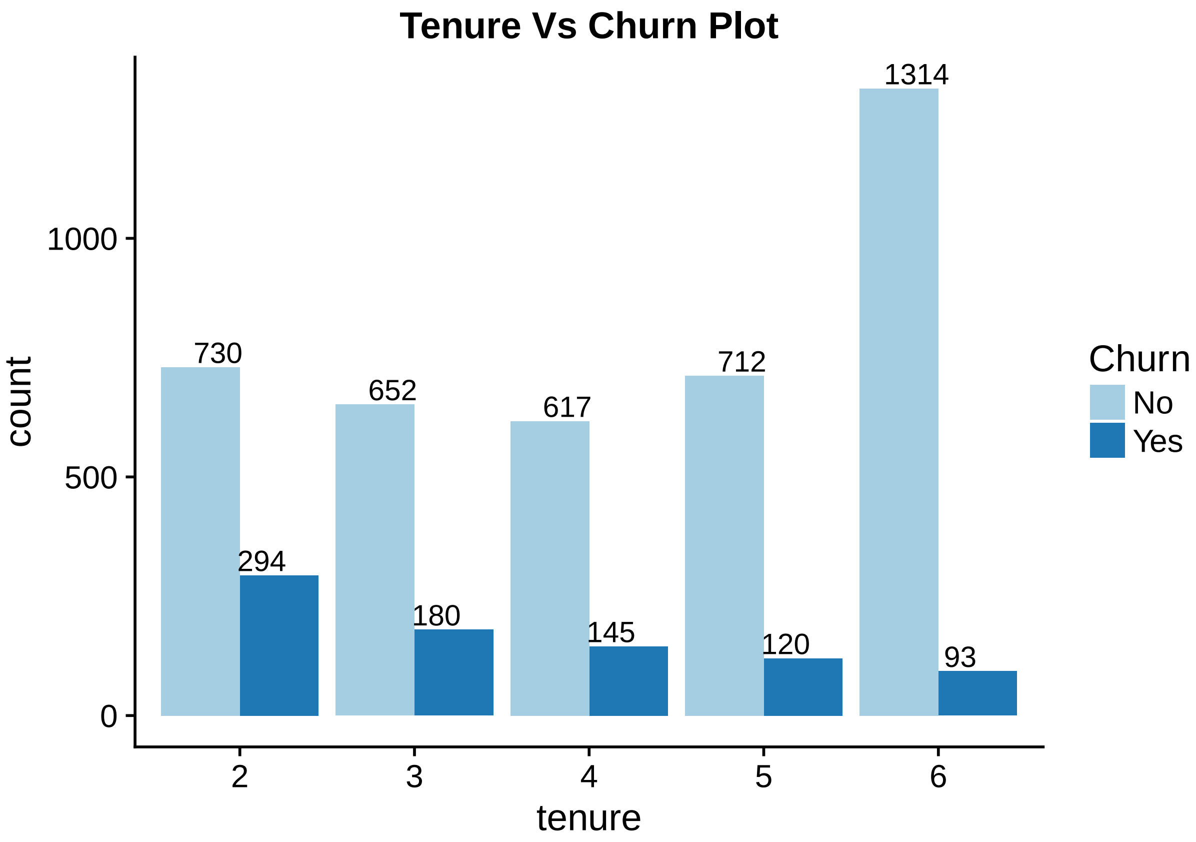
This visualization shows that customer having partner have the lesser tendency of leaving Telco Telecom Services.

```
# Analyse Dependents variable

ggplot(telcom_churn, aes(Dependents,fill = Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Dependents Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



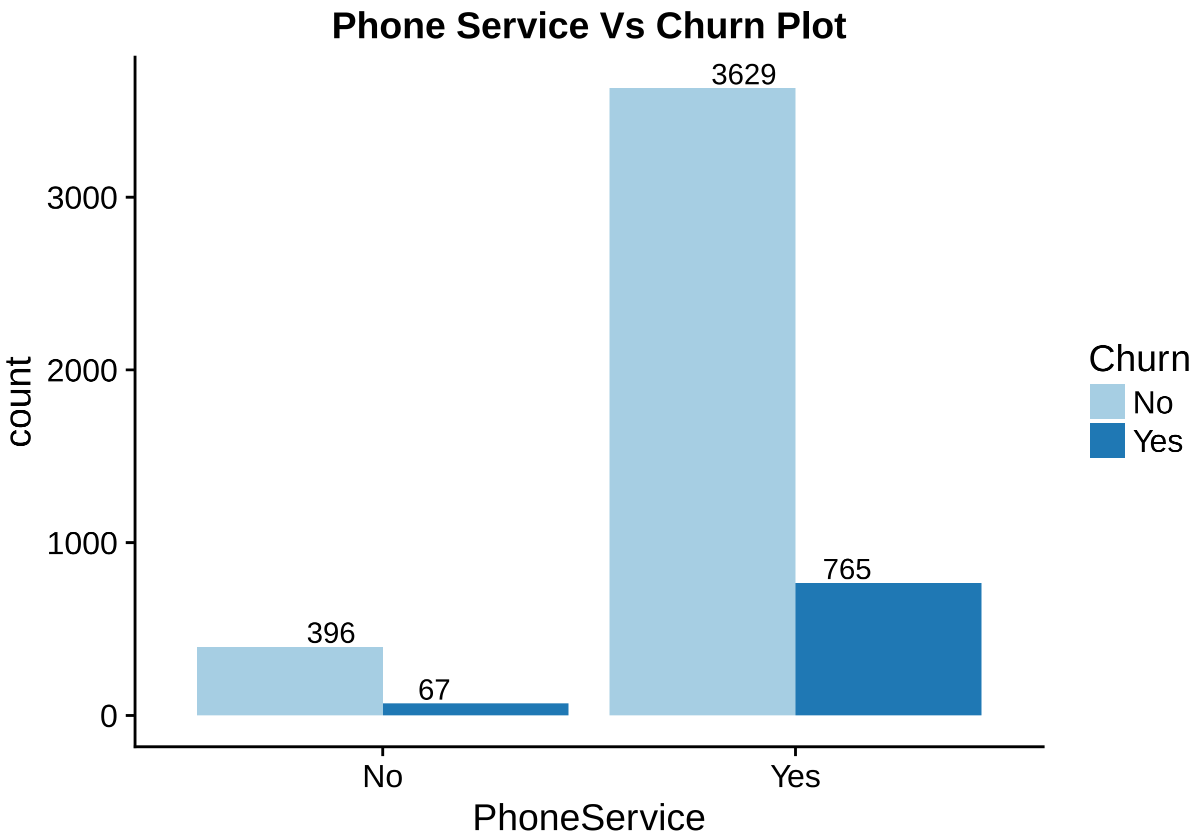
Above graph shows indicates that the number of customers having dependents is less hence the proportion of churning is as well.

```
# Analyse tenure variable

ggplot(telcom_churn, aes(tenure,fill = Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Tenure Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



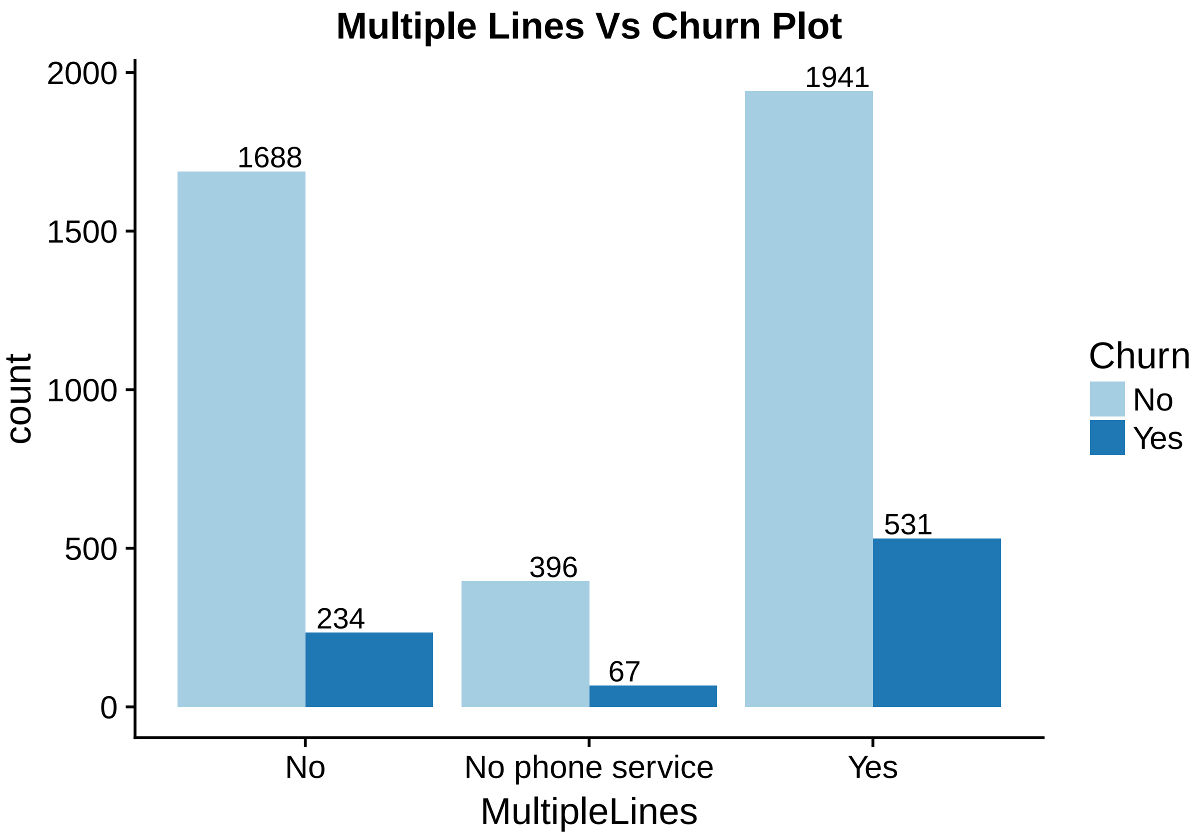
Above plot represents that tenure in the year and churning customers are inversely proportional. Which implies that older customers have a lesser tendency of churning.

```
# Analyse PhoneService variable

ggplot(telcom_churn, aes(PhoneService,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Phone Service Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



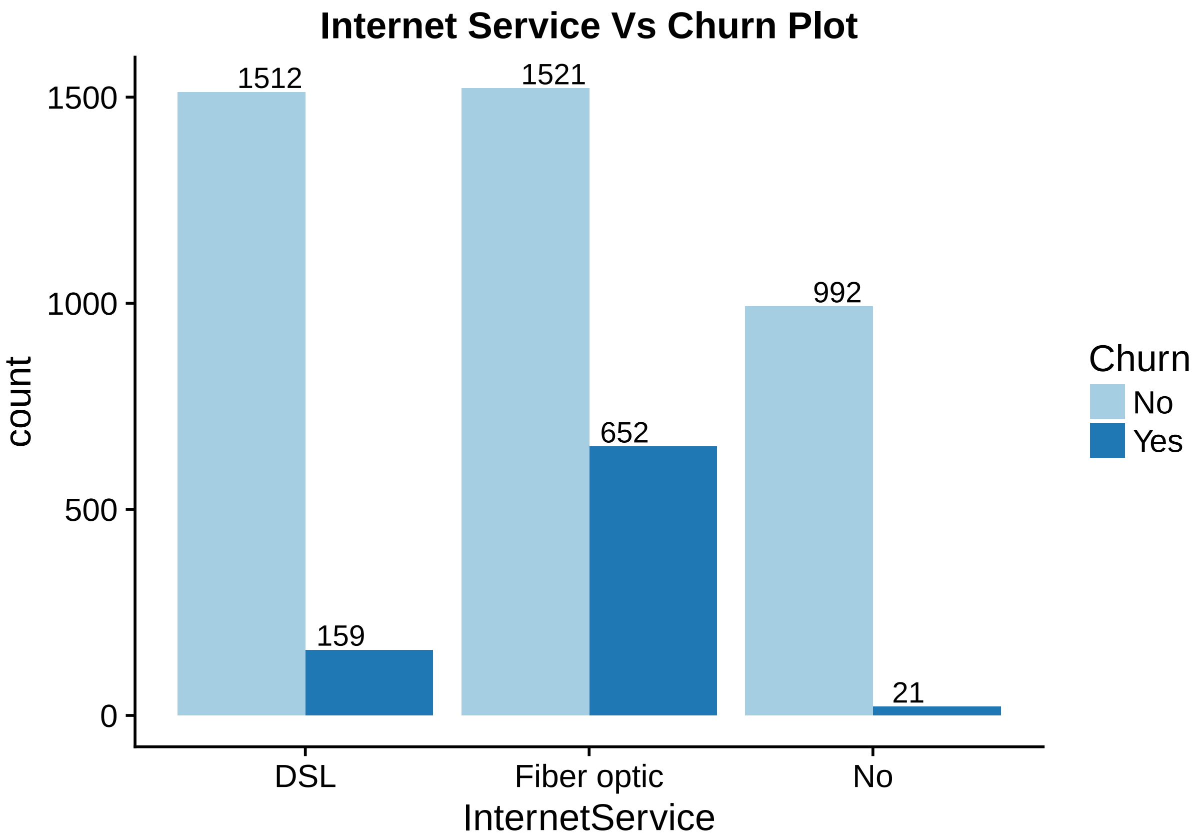
This plot shows that in the examined data set very few customers don't use phone services.

```
# Analyse MultipleLines variable

ggplot(telcom_churn, aes(MultipleLines,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Multiple Lines Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



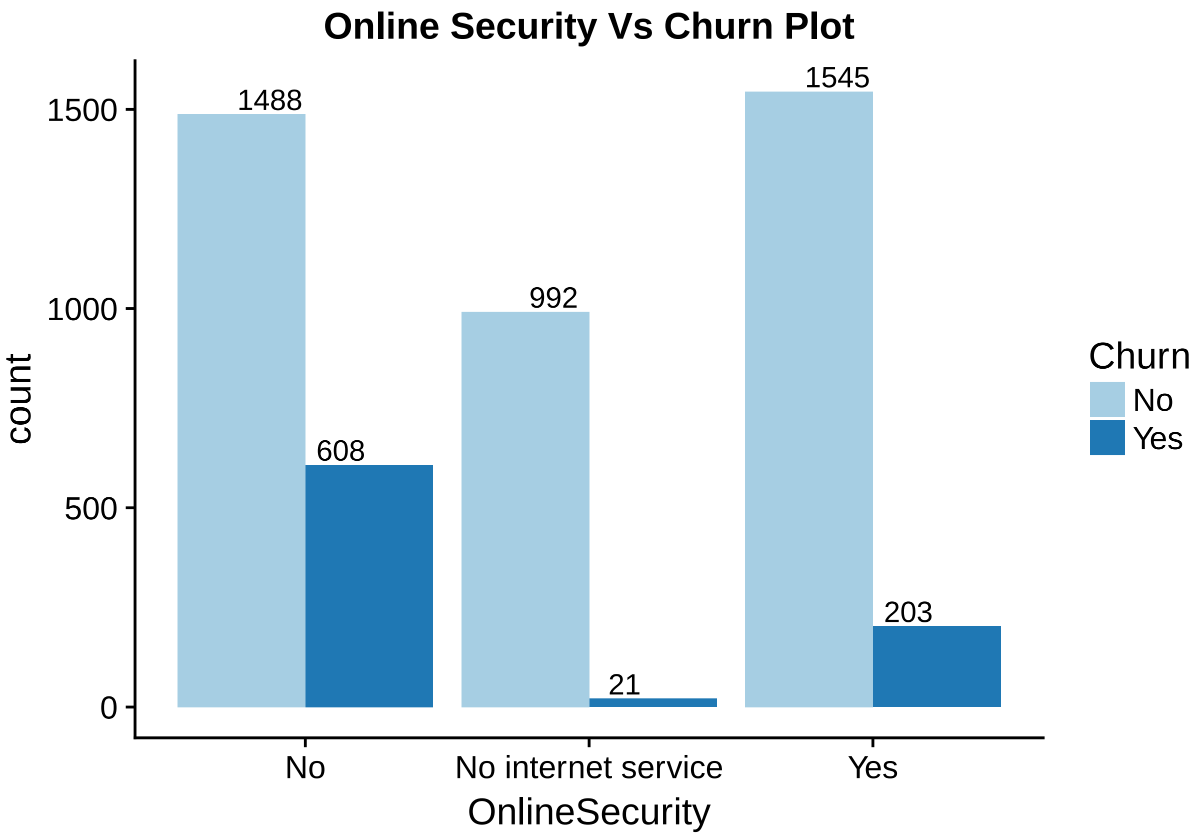
Above plot clearly, suggest that the number of customers with multiple lines are higher and churning tendency is higher as well.

```
# Analyse InternetService variable

ggplot(telcom_churn, aes(InternetService,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Internet Service Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



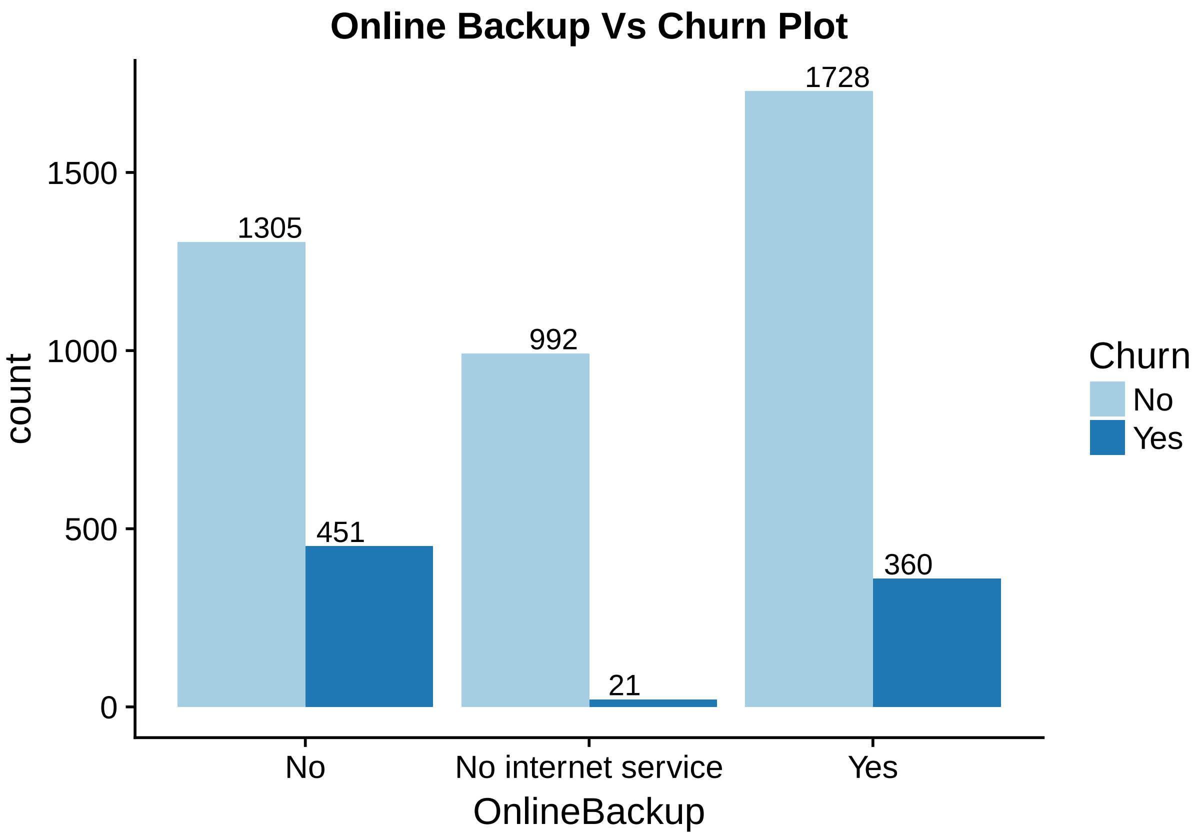
Above graph shows that number of customers using DSL internet service and Fiber Optic internet service are almost same. But customers using Fiber Optic internet service are more likely to leave Telco Telcom Service.

```
# Analyse OnlineSecurity variable

ggplot(telcom_churn, aes(OnlineSecurity,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Online Security Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



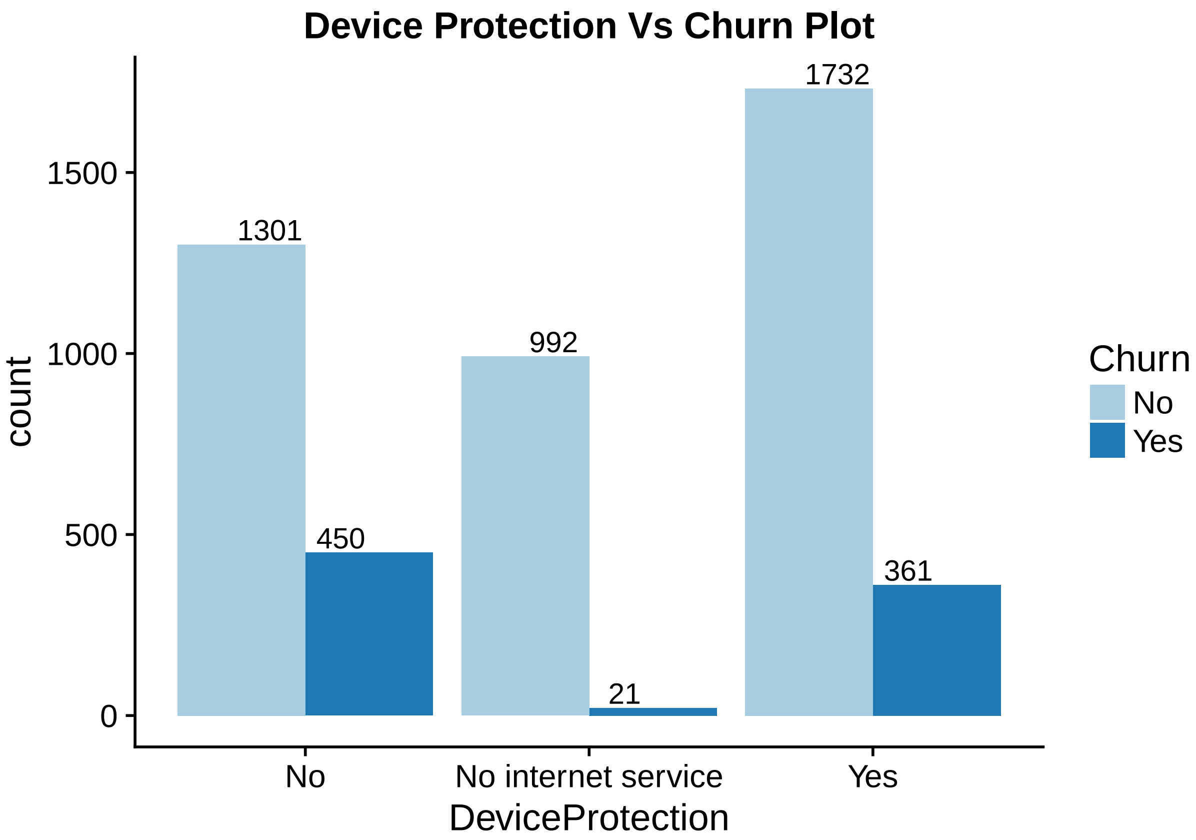
This Visualization shows that customers with no online security services are more likely to leave.

```
# Analyse OnlineBackup variable

ggplot(telcom_churn, aes(OnlineBackup,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Online Backup Vs Churn Plot")+
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



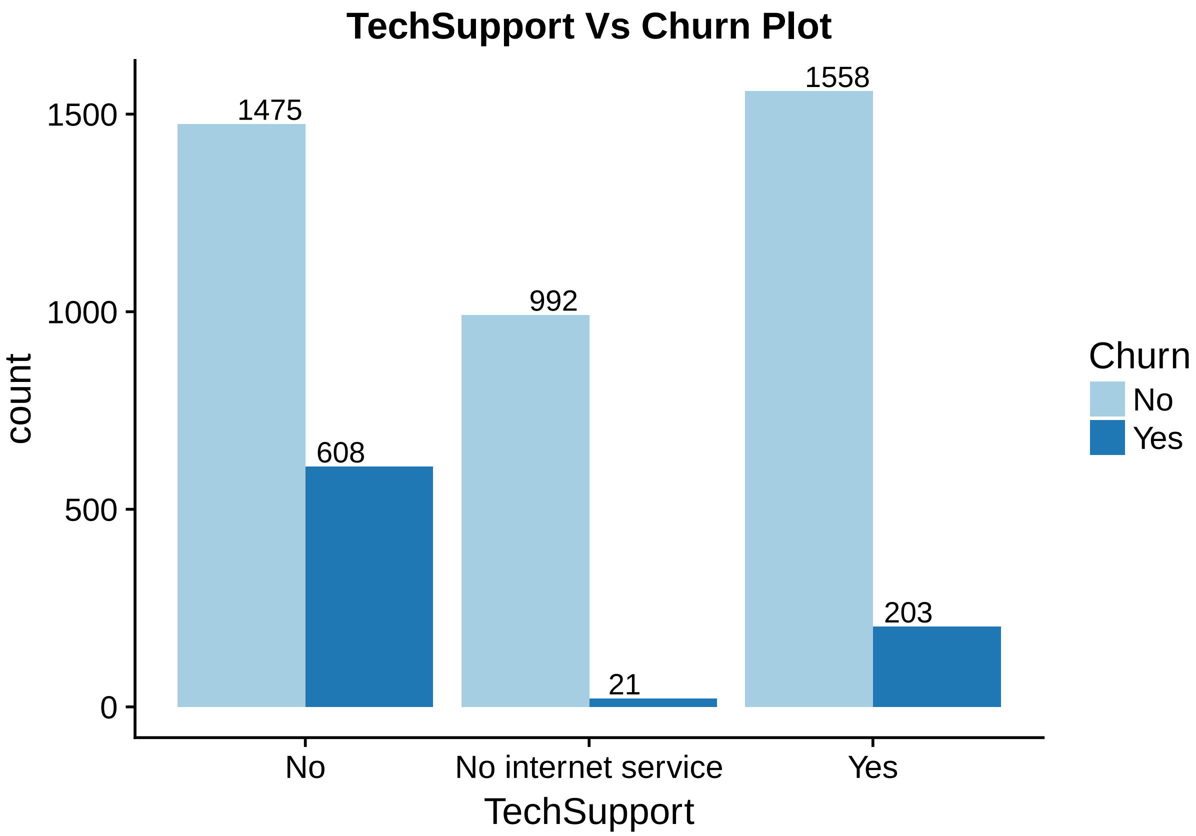
Above graph shows that the number of customers having an online backup is higher and customers without online backup are more likely to churn.

```
# Analyse DeviceProtection variable

ggplot(telcom_churn, aes(DeviceProtection,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Device Protection Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



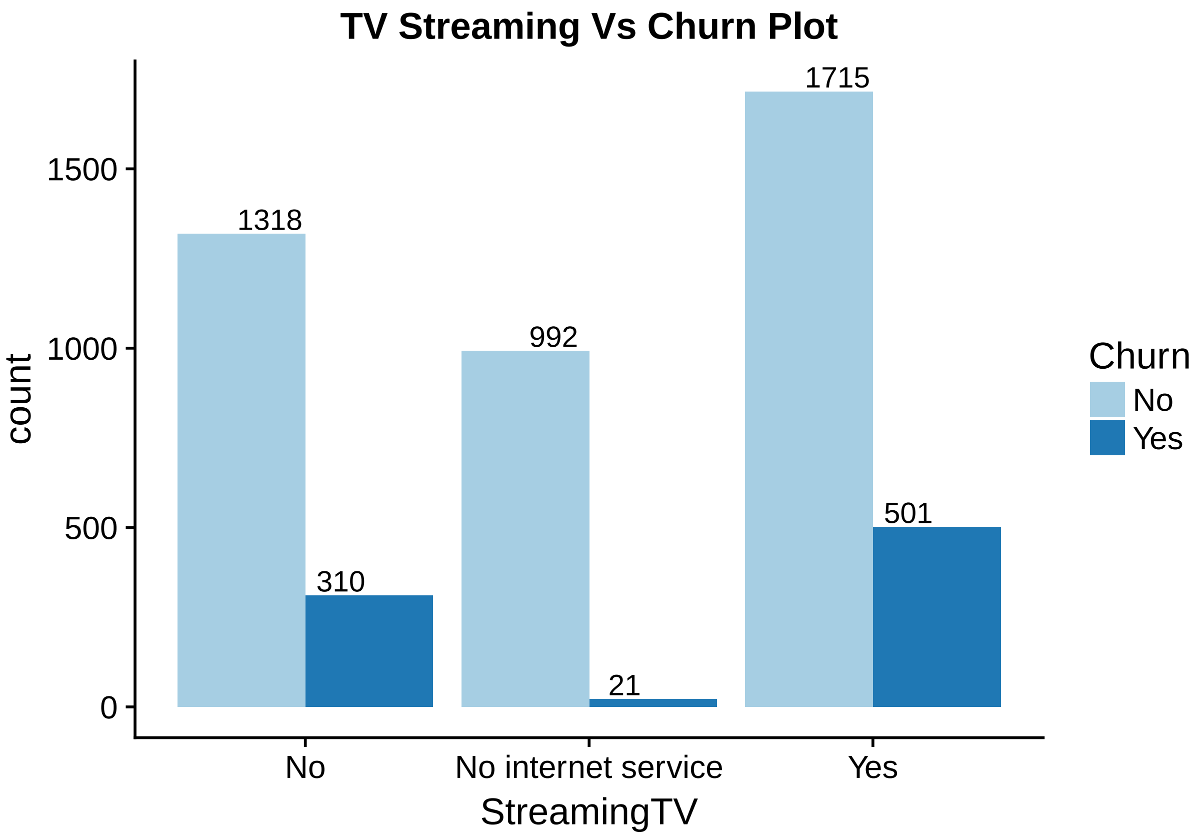
Above visualization represents that the number of customers having Device Protection is higher and customers without Device Protection are more likely to churn.

```r
# Analyse TechSupport variable

ggplot(telcom_churn, aes(TechSupport, fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("TechSupport Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



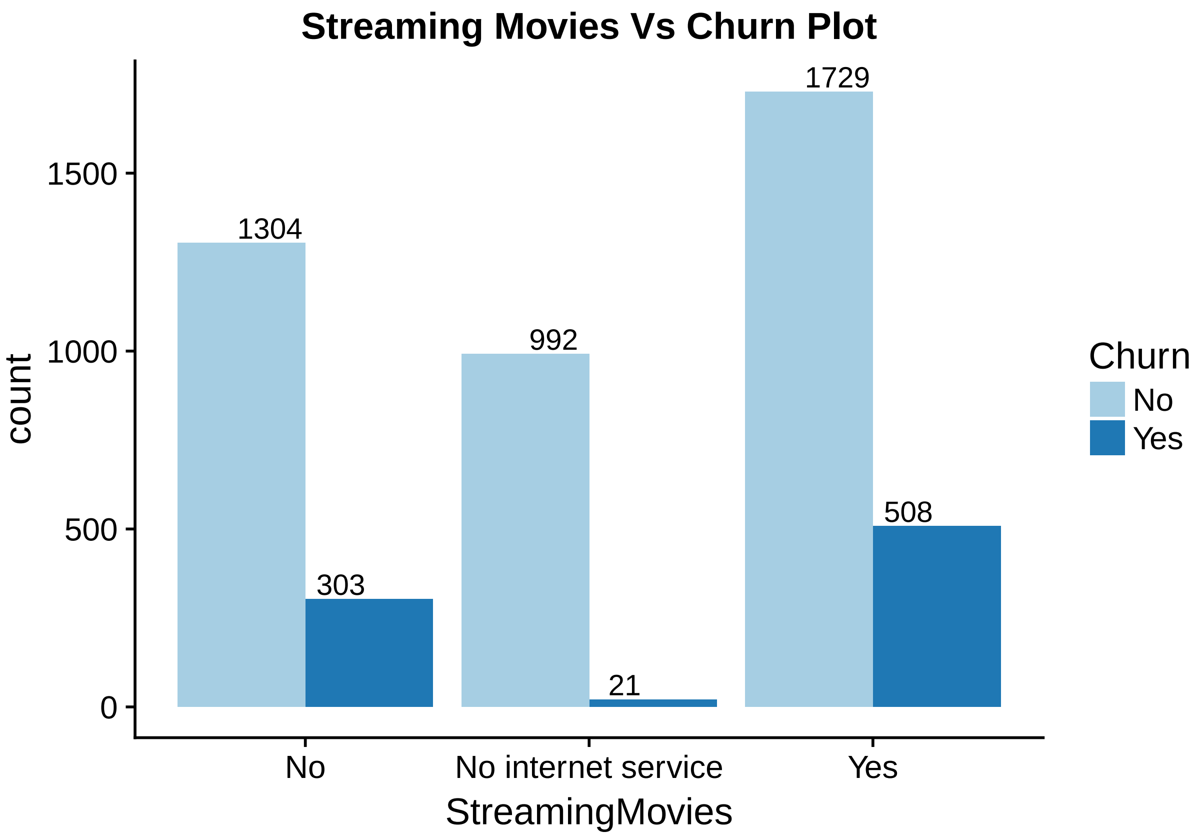
This graph clearly shows that the number of customers with tech support is higher and customers without online Tech support are more likely to churn.

```
# Analyse StreamingTV variable

ggplot(telcom_churn, aes(StreamingTV,fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("TV Streaming Vs Churn Plot")+
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



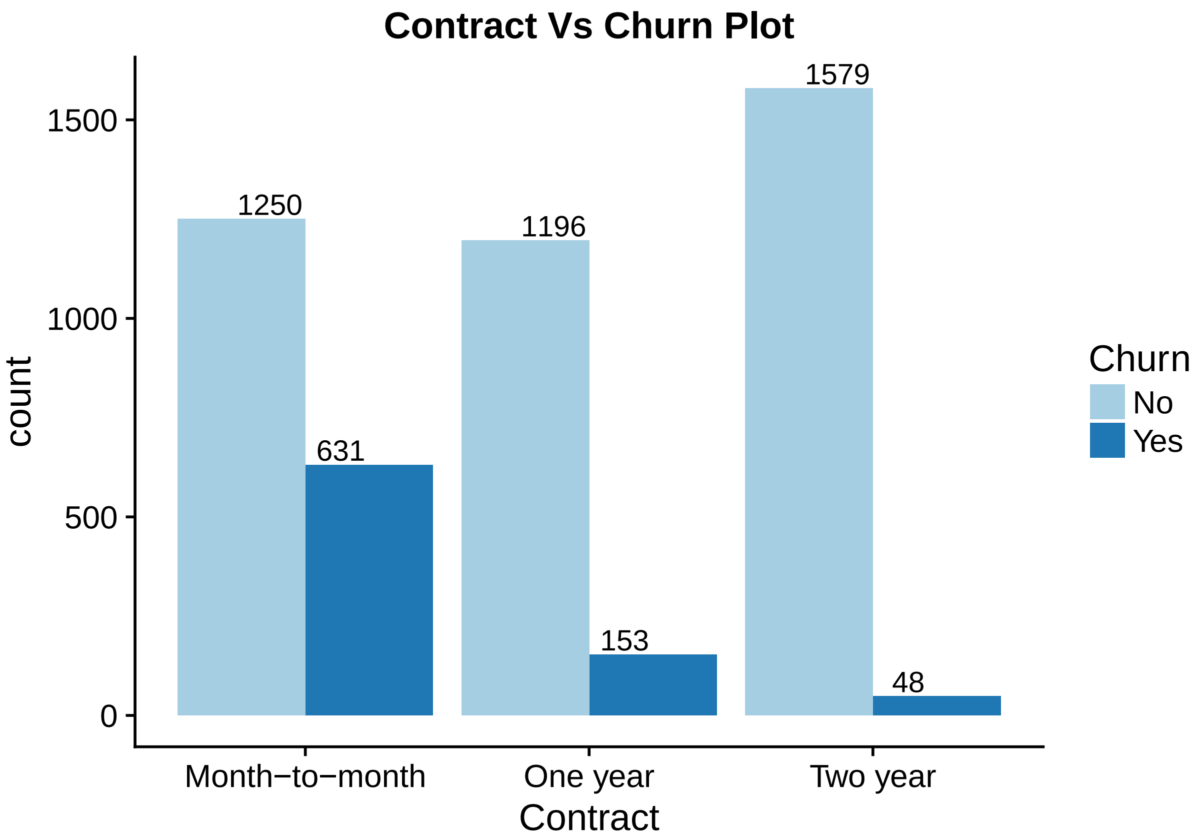
This graph clearly shows that the number of customers with TV Streaming services are higher so the number of customers to churn is also higher.

```
# Analyse Streaming Movies variable

ggplot(telcom_churn, aes(StreamingMovies, fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Streaming Movies Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



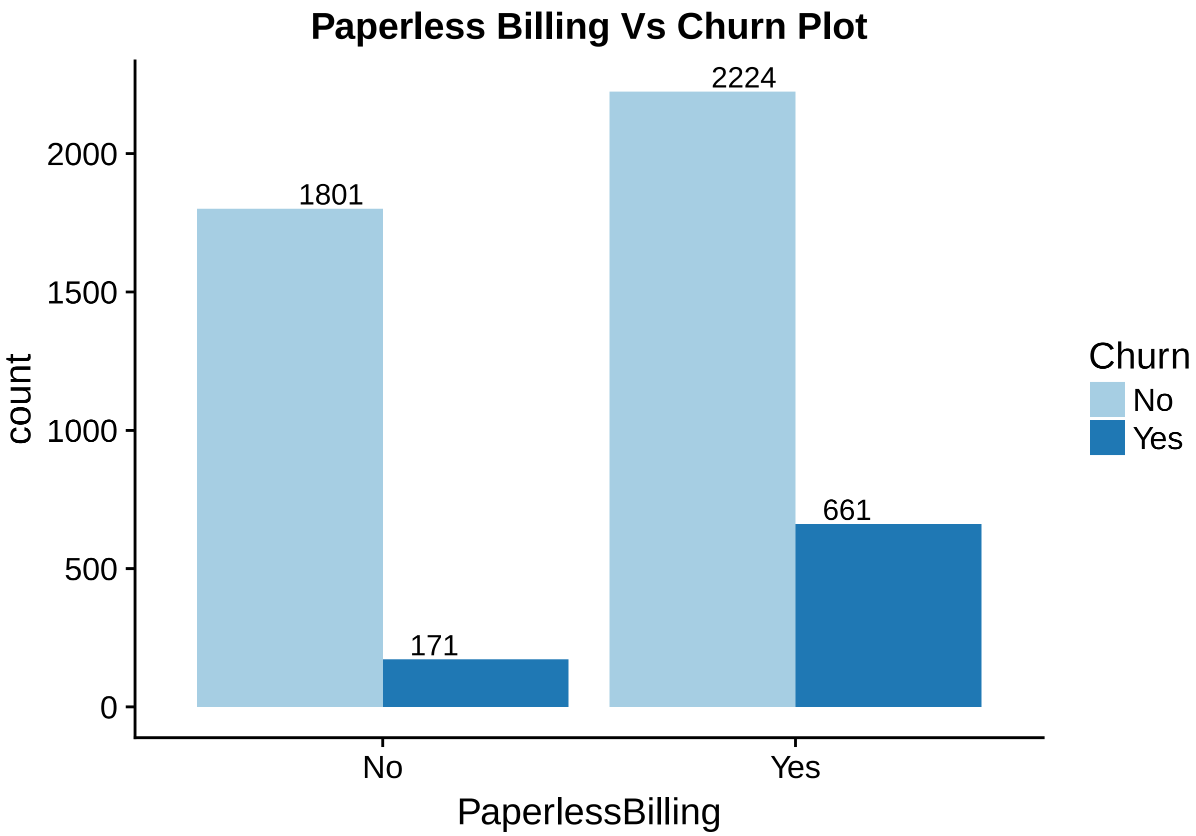
This graph represents that the number of customers with Movie Streaming services are higher so the number of customers to churn is also higher.

```
# Analyse Contract variable

ggplot(telcom_churn, aes(Contract, fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Contract Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5))+
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



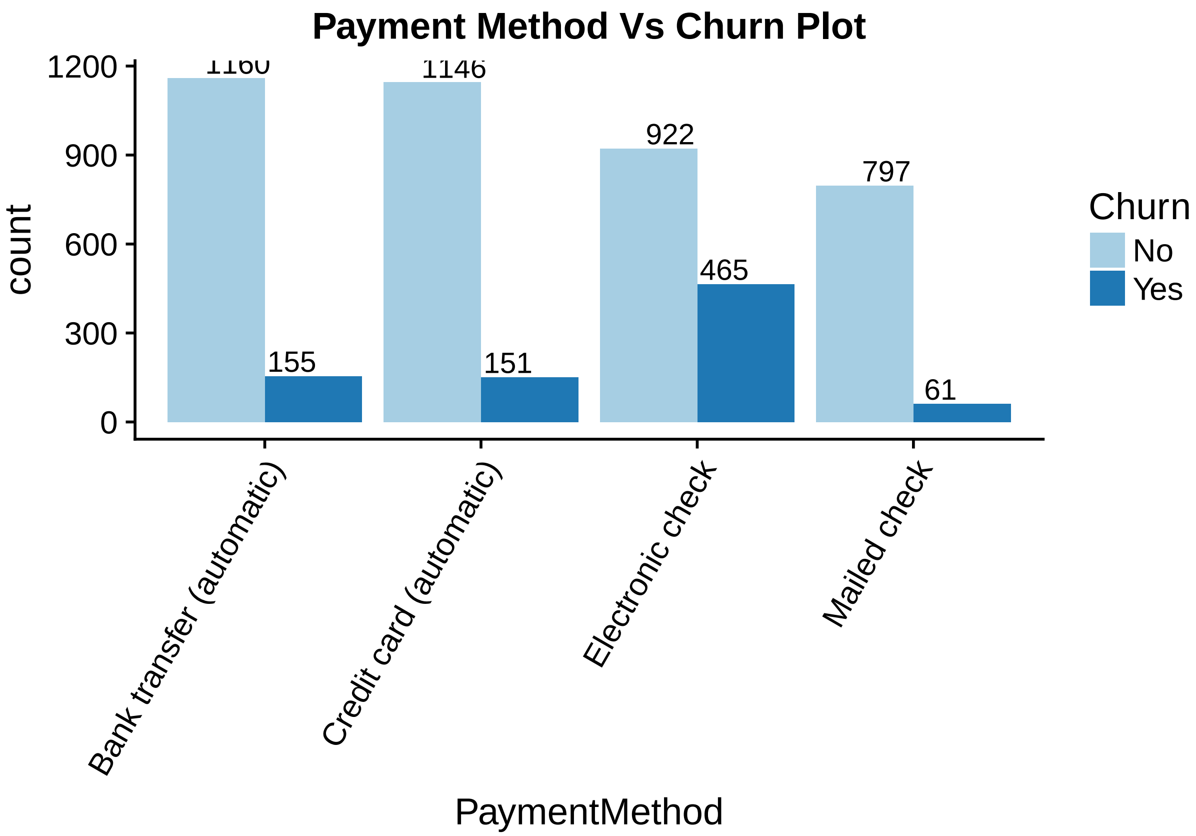
Above graph shows that the number of customers with two years of the contract are higher whereas customer with the month to month contract has a higher chance of leaving.

```r
# Analyse PaperlessBilling variable

ggplot(telcom_churn, aes(PaperlessBilling, fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Paperless Billing Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2)
```



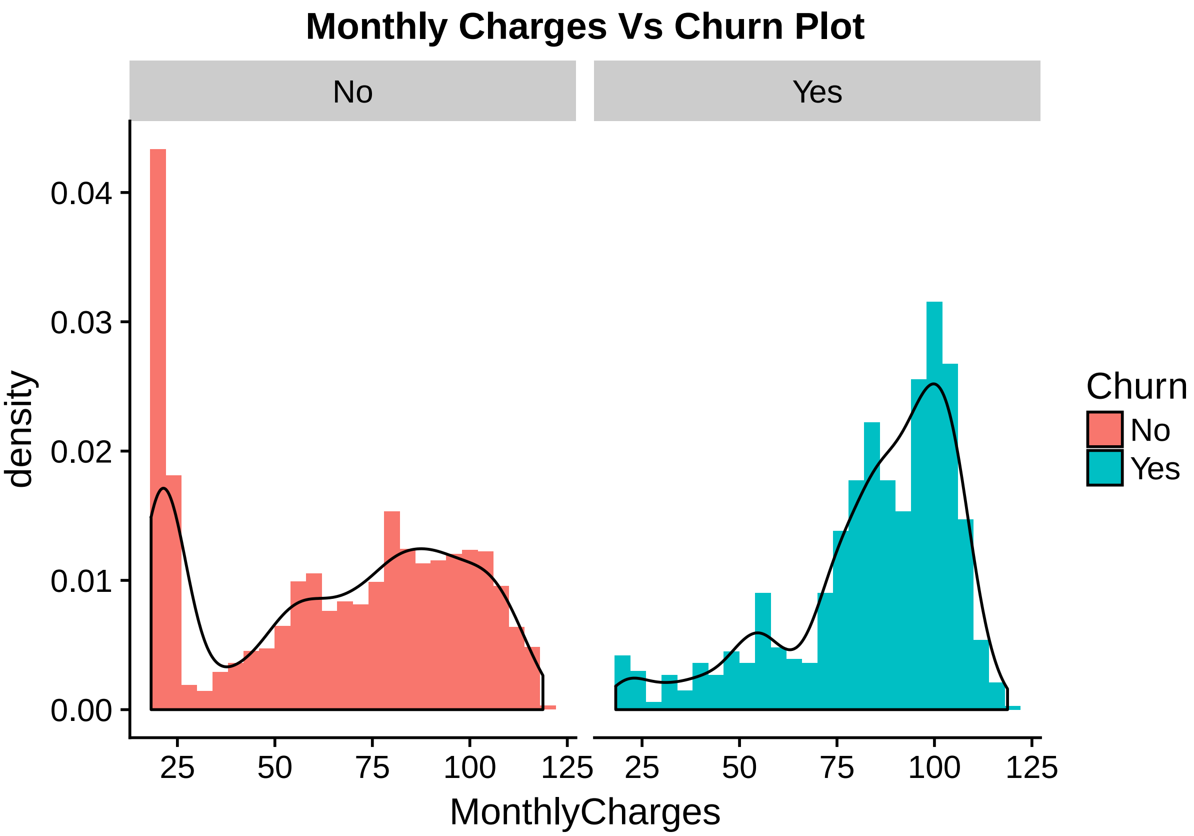
This graph shows that the number of customers with Paperless Billing are higher so the proportion of churning is higher as well.

```
# Analyse PaymentMethod variable

ggplot(telcom_churn, aes(PaymentMethod, fill=Churn)) +
  geom_bar(stat="count", position = "dodge") +
  scale_fill_brewer(palette="Paired") +
  ggtitle("Payment Method Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5)) +
  geom_text(aes(label=..count..),
            stat="count",position=position_dodge(0.5),vjust=-0.2) +
  theme(axis.text.x = element_text(angle = 60, hjust = 1))
```



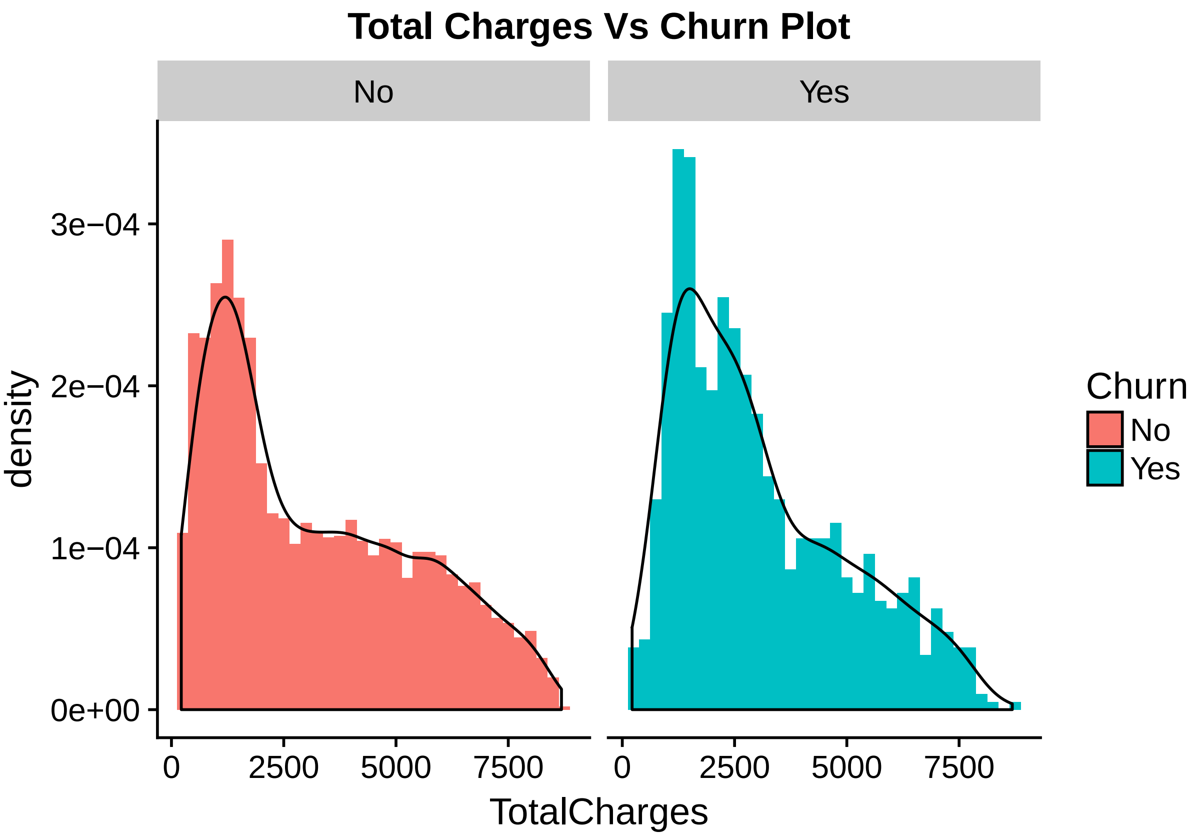
This visualization shows that majority of customers prefers Bank Transfer or Credit Card payment method and customers using electronic check payment method are more likely to churn.

```
# Analyse MonthlyCharges variable
ggplot(telcom_churn, aes(x=MonthlyCharges)) +
  geom_histogram(aes(y=..density..,fill=Churn),binwidth = 4) +
  geom_density(alpha=.05)+
  facet_grid(. ~ Churn)+ggtitle("Monthly Charges Vs Churn Plot") +
  theme(plot.title = element_text(hjust = 0.5))
```



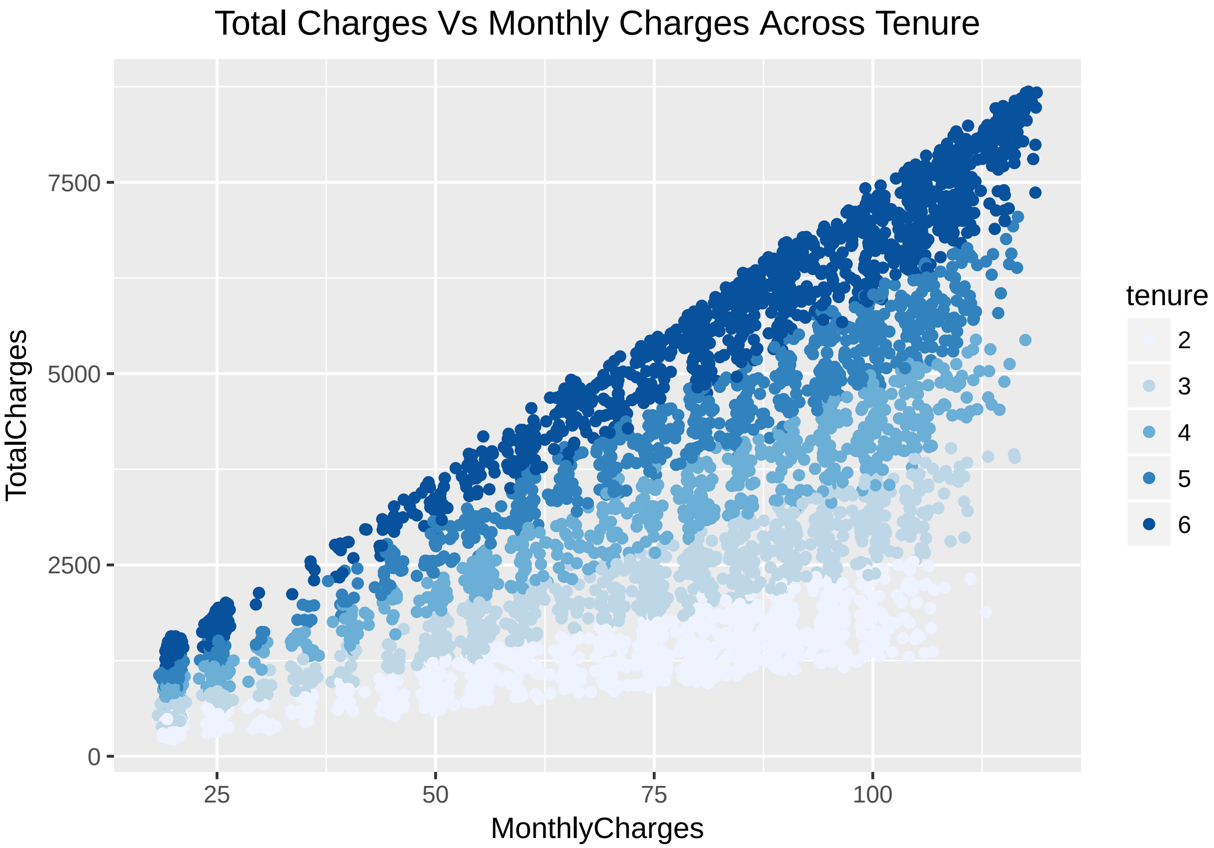
This plot shows that monthly charges are not monthly distributed for churned or not churned customers. Customers paying higher monthly charges are more so number of customers churning are more as well.

```
# Analyse TotalCharges variable
ggplot(telcom_churn, aes(x=TotalCharges)) +
  geom_histogram(aes(y=..density..,fill=Churn),binwidth = 250) +
  geom_density(alpha=.05)+
  facet_grid(. ~ Churn)+
  ggtitle("Total Charges Vs Churn Plot")+
  theme(plot.title = element_text(hjust = 0.5))
```



This graph clearly shows that total charges are right-skewed for churned or not churned customers.

```r
ggplot(telcom_churn, aes(MonthlyCharges, TotalCharges, color = tenure)) +
  geom_point() +
  scale_colour_brewer()+theme_gray()+
  ggtitle("Total Charges Vs Monthly Charges Across Tenure")+
  theme(plot.title = element_text(hjust = 0.5))
```



Above plot indicates that there is a strong positive correlation between monthly charges and total charges across all the Tenure levels.

5 Results and Discussion

The data is highly imbalanced. In the dataset, there are only 17% of the customers who are churned. It is seen that gender doesn't have any effect on the deciding whether a customer is churned or not. As the tenure of the customer increases, the probability of the customer getting churned decreases. Out of all those people who don't use internet service at all, only 2% of the people got churned. This variable will have a strong relation effect in the model. Even though there is the highest number of people who take 2-year contracts and only 3% of them got churned. This is also an important variable in deciding if a customer gets churned or not. The proportion of churned customer is more in case of a month to month contract customers. Almost half of them got churned in the given dataset. People who are using the electronic check as the payment option has more chance of getting churned than people using any other payment options. We have seen from the multivariable scatter plot between total charges and monthly charges across the tenures that monthly charges are linearly and positively proportional to total charges across any particular tenure. So total charges may be a redundant variable and can create issues in the model creation unless removed. It is also interesting to note that as the tenure increases, the total charges and the monthly charges increases. This can mean that people may like to go for a higher payment plan over the years.

6 Conclusion

The data had several issues including missing values. Some of the factor variables were denoted as numerical and has to be converted into factors. Attributes were examined using graphs and numerical values. Several relations among variables were identified and some of the important variables were found. The data has been cleaned and processed for model creation.