project1

zbuckley, pedrouria, seanpili

October 13, 2018

Contents

Load Telco Data **1**

Research Question **2**
Null Hypotheses . . . 2
Alternative Hypotheses . . . 3

Conducting Exploratory Data Analysis **3**
Corplot and Descriptives . . . 3
Numeric Variables . . . 3
Categorical Variables . . . 9
Literature-Focused EDA . . . 10
Churn . . . 10
Contract Type . . . 11
Tenure . . . 17
Monthly Charges . . . 24
Individual Variables . . . 29
Online Security . . . 29
Tech Support . . . 32
Online Backup . . . 35
Paperless Billing . . . 38
Device Protection . . . 41
Dependents . . . 44
Partner . . . 47
Senior Citizen . . . 50
Gender . . . 53
Phone Service . . . 56
Multiple Lines . . . 59
Internet Service . . . 62
Streaming TV . . . 65
Streaming Movies . . . 68
Payment Method . . . 71

Models/Analysis **74**
Tenure T-test . . . 74
Monthly Charges T-test . . . 77
Chisq Test for Contract Type and Churn . . . 81

Answering our Question. **83**

Load Telco Data

The following code will load in the Telco Dataset, and setup the variable types appropriately.

```r
#https://github.com/tidyverse/readr/issues/530
telco <- read_csv('WA_Fn-UseC_-Telco-Customer-Churn.csv',
                  col_types = cols()) %>%
  # Change all character-type variables to factor variables
  # Implementation Inspired by:
  #   https://stackoverflow.com/questions/27668266/dplyr-change-many-data-types
  #   https://bit.ly/2qeo4me
  mutate_if(sapply(., is.character), as.factor) %>%
  mutate(
    # Senior Citizen is coded numerically (0,1), and indicates if someone is
    # a senior citizen, so we switched it's type to factor, and labelled it accordingly.
    SeniorCitizen = factor(SeniorCitizen, labels = c('No', 'Yes')),
    # CustomerID was converted to factor by the blanket mutate_if above,
    # but has too many levels to be a factor variable, so we'll convert it
    # back to a character variable.
    customerID = as.character(customerID)
  )

str(telco)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : chr  "7590-VHVEG" "5575-GNVDE" "3668-QPYBK" "7795-CFOCW" ...
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
```

Research Question

Is a customer's decision to Churn independent of the type of contract a customer holds, the customer's tenure, and the customers monthly charges?

Null Hypotheses

1. Customer's Decision to Churn and the type of contract a customer holds are independent.

2. There is no difference in the mean tenture for Customers who decided to Churn, and those who did not decide to Churn. $\mu_{Churned} - \mu_{NotChurned} = 0$
3. There is no difference in the mean monthly charges for Customers who decided to Churn, and those who did no decide to Churn. $\mu_{Churned} - \mu_{NotChurned} = 0$

Alternative Hypotheses

1. Customer's Decision to Churn and the type of contract a customer holds are not independent.
2. The mean tenure for Customers who decided to Churn is less than the mean tenure of and those who did not decide to Churn. $\mu_{Churned} - \mu_{NotChurned} < 0$
3. The mean monthly charges for Customers who decided to Churn is greater than the mean monthly charges of the customers who did not decide to Churn. $\mu_{Churned} - \mu_{NotChurned} > 0$

Conducting Exploratory Data Analysis

See "Literature-Focused EDA" for more comprehensive EDA on how the variables in our research question relate to Churn.

a. Conduct Exploratory Data Analysis that will begin to answer your question, this can include for example:
b. Summary of the dataset/Descriptive Statistics

ii. Graphical representations of the data

Corplot and Descriptives

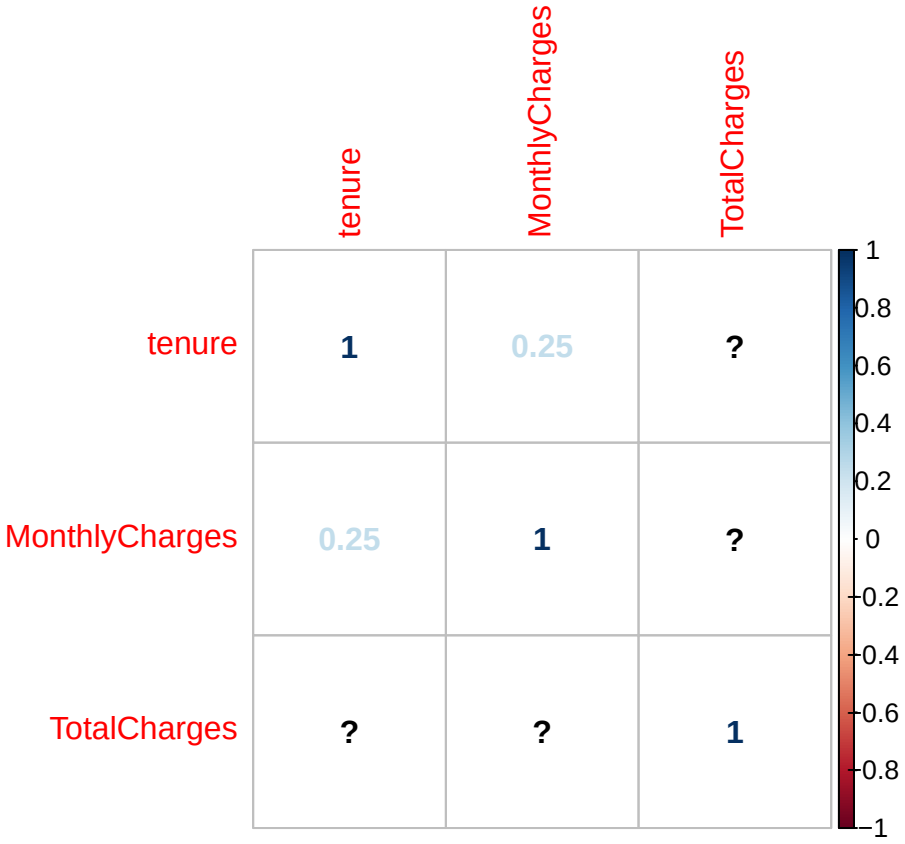
First, let's start with some general descriptive statistics and broad comparisons of the variables:

Numeric Variables

```
telcon <- telco %>%
  select_if(
    sapply(telco, is.numeric))
describe(telcon)
```

```
##                vars    n    mean      sd  median trimmed     mad   min
## tenure            1 7043   32.37   24.56   29.00   31.43   32.62  0.00
## MonthlyCharges    2 7043   64.76   30.09   70.35   64.97   35.66 18.25
## TotalCharges      3 7032 2283.30 2266.77 1397.47 1970.14 1812.92 18.80
##                    max  range  skew kurtosis    se
## tenure           72.00   72.0  0.24    -1.39  0.29
## MonthlyCharges  118.75  100.5 -0.22    -1.26  0.36
## TotalCharges   8684.80 8666.0  0.96    -0.23 27.03
```

```
# corplot of numeric variables
M <- cor(telcon)
corrplot(M, method='number')
```



Interestingly, all of our numeric variables have a rather large standard deviation compared to their respective means and ranges, especially for the Total Charges variable (additionally, its mean is roughly 1.6x larger than its median.)

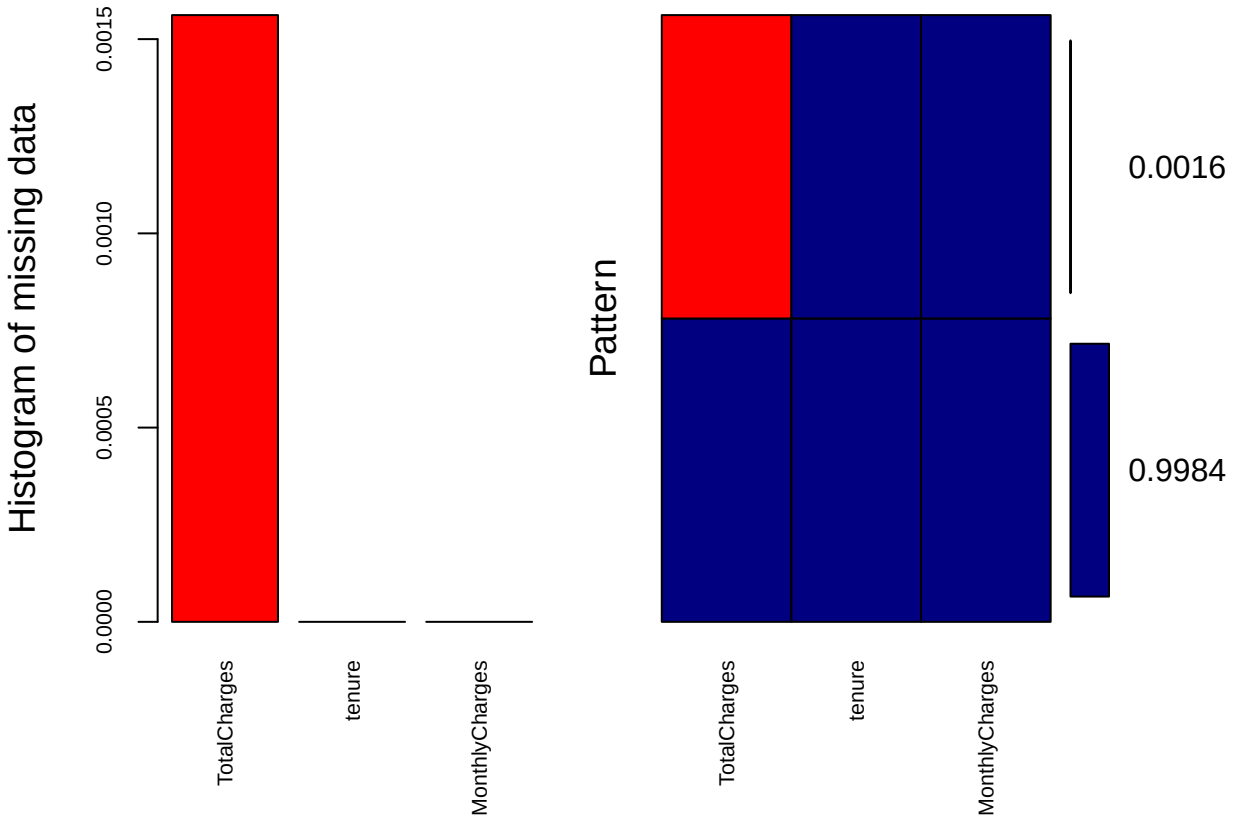
This makes sense though, because customers have different combinations of services, leading to custom variations in each customers Monthly Charges. The longer customers stay with the company, the more the variations in Monthly Charges expand on the Variance in Total Charges. In fact, it's quite likely that the variance linearly increases as the tenure of customers increase.

Tenure should probably have a high standard deviation because our data only contains the records of customers who haven't churned, which we beleive explains the right skewed-ness we see represented by the `describe` output.

MonthlyCharges and Tenure are skewed slightly left and right respectively, whereas TotalCharges is heavily right skewed as we mentioned earlier.

The corrplot of tenure, MonthlyCharges and TotalCharges, shows that there may be a weak, positive linear relationship between tenure and MonthlyCharges. The `?`s indicate that there are missing TotalCharge values. Let's see if we can impute the missing values and try building the corrplot again. First, we'll need a better understanding of where the missing values are, as it will inform our imputation method.

```
# https://www.r-bloggers.com/imputing-missing-data-with-r-mice-package/
aggr_plot <- aggr(telcon, col=c('navyblue','red'), numbers=TRUE,
                  sortVars=TRUE, labels=names(data), cex.axis=.7,
                  gap=3, ylab=c("Histogram of missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##        Variable       Count
##    TotalCharges 0.001561834
##          tenure 0.000000000
##  MonthlyCharges 0.000000000
```

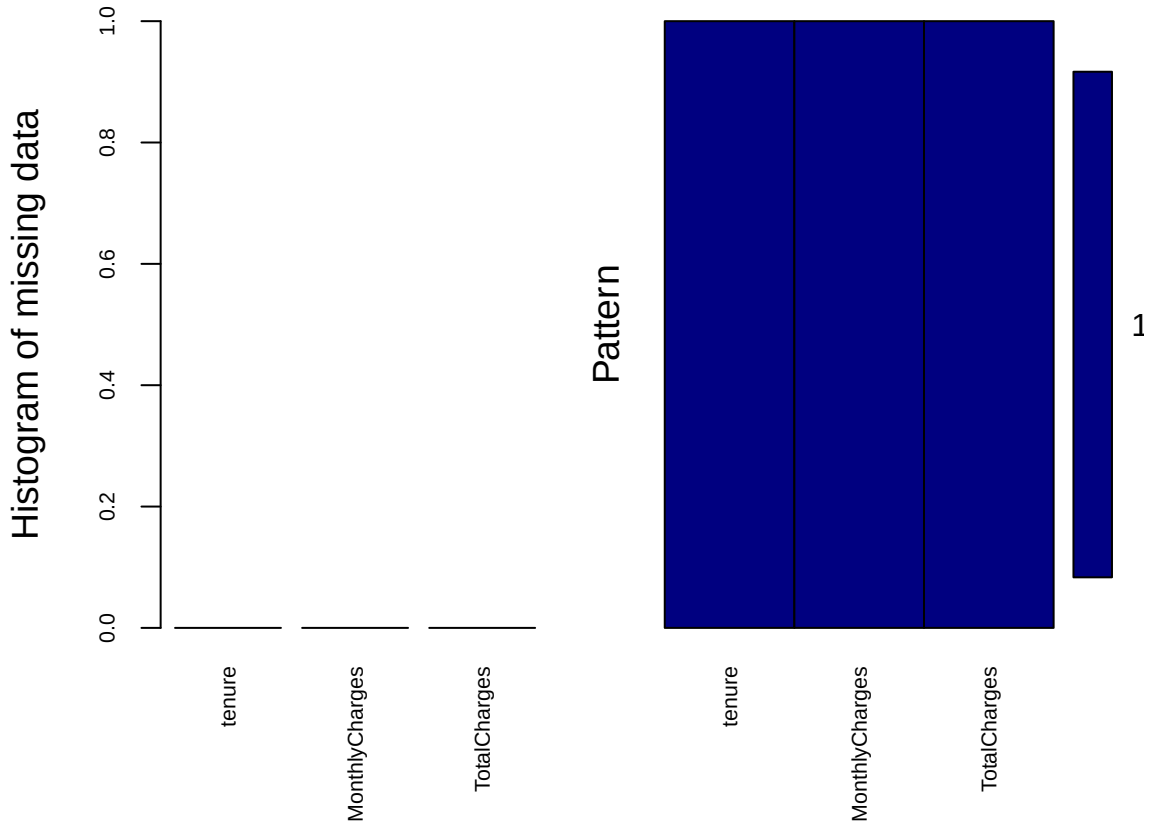
So we can see that the missing values are restricted to TotalCharges, and aren't exceptionally dense. Let's impute the values using mice, and recompute the corrplot.

```
# Since the missing data isn't very dense:
#  Using default parameters for mice
#  Immediately completing the dataframe
telcon.imputed <- complete(mice(data = telcon))
```

```
##
##  iter imp variable
##   1   1  TotalCharges
##   1   2  TotalCharges
##   1   3  TotalCharges
##   1   4  TotalCharges
##   1   5  TotalCharges
##   2   1  TotalCharges
##   2   2  TotalCharges
##   2   3  TotalCharges
##   2   4  TotalCharges
##   2   5  TotalCharges
##   3   1  TotalCharges
##   3   2  TotalCharges
```

```
##    3   3  TotalCharges
##    3   4  TotalCharges
##    3   5  TotalCharges
##    4   1  TotalCharges
##    4   2  TotalCharges
##    4   3  TotalCharges
##    4   4  TotalCharges
##    4   5  TotalCharges
##    5   1  TotalCharges
##    5   2  TotalCharges
##    5   3  TotalCharges
##    5   4  TotalCharges
##    5   5  TotalCharges
```

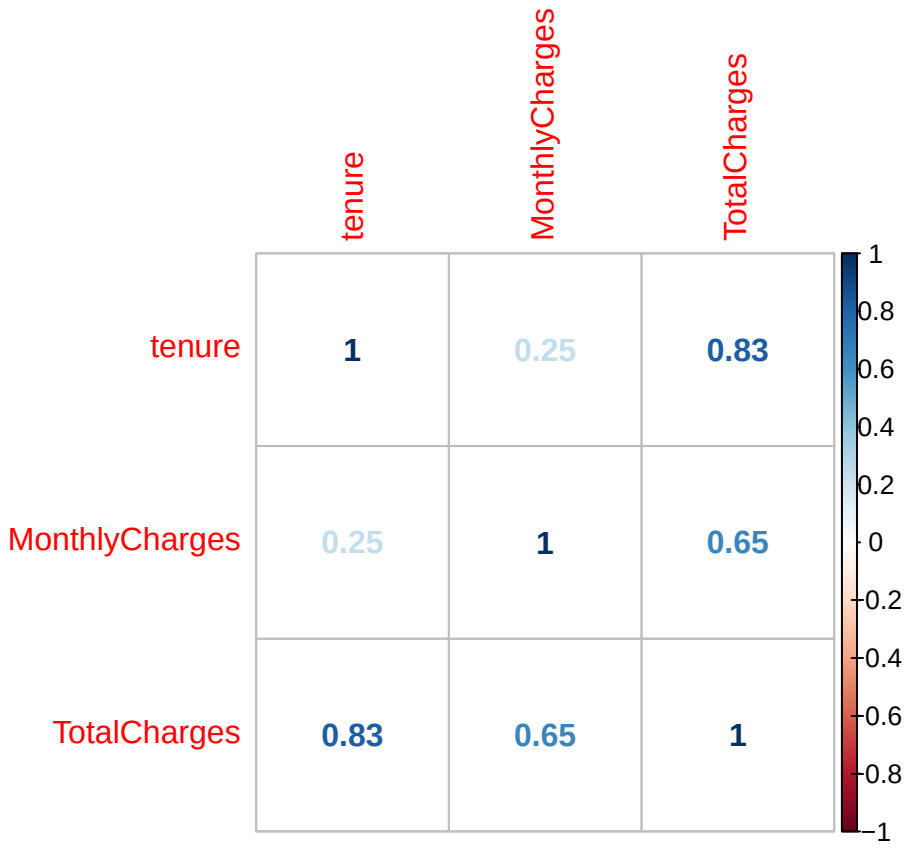
```
# recompute aggr_plot
aggr_plot.imputed <- aggr(telcon.imputed, col=c('navyblue','red'), numbers=TRUE,
                          sortVars=TRUE, labels=names(data), cex.axis=.7, gap=3,
                          ylab=c("Histogram of missing data","Pattern"))
```




```
##
##  Variables sorted by number of missings:
##        Variable Count
##          tenure     0
##  MonthlyCharges     0
##    TotalCharges     0
```

Excellent. telcon.imputed has no missing values.

```
M.imputed <- cor(telcon.imputed)
corrplot(M.imputed, method = "number")
```

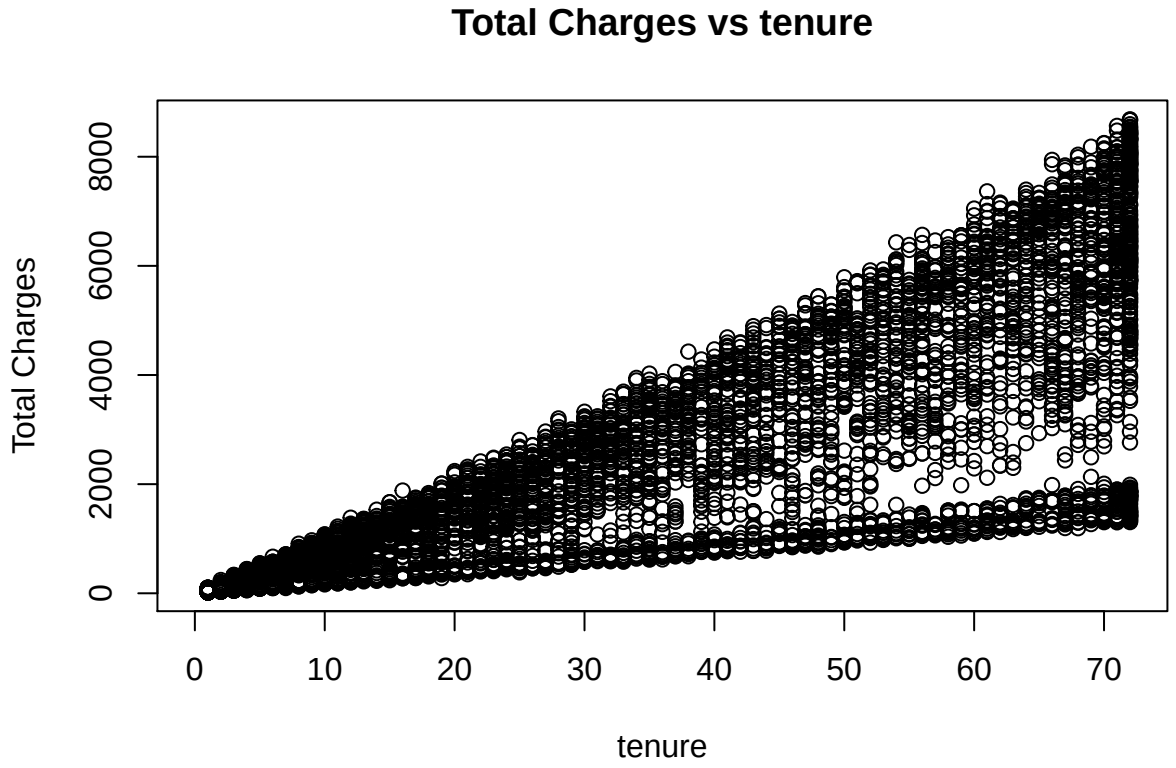


Now with the completed corrplot of the numeric values in our data, we see that there is a strong positive correlation between tenure, and TotalCharges, which is intuitive, as we'd expect both fields to increase every month for any customer.

Again, the same weak positive correlation between Monthly Charges and TotalCharges, which fits our assumptions as to how TotalCharges would nominally be calculated.

Let's take a look at how tenure and totalCharges are related graphically. We know from the corplot above, that they are strongly correlated:

```
plot(telco$tenure,telco$TotalCharges,
     main = "Total Charges vs tenure",
     ylab = "Total Charges",
     xlab = "tenure")
```

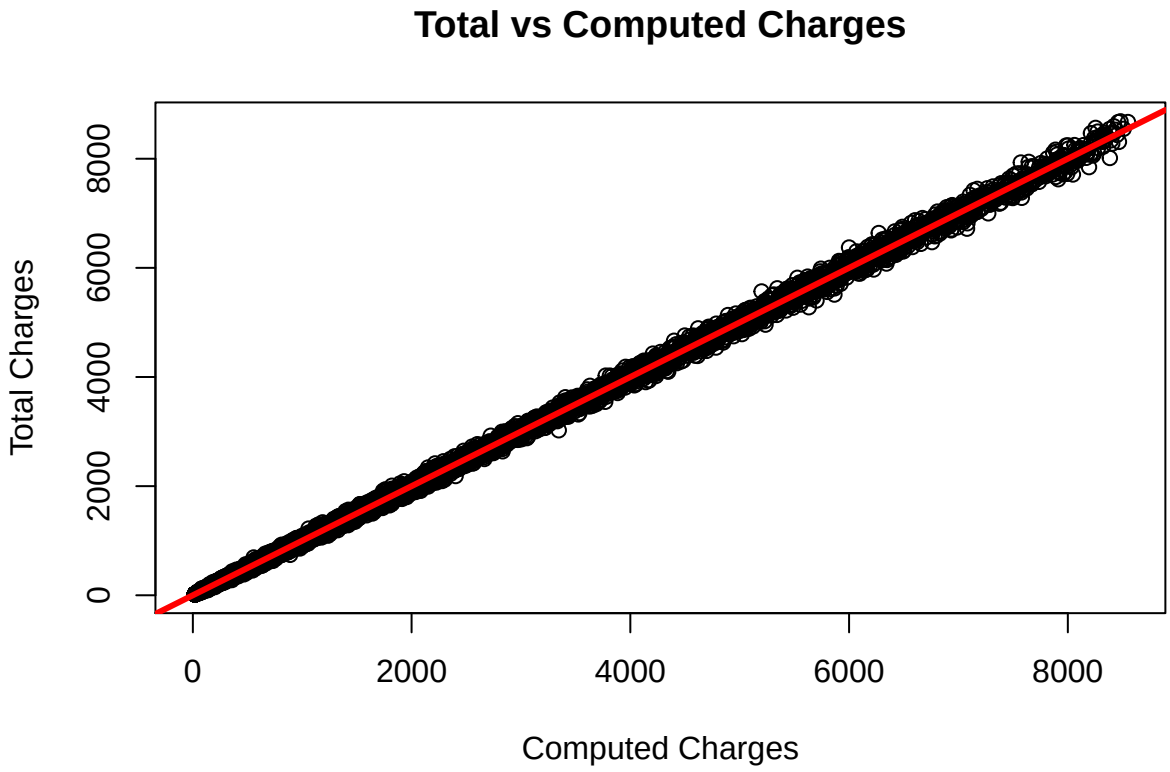


We can clearly see from this plot why the corplot above called out TotalCharges and tenure as being related. Visually, you see the lower and upper bounding of the monthly charges a customer of the company gets. Most likely due to the variety of service configuration options being offered (which makes us believe that maybe the relationship isn't quite linear)

A simple approximation of TotalCharges should be ($TotalCharges = tenure * MonthlyCharges$), but it is only an approximation because we can't assume that the customers monthly charges stay constant over time. We have to remember we're looking at these values at a given snapshot in time.

Lets apply the equation and plot this:

```
telco.computedCharges <- telco %>%
  mutate(
    computedCharges = tenure*MonthlyCharges
  )
plot(x=telco.computedCharges$computedCharges, y=telco.computedCharges$TotalCharges,
     xlab = "Computed Charges", ylab = "Total Charges",
     main = "Total vs Computed Charges")
abline(0,1,col = 'Red', lwd=3)
```



As expected, this is pretty good approximation, but there is some variance, which is likely because we only have access to data given at a snapshot in time.

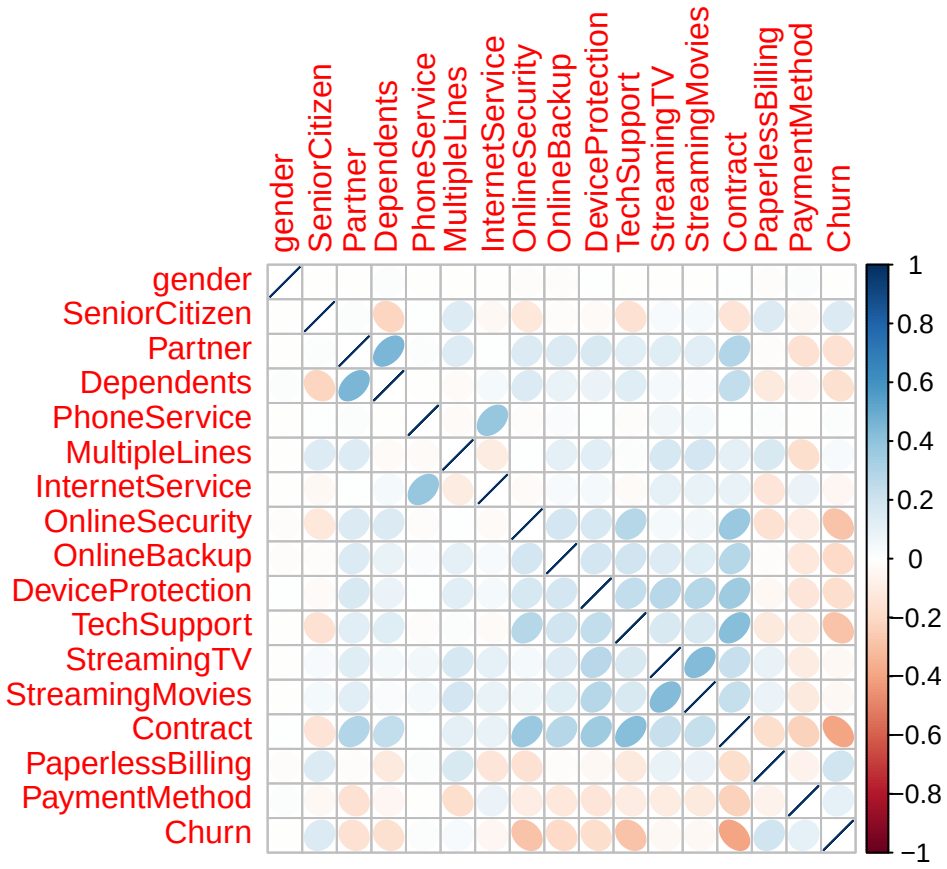
Categorical Variables

```
telcoc <- telco %>% select_if(
  sapply(., is.factor)
)
str(telcoc)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    7043 obs. of  17 variables:
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
##  $ 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 ...
##  $ 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 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

We later realized that this corrplot cannot be accurately interpreted because categorical variables cannot co-vary, we but left it (and the other two) in our markdown document to show our process.

```
# We'll need to force the factor variables into numeric variables temporarily
telcoc.num <- telcoc %>% mutate_if(sapply(., is.factor), as.numeric)
M <- cor(telcoc.num)
corrplot(M, method = "ellipse")
```



Literature-Focused EDA

First, let's create dataframes for customers who have and have not churned as we will be splitting the data often in our EDA.

```
telco_yes <- telco %>% filter(Churn == "Yes")
telco_no <- telco %>% filter(Churn == "No")
nrow(telco_yes)
```

```
## [1] 1869
```

```
nrow(telco_no)
```

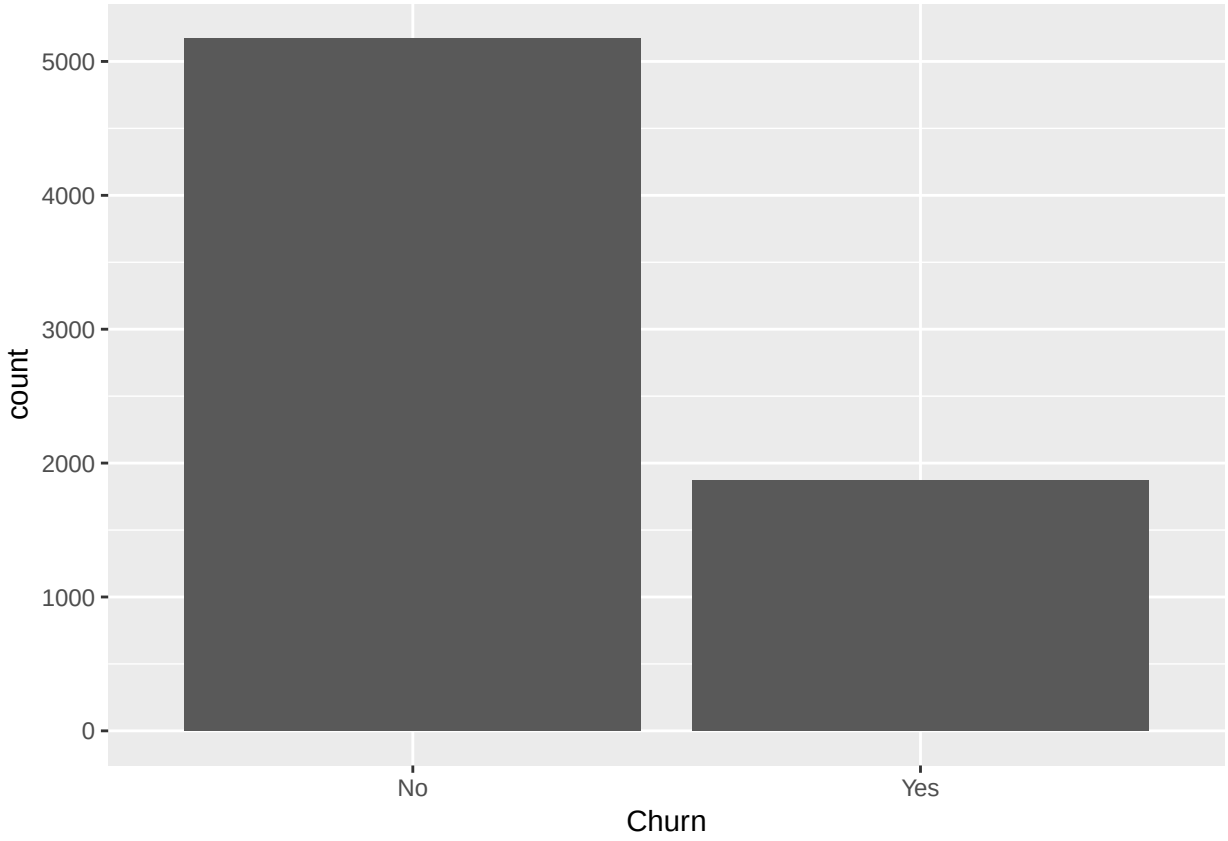
```
## [1] 5174
```

Churn

Churn indicates whether a given customer has decided to leave the company this month, or not.

How many of the customers churned in the given month:

```
ggplot(data=telco)+geom_bar(aes(x=Churn))
```




```
# calculate percentage churn
t <- table(telco[,'Churn'])
t[[2]]/(t[[1]]+t[[2]])
```

```
## [1] 0.2653699
```

From the graph we can see that the majority of customers aren't churning, but at the same time we've illustrated why the Churn factors are important to telecom companies. 26% of their **total** customers are churning in the given month.

Contract Type

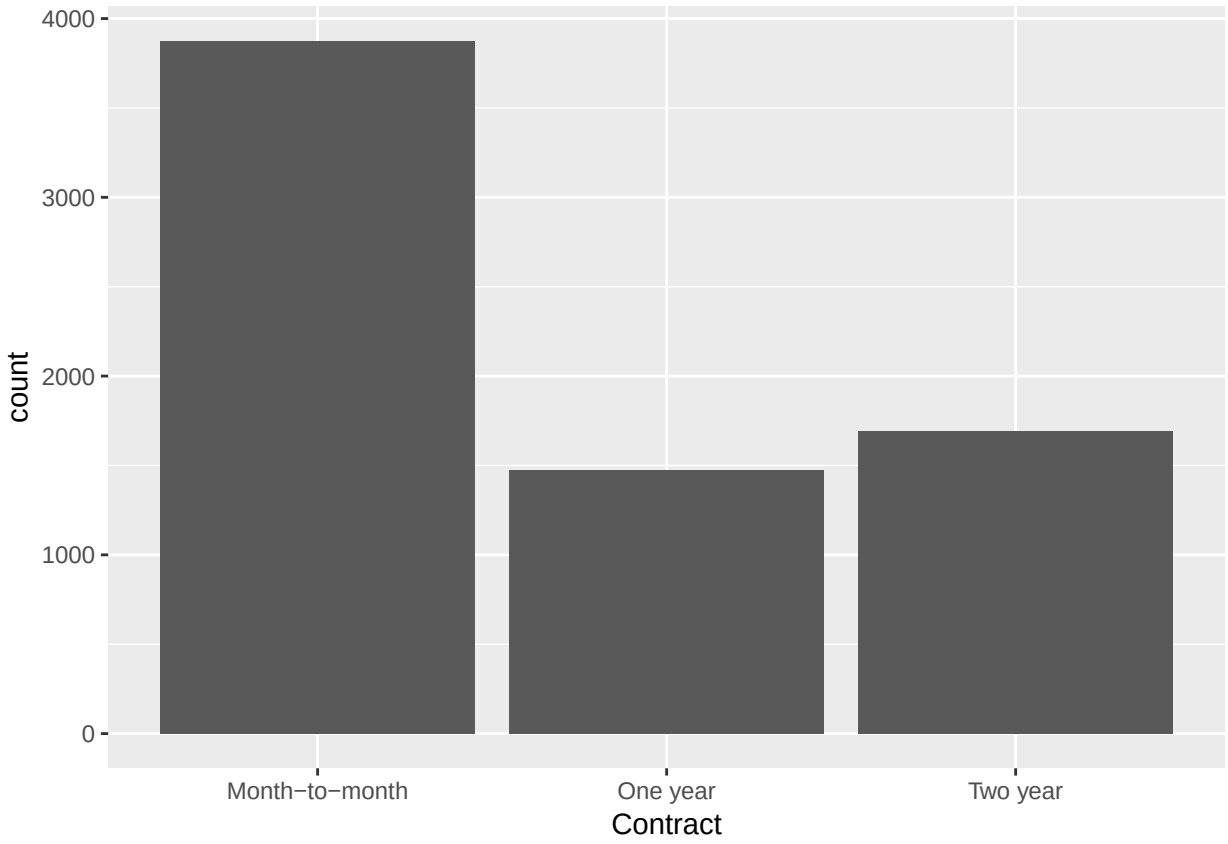
There are three kinds of contracts a customer can have, that are mentioned in the telco data.

```
levels(telco$Contract)
```

```
## [1] "Month-to-month" "One year"       "Two year"
```

So how many customers in the data are using each contract type?

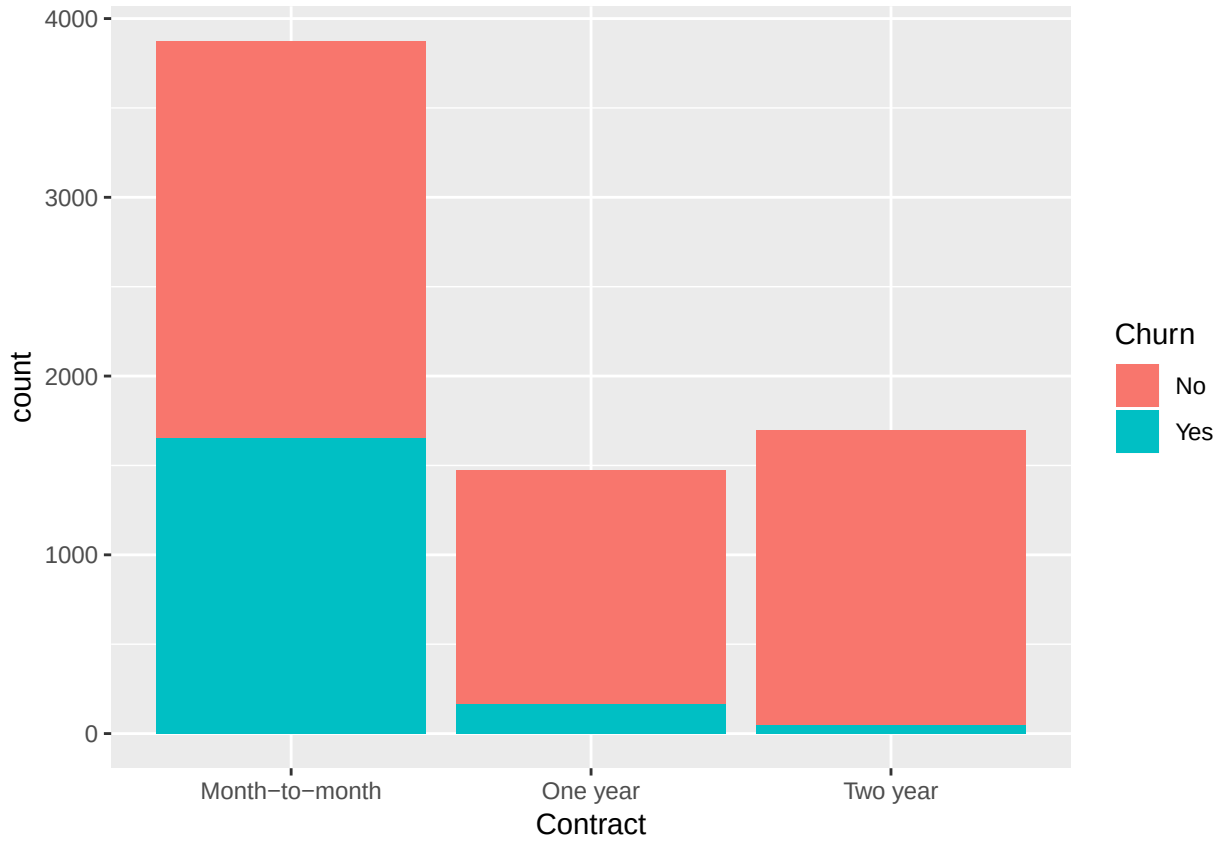
```
ggplot(data=telco) + geom_bar(aes(x=Contract))
```



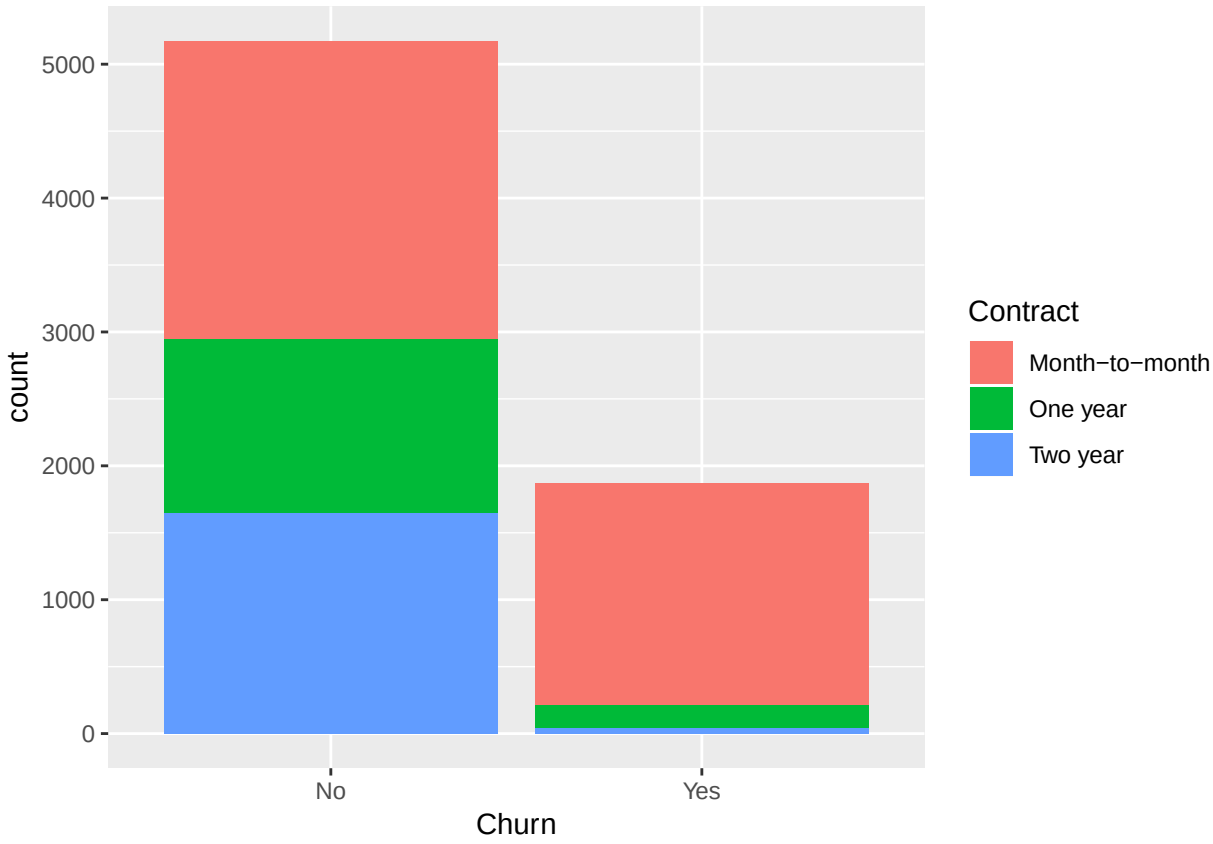
We can see from the graph that the month-to-month contracts are far more popular than either the one year or two year contracts. But we should also take a look at how many customers using the each contract type are churning in the given month, vs. how many customers using the other contract types are churning.

Below we provide two charts for doing that:

```
churn_vs_contract_plot = ggplot(data=telco,aes(x=Contract))
churn_vs_contract_plot + geom_bar(aes(fill = Churn))
```




```
contract_vs_churn_plot = ggplot(data=telco,aes(x=Churn))
contract_vs_churn_plot + geom_bar(aes(fill=Contract))
```



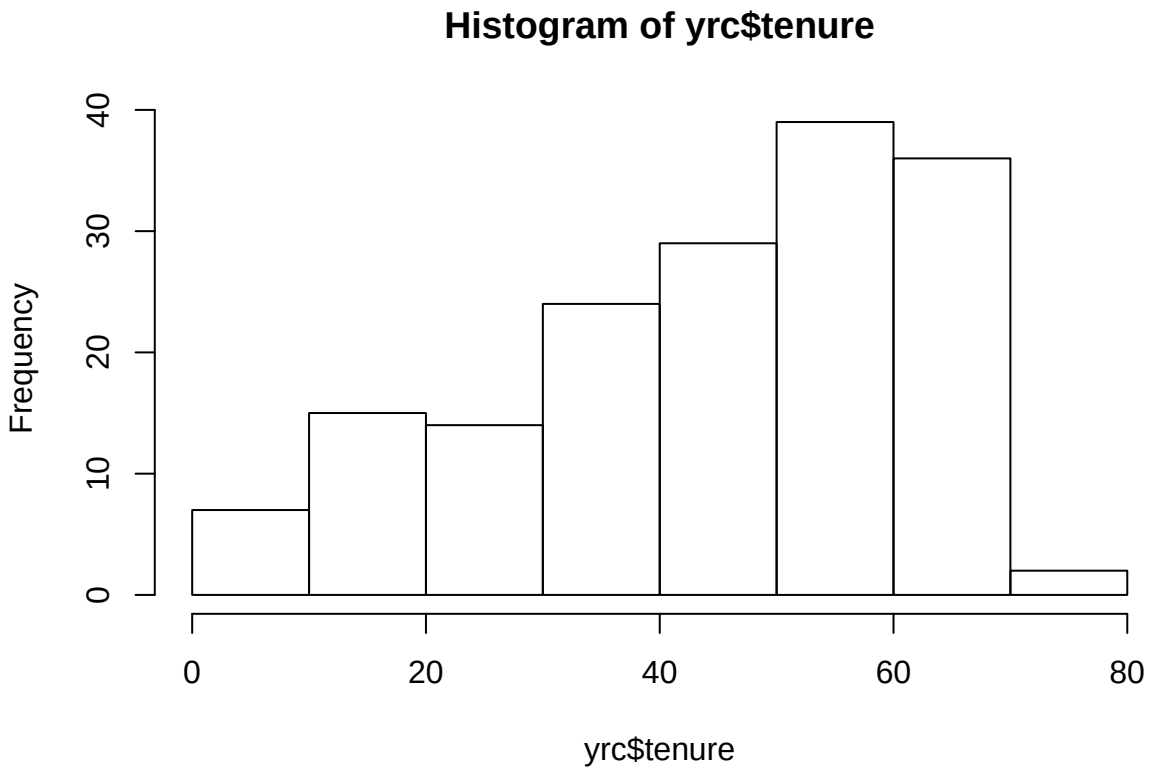
Visually, the Contract and Churn variables don't seem to be independent, which casts doubt our null hypothesis, but we'll revisit this again with a more rigorous analysis later on.

It may also be interesting to see the distribution of the tenure of the few customers who had one or two year contracts who churned.

```
# Filter for customers who have one or two year contracts
yrc <- telco %>%
  filter(
    Contract == "One year" |
      Contract == "Two Year",
    Churn == "Yes"
  )
unique(yrc$tenure)
```

```
##  [1] 53 38 54 68 22 56  2 33 46 62 58 60 50 55 12 59 25 19 70 39  5 41 64
## [24] 61  8 63 17 67 48 66 18 69 49 37 40 51 34 42 28 45 52  9 23 31 36 57
## [47]  4 35 21 72 13 43 32 44 65 30  7 11 47 14
```

```
hist(yrc$tenure)
```




```
nrow(yrc)
```

```
## [1] 166
```

```
skewness(yrc$tenure)
```

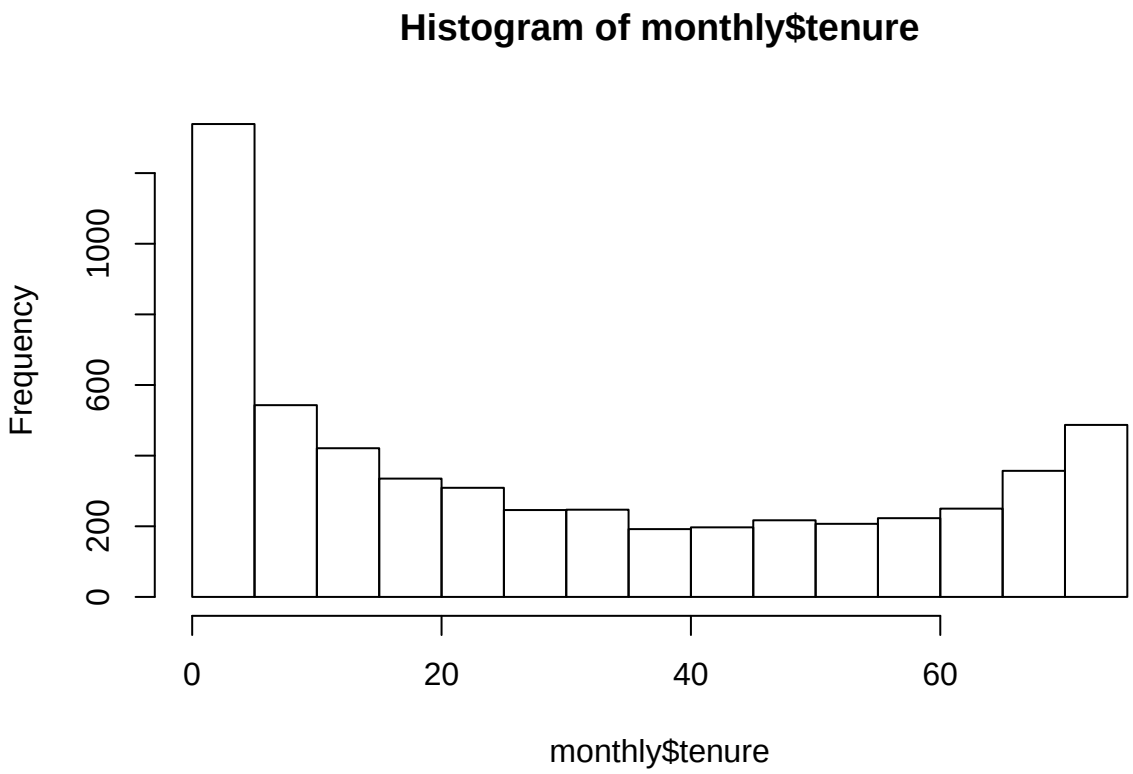
```
## [1] -0.5482024
```

Our data is fairly skewed to the left, showing that the majority of customers with one and two year contracts who churned have been with the company for a relatively long period of time, which makes sense.

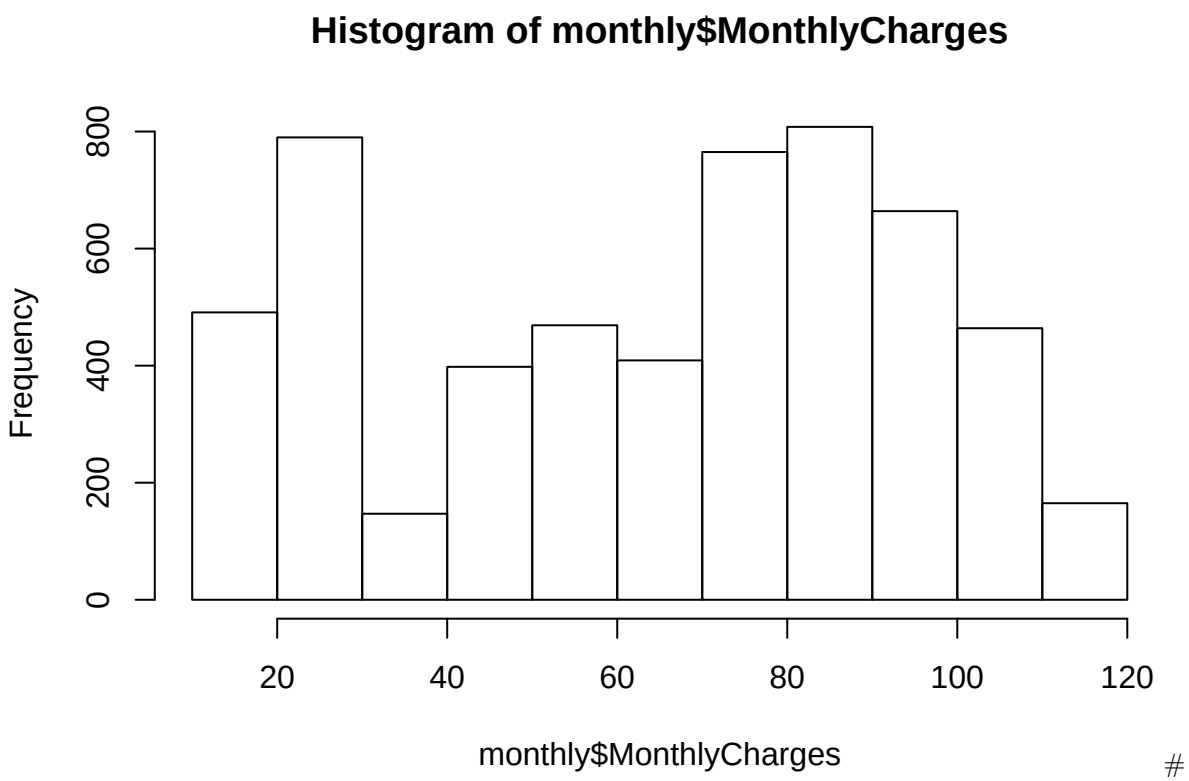
It is important to note that customers who have been with the company for less than 1 year have churned, i.e. they aren't locked in, so it allows us to compare the churn rates of customers who churned in any one month, which we couldn't do if only 1/12 of the customers who had yearly contracts, (and 1/24 of customers who had two year contracts) could churn.

Interestingly, the churn rates aren't highest after one or two years, rather near roughly 4.5 and 5.5 years (i.e. customers with longer term contracts typically surive one or two contracts before terminating their service.)

```
# Create a dataframe that only has month to month contract customers.
monthly <- telco %>%
  filter(
    Contract != "One year",
    Contract != "Two Year"
  )
hist(monthly$tenure)
```




```
hist(monthly$MonthlyCharges)
```



#

Surpisingly, there are more customers left with the company who have a tenure greater than 60 months than there are customers left with the company who have a tenure between 35 and 60 months. We would have expected to see a negative pattern across the graph.

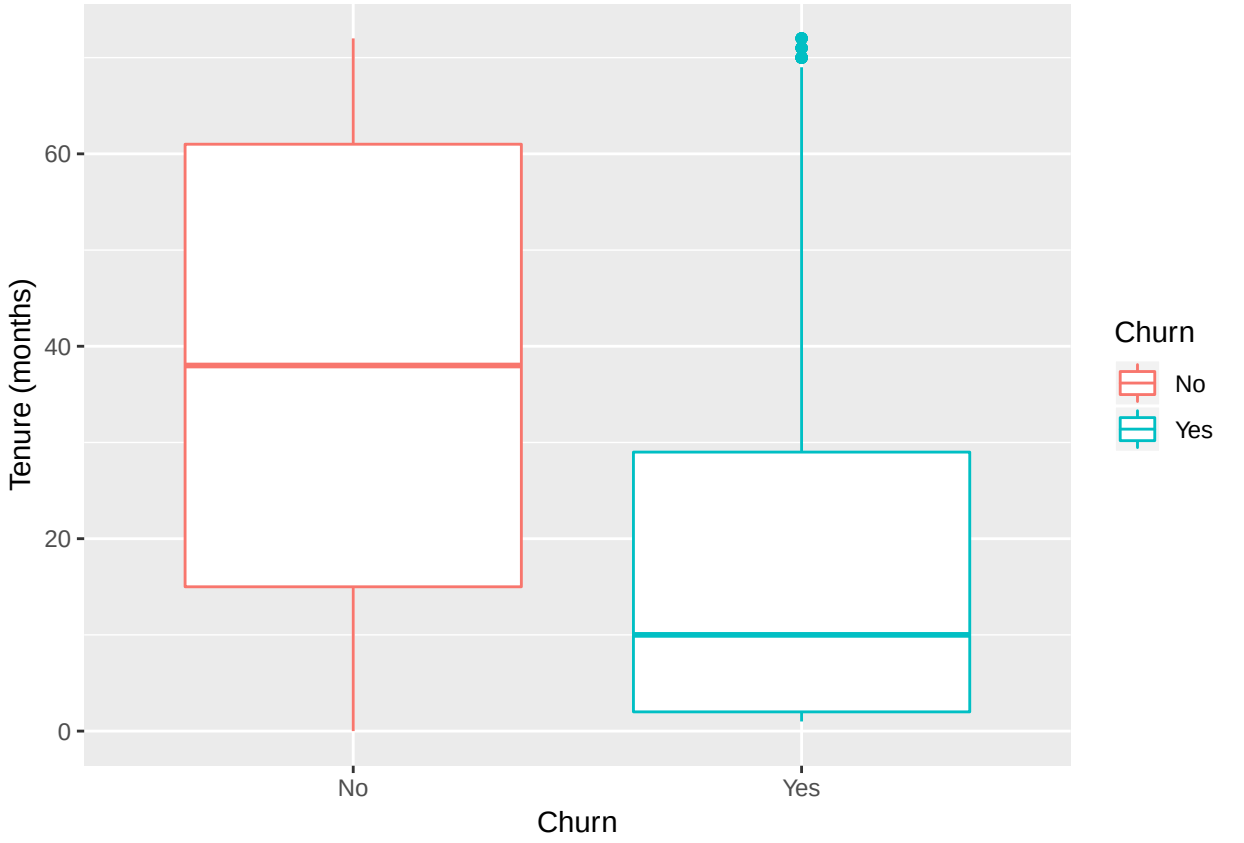
The customers who did not churn appear to have a wide range of Monthly Charges.

Tenure

Tenure is a continuous variable representing the time (in months) that a customer has been with the company.

Have churning customers been with the company for a long time, or are they new?

```
cten_box<-ggplot(telco, aes(x=Churn, y=tenure, color=Churn)) +
  geom_boxplot() + ylab("Tenure (months)")
cten_box
```




```
median(telco_yes$tenure)
```

```
## [1] 10
```

```
median(telco_no$tenure)
```

```
## [1] 38
```

```
mean(telco_yes$tenure)
```

```
## [1] 17.97913
```

```
mean(telco_no$tenure)
```

```
## [1] 37.56997
```

We can see that most of the customers that churned had less tenure than the customers who did not churn this month. The median tenure of the churning customers was 10 months, almost 1/4 the median tenure of

the non-churning customers, 38 months. The means were slightly closer because of outliers, but churning customers had a mean tenure roughly 1/2 as high as non churning customers.

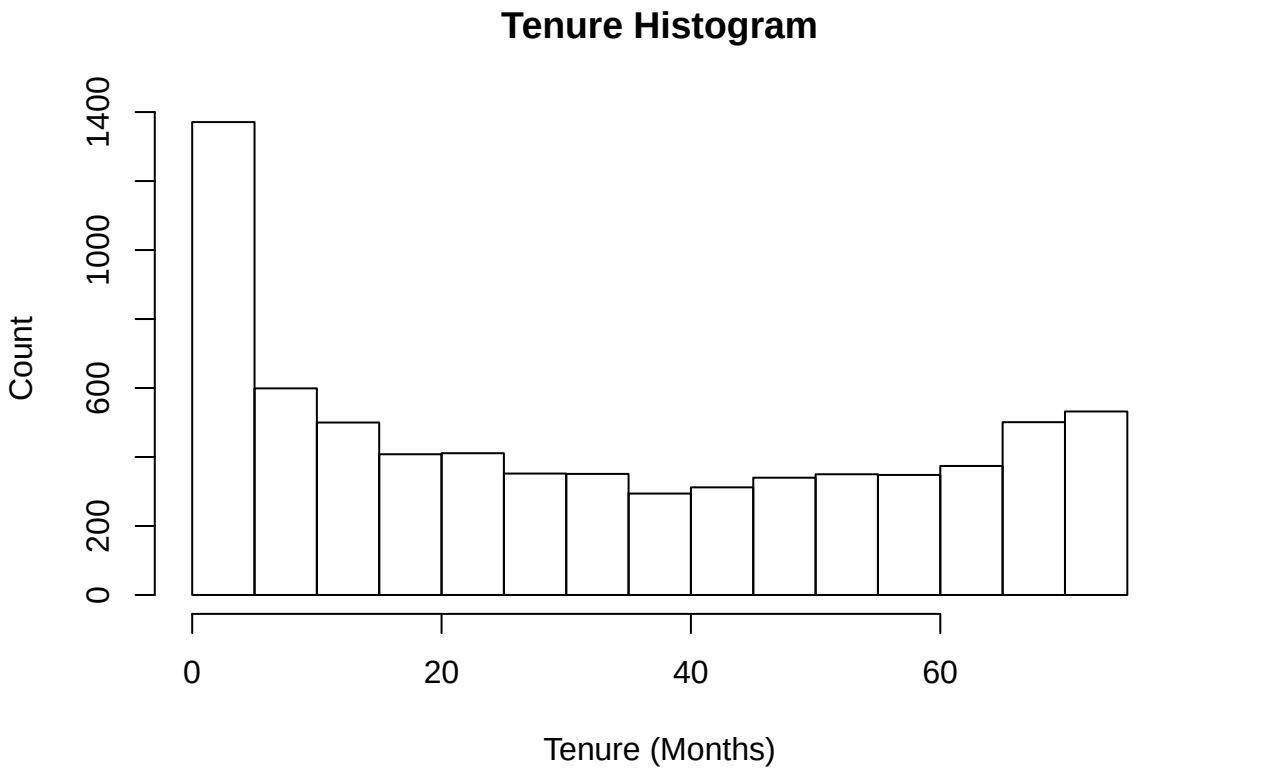
This makes us believe that we should conduct a one-sided, unpaired, two sample Welch's T-test to determine if the mean tenure of non-churning customers is greater than the mean tenure of churning customers. We chose that test because our sample standard deviations are not equal shown by the spread in our boxplots and our sample sizes are uneven.

To conduct our t-test, we will need to determine if our samples are independent and if our continous variable is (nearly) normally distrubted.

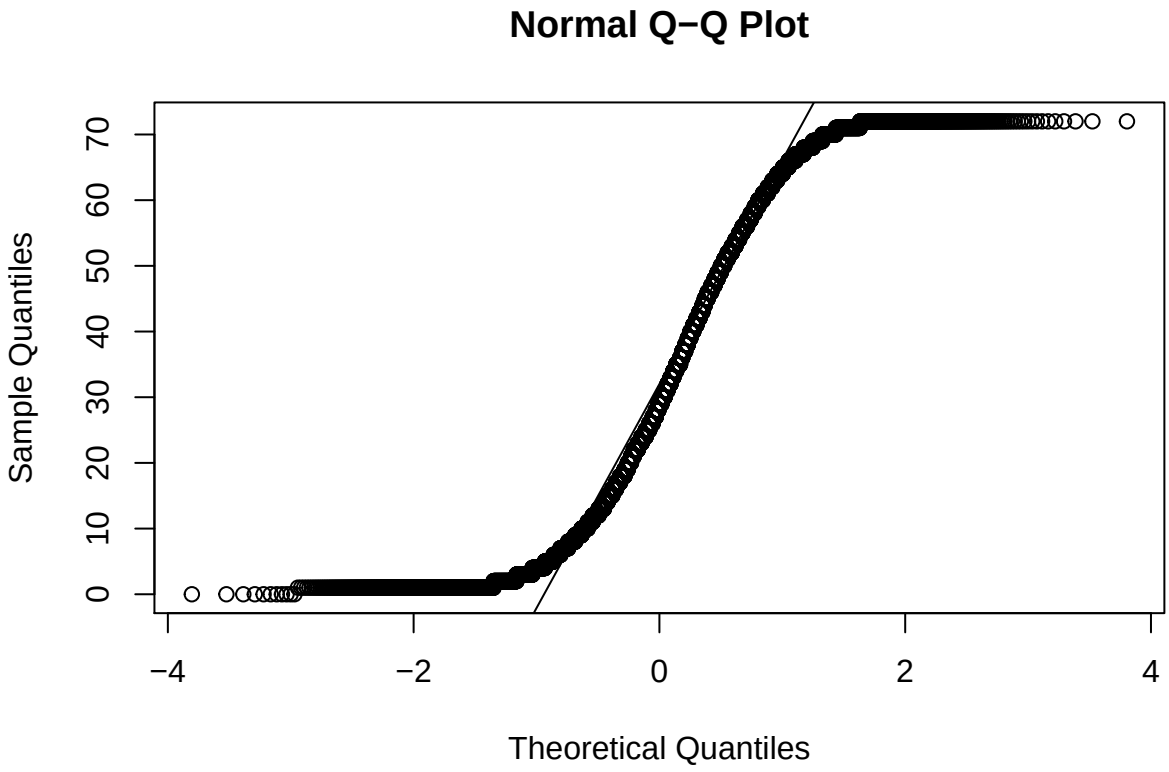
We aren't sure because we don't have a lot of metadata but have no reason to believe that the customers aren't churning independently from each other and t-tests are robust to normality if our sample size is large enough (there are 1869 customers who churned and 5174 who did not churn, both of which are much greater than 30) due to the Central Limit Theorem (More on this in our Models/Analysis Section)

Is it normally distributed?

```
hist(telco$tenure, main = "Tenure Histogram", xlab = "Tenure (Months)", ylab = "Count")
```




```
qqnorm(telco$tenure)
qqline(telco$tenure)
```



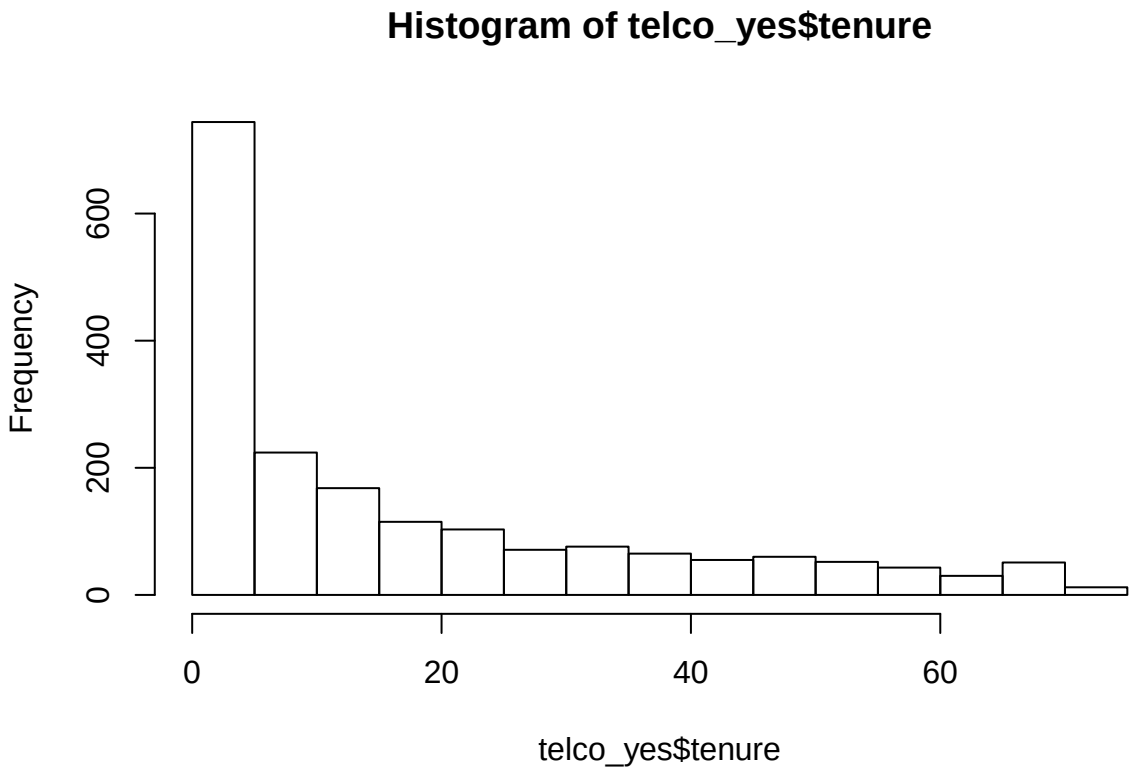
This is clearly not a normally distributed variable.

The histogram of the tenure data seems to indicate a large amount of newer customers relative to the number of customers that have long-running accounts. But it's hard to say if this is a recent trend or caused by the ongoing churn that the company is interesting in predicting.

We expected the graph to show a negative linear pattern where the number of customers who have stayed with telco has steadily declined over the years. There are, however more customers that joined in the last 5 months than any other 5 month interval in the data, but it appears to level out after about 25 months and then rises again after 5 years, which suggests other variables are influencing this, or the customers acquired in those months are outliers. Regardless of the cause, this is good news for the company, as there appears to be a solid core of happy long-term customers.

Lets look at tenure filtered based on Churn.

```
hist(telco_yes$tenure)
```



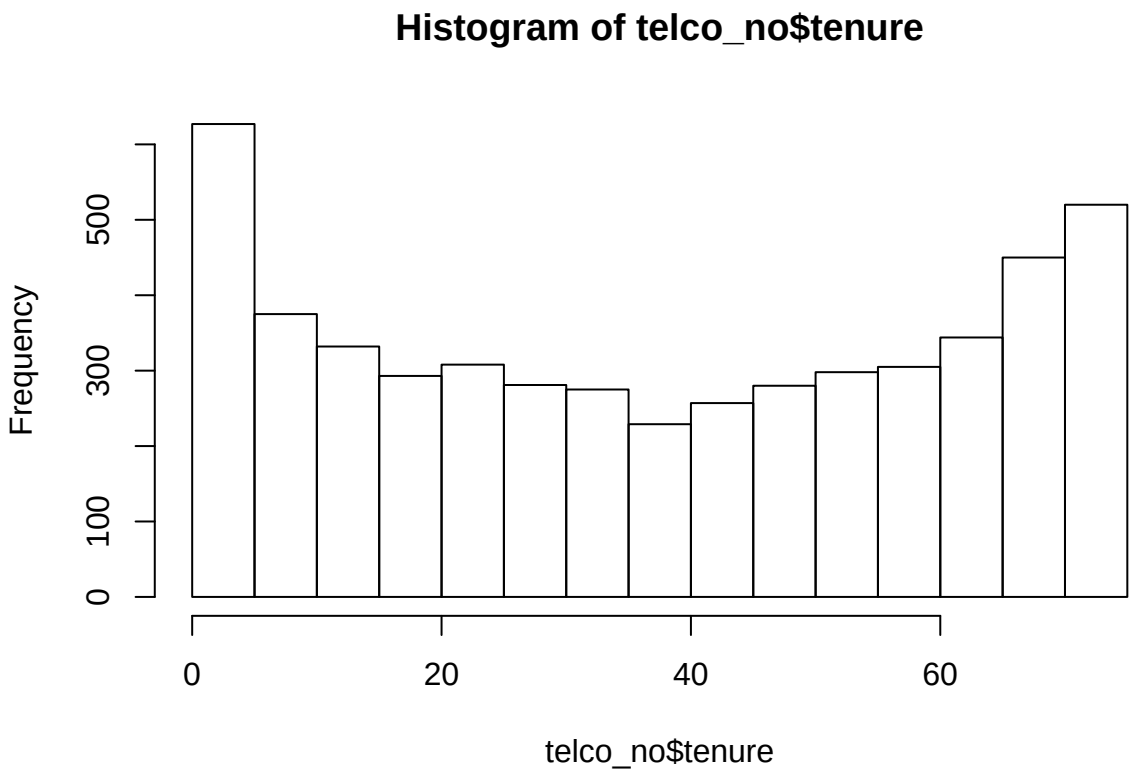
It appears that more of the Churning behavior is occuring in the most recently added customers.

```
# Check the percentage of churned customers who started their subscriptions
# within the last 10 months in the dataset.
telco_yes_less <- telco_yes %>% filter(tenure <= 10)
nrow(telco_yes_less)/nrow(telco_yes)
```

```
## [1] 0.517924
```

Indeed, over 51% of the customers that churned in the given month started their subscriptions within the previous 10 months, which makes sense given how the histogram of the tenure of churned customers looked.

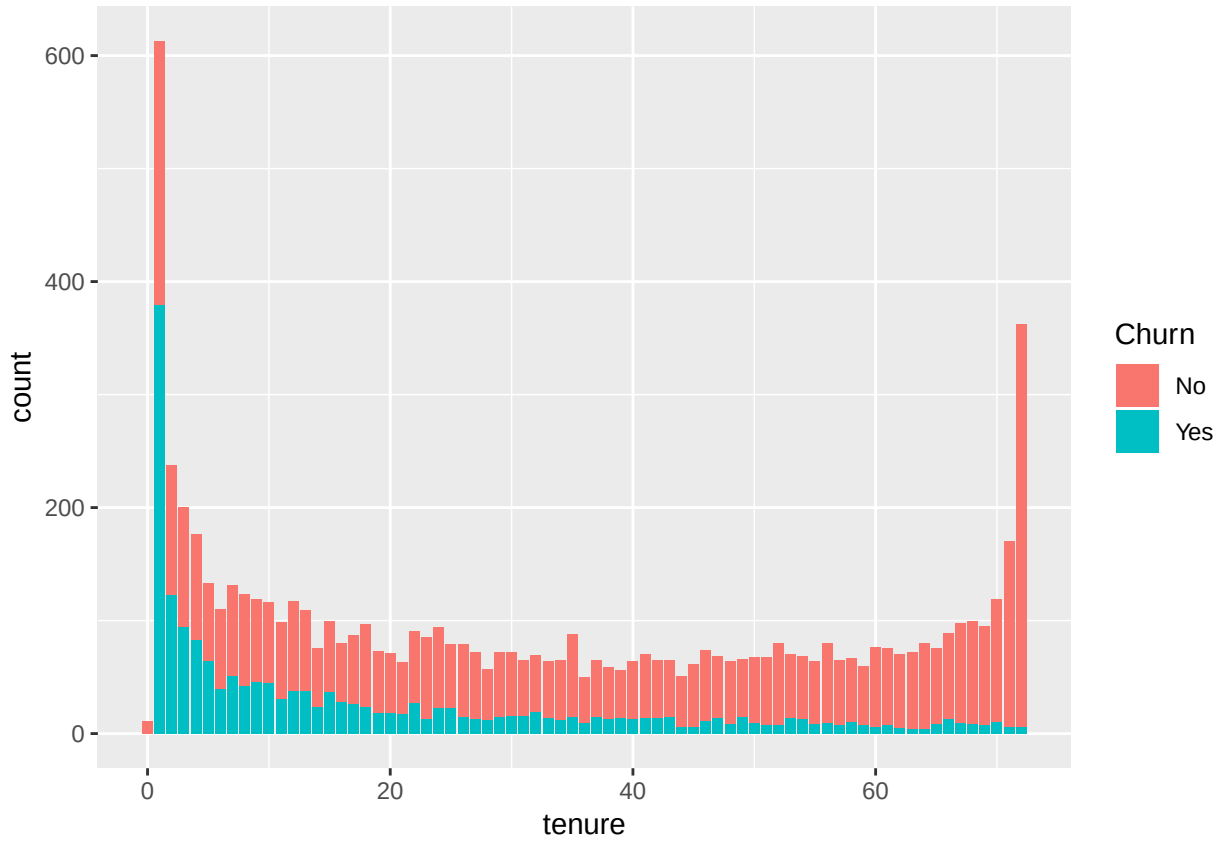
```
hist(telco_no$tenure)
```



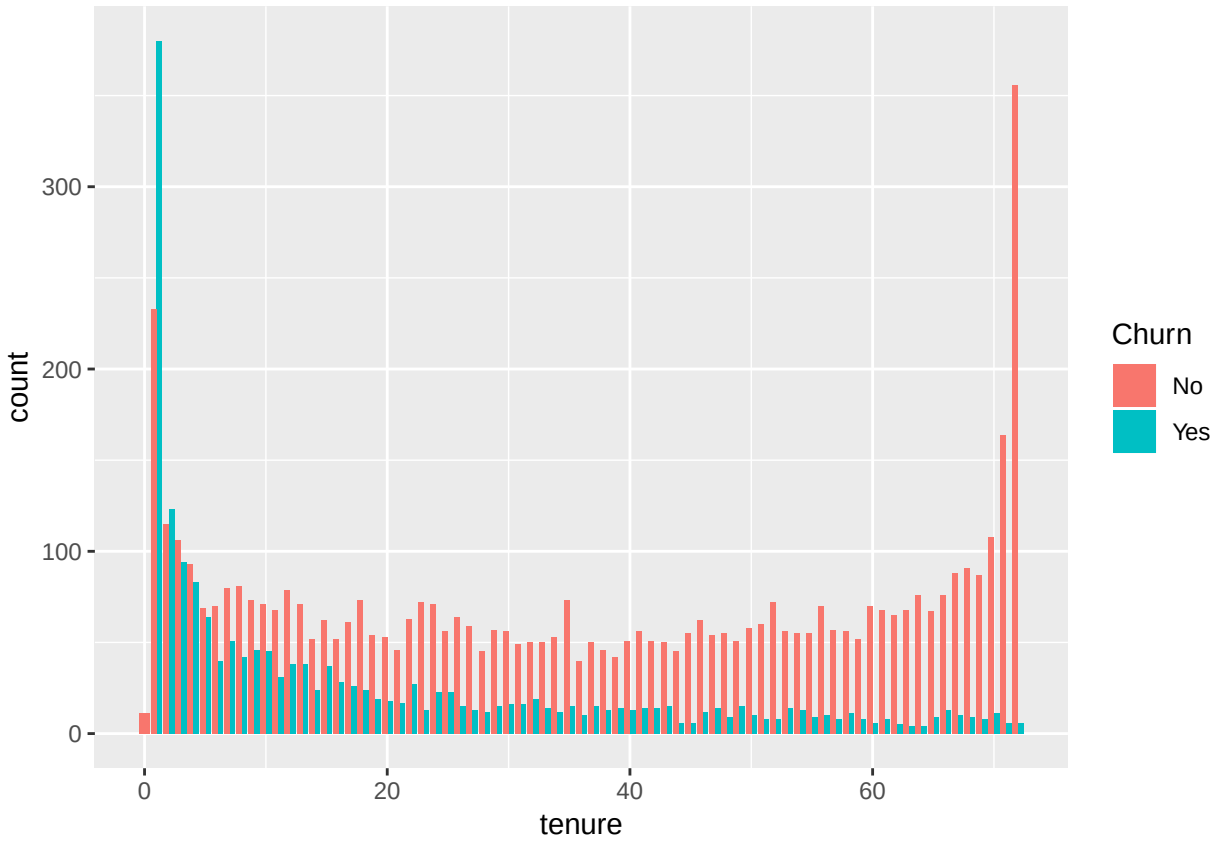
Relative to the overall histogram for tenure above, we can see the spike of new customers has dissapeared. It is also extremely interesting that the chart above has peaks at the x-limits of the graph and is lowest near the middle. This seems to imply that either a much smaller number of customers was joining the service about 40 months ago (compared to now and 6 years ago), or that the company may have experienceing a higher higher rate of churn during that period, if we were to assume the company obtains roughly the same number of new customers each month. It is hard to say because we don't have information on the customers who churned more than one month ago.

It may also be interesting to see 'Not Churning', and 'Churning' customers combined in the same graph vs tenure. Here are two variations of that graph.

```
ggplot(data=telco)+geom_bar(aes(x=tenure, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=tenure, fill=Churn),position='dodge')
```



Again, it appears that most customers who churn do it early, meaning generally that customers who have been with the company longer are less likely to churn than newer customers.

The tiny bar where tenure equals 0 appears to represent customers who haven't been with the company for a full month yet, we aren't certain why it's so much smaller than everything else. Speculating, it makes sense that none of these customers have churned because they technically haven't been billed yet. We think that the blib on the chart is due to the 'new' contracts in the first few days of the month month during which the data was recorded and were meant to be excluded from churn analysis, but since we couldn't be sure, we didn't remove them from our analysis.

It is important to note that the bars in this chart are calculated by adding up all of the customers who did not churn last month from each tenure point. In the process of doing this EDA, we found ourselves accidentally confusing total number of customers x number of months ago, with *the number of customers that joined x number of months ago that haven't churned during month 1*, which is what a single bar at a given tenure actually tells you.

We can clearly see that the majority of customers who opened their accounts last month have already Churned.

```
month1.no_telco <- telco %>% filter(tenure <= 1, Churn == "No")
month1.all <- telco %>% filter(tenure <= 1)
nrow(month1.no_telco)/nrow(month1.all)
```

```
## [1] 0.3910256
```

```
month2.no_telco <- telco %>% filter(tenure == 2, Churn == "No")
month2.all <- telco %>% filter(tenure == 2)
nrow(month2.no_telco)/nrow(month2.all)
```

```
## [1] 0.4831933
```

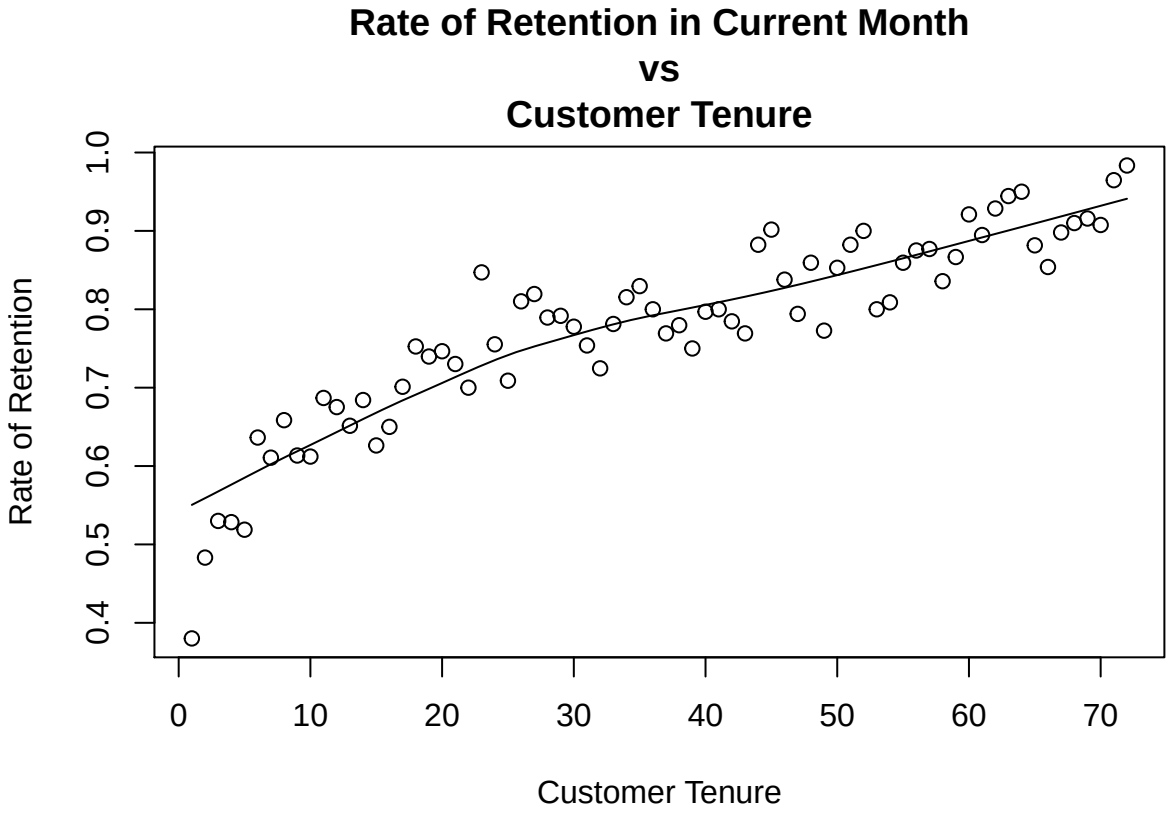
During the current month, the company only has ~39% of the customers that started using it's services

during the previous month.

Of those customers that started using the companies services 2 months ago, and were still using their services through the previous month, the company has only retained ~48%.

There appears to be a trend of sorts, lets look at %retention after the previous month / months of service.

```
ret <- telco %>% group_by(tenure, Churn) %>%
  summarize(n = n()) %>%
  mutate(freq = n/sum(n)) %>%
  filter(Churn == "No")
ret.final = ret[-1,]
scatter.smooth(y=ret.final$freq,
               x= ret.final$tenure,
               xlab = "Customer Tenure",
               ylab = "Rate of Retention",
               main = "Rate of Retention in Current Month\nvs\nCustomer Tenure")
```

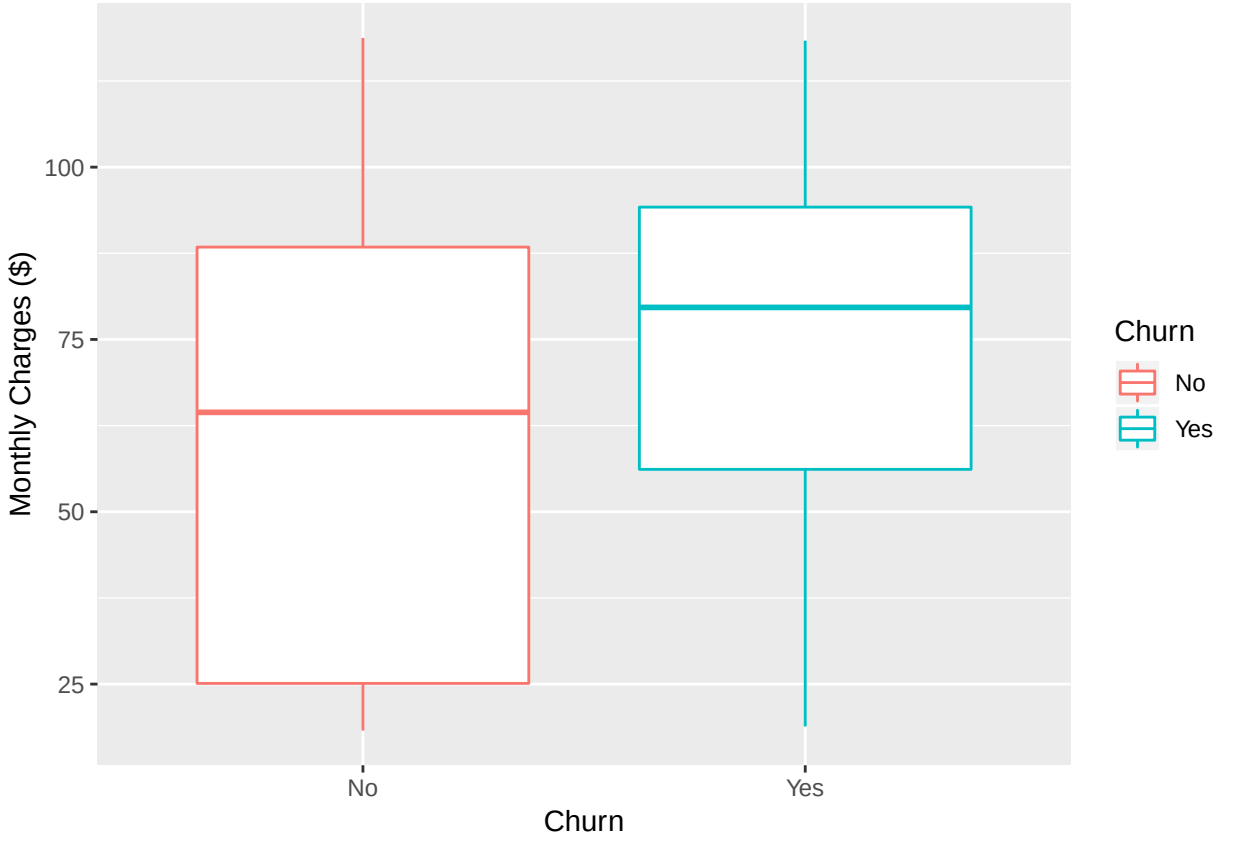


This shows a trend that seems to make intuitive sense... *Customers that have been with the company longer appear to be more likely to stay with the company through the current month.*

Monthly Charges

Side by side boxplots of Monthly charges, separated by Churn.

```
ggplot(telco, aes(x=Churn, y=MonthlyCharges, color=Churn)) +
  ylab("Monthly Charges ($)") + geom_boxplot()
```




```
median(telco_yes$MonthlyCharges)
```

```
## [1] 79.65
```

```
median(telco_no$MonthlyCharges)
```

```
## [1] 64.425
```

```
mean(telco_yes$MonthlyCharges)
```

```
## [1] 74.44133
```

```
mean(telco_no$MonthlyCharges)
```

```
## [1] 61.26512
```

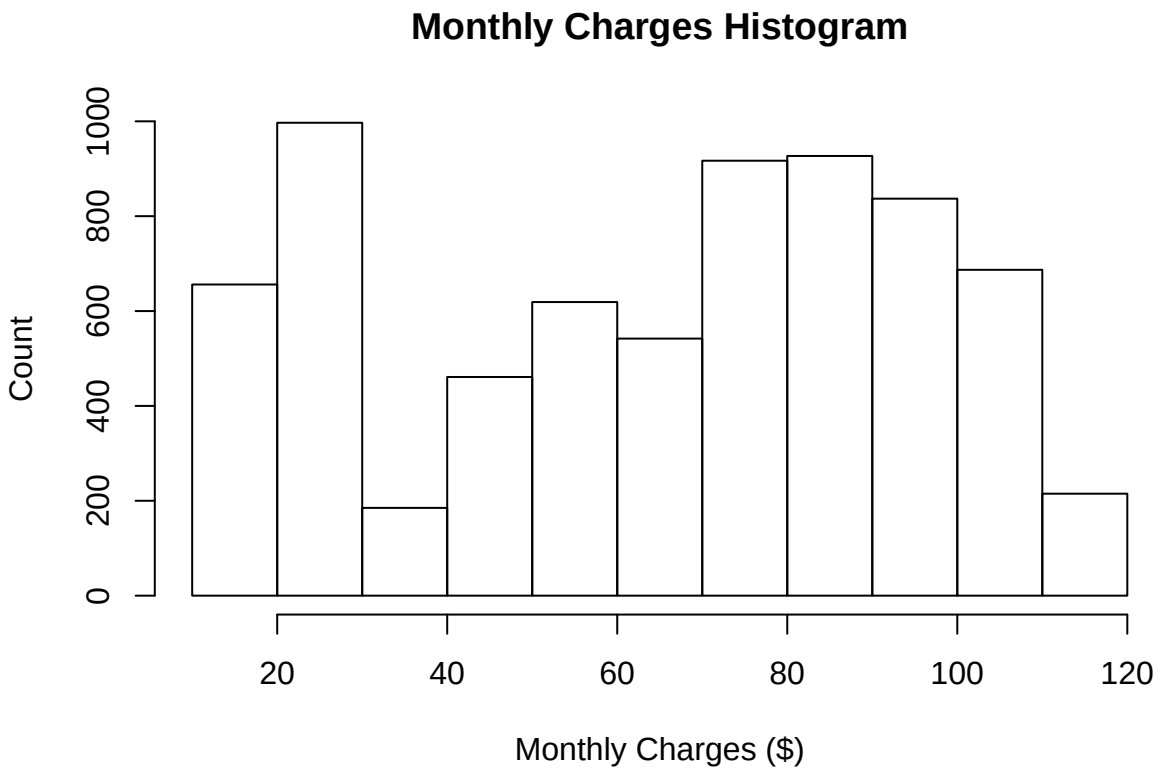
Both the median MonthlyCharges incurred by the churning customers is slighly higher than the median MonthlyCharges incurred by the non-churning customers, but what interested us is that the IQR of the boxplot of the churning customers is much smaller than the IQR of the boxplot of the non-churning customers suggesting that the middle 50 percent of the customers who churned are paying more than the customers who did not churn.

This makes us beleive that we should conduct a one-sided, unpaired, two sample Welch's T-test to determine if the mean MonthlyCharges incurred by churning customers is greater than the mean MonthlyCharges incurred by non-churning customers. (More on this in our Models/Analysis section)

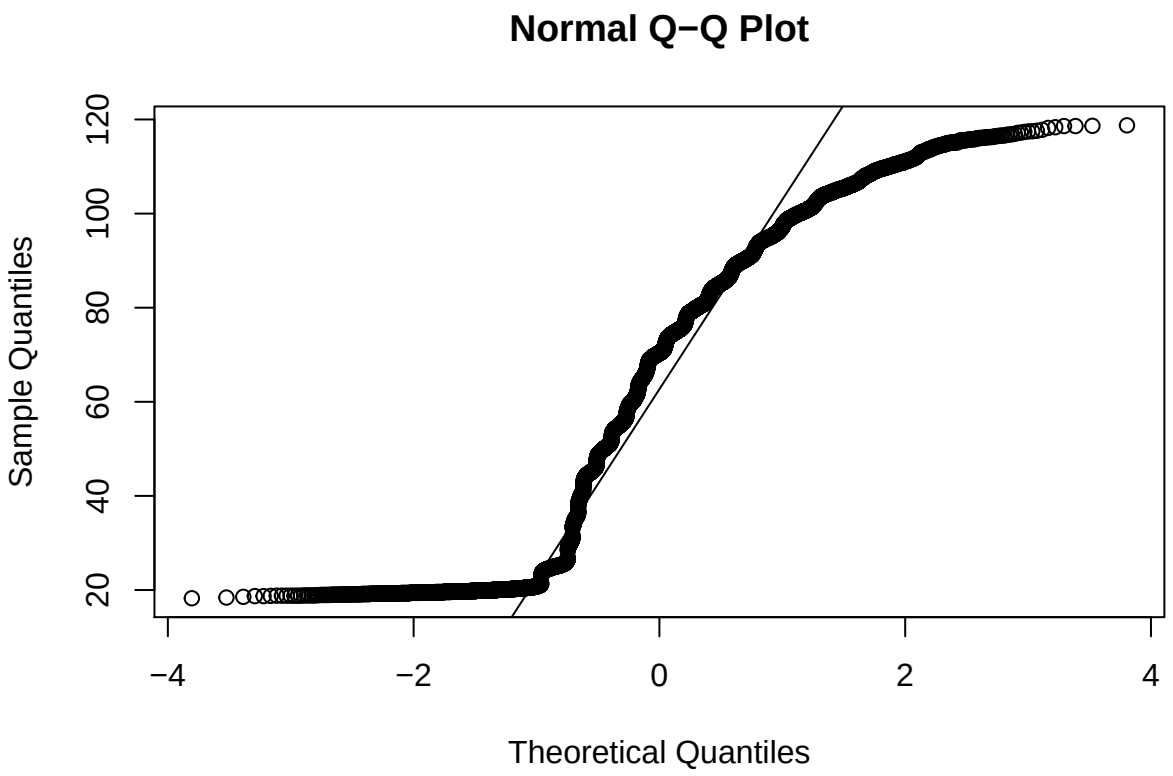
Again, we beleive that customers are churning independently of each other and that our sample size should allow us to wave the normality assumption if it happens to fail.

Let's check for normality through the histogram and qqplots.

```r
hist(telco$MonthlyCharges, main = "Monthly Charges Histogram",
     xlab = "Monthly Charges ($)", ylab = "Count")
```

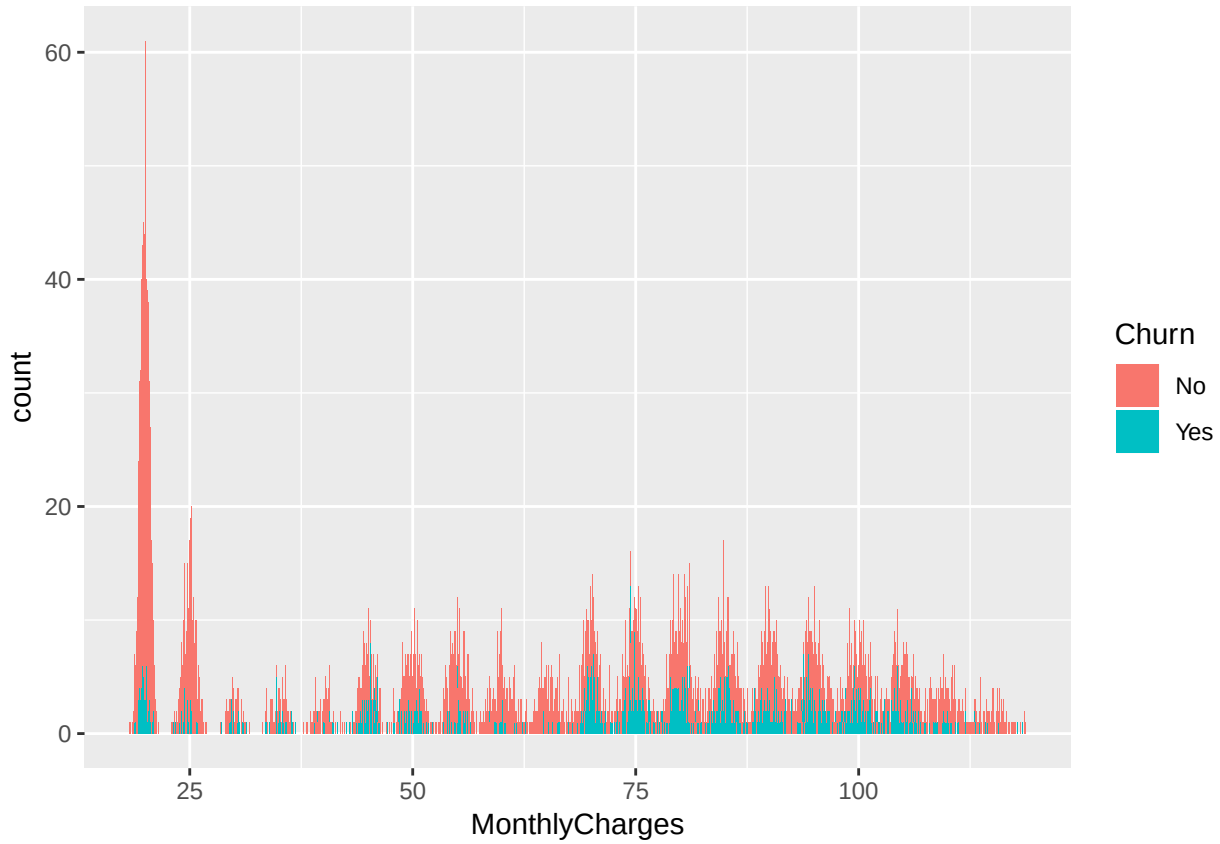



```r
qqnorm(telco$MonthlyCharges)
qqline(telco$MonthlyCharges)
```

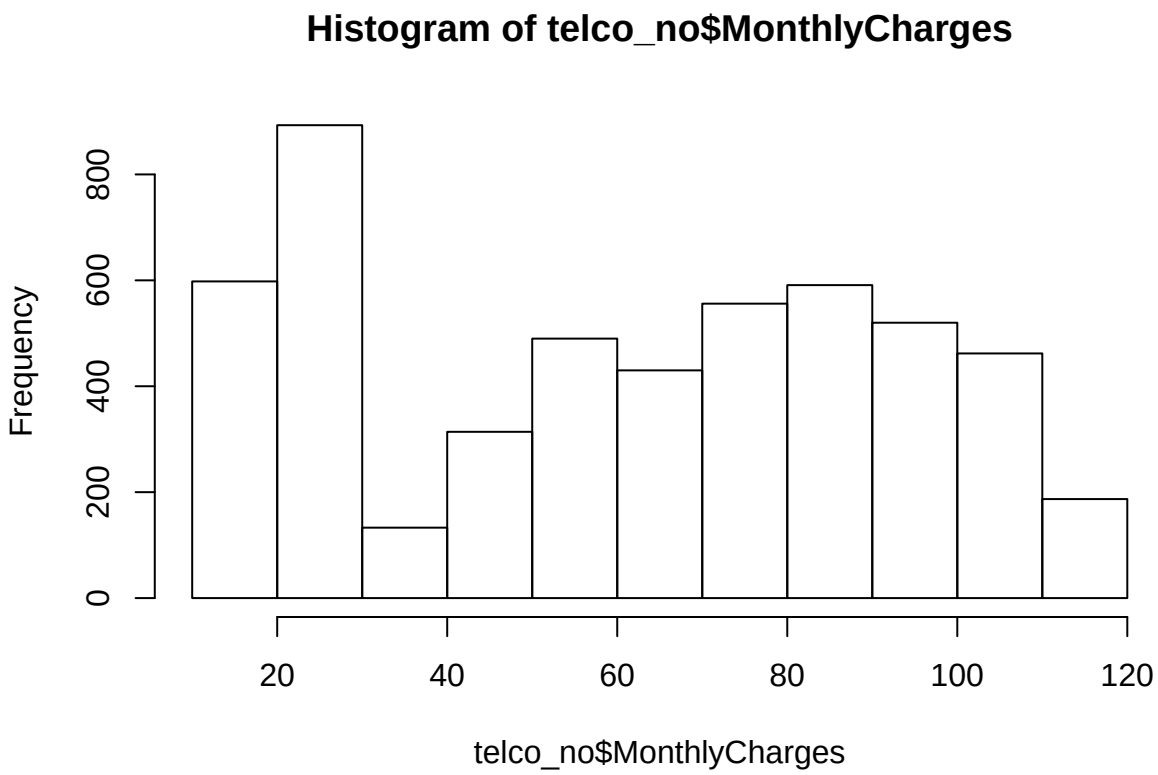


From visual inspection of the histogram and qqplots provided above, we can see that Monthly Charges is not normally distributed. Histograms of the customers monthly charges filtered by churn.

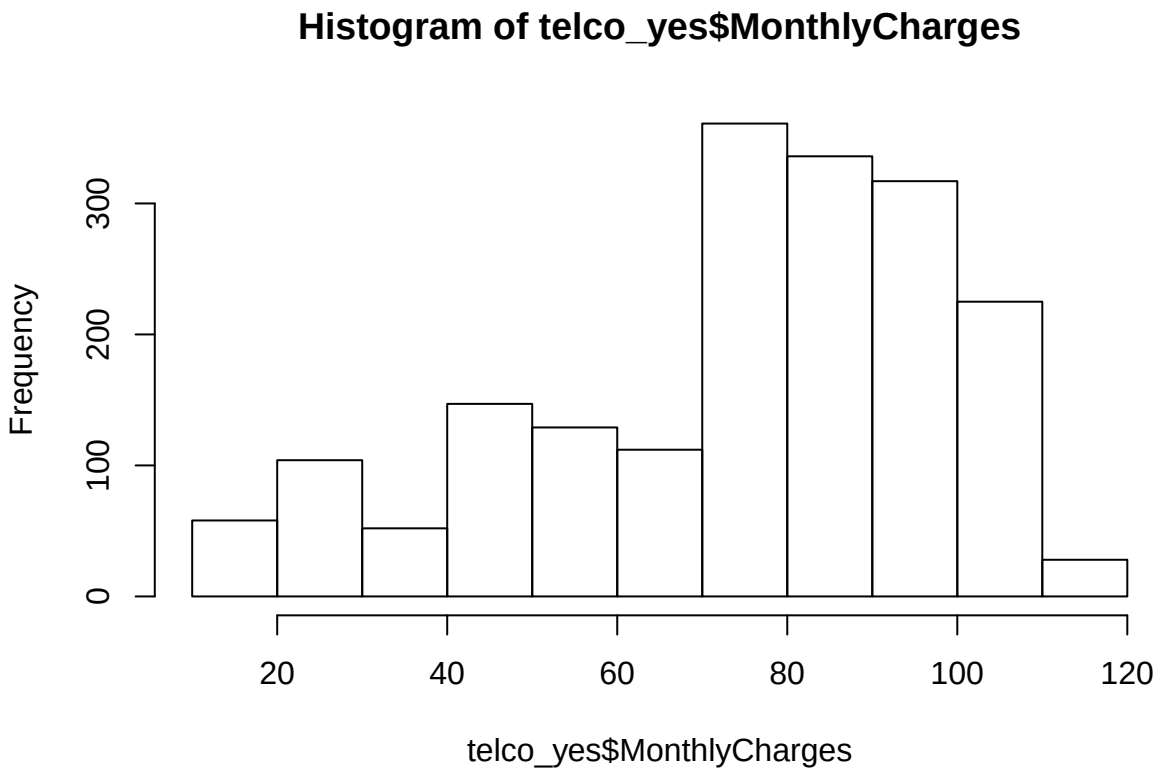
```
ggplot(data=telco)+geom_bar(aes(x=MonthlyCharges, fill=Churn))
```




```
hist(telco_no$MonthlyCharges)
```



```
hist(telco_yes$MonthlyCharges)
```



It's hard to interperet the stacked histogram, but the separate histograms show that most of their customers paid relatively high charges this month. The customers who did not churn has a more even spread of monthly charges, but there appear to be many more customers with relatively low monthly charges in this group.

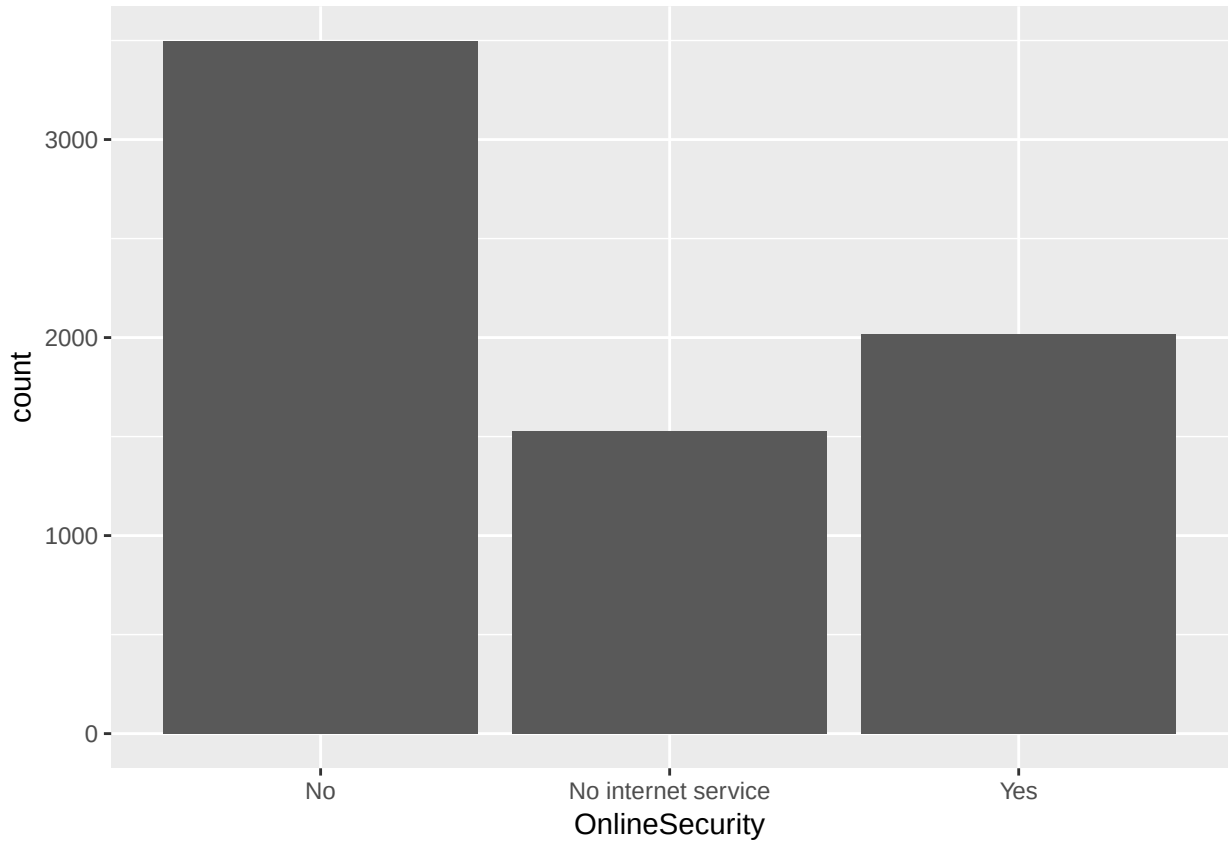
Individual Variables

Now we'll dig through individual variables and explore how they relate to Churn, starting with the variables in our research question.

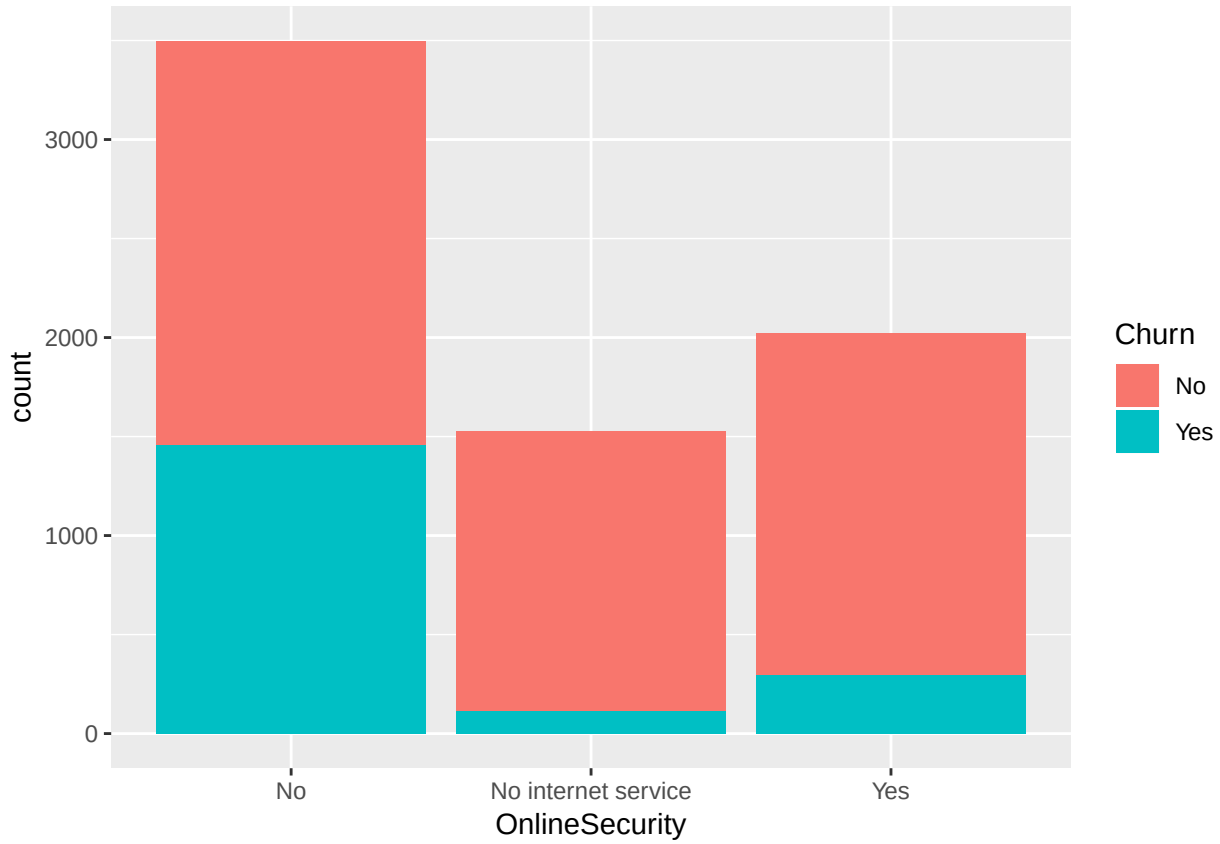
Online Security

Online Security is a categorical variable (3 levels), that indicates whether the customer has subscribed to the company's online security, has no internet service, or hasn't subscribed.

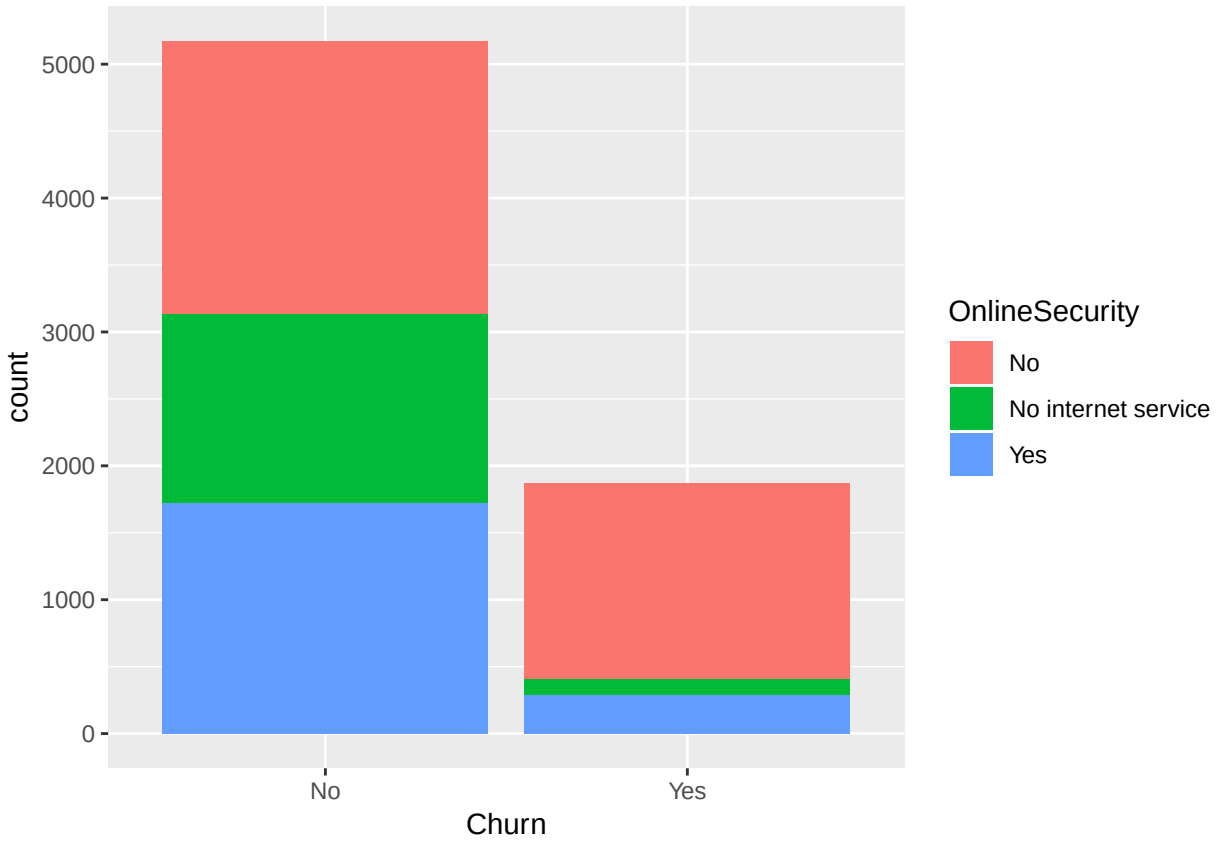
```
ggplot(data=telco)+geom_bar(aes(x=OnlineSecurity))
```




```
ggplot(data=telco)+geom_bar(aes(x=OnlineSecurity, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=OnlineSecurity))
```



Online Security shows an interesting relationship with Churn, as it appears that counter to our initial guess, those customers without internet service are less likely to churn. There visually appears that the majority of churning customers elect not to have online security.

```
count(telco %>% filter(OnlineSecurity == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes",OnlineSecurity !="No internet service"))
```

```
##           n
## 1 0.8320046
```

```
count(telco %>% filter(OnlineSecurity == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

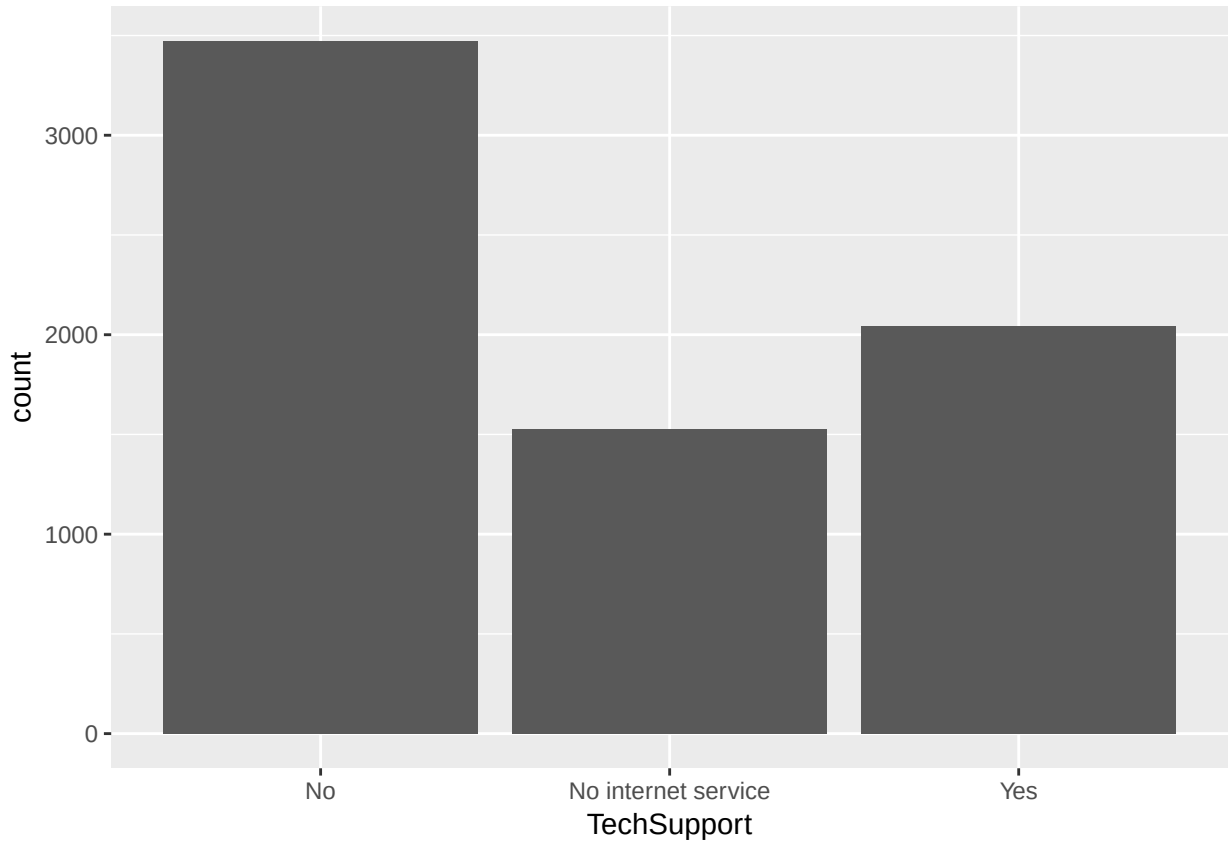
```
##           n
## 1 0.7817014
```

As we can see, ~83% of the customers with internet service that churned did not have OnlineSecurity, (~78% overall.)

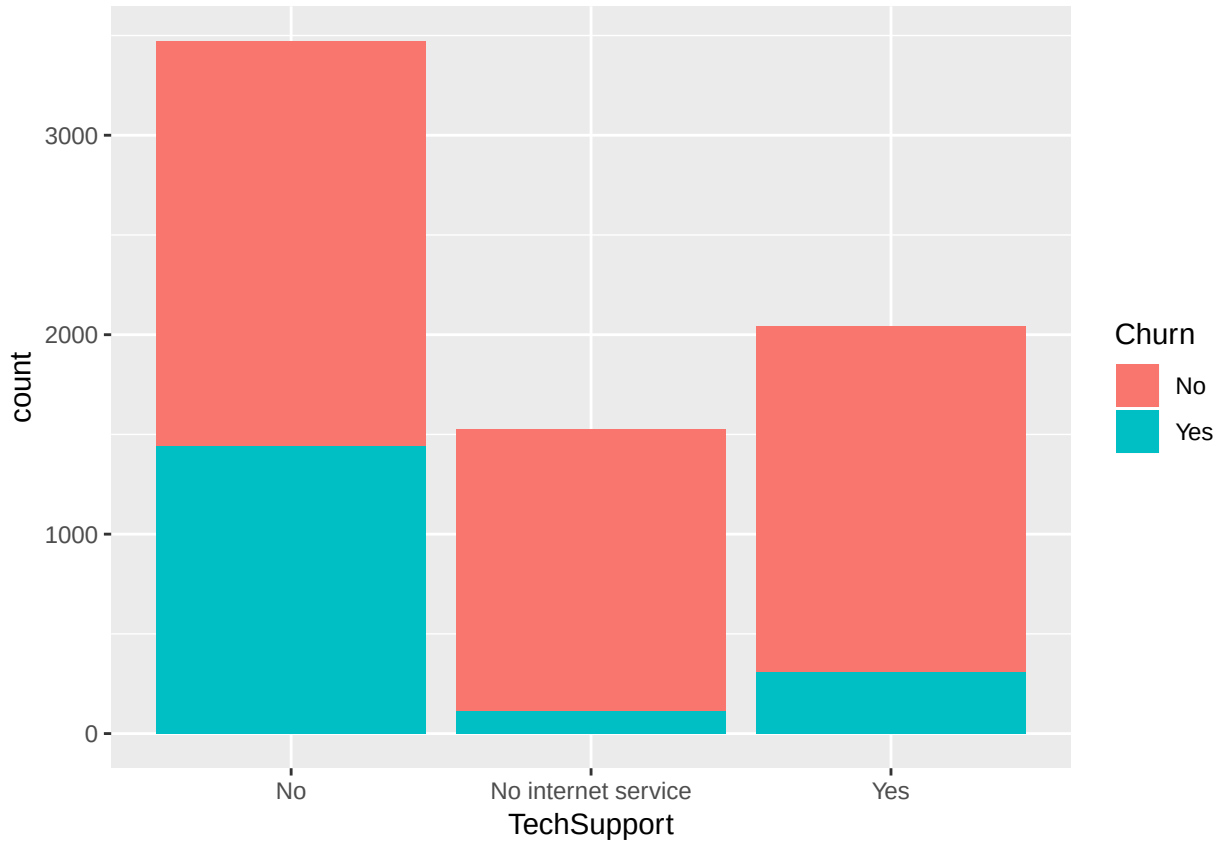
Tech Support

Tech support is a categorical variable (3 levels) that indicates whether a customer has subscribed to the company's technical support service, has no internet access, or has opted out of the technical support service.

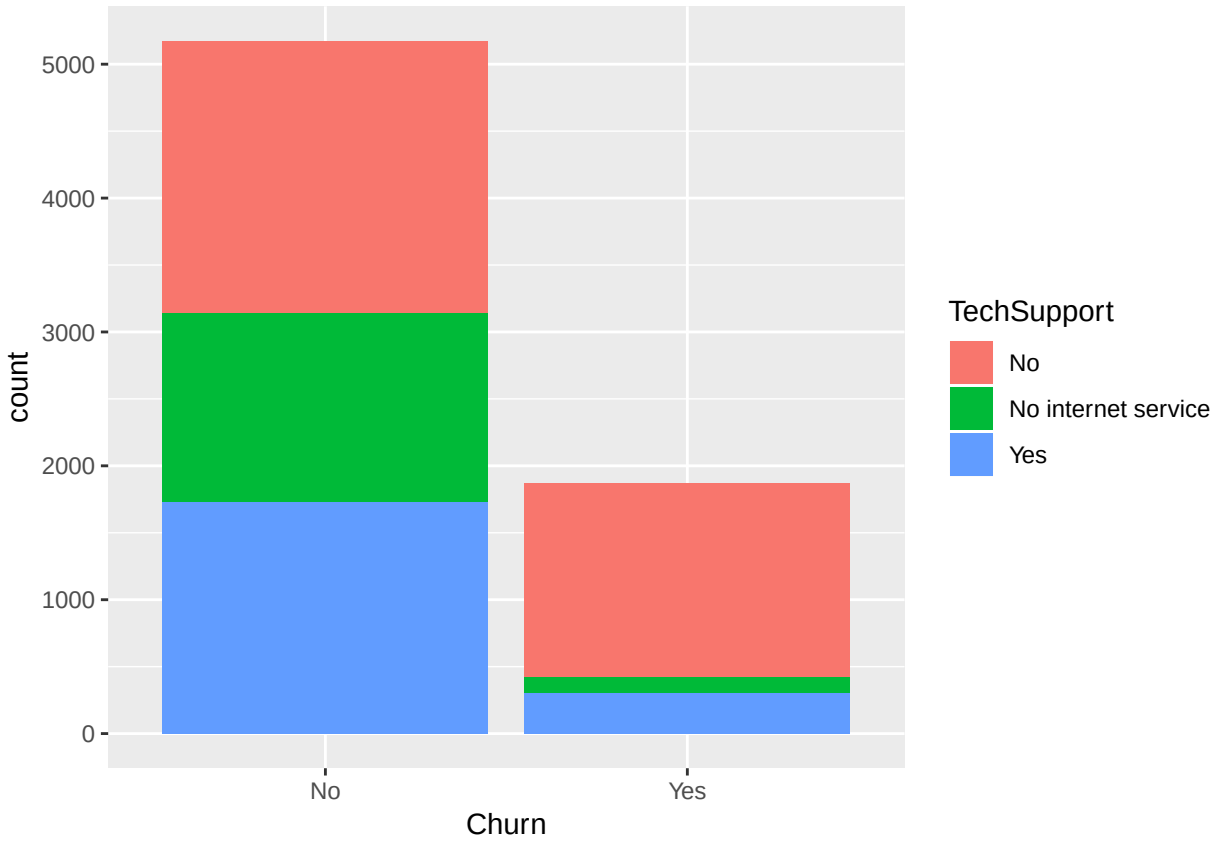
```
ggplot(data=telco)+geom_bar(aes(x=TechSupport))
```




```
ggplot(data=telco)+geom_bar(aes(x=TechSupport, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=TechSupport))
```



The relationship between Tech Support Churn does appear to visually be significant.. in that the majority of customers that churned appear to opted out of the company's Tech Support services.

```
count(telco %>% filter(TechSupport == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes", TechSupport != "No internet service"))
```

```
##           n
## 1 0.8234624
```

```
count(telco %>% filter(TechSupport == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

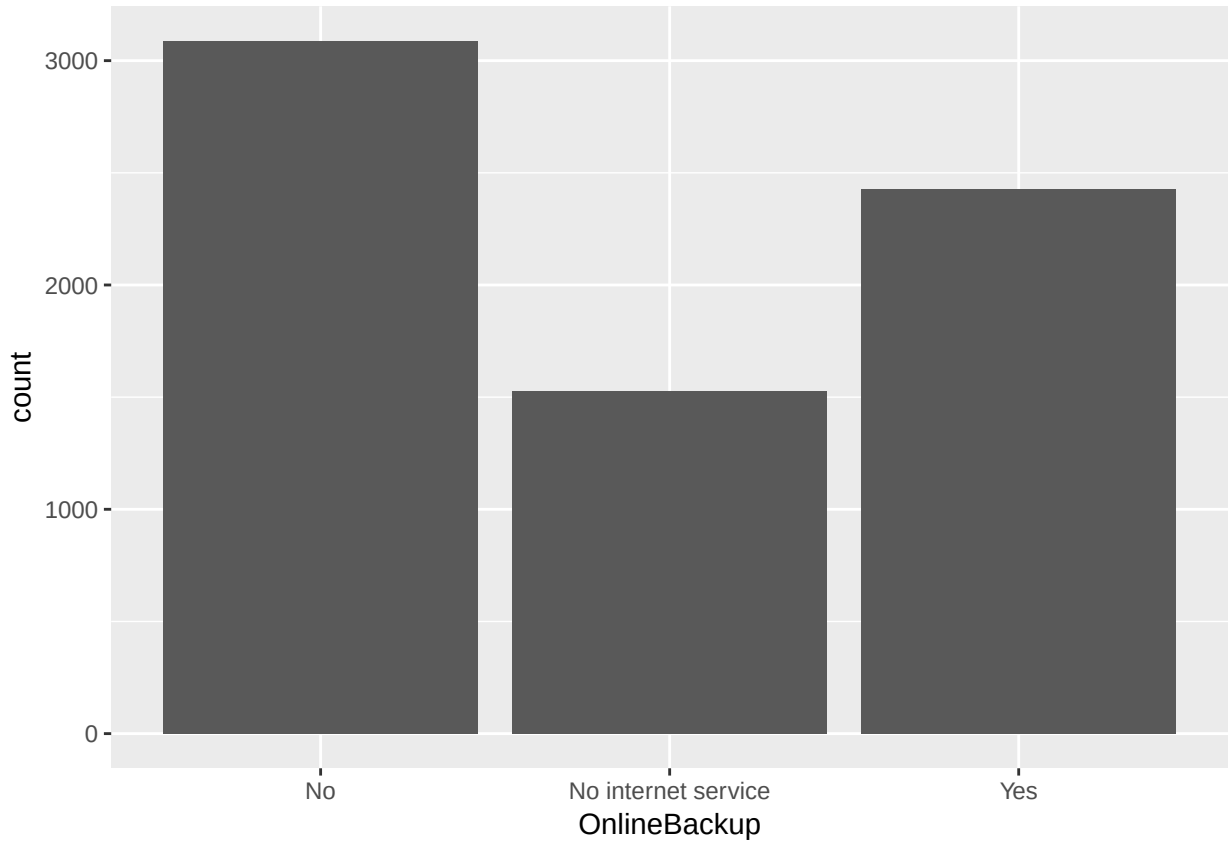
```
##           n
## 1 0.7736758
```

As we can see ~82% of the customers with internet service that churned this month, had opted out of the company's tech support services (~77% overall.)

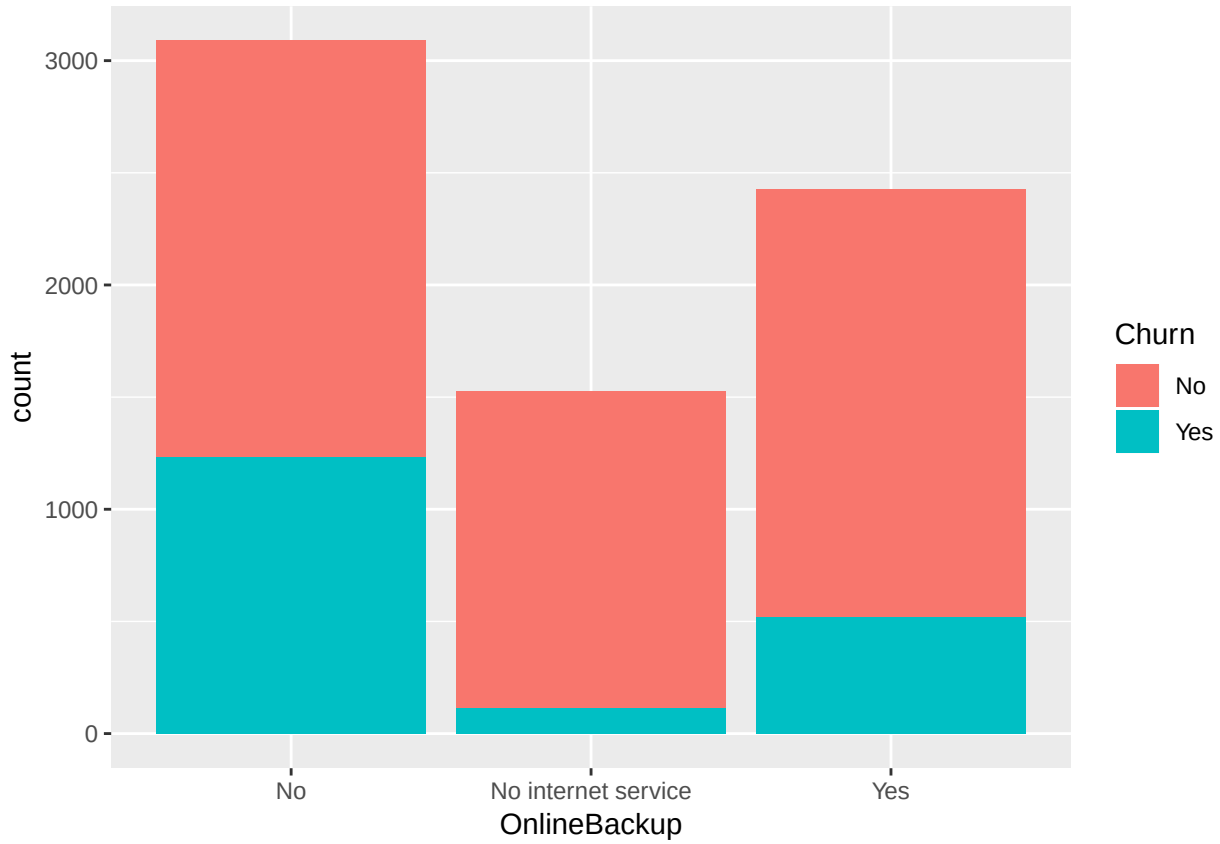
Online Backup

Online Backup is a categorical variable (3 levels), indicating whether a customer has subscribed to the company's online backup services, has no internet service, or has opted out.

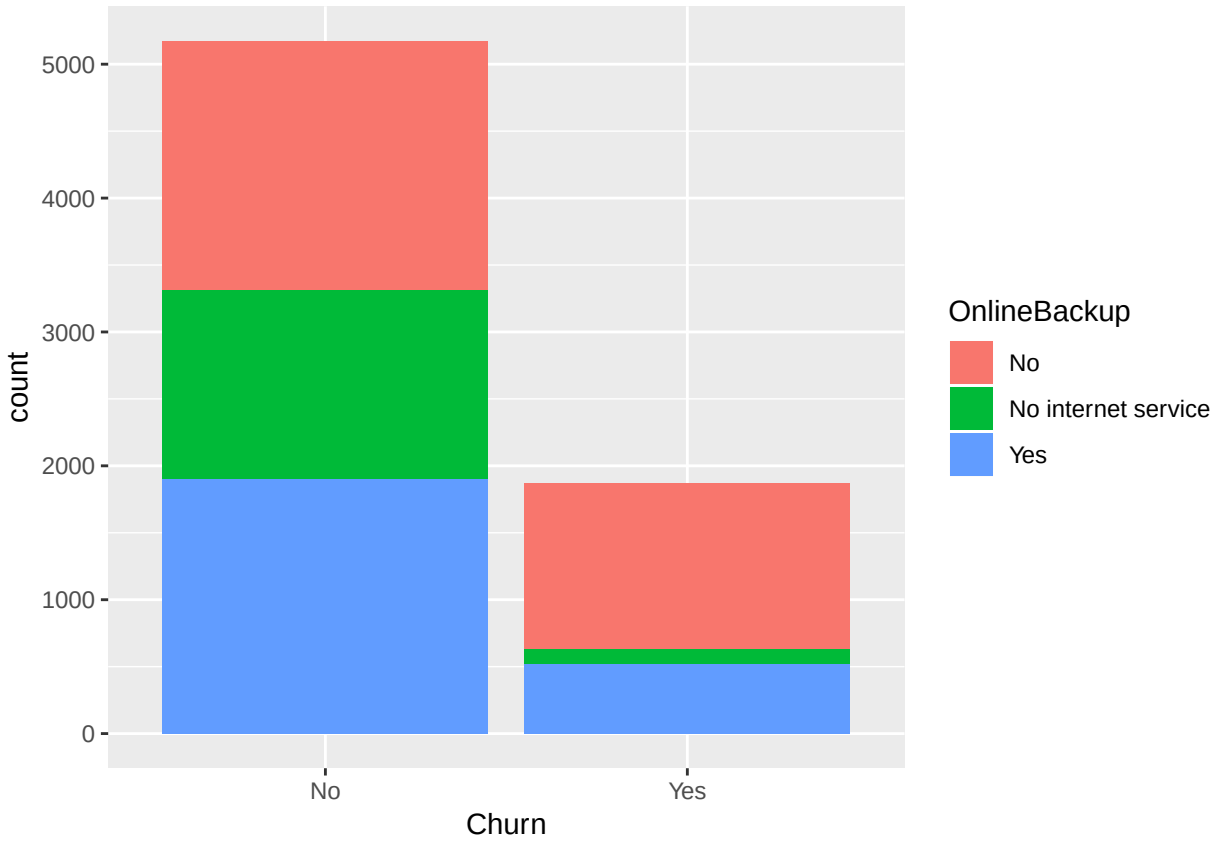
```
ggplot(data=telco)+geom_bar(aes(x=OnlineBackup))
```




```
ggplot(data=telco)+geom_bar(aes(x=OnlineBackup, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=OnlineBackup))
```



We see here a similar trend to that found with Tech Support and Online Security related to Churn, in that visually it appears that the majority of customers who churned in the previous month had opted out of the service.

```
count(telco %>% filter(OnlineBackup == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes", OnlineBackup != "No internet service"))
```

```
##          n
## 1 0.702164
```

```
count(telco %>% filter(OnlineBackup == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

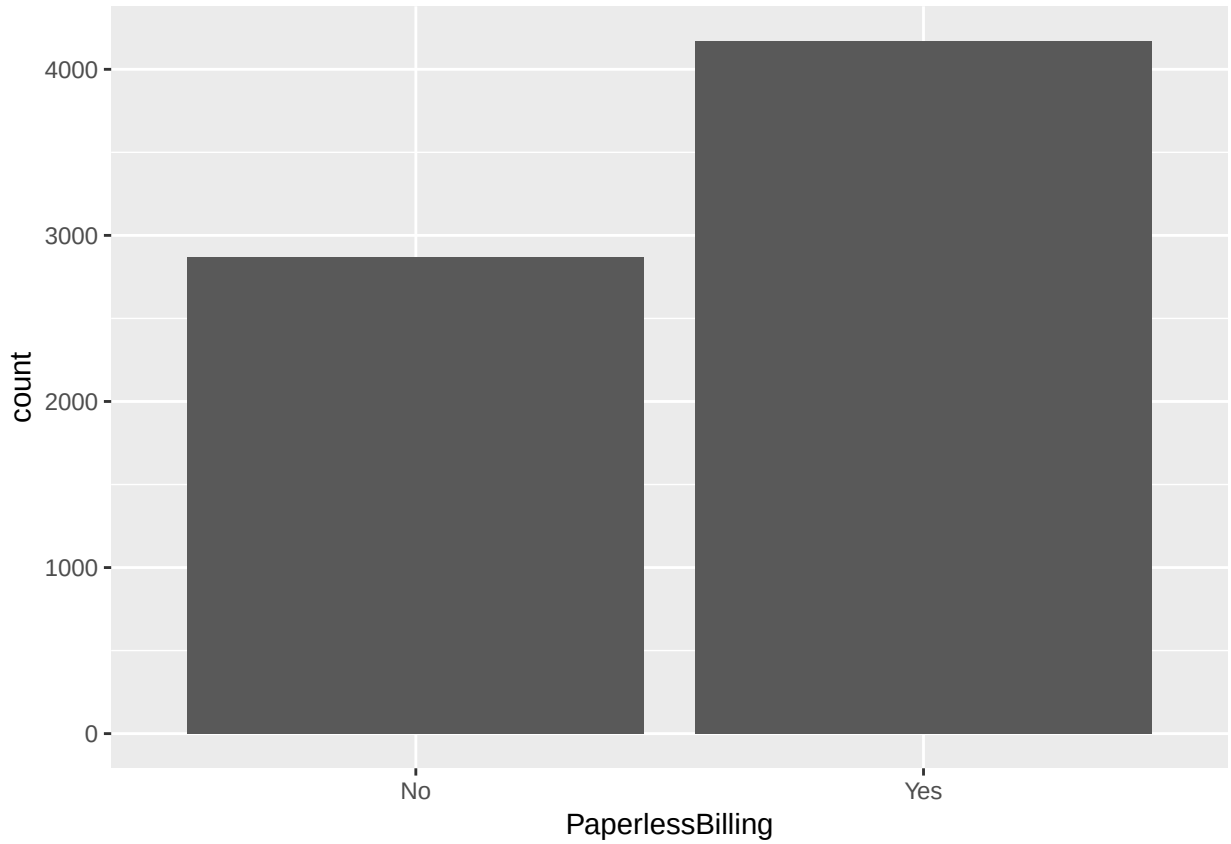
```
##           n
## 1 0.6597111
```

As we can see ~70% of the customers with internet service that churned this month, had opted out of the company's online backup services (~66% overall.)

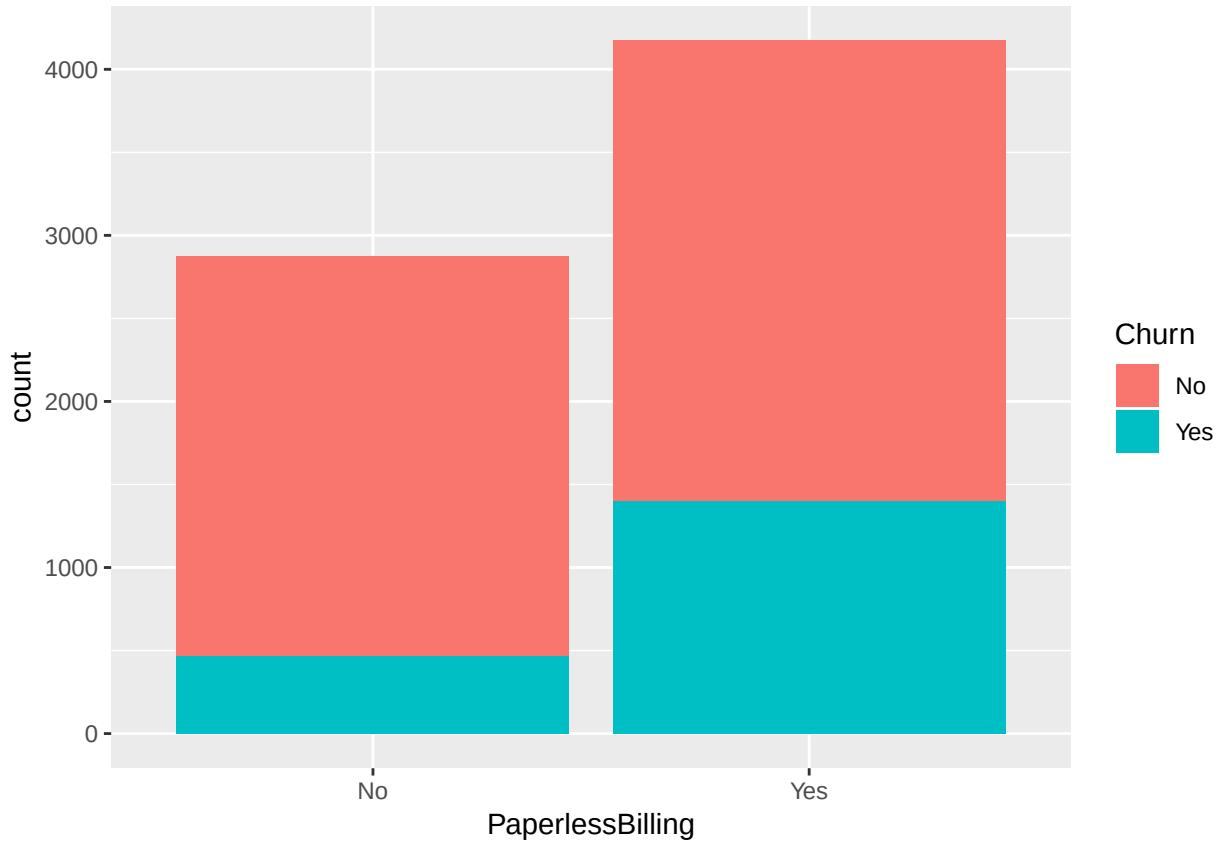
Paperless Billing

Paperless Billing is a categorical variable (2 levels), indicating whether a customer is using Paperless Billing, or not.

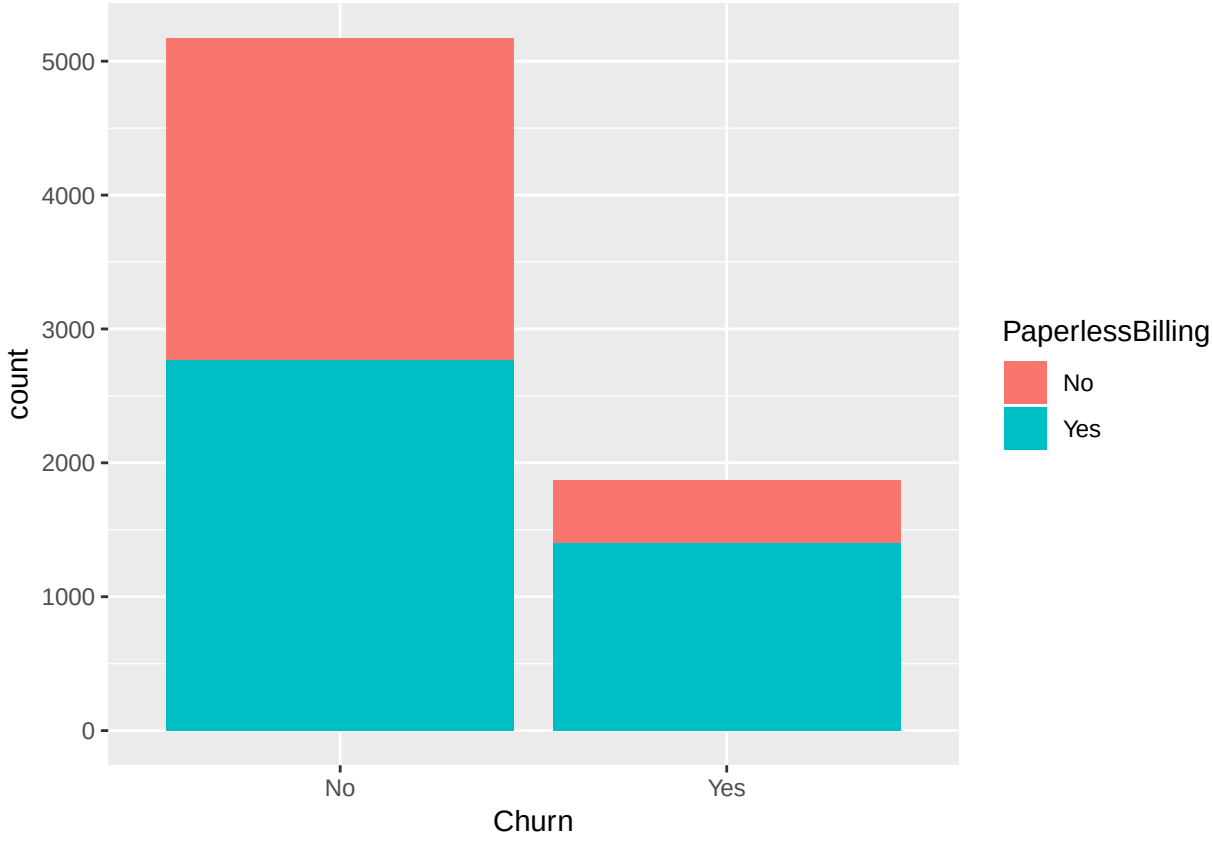
```
ggplot(data=telco)+geom_bar(aes(x=PaperlessBilling))
```




```
ggplot(data=telco)+geom_bar(aes(x=PaperlessBilling, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=PaperlessBilling))
```



There is still a trend here in that customers with paperless billing appear to be more likely to churn.

```
count(telco %>% filter(PaperlessBilling == "Yes", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

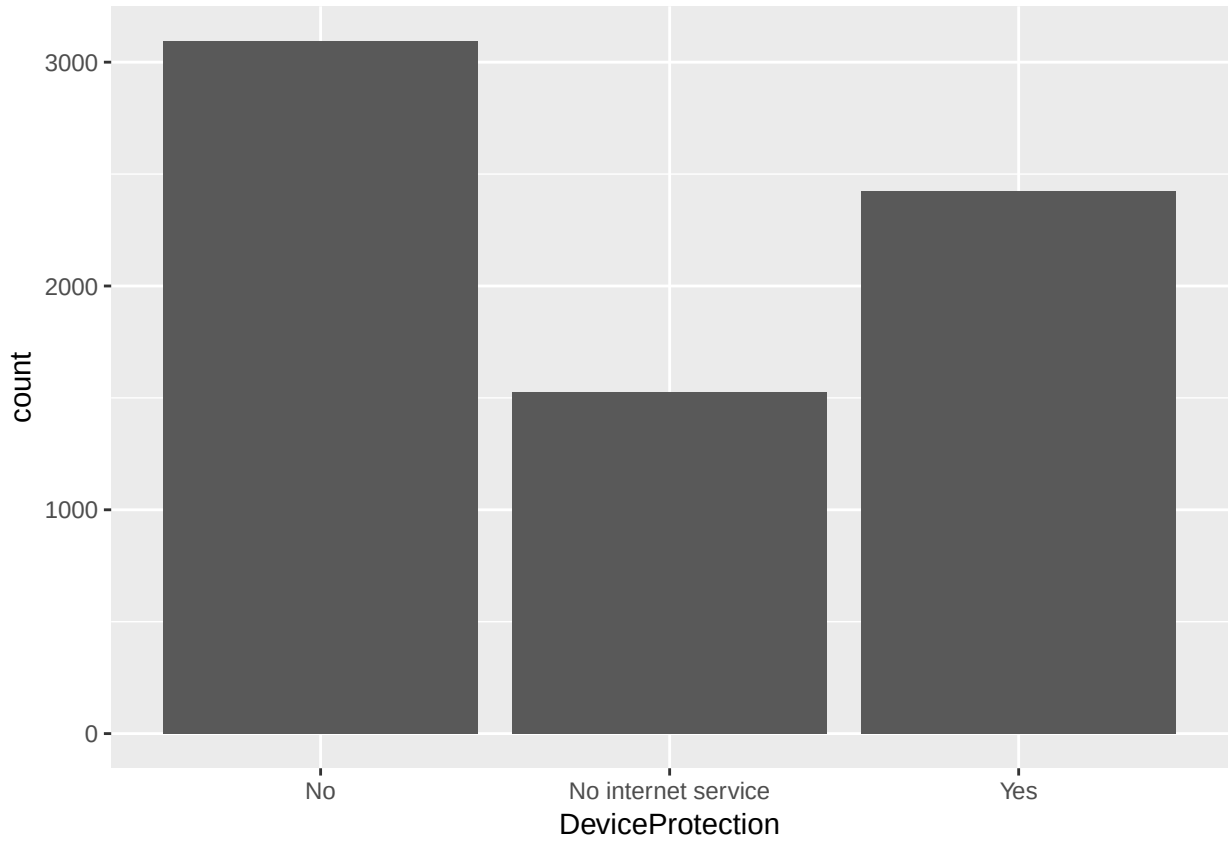
```
##           n
## 1 0.7490637
```

As we can see ~75% of the customers that churned this month, had opted into of the company's paperless billing option.

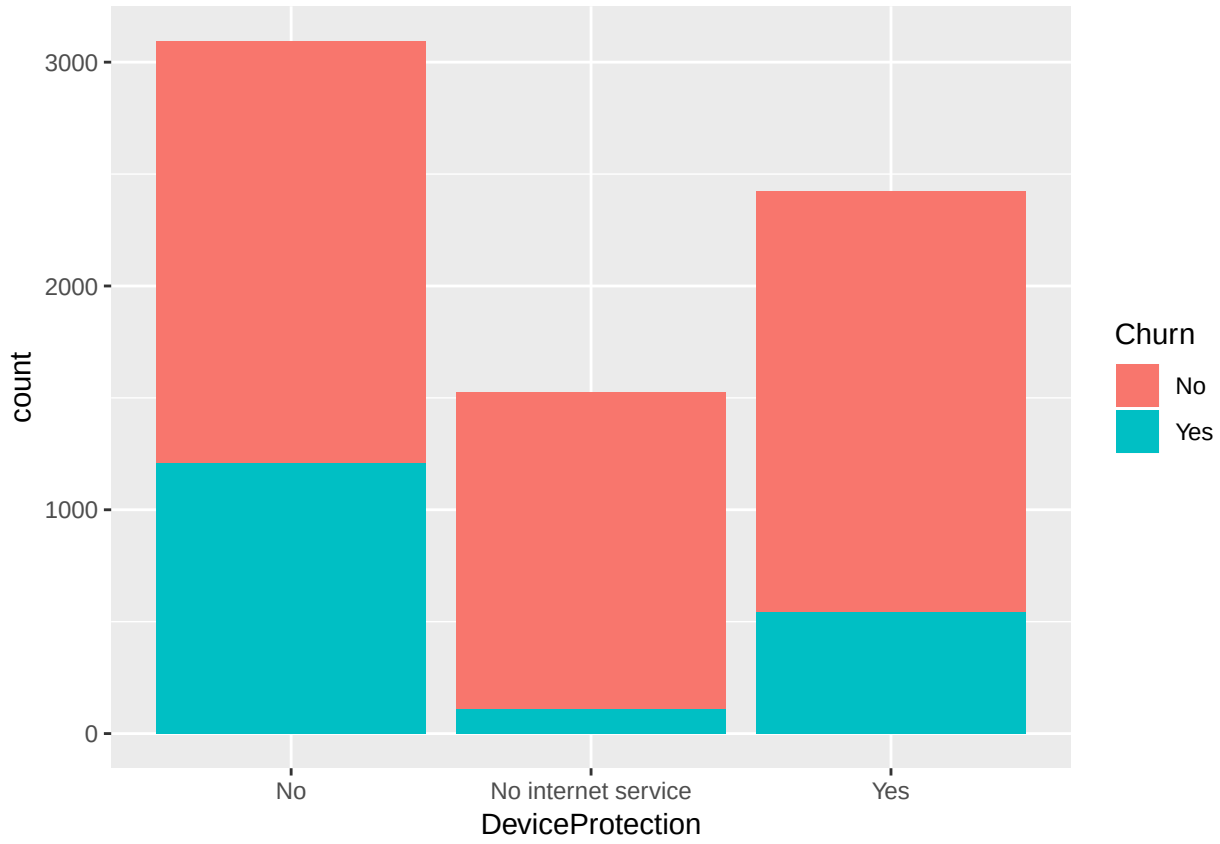
Device Protection

Device Protection is a categorical variable (3 levels), indicating whether a customer has opted into the company's Device Protection plans, has no internet service, or has opted out of the device protection plan.

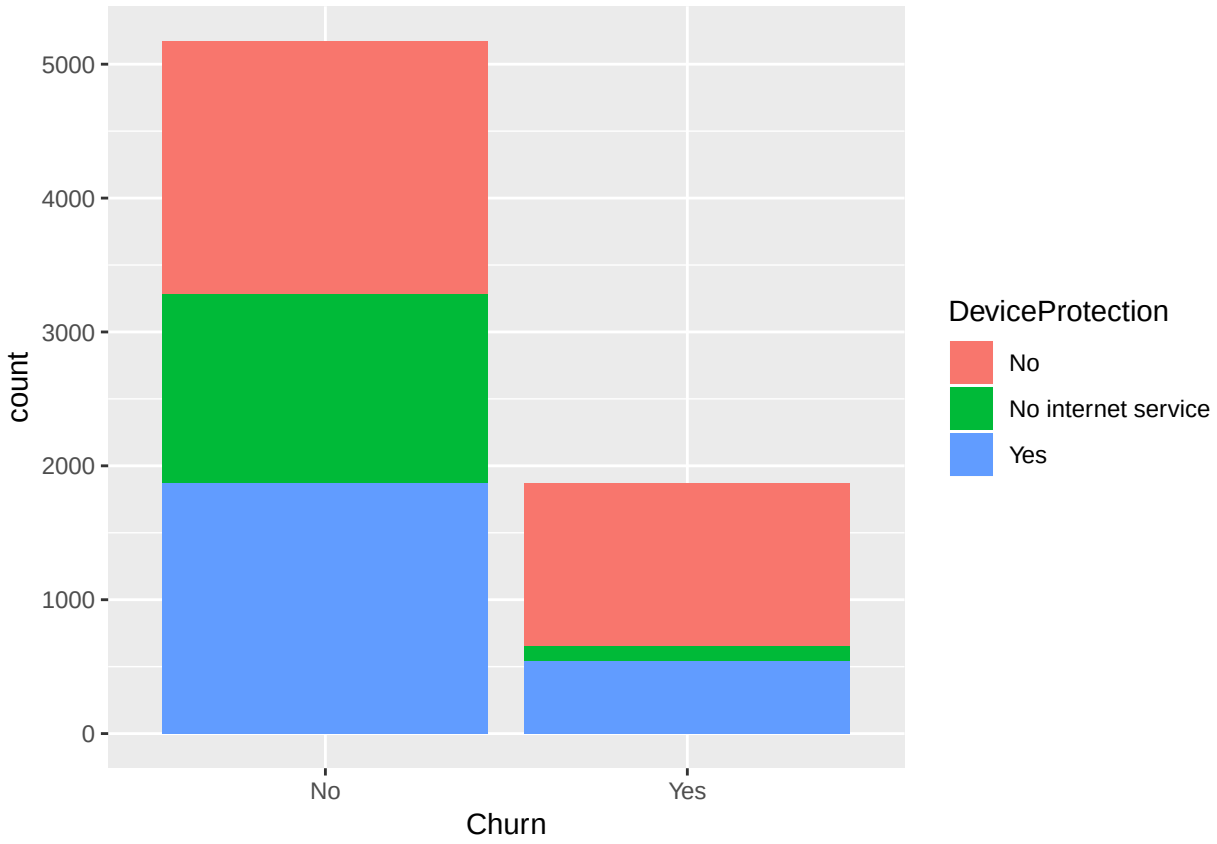
```
ggplot(data=telco)+geom_bar(aes(x=DeviceProtection))
```




```
ggplot(data=telco)+geom_bar(aes(x=DeviceProtection, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=DeviceProtection))
```



Visually we can tell that most of the customer's that are churning have opted out of the company's device protection plan.

```
count(telco %>% filter(DeviceProtection == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes", DeviceProtection != "No internet service"))
```

```
##           n
## 1 0.6896355
```

```
count(telco %>% filter(DeviceProtection == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

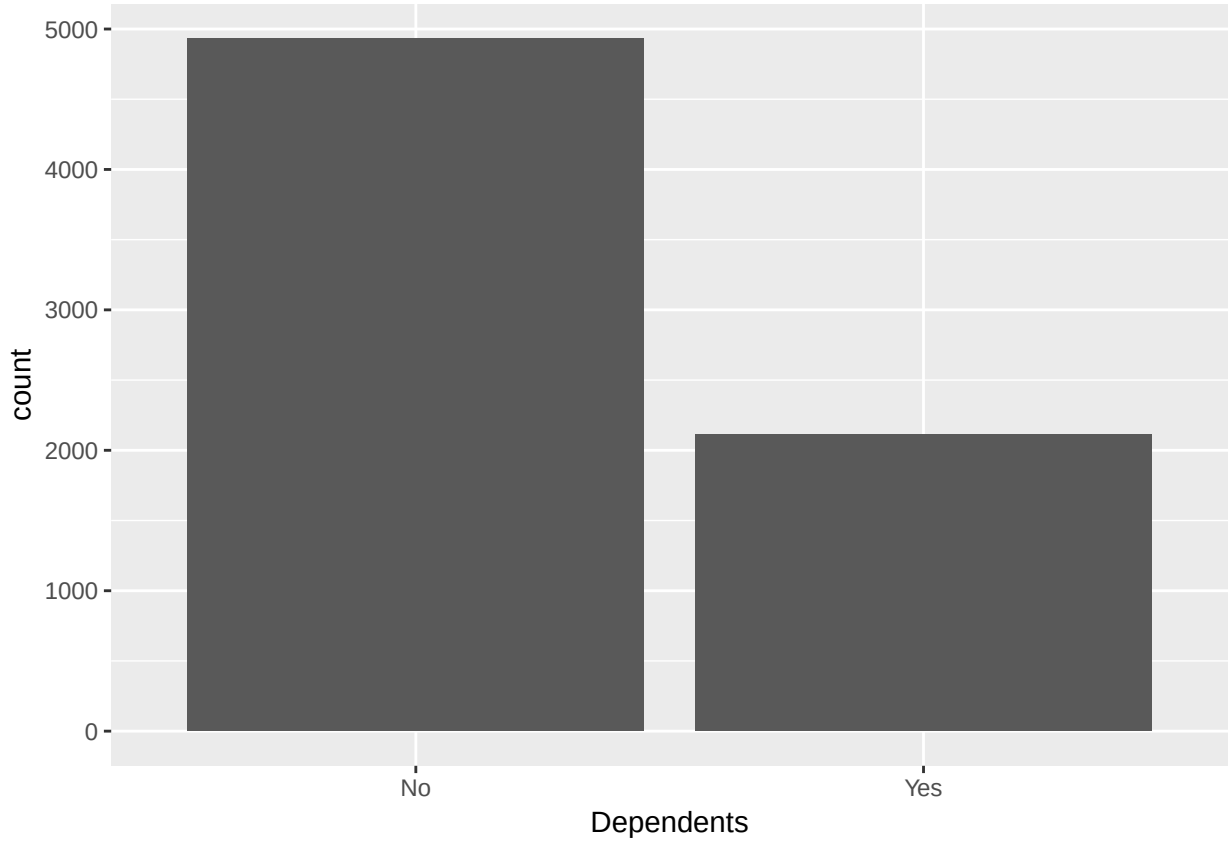
```
##           n
## 1 0.6479401
```

As we can see ~69% of the customers with internet service that churned this month, had opted out of the company's device protection plan (~65% overall.)

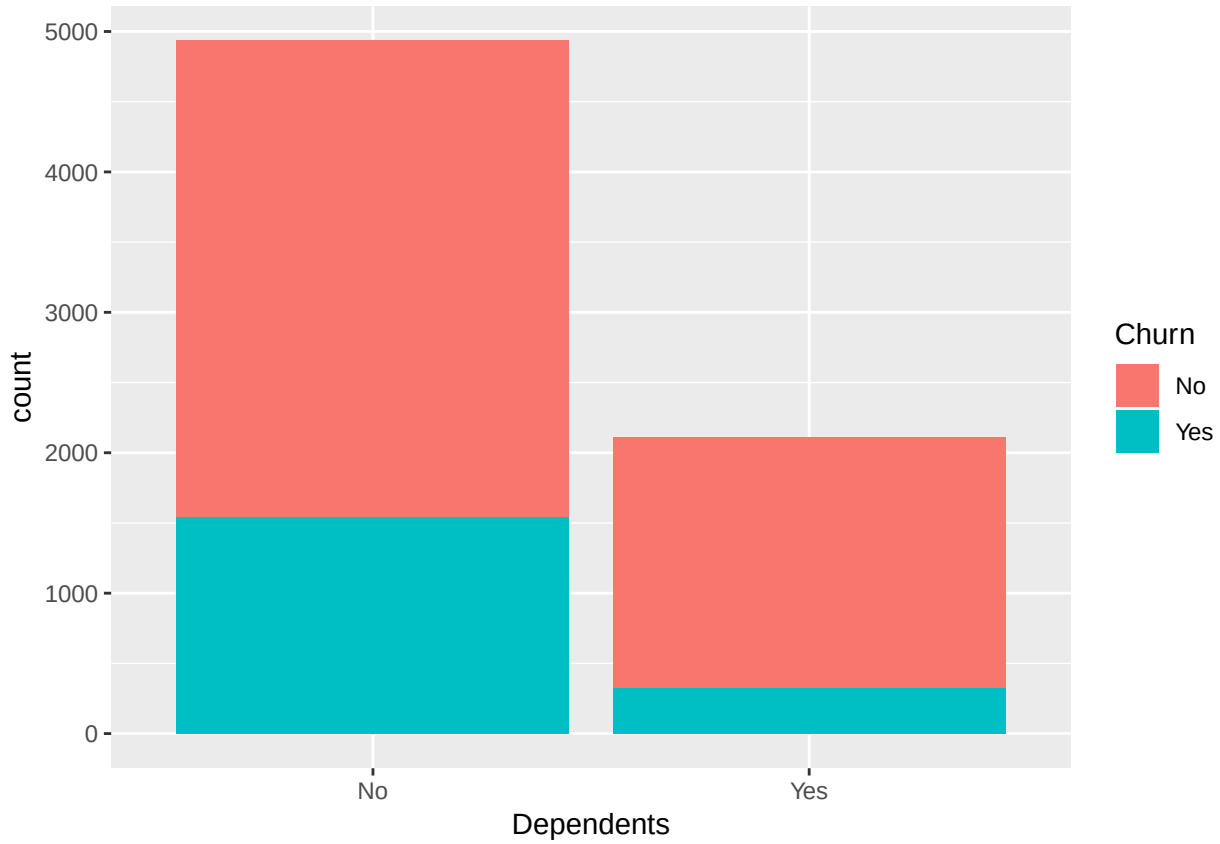
Dependents

Dependents is a categorical variable (2 levels), indicating whether a customer has dependents or not.

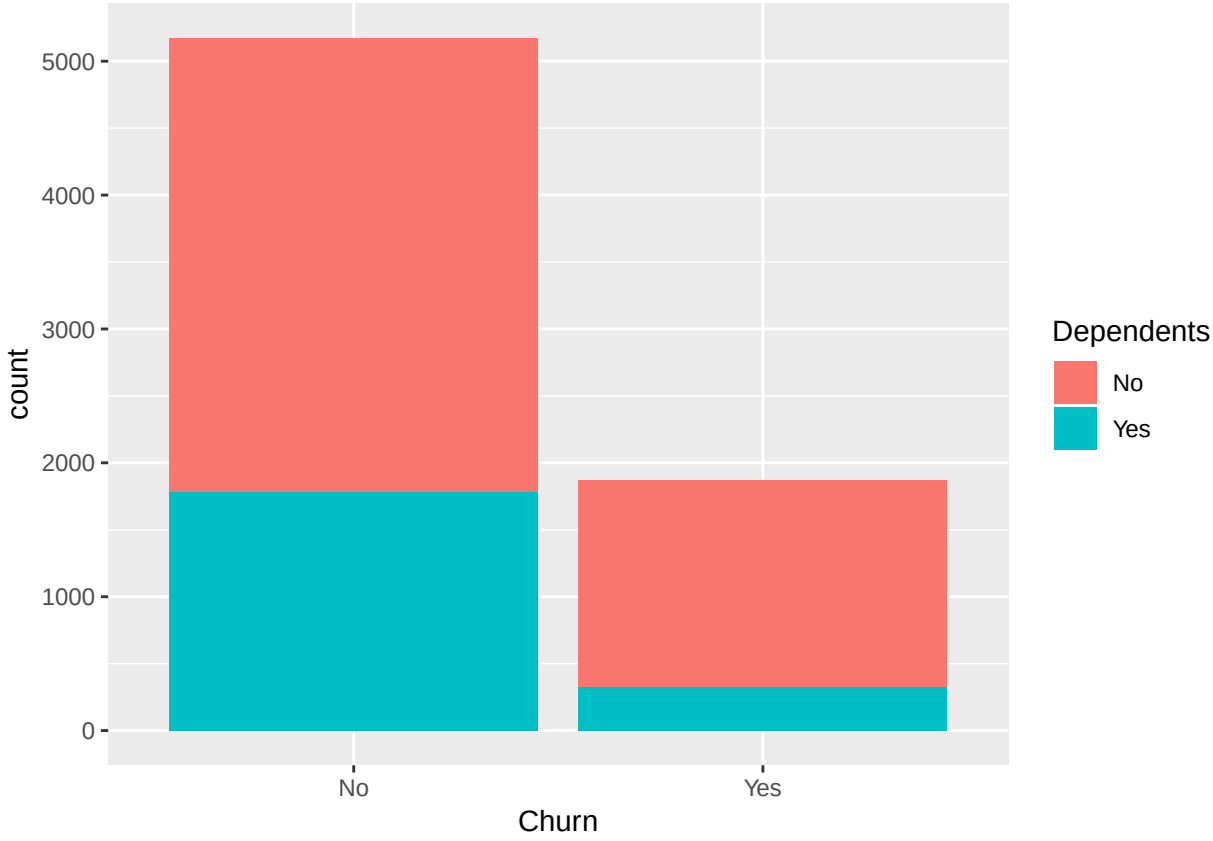
```
ggplot(data=telco)+geom_bar(aes(x=Dependents))
```




```
ggplot(data=telco)+geom_bar(aes(x=Dependents, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=Dependents))
```



Visually, we can see that the majority of customers that churned have no dependents.

```
count(telco %>% filter(Dependents == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

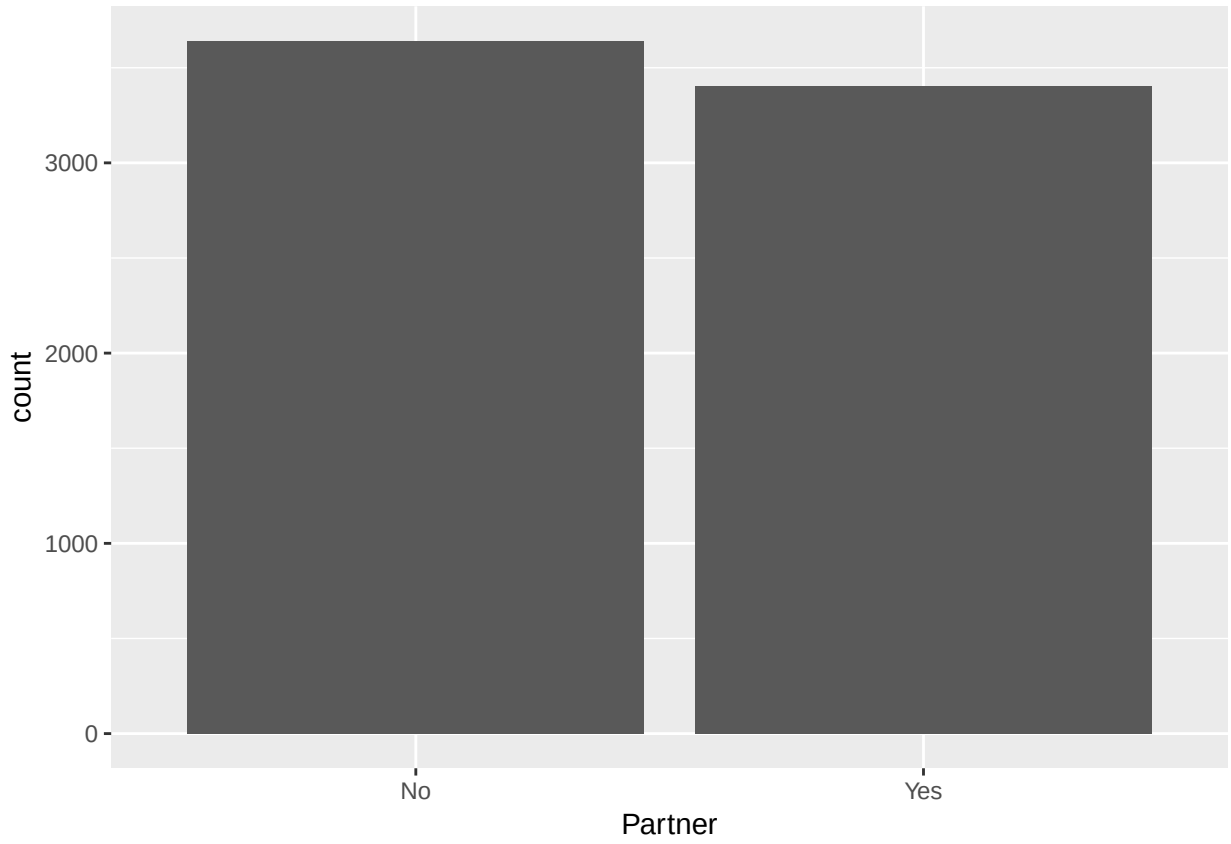
```
##           n
## 1 0.8255752
```

As we can see ~82% of the customers that churned this month, have no dependents.

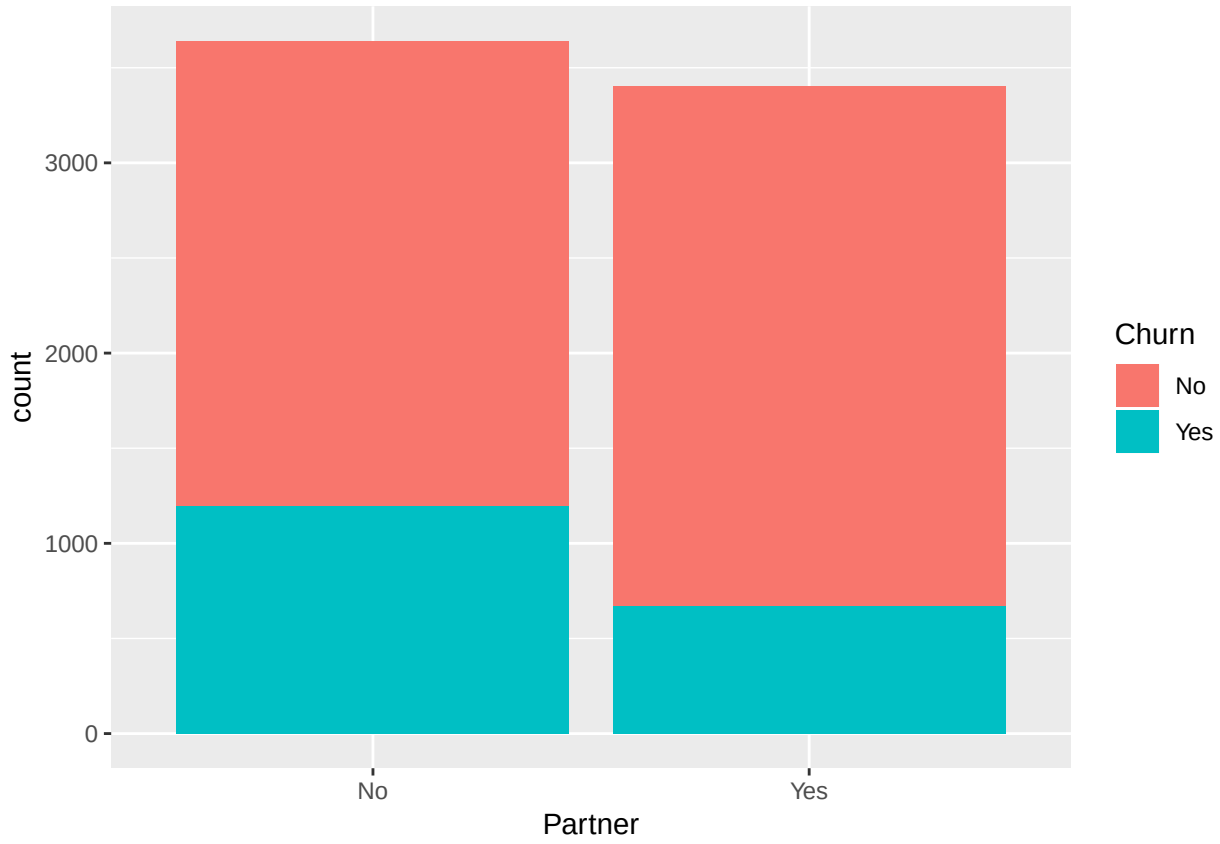
Partner

Partner is a categorical variable (2 levels), indicating whether a customer has a partner or not.

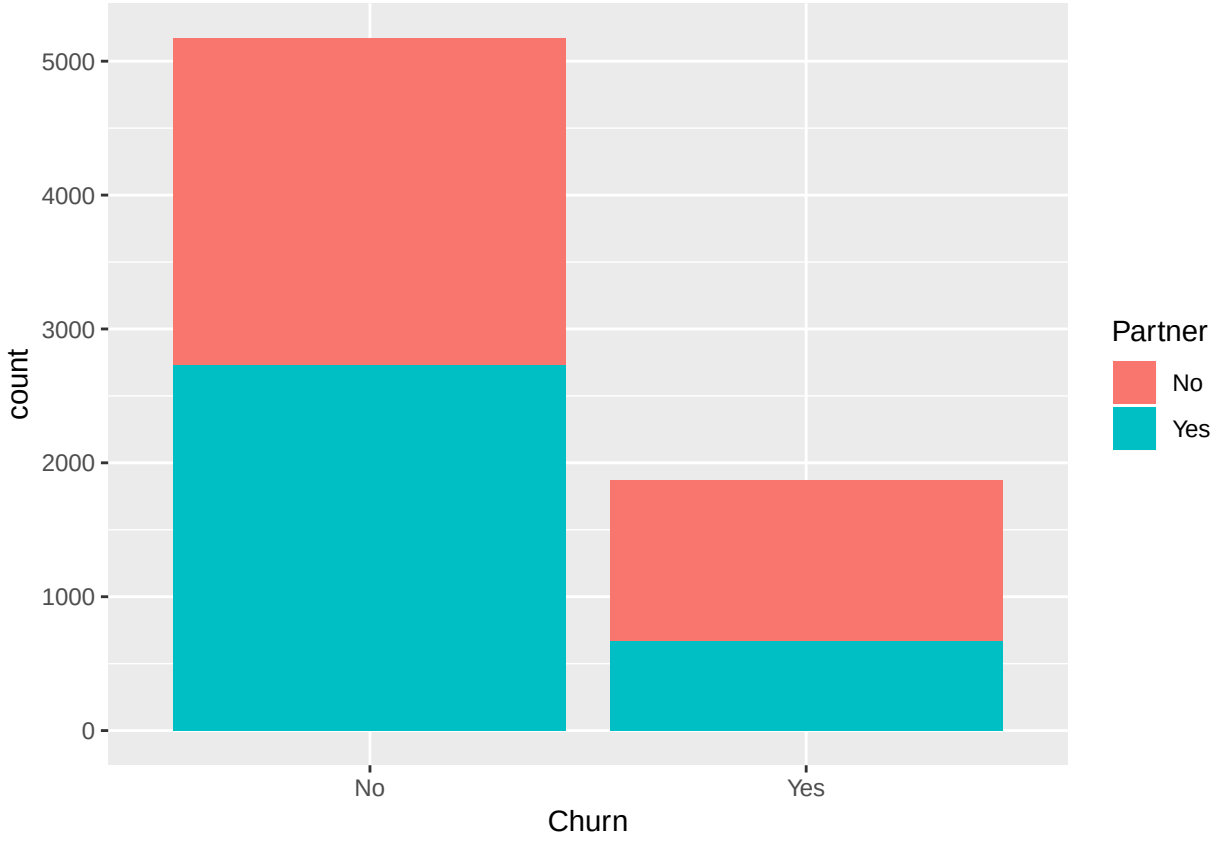
```
ggplot(data=telco)+geom_bar(aes(x=Partner))
```




```
ggplot(data=telco)+geom_bar(aes(x=Partner, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=Partner))
```




```
count(telco %>% filter(Partner == "Yes"))/count(telco)
```

```
##           n
## 1 0.4830328
```

Interestingly ~48% of our sample customers have Partners, which appears to be the most evently distributed categorical variable we've seen yet. We continue to see a strong relationship between the variable and Churn.

```
count(telco %>% filter(Partner == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

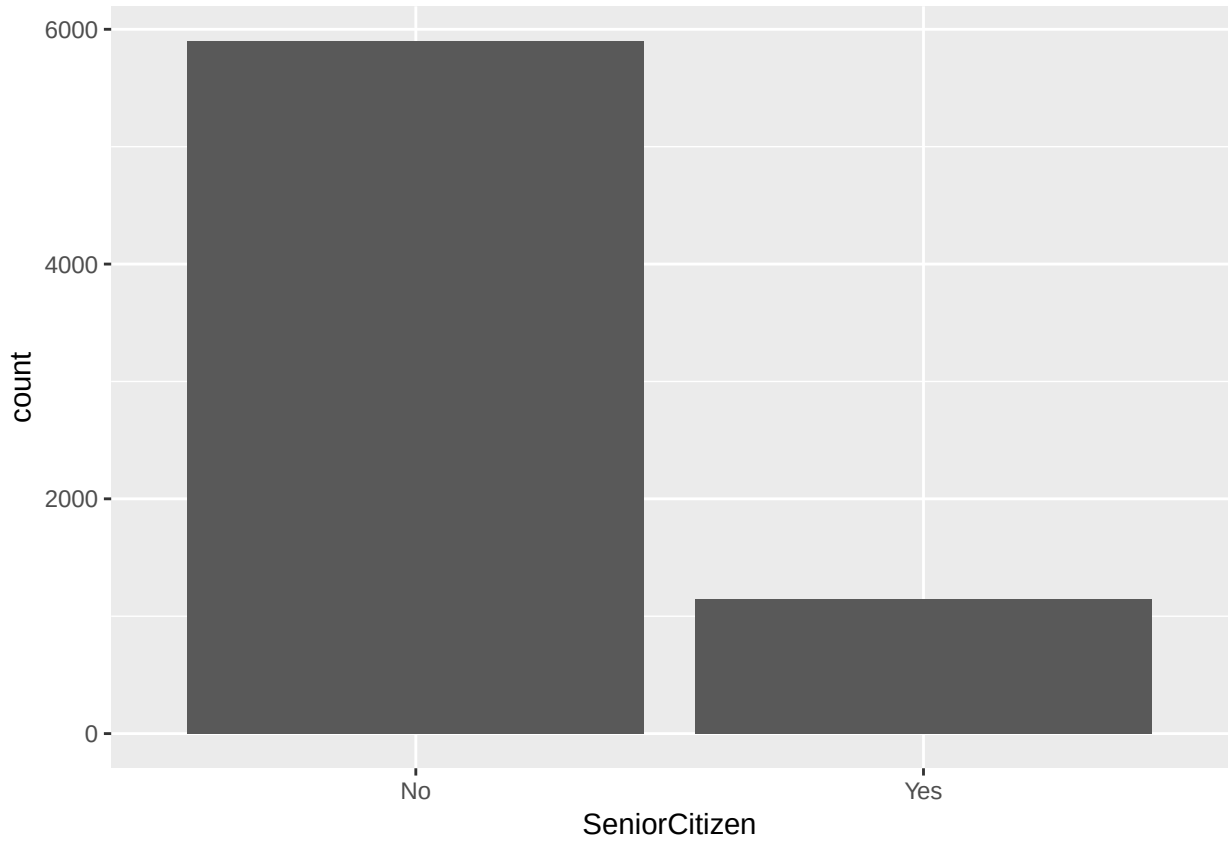
```
##           n
## 1 0.6420546
```

As we can see ~64% of the customers that churned this month were single.

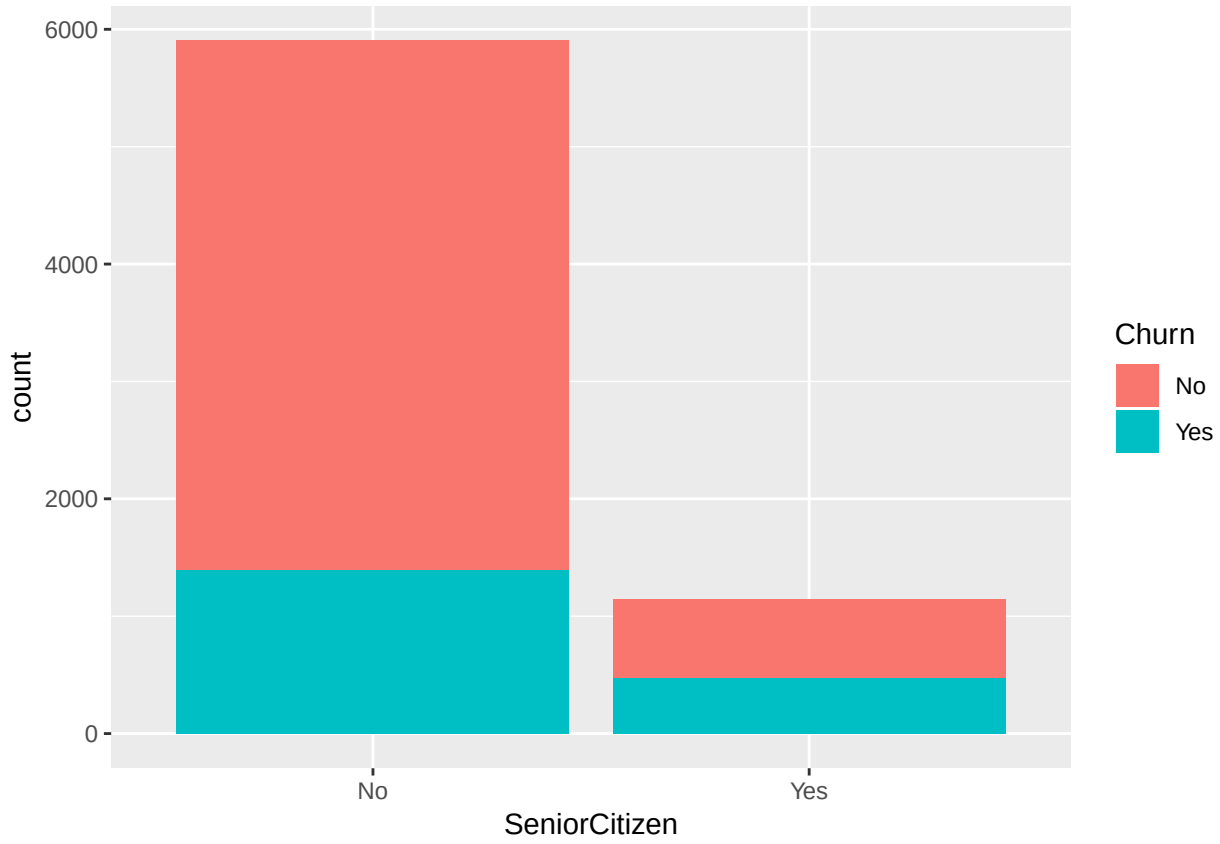
Senior Citizen

Senior Citizen is a categorical variable (2 levels), indicating whether a customer is a senior citizen or not.

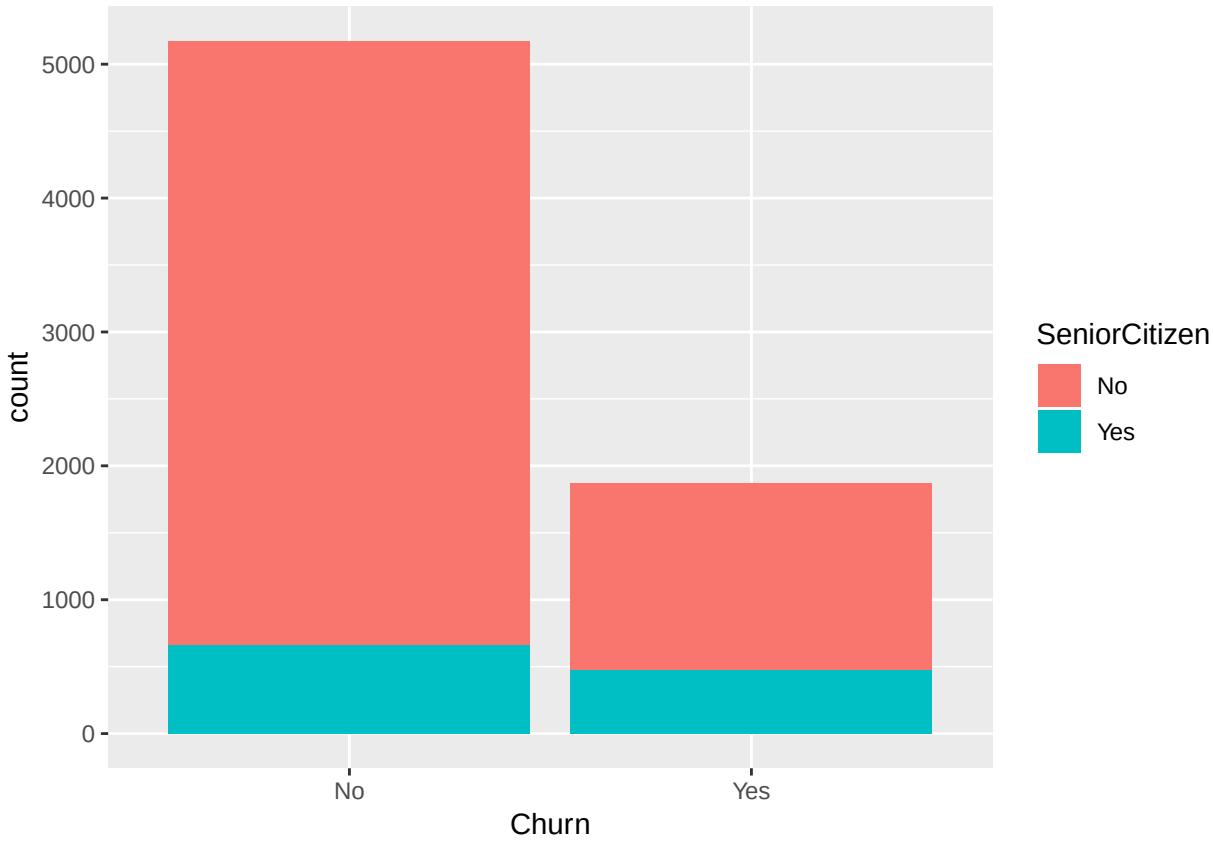
```
ggplot(data=telco)+geom_bar(aes(x=SeniorCitizen))
```




```
ggplot(data=telco)+geom_bar(aes(x=SeniorCitizen, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=SeniorCitizen))
```



There does appear to be a relationship between SeniorCitizen and Churn, though it isn't as pronounced as the previous visual trends were.

```
count(telco %>% filter(SeniorCitizen == "No", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

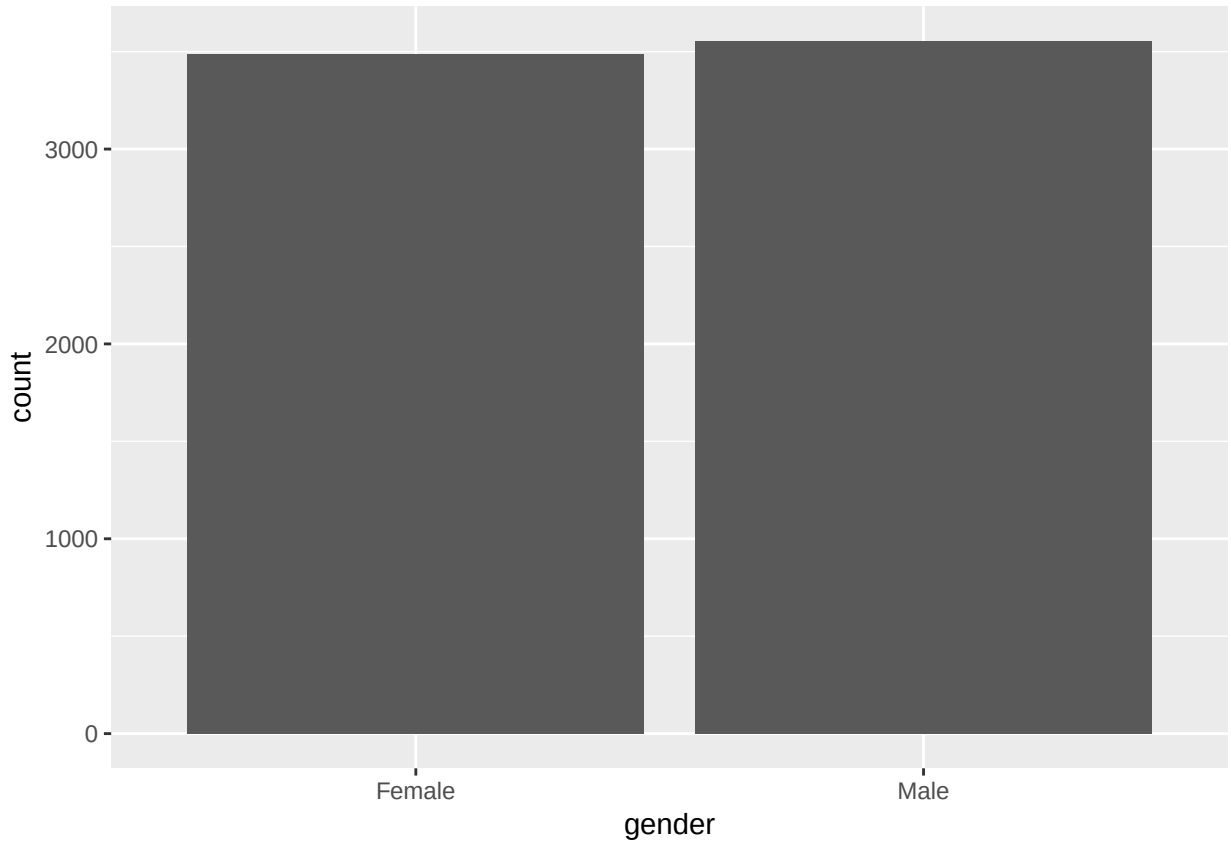
```
##           n
## 1 0.7453184
```

As we can see ~75% of the customers that churned this month, are not senior citizens.

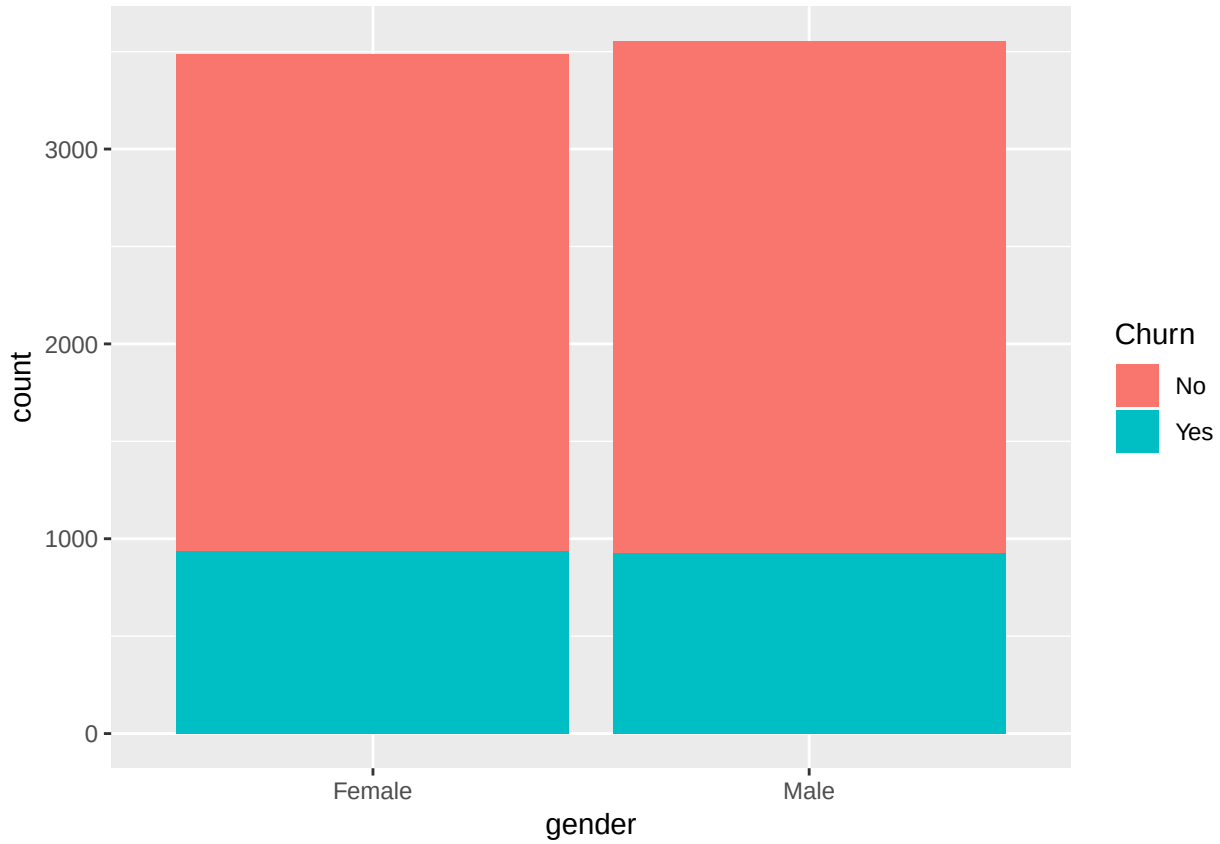
Gender

Gender is a categorical variable (2 levels), indicating whether a customer is

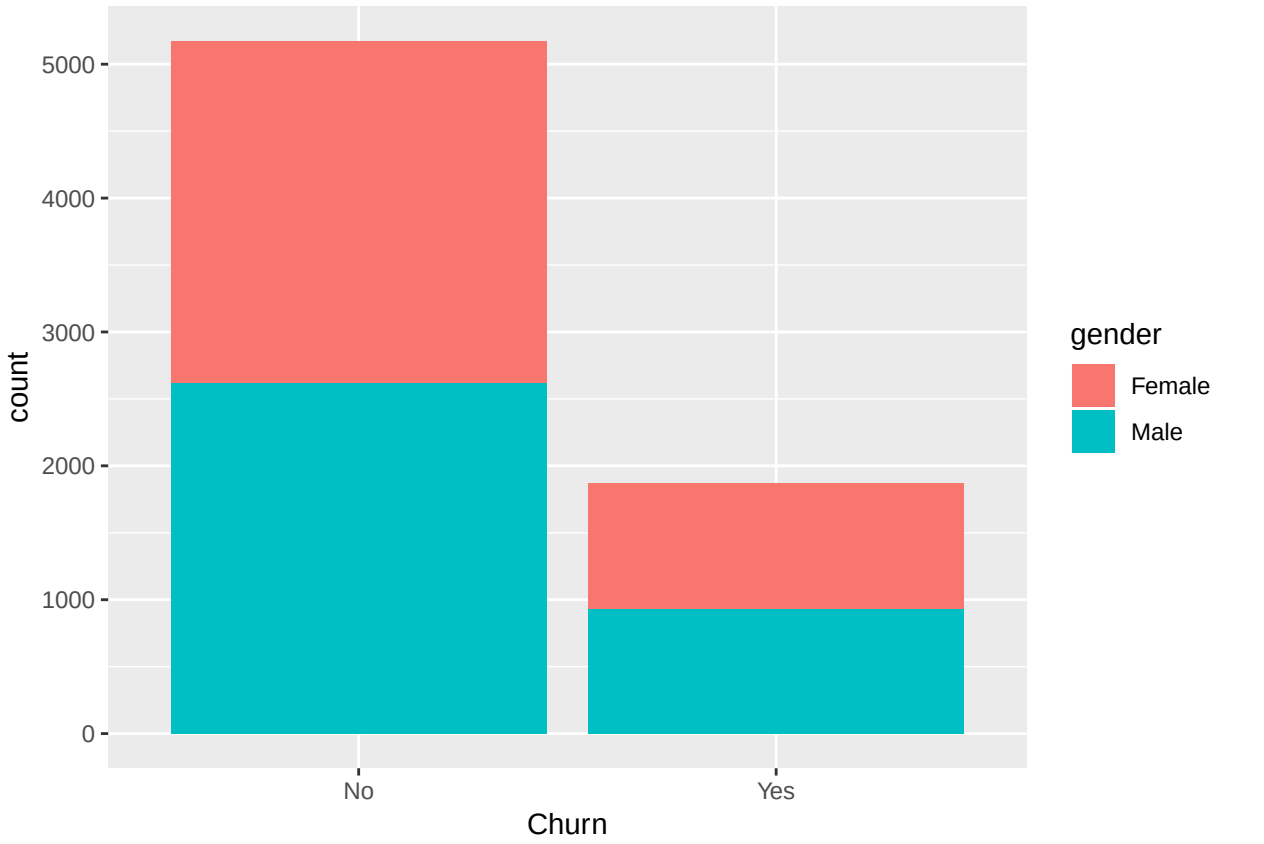
```
ggplot(data=telco)+geom_bar(aes(x=gender))
```




```
ggplot(data=telco)+geom_bar(aes(x=gender, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=gender))
```




```
count(telco %>% filter(gender=="Male"))/count(telco)
```

```
##           n
## 1 0.5047565
```

Gender appears to also be fairly evenly distributed with ~50.5% of the customers in our sample being Male. Visually there appears to be no relationship between the gender and Churn variables.

```
count(telco %>% filter(gender == "Male", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

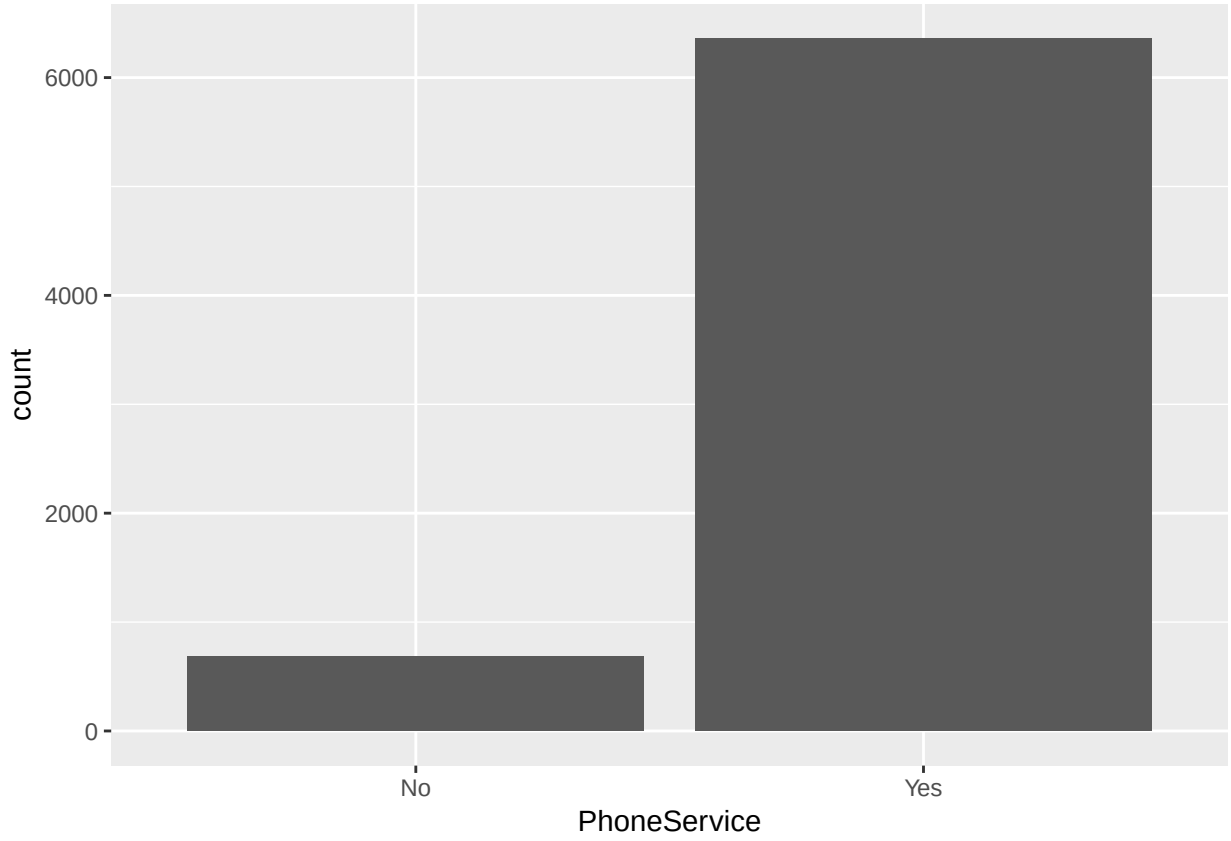
```
##           n
## 1 0.4975923
```

Confirming our visual inspection, ~49.8% of churning customers are Male, since this is aligned with the population distribution, there is no statistically relevent trend between gender and Churn.

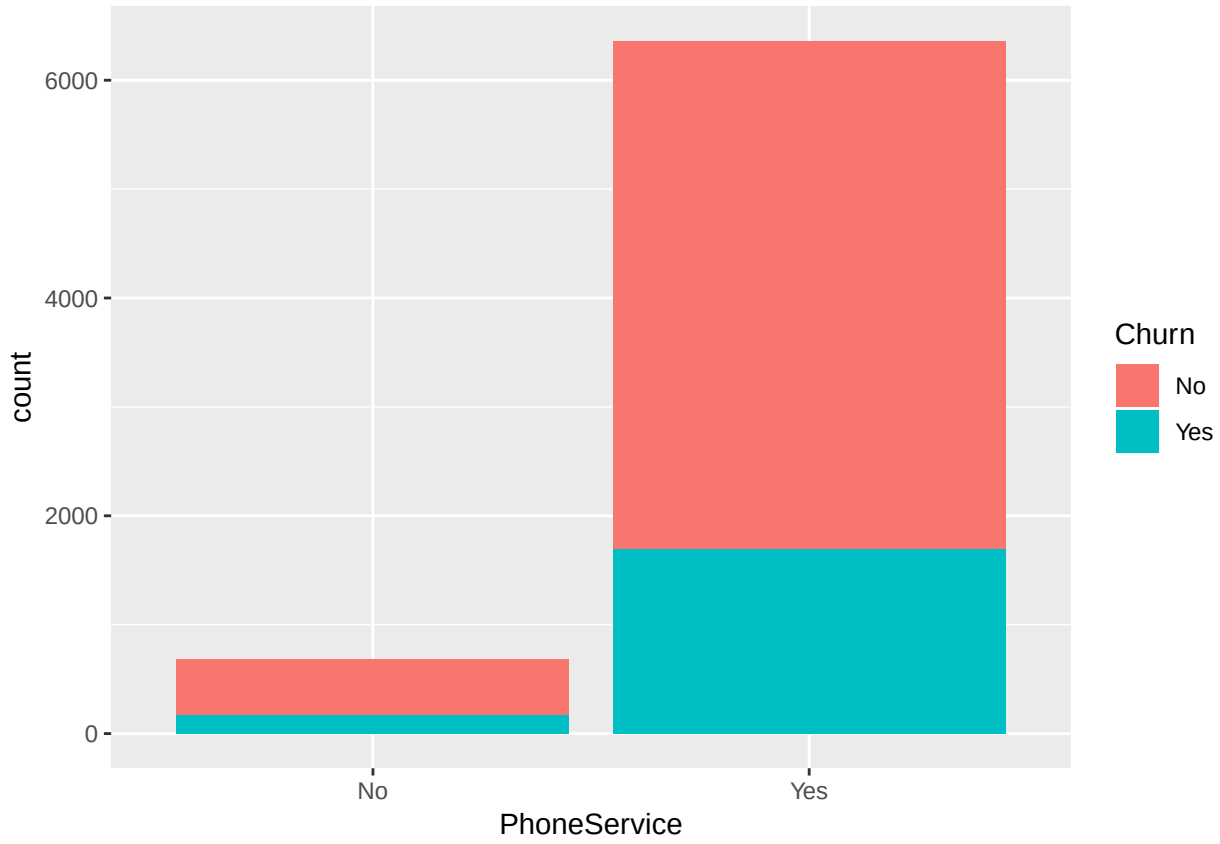
Phone Service

Phone Service is a categorical variable (2 levels), indicating whether a customer has to the phone service provided by the company or not.

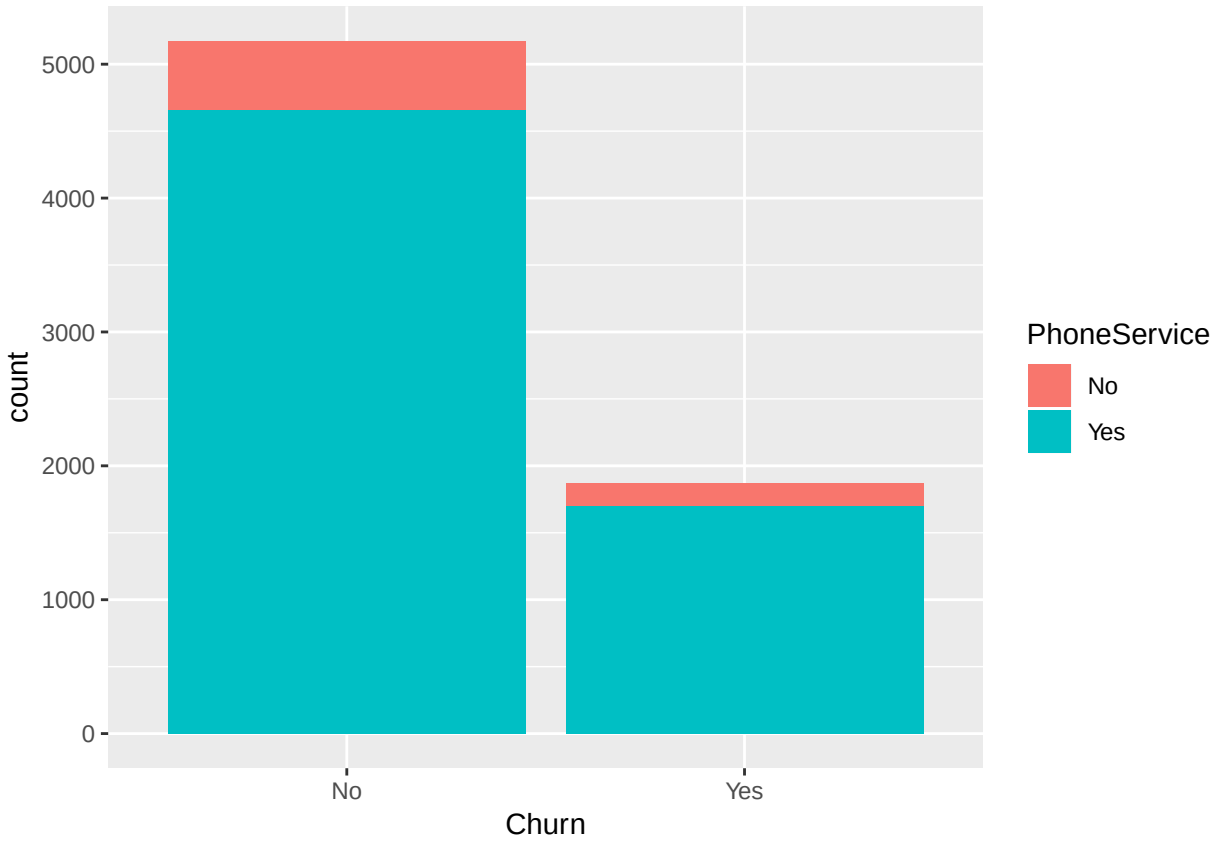
```
ggplot(data=telco)+geom_bar(aes(x=PhoneService))
```




```
ggplot(data=telco)+geom_bar(aes(x=PhoneService, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=PhoneService))
```



From the graph we see that the PhoneService categorical variable is skewed heavily toward having the service.

```
count(telco %>% filter(PhoneService == "Yes", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

```
##           n
## 1 0.9090423
```

As we can see ~91% of the customers that churned this month, have the company's phone service. We must be mindful here, that this may simply be reflecting the distribution of the Phone Service variable itself.

```
count(telco %>% filter(PhoneService == "Yes"))/count(telco)
```

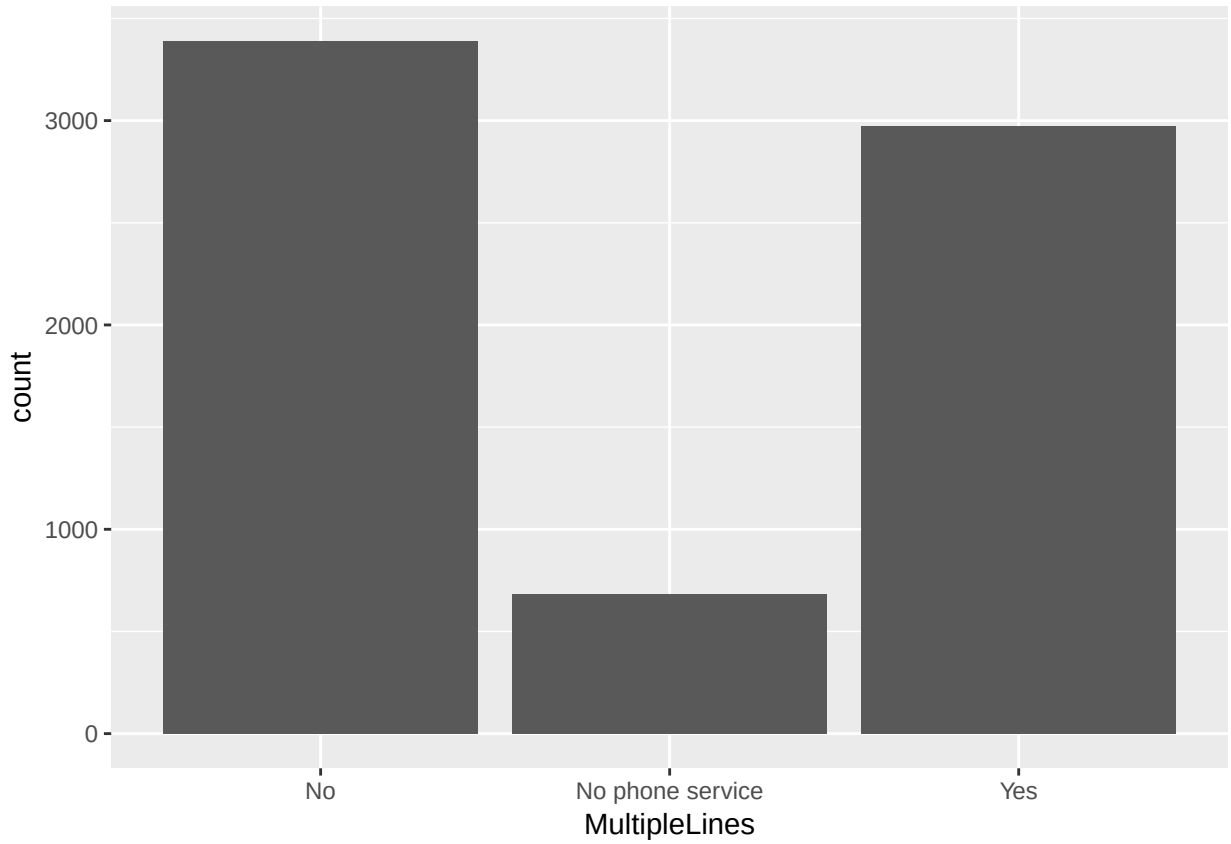
```
##           n
## 1 0.9031663
```

So ~90% of the population have the phone service, and ~91% of the churning customers have the phone service, so there isn't a significant relationship between PhoneService and Churn.

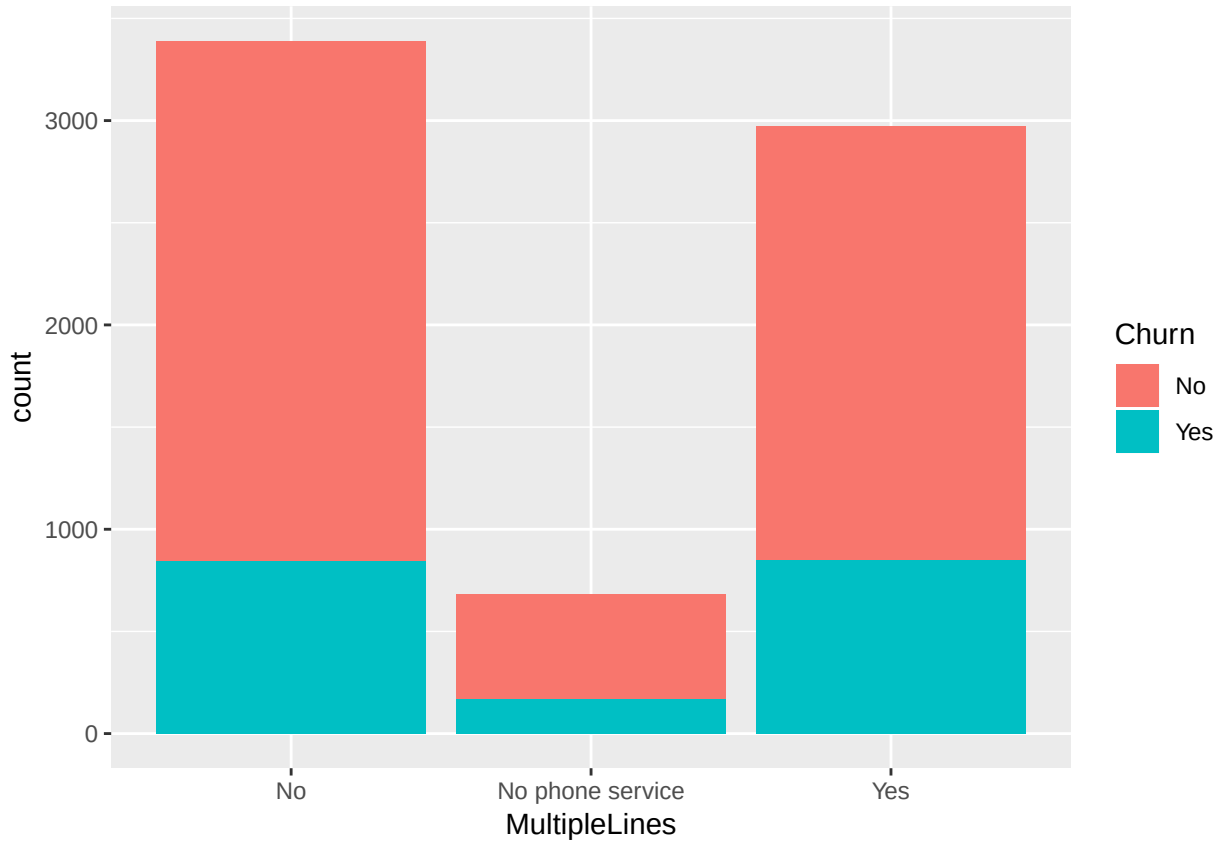
Multiple Lines

Multiple Lines is a categorical variable (3 levels), indicating whether a customer has multiple lines, no phone service, or does not have multiple lines.

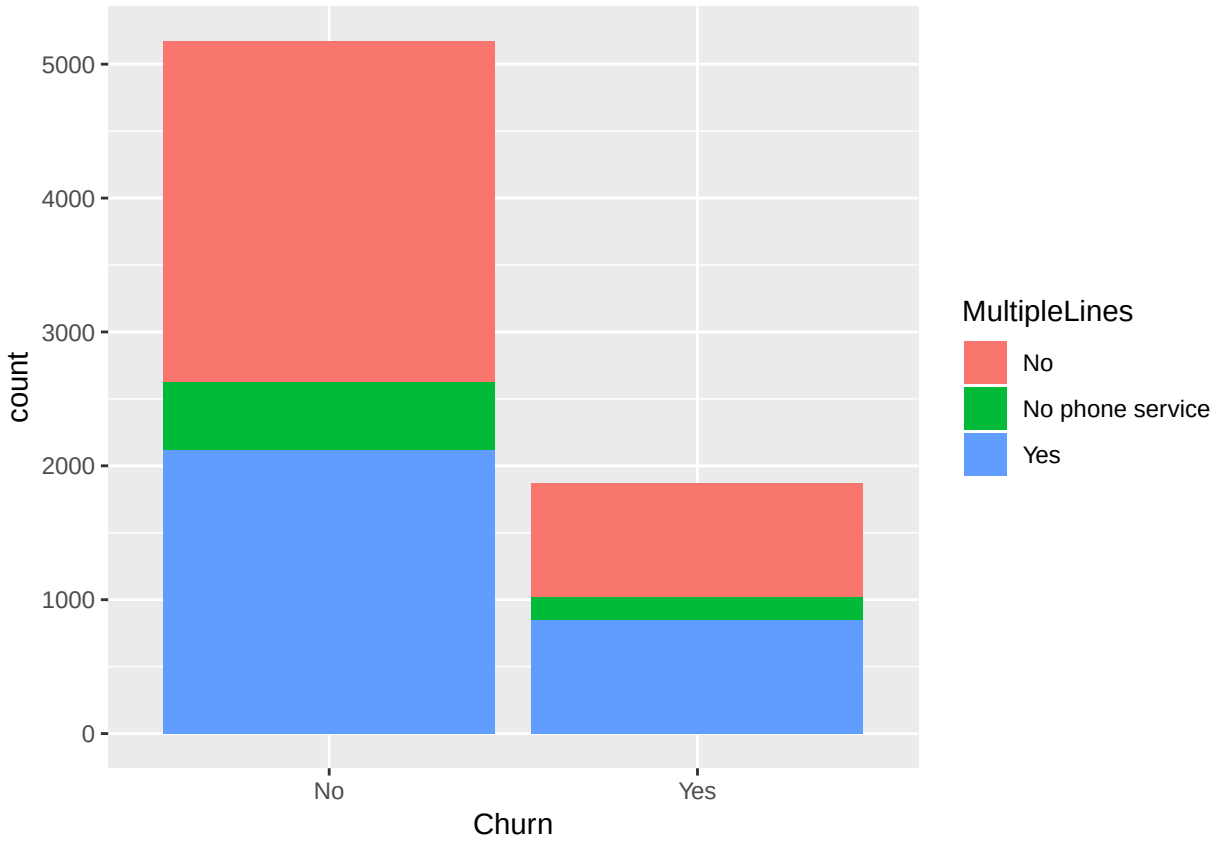
```
ggplot(data=telco)+geom_bar(aes(x=MultipleLines))
```




```
ggplot(data=telco)+geom_bar(aes(x=MultipleLines, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=MultipleLines))
```

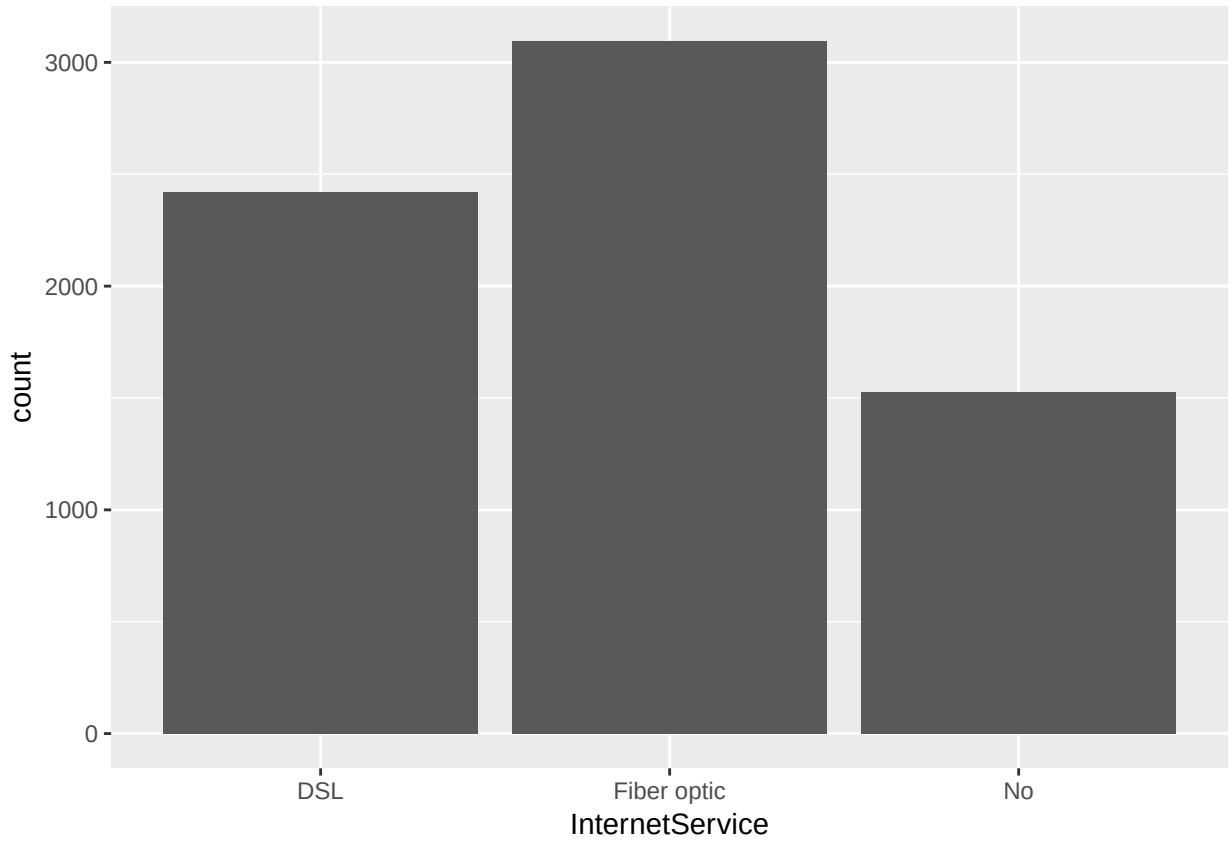


Visually there doesn't appear to be a very strong trend here, as the ratio's between the categorical variables values seem to remain similar between the customers that Churn, and the customers that don't Churn.

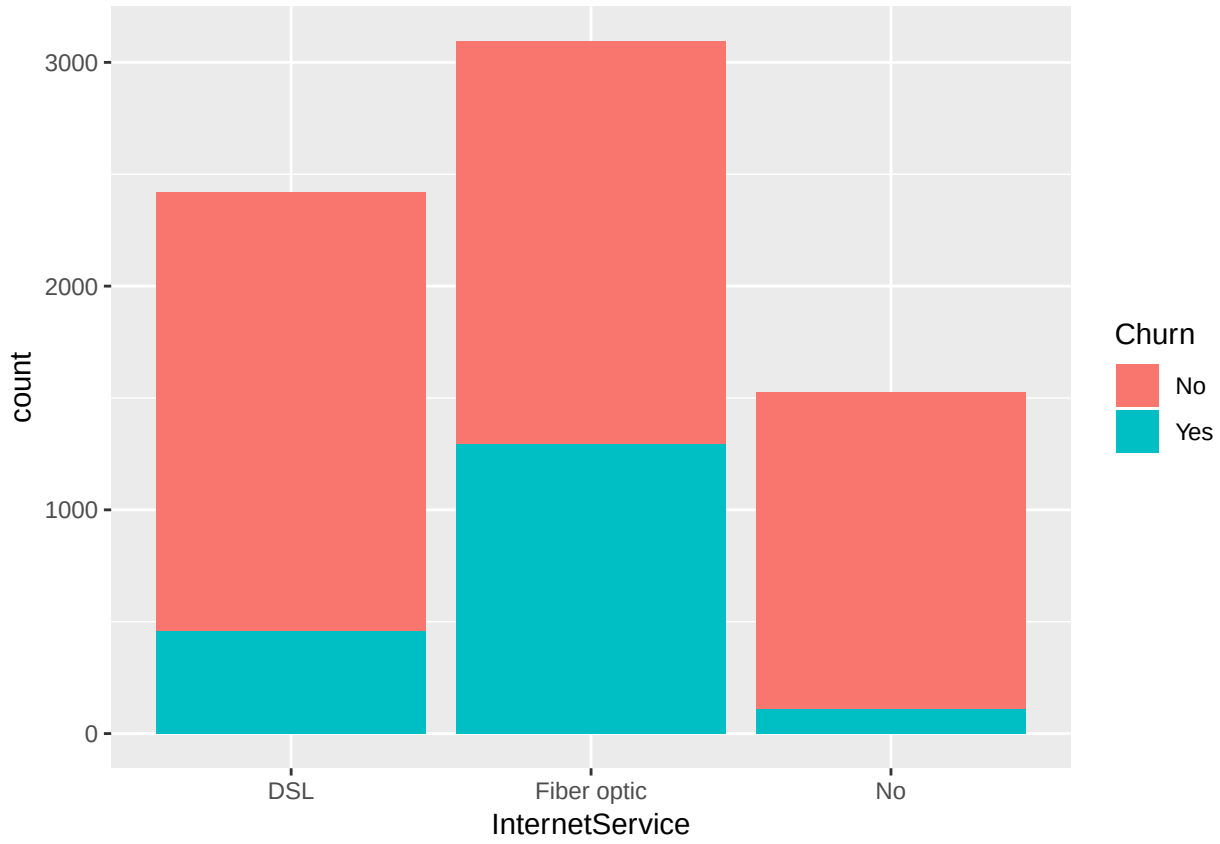
Internet Service

Internet Service is a categorical variable (3 levels), indicating whether a customer is DSL, Fiber optic, or no Internet Service.

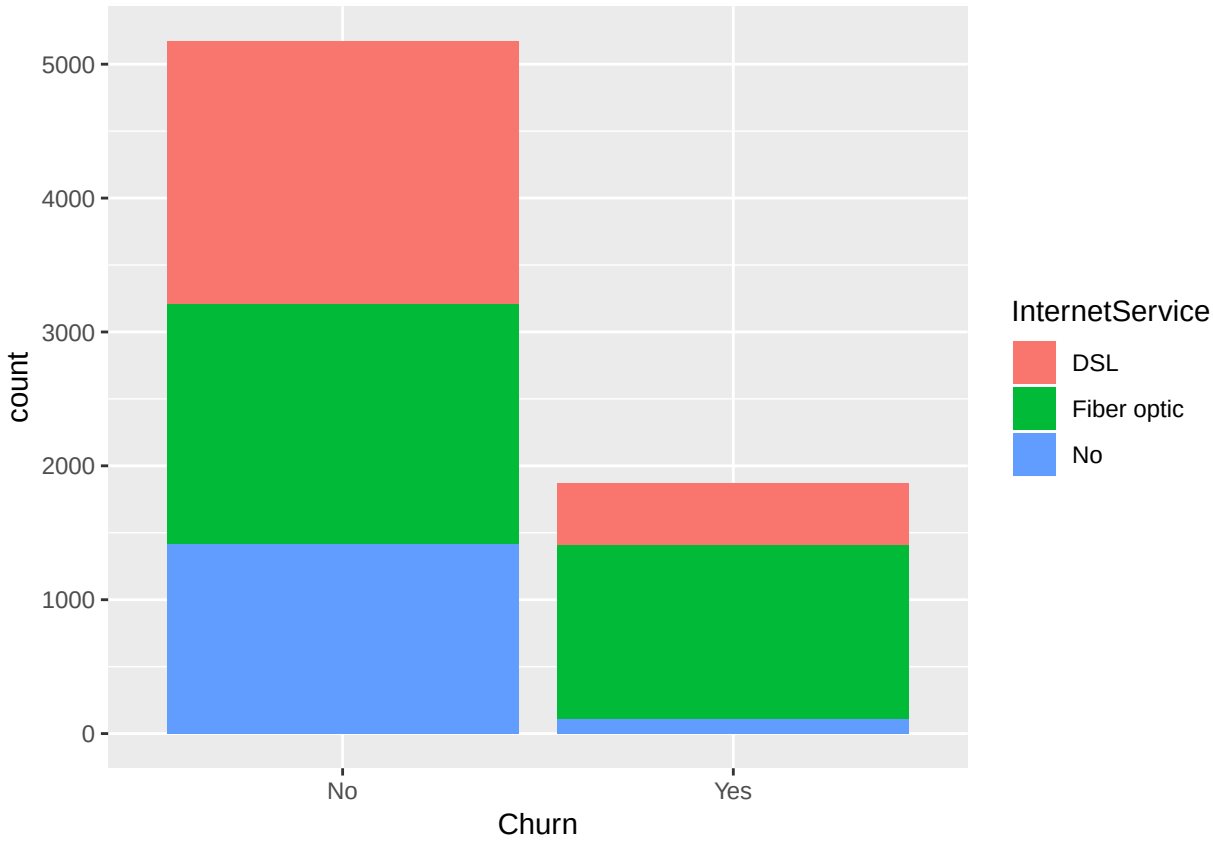
```
ggplot(data=telco)+geom_bar(aes(x=InternetService))
```




```
ggplot(data=telco)+geom_bar(aes(x=InternetService, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=InternetService))
```



There appears to be a relationship between the InternetService and Churn variables, as the majority of churning customers are using Fiber optic service.

```
count(telco %>% filter(InternetService == "Fiber optic", Churn == "Yes"))/
  count(telco %>% filter(Churn == "Yes"))
```

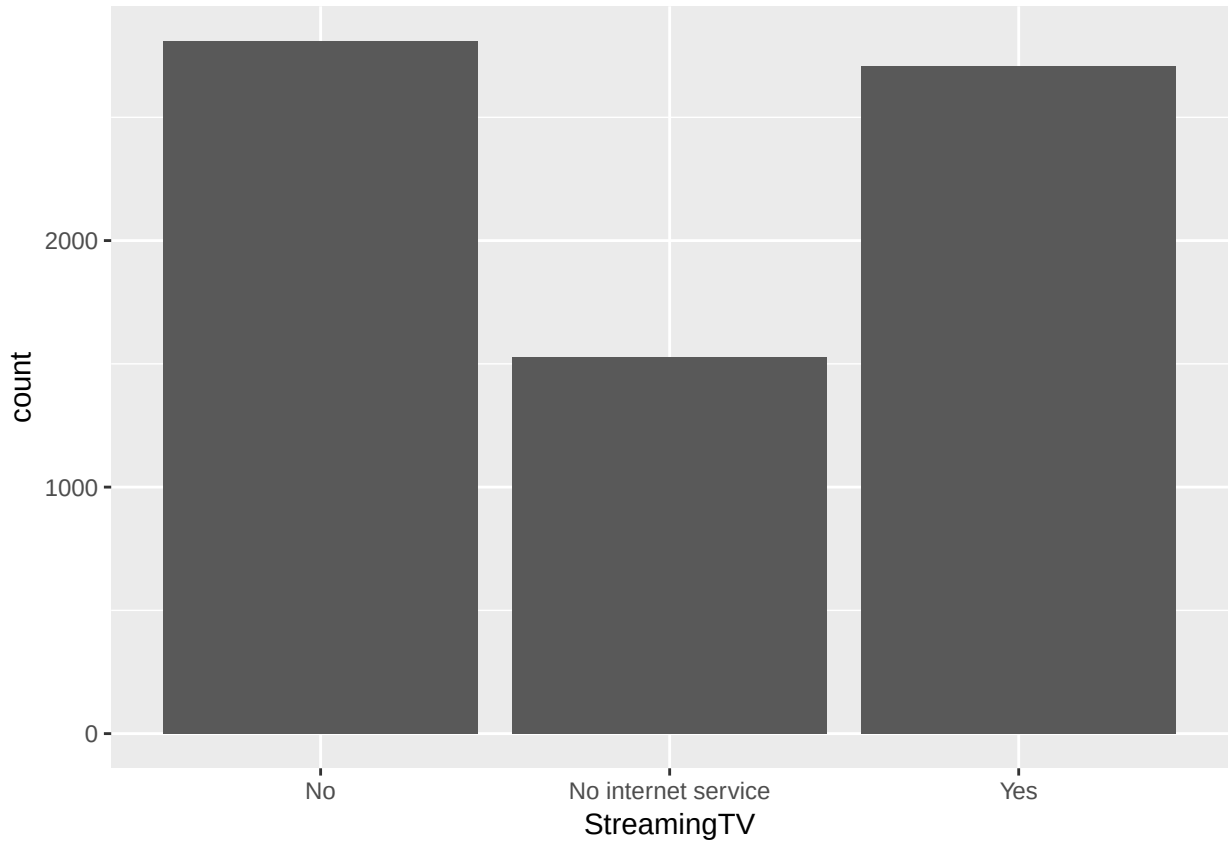
```
##          n
## 1 0.693954
```

As we can see ~70% of the customers that churned this month, were using the company's Fiber Optic internet service.

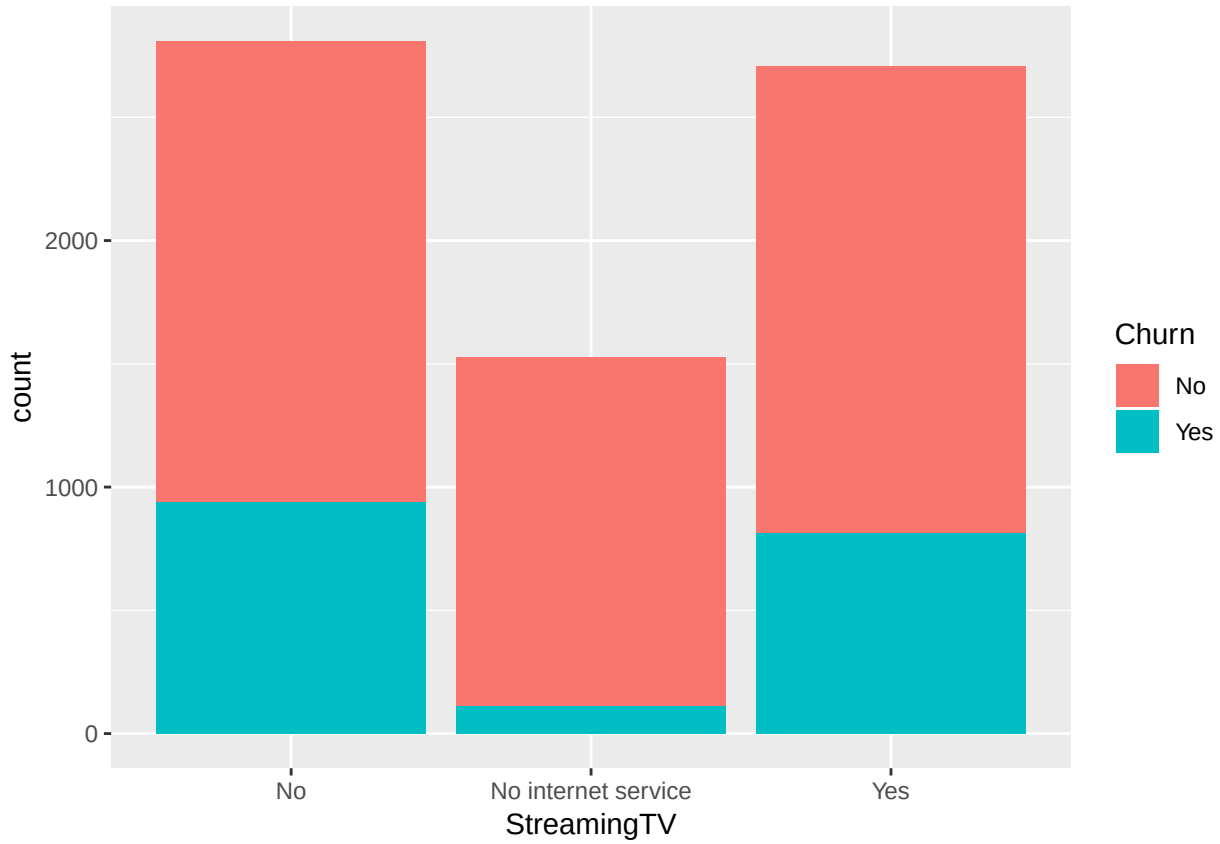
Streaming TV

StreamingTV is a categorical variable (3 levels), indicating whether a customer has the company's TV Streaming Service, has no internet service, or has opted out of the TV streaming service.

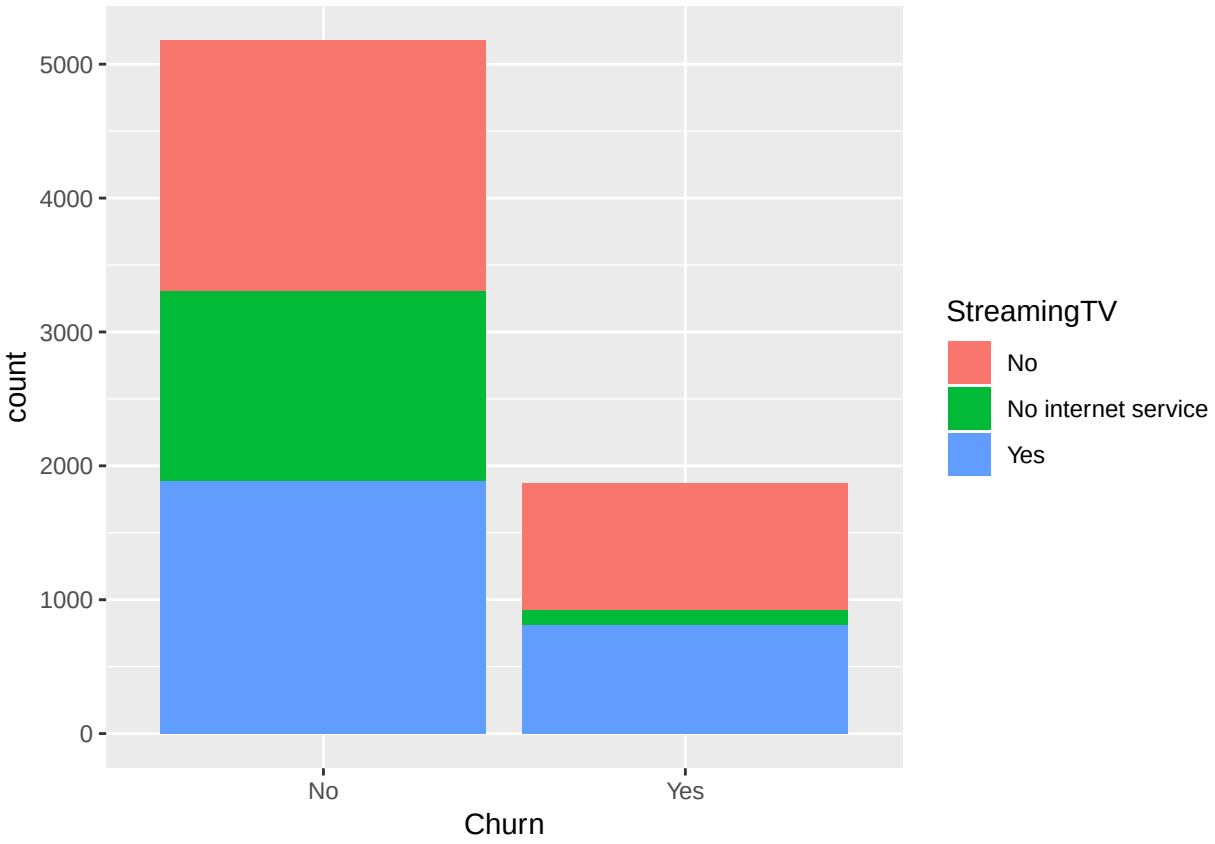
```
ggplot(data=telco)+geom_bar(aes(x=StreamingTV))
```




```
ggplot(data=telco)+geom_bar(aes(x=StreamingTV, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=StreamingTV))
```



There does appear to be a releationship here, though simply be a reflection of trends seen earlier, as the ratios between having StreamingTV and not having it don't seem to change when comparing a sample of only customers that churn versus a sample of only customers that aren't churning. The trend we do see, it that it appears that customers with No Internet Service are less likely to Churn.

```
count(telco %>% filter(StreamingTV == "No internet service", Churn == "Yes"))/
  count(telco %>% filter(StreamingTV == "No internet service"))
```

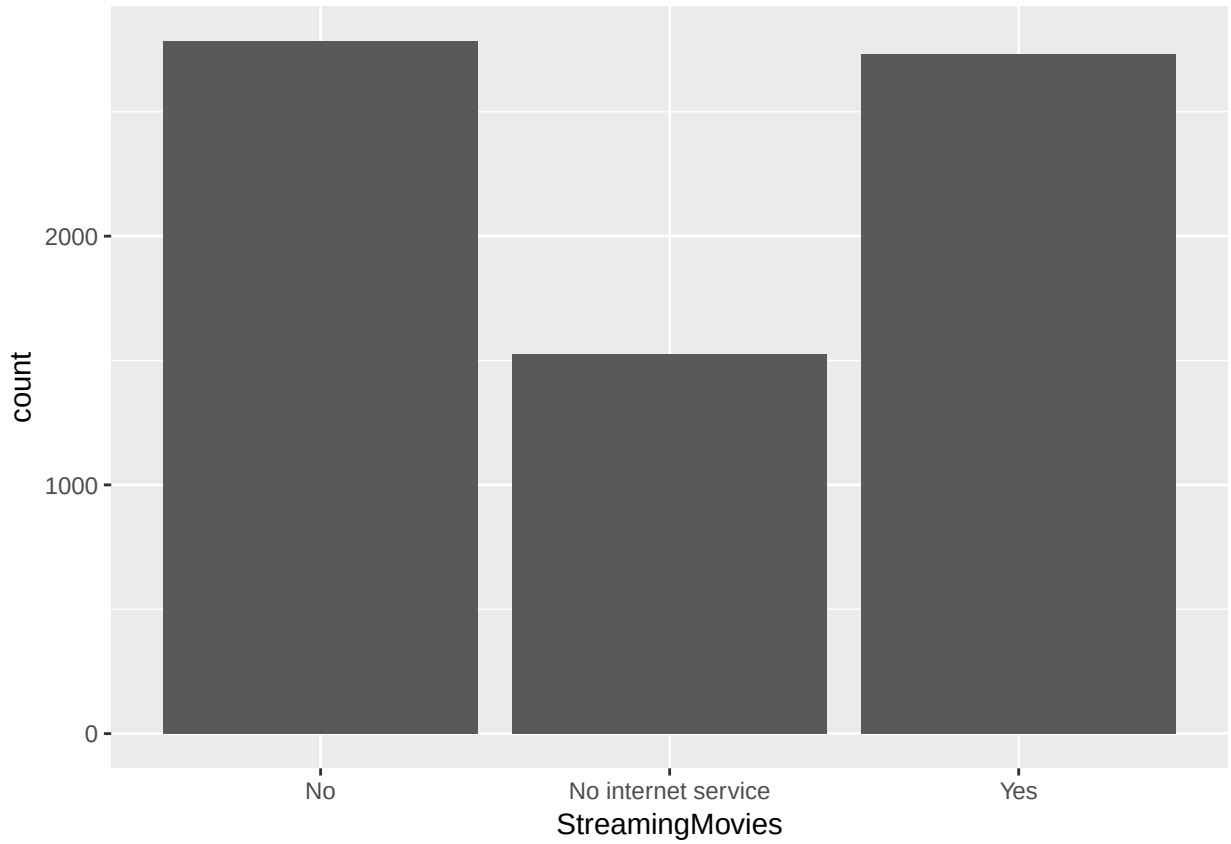
```
##           n
## 1 0.0740498
```

Of those customers that have No internet service, only ~7% churned.

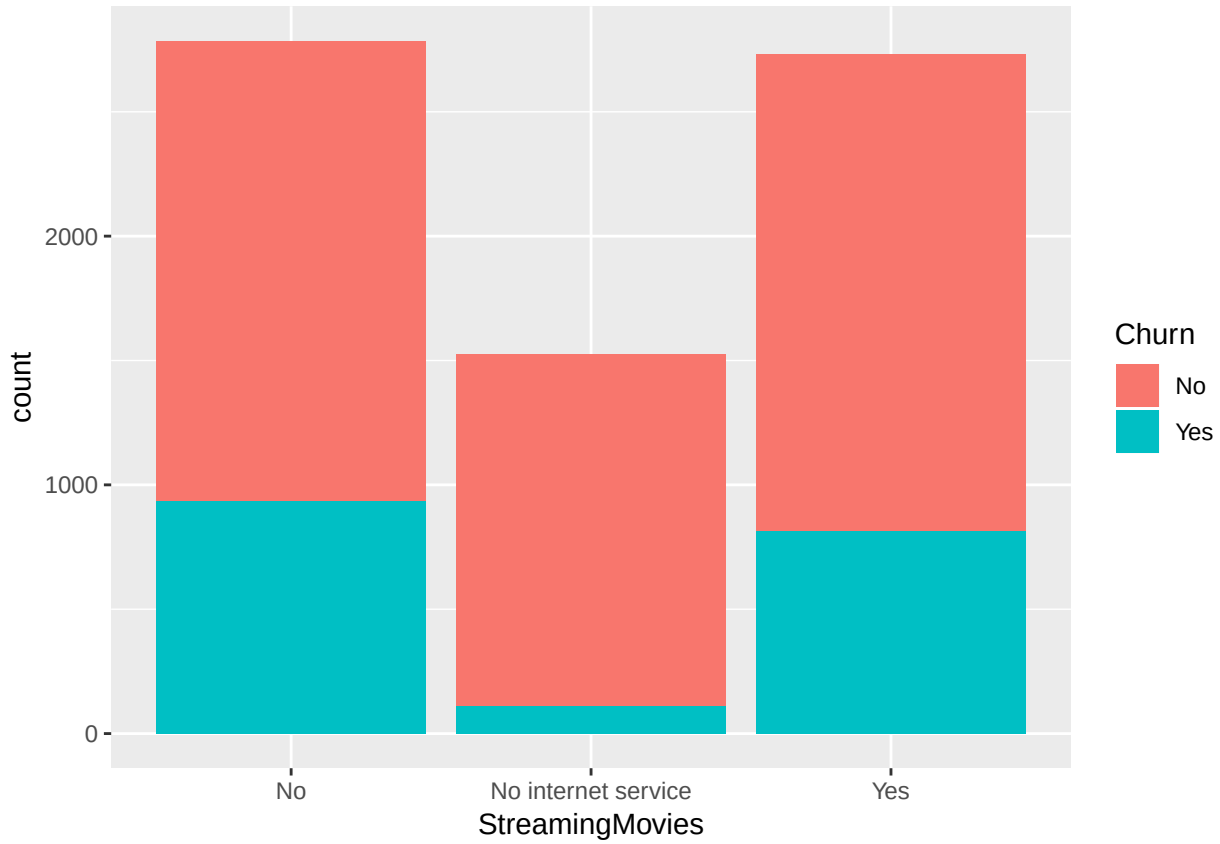
Streaming Movies

Streaming Movies is a categorical variable (3 levels), indicating whether a customer has the company's movie streaming service, has no internet service, or has opted out of the movie streaming service.

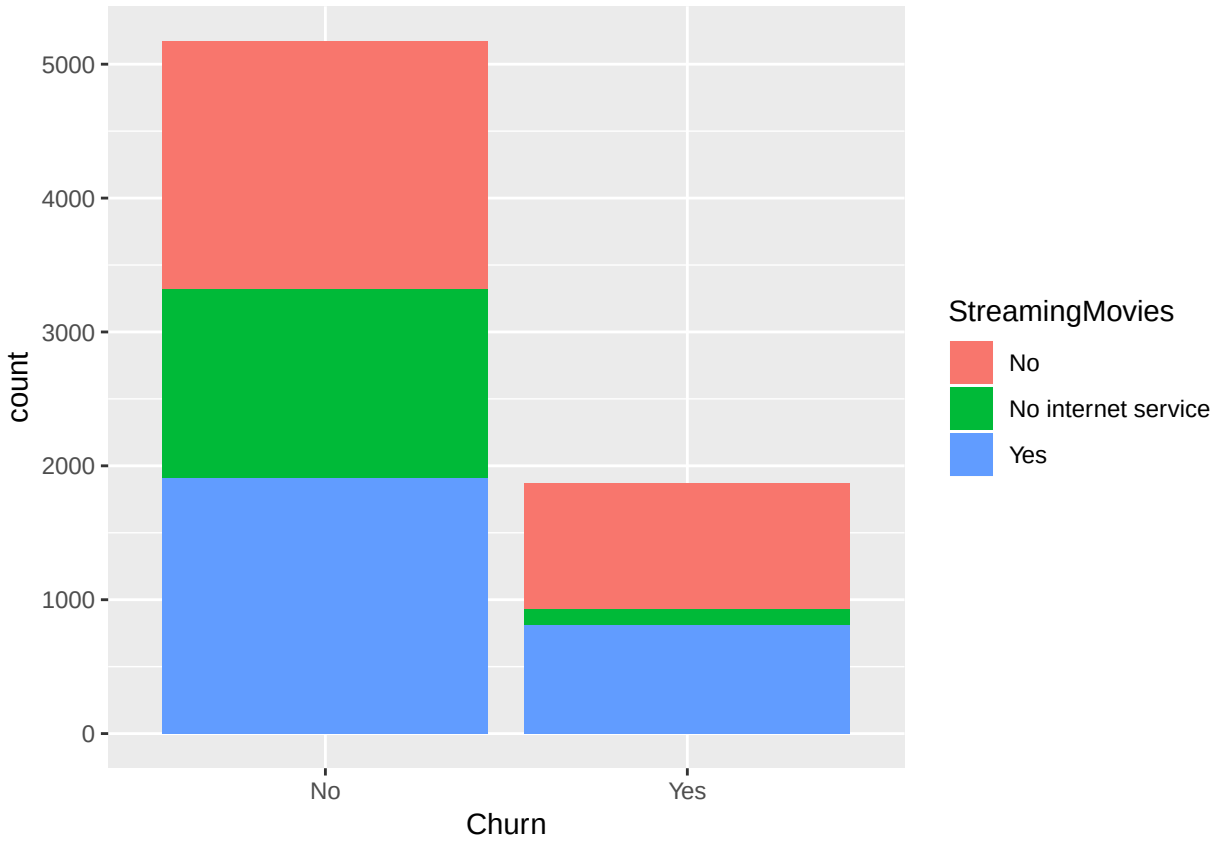
```
ggplot(data=telco)+geom_bar(aes(x=StreamingMovies))
```




```
ggplot(data=telco)+geom_bar(aes(x=StreamingMovies, fill=Churn))
```




```
ggplot(data=telco)+geom_bar(aes(x=Churn, fill=StreamingMovies))
```

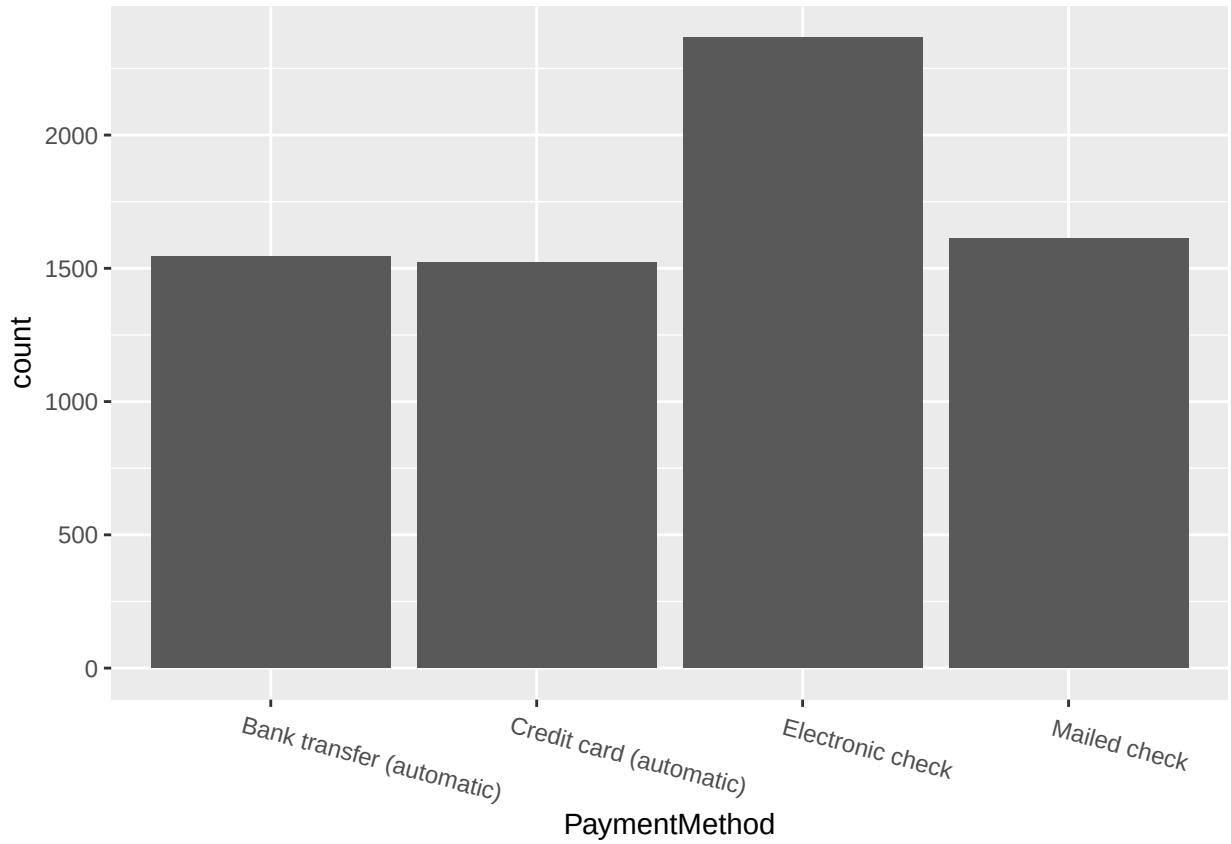


Similarly to the StreamingTV variable above, we can tell visually that the only trend here is related having internet service, it doesn't appear a relationship between StreamingTV and Churn exists outside of that.

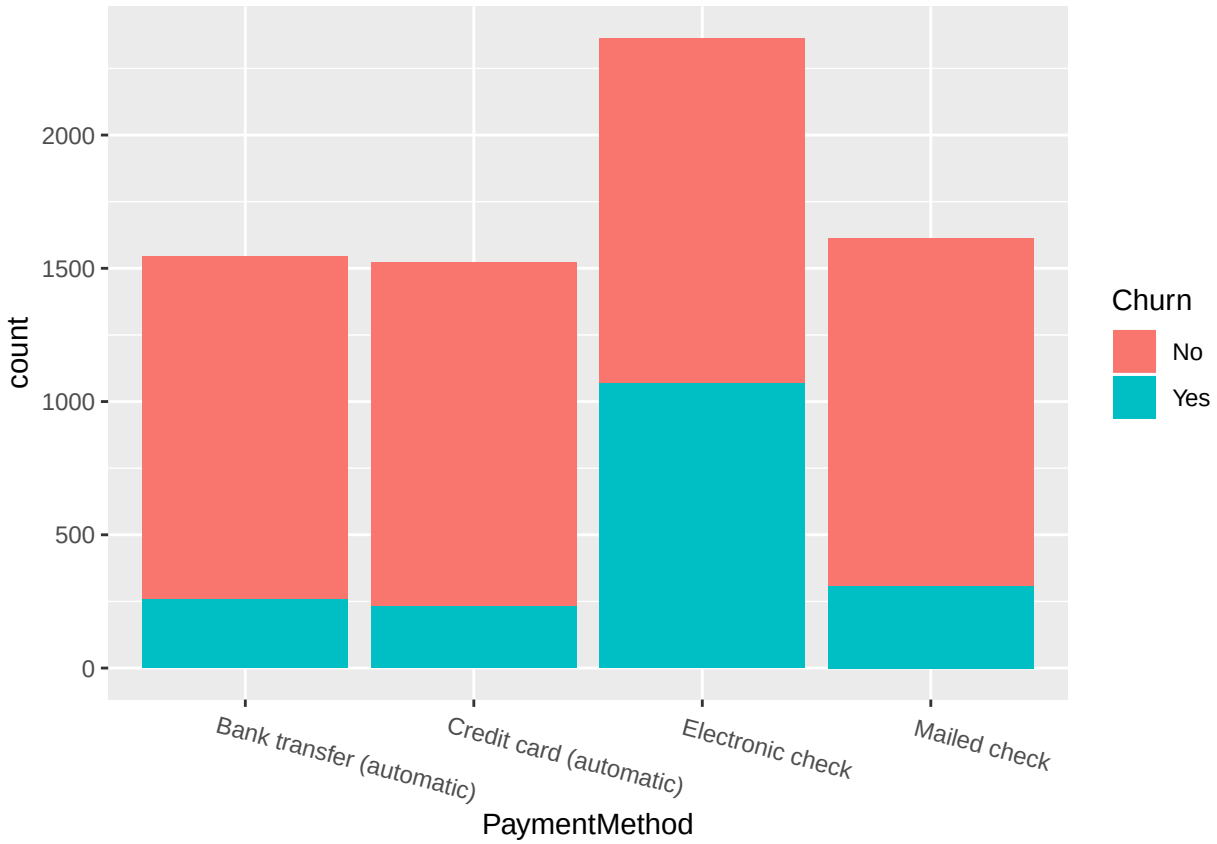
Payment Method

Payment Method is a categorical variable (2 levels), indicating whether a customer is using Paperless Billing, or not.

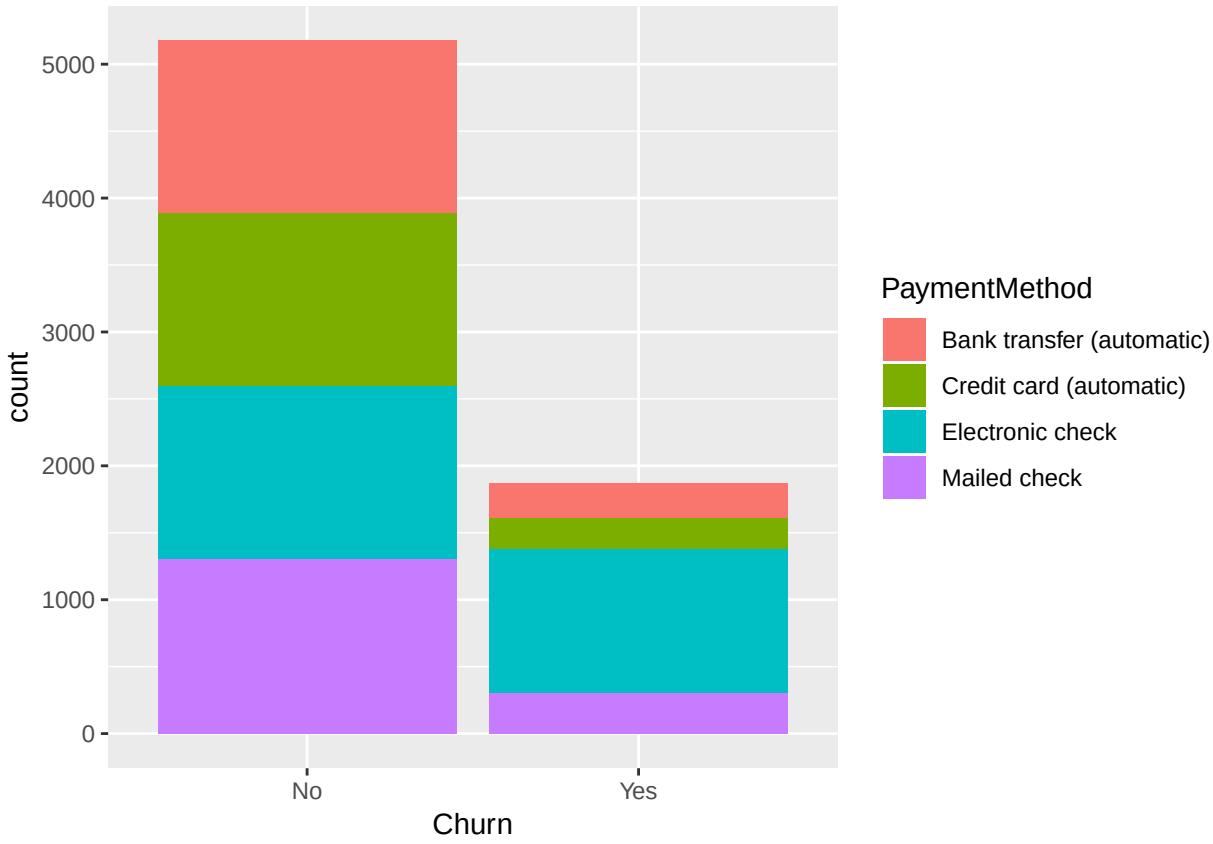
```
ggplot(data=telco) +
  geom_bar(aes(x=PaymentMethod)) +
  theme(axis.text.x=element_text(angle=-15,hjust=.1,vjust=1))
```




```
ggplot(data=telco) +
  geom_bar(aes(x=PaymentMethod, fill=Churn)) +
  theme(axis.text.x=element_text(angle=-15,hjust=.1,vjust=1))
```




```
ggplot(data=telco) +
  geom_bar(aes(x=Churn, fill=PaymentMethod))
```



The relationship between these appears to be fairly complicated. But visually a relationship does exist, for example it appears that ~50% of those customers who are churning are paying with electronic check, while only about 25% of the customer who are NOT churning are paying with electronic check.

Models/Analysis

Tenure T-test

Assumptions:

Independence: Although we are not sure if our Telco data is synthetically generated or from real customers who might communicate with each other about their service, we still think that it is reasonable to assume that customers are churning independently of each other.

Normality

```
skewness(telco$tenure)
```

```
## [1] 0.2394377
```

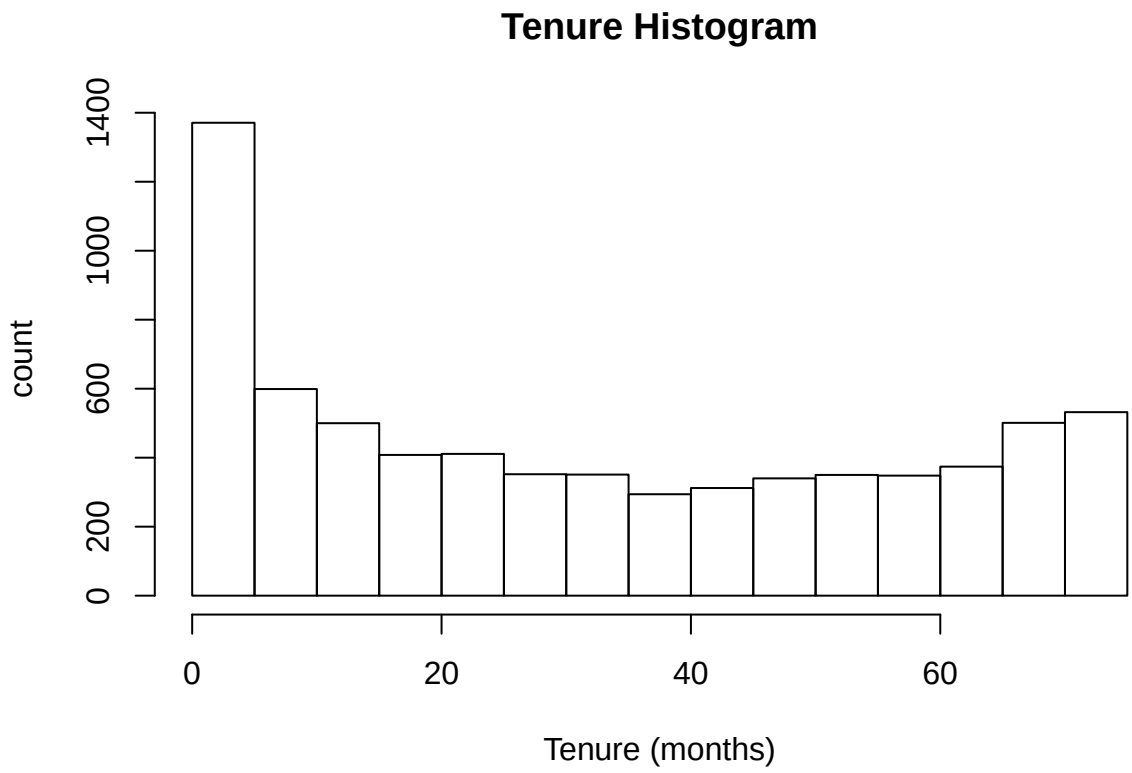
```
skewness(telco_no$tenure)
```

```
## [1] -0.03170148
```

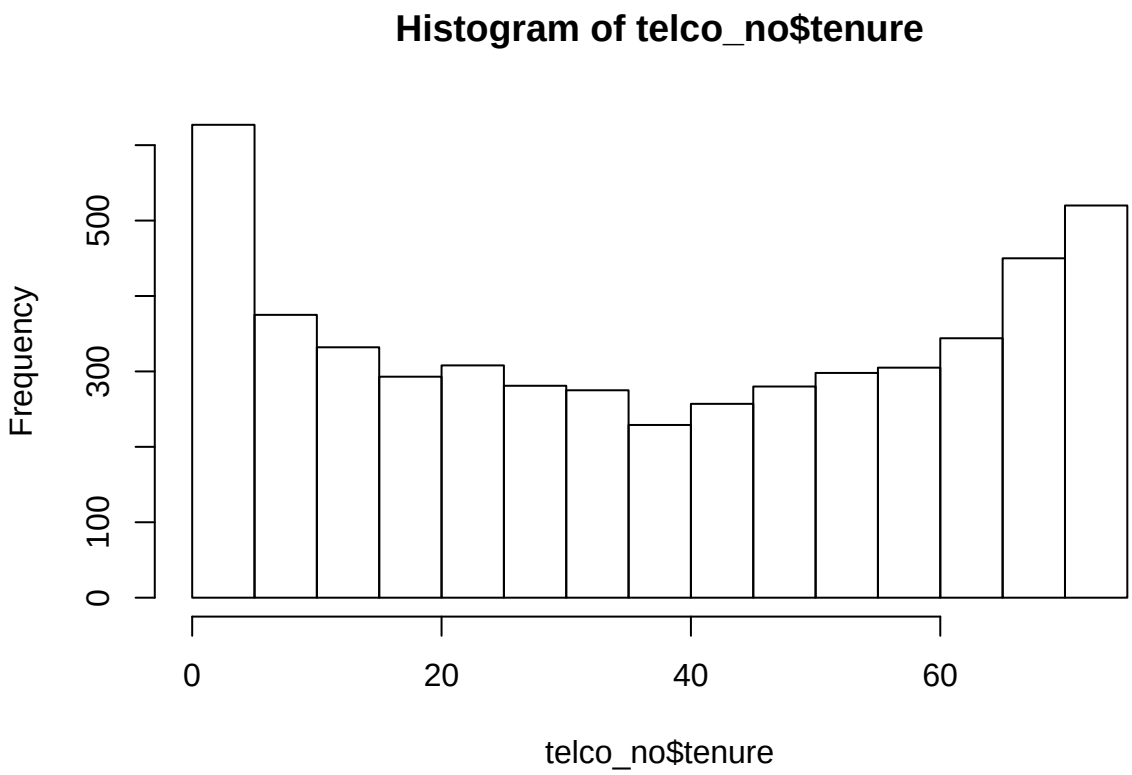
```
skewness(telco_yes$tenure)
```

```
## [1] 1.147436
```

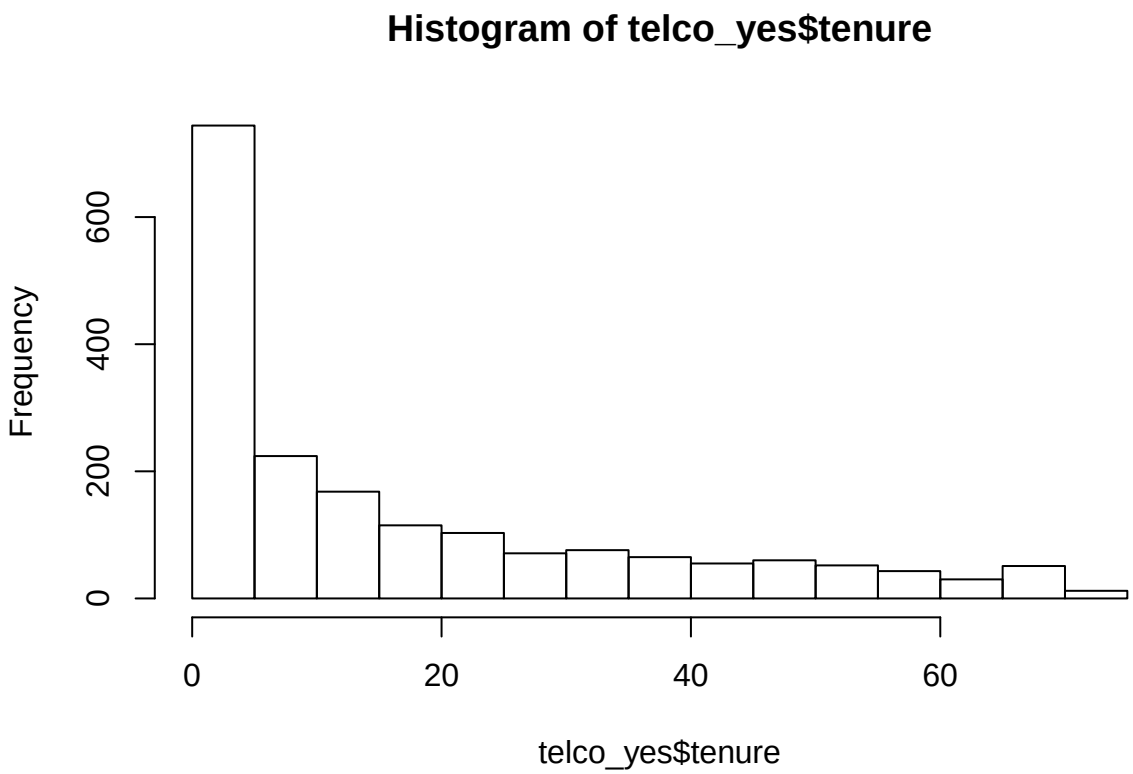
```r
hist(telco$tenure, ylab = "count", xlab = "Tenure (months)",
     main = "Tenure Histogram")# not normal
```



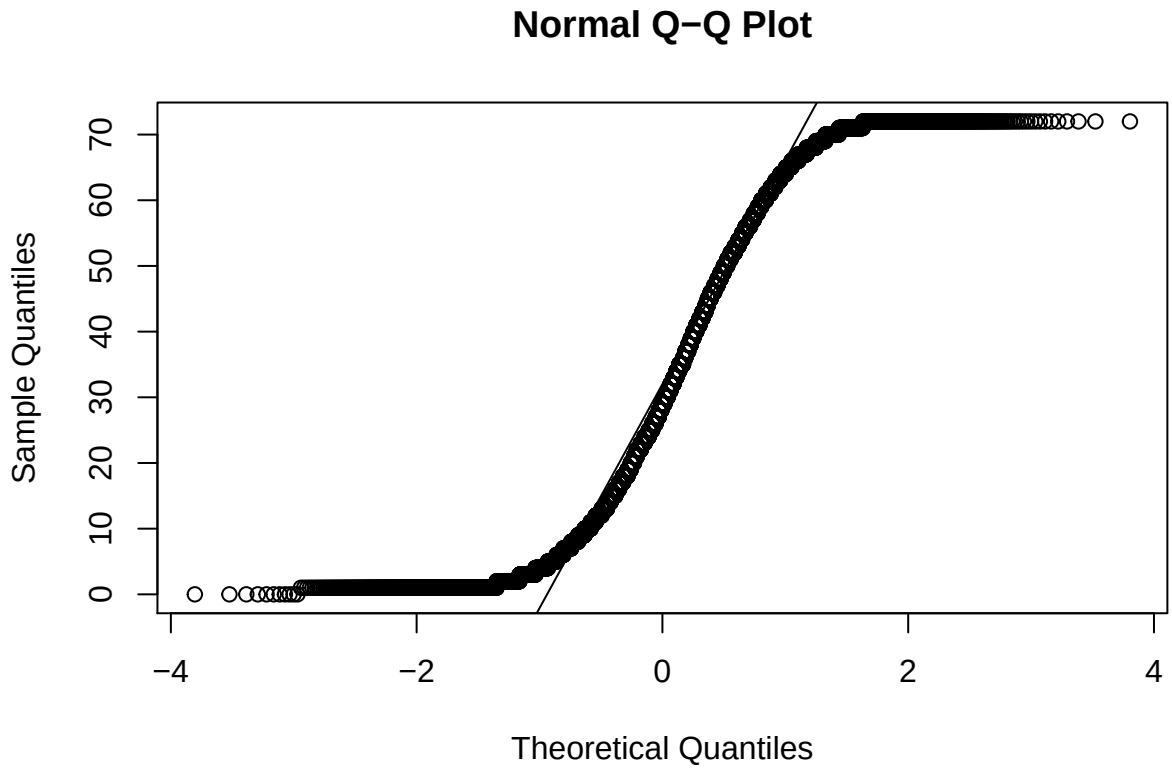

```r
hist(telco_no$tenure) # not normal
```



```r
hist(telco_yes$tenure)# not normal and definitely skewed right,
```




```r
qqnorm(telco$tenure)
qqline(telco$tenure)
```



Visually, our tenure (overall or for each sample) is not normal, and the tenure of churning customers is strongly right skewed, but since our smallest sample contains 1869 observations, we believe that we can waive the normality assumption because t-tests are robust to the normality assumption, even if the data is strongly skewed, when n is large, which it is, according to the central limit theorem.

Check if we should use pooled variances or Welch's version of t-test.

```
sd(telco_yes$tenure)
```

```
## [1] 19.53112
```

```
sd(telco_no$tenure)
```

```
## [1] 24.11378
```

```
leveneTest(tenure~Churn,data=telco)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  417.17 < 2.2e-16 ***
##       7041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sample std devs are not = and we reject H0 that the variances of each group are not equal.

Conduct our t-test for tenure.

```
t.test(telco_yes$tenure,telco_no$tenure,alternative = "less",var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
## data:  telco_yes$tenure and telco_no$tenure
## t = -34.824, df = 4048.3, p-value < 2.2e-16
## alternative hypothesis: true difference in means is less than 0
## 95 percent confidence interval:
##       -Inf -18.66528
## sample estimates:
## mean of x mean of y
##  17.97913  37.56997
```

Since our p-value is much less than .05, we reject H0, the mean tenure of telecom customers who churned (for this company) is equal to the mean tenure of the telecom customers who did not churn, at the alpha = .05 level of significance and consider our alternative hypothesis: the mean tenure of the customers who churned is less than the mean tenure of the customers who did not churn.

We now have reason to beleive that there is a relationship between whether a customer churns and their tenure, on average. Therefore it makes sense that researchers studying churn rates typically include a customer's tenure in a logistic regression or classification model and that it showed predictive power.

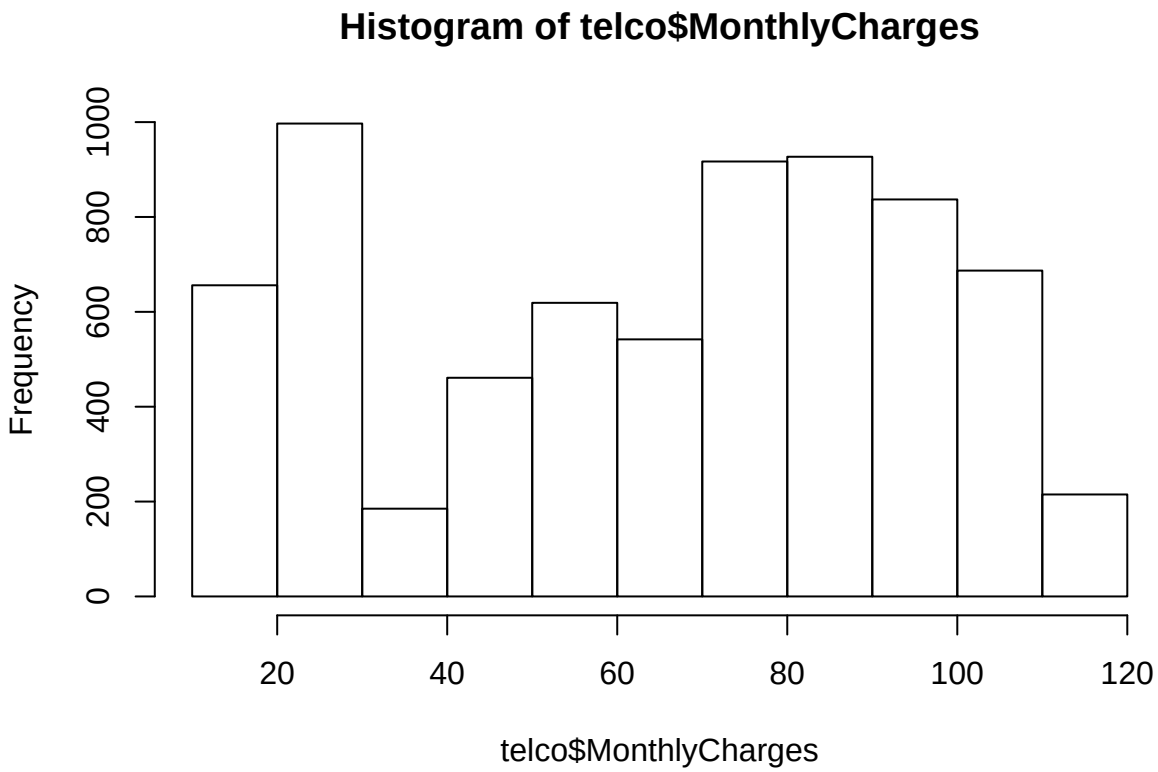
Monthly Charges T-test

Assumptions:

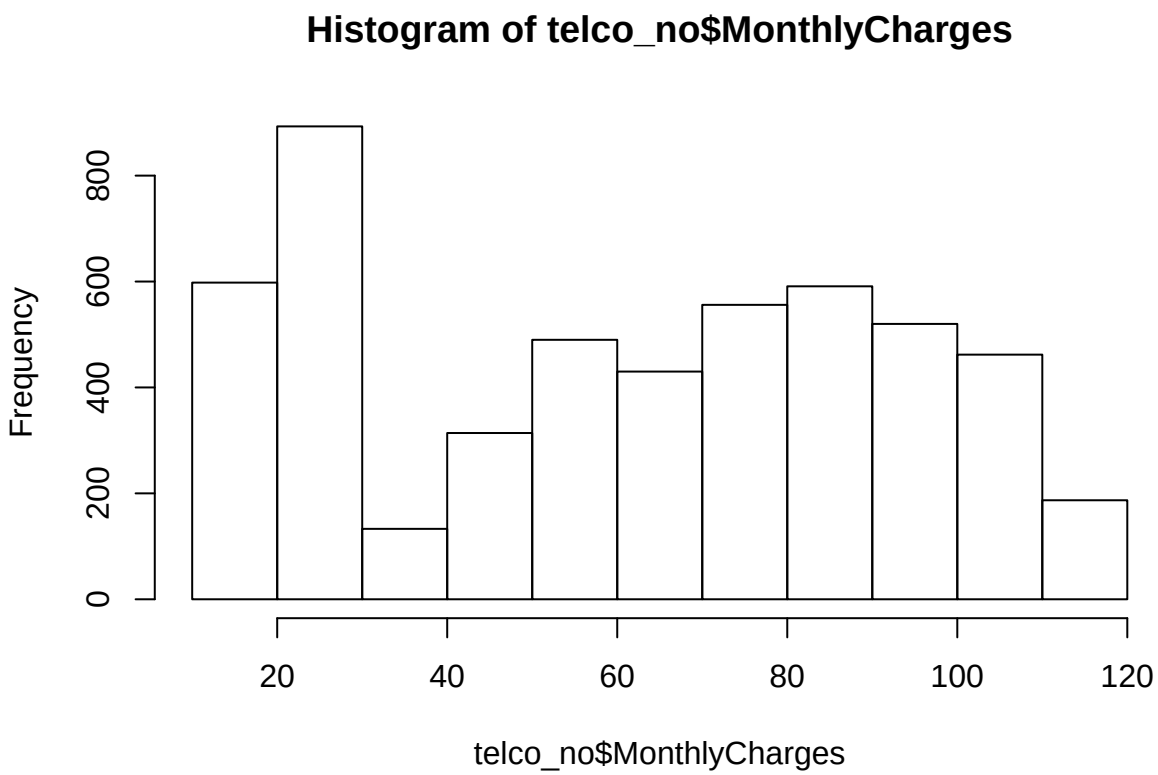
Independence: Although we are not sure if our Telco data is synthetically generated or from real customers who might communicate with each other about their service, we still think that it is reasonable to assume that customers are churning independently of each other.

Normality

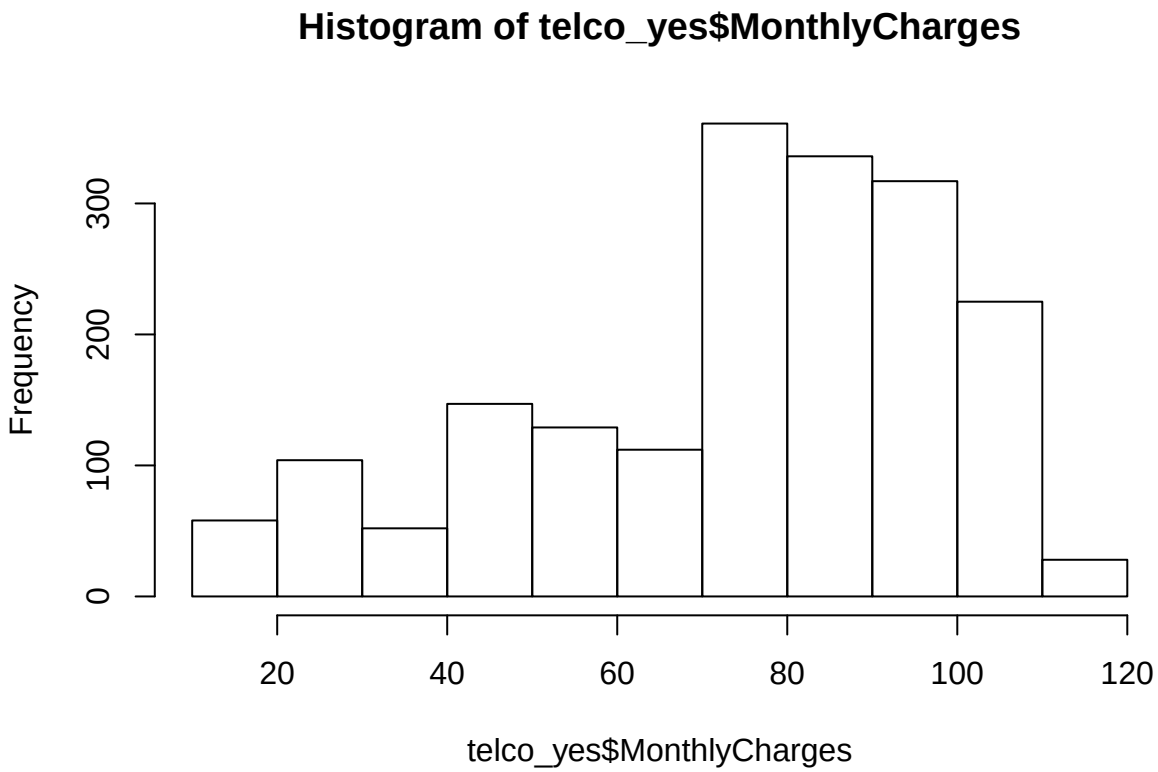
```r
hist(telco$MonthlyCharges)# not normal
```




```r
hist(telco_no$MonthlyCharges)# not normal
```



```r
hist(telco_yes$MonthlyCharges)# not normal
```




```r
skewness(telco$MonthlyCharges)# skewed slightly left.
```

```
## [1] -0.2204305
```

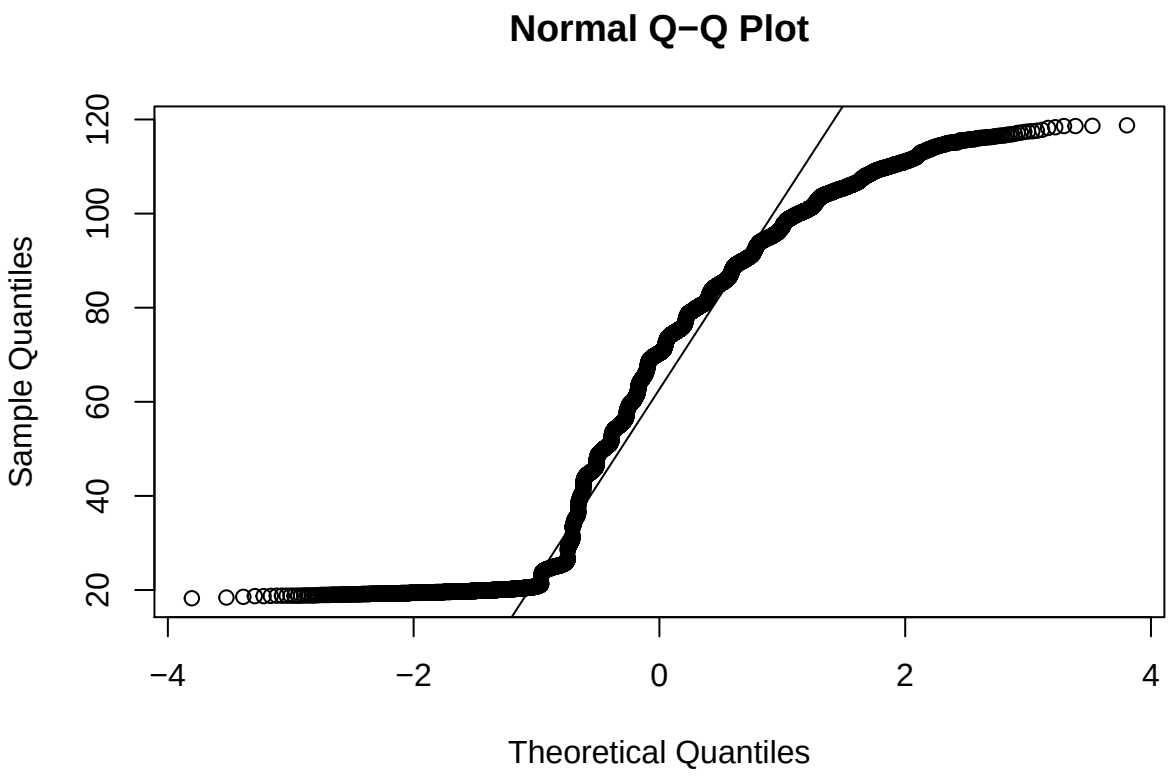
```r
skewness(telco_no$MonthlyCharges)# skewed slightly left
```

```
## [1] -0.02500504
```

```r
skewness(telco_yes$MonthlyCharges)# moderately to strongly skewed left
```

```
## [1] -0.7282035
```

```r
qqnorm(telco$MonthlyCharges)
qqline(telco$MonthlyCharges)
```



Visually, MonthlyCharges (overall or for each sample) is not normal, and the MonthlyCharges of churning customers is moderately left skewed, but since our smallest sample contains 1869 observations, we believe that we can waive the normality assumption because t-tests are robust to the normality assumption (even if the data is strongly skewed) when n is large, which it is, according to the central limit theorem.

Check if we should use pooled variances or Welch's version of t-test.

```
sd(telco_no$MonthlyCharges)
```

```
## [1] 31.09265
```

```
sd(telco_yes$MonthlyCharges)
```

```
## [1] 24.66605
```

```
leveneTest(MonthlyCharges~Churn,data=telco)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  361.84 < 2.2e-16 ***
##       7041
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Sample std devs are not = and we reject H0 that the variances of each group are not equal.

Conduct our t-test for tenure.

```
t.test(telco_yes$MonthlyCharges,telco_no$MonthlyCharges,
       alternative = "greater",var.equal = FALSE)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  telco_yes$MonthlyCharges and telco_no$MonthlyCharges
## t = 18.408, df = 4135.8, p-value < 2.2e-16
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  11.99855      Inf
## sample estimates:
## mean of x mean of y
##  74.44133  61.26512
```

Since our p-value is much less than .05, we reject H0, [the mean monthly charges of customers who churned is equal to the monthly charges of telecom customers who did not churn for this company] at the alpha = .05 level of significance and consider the alternative, the mean monthly charges of the customers who churned are greater than the mean monthly charges of the customers who did not churn.

We now have reason to believe that there is a relationship between whether a customer churns and their tenure, on average. Therefore it makes sense that researchers studying churn rates typically include a customer's tenure in a logistic regression or classification model.

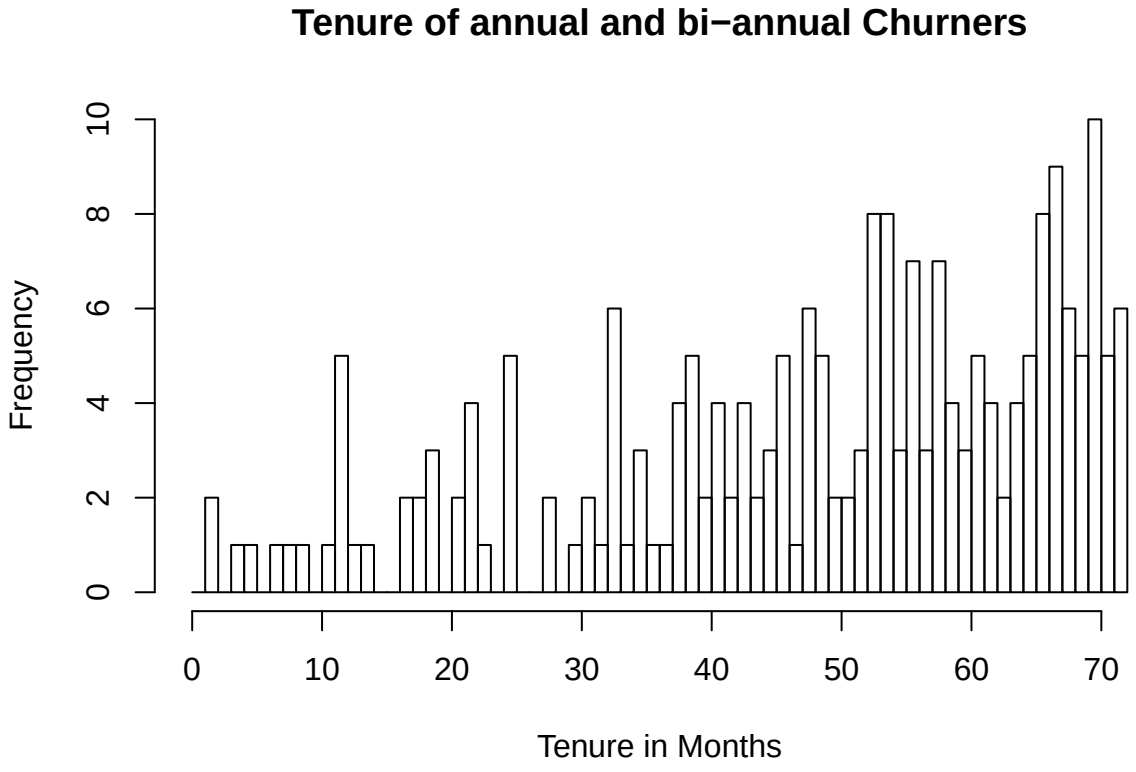
Chisq Test for Contract Type and Churn

Subset the data into customers who had contracts longer than one year just to verify that customers with contracts of 1 year or longer aren't "locked in" and can cancel their service in any month.

```
yr = telco[which(telco$Contract=="One year"|telco$Contract=="Two year"),]
yrc = yr[yr$Churn == "Yes",]
unique(yrc$tenure)
```

```
##  [1] 68 53 59 38 54 43 22 56  2 33 46 62 58 70 60 50 55 12 71 25 19 39  5
## [24] 41 64 61  8 63 17 67 48 66 18 65 69 57 49 37 40 51 34 42 28 45 52  9
## [47] 23 72 31 36  4 35 21 13 32 44 30  7 11 47 14
```

```
# make a separate bar for each month
hist(yrc$tenure, xlab = "Tenure in Months",
     main = "Tenure of annual and bi-annual Churners",
     breaks = c(0:72)*1 )
```



As we can see, there are people churing in every month.

Although there does appear to be a spike in the 12th and 24th months, there appear to be more people churning in every month when tenure >50 than tenure < 10, meaning that many long-tenured customers with long term contracts are churning whether they have time left on their contracts or not.

Conduct Our Chisq test.

```
contract_tab = table(telco$Contract,telco$Churn)
contract_test = chisq.test(contract_tab)
contract_test
```

```
##
##  Pearson's Chi-squared test
##
## data:  contract_tab
## X-squared = 1184.6, df = 2, p-value < 2.2e-16
```

Since our p-value (<2.2e-16) < .05, we reject H0, that churn and contract type are independent at the alpha = .05 level of significance and consider the alternative hypothesis: contract type and churn are NOT independent.

Let's check our observed and expected values to see where the biggest differences occur.

```
  contract_test$observed
```

```
##
##                    No  Yes
##   Month-to-month 2220 1655
##   One year       1307  166
##   Two year       1647   48
```

```
  contract_test$expected
```

```
##
##                         No       Yes
##   Month-to-month 2846.692 1028.3082
##   One year       1082.110  390.8898
##   Two year       1245.198  449.8019
```

Based on the marginal distributions of churn and contract type, we expected to see approximately 1028 monthly contract owners churn, but we ended up observing 1655 monthly contract owners churn, which is roughly 61% as many as we expected, which is interesting, but what's even more surprising is that we expected approximately 391 customers with one year contracts to churn, but we only observed 166 of those customers churh, which is approximately 42.4% of what we expected, which makes us believe that the two variables are related. Most convincingly, we expected to see approximately 450 customers with two year contracts churn, and we only noticed that approximately 10.7% of them churned, which definitely makes us think that churn and contract type are related.

Of course, what we cannot see, is the number of months every user with an one/two year contract has left before their which may be a confounding variable, but, we don't have that information, and it appears that there are people with one/two year contracts in every month, we still think a chi-squared test is valid based on the data we have, although this is something we could check if we find data with more fields.

Answering our Question.

Through our hypothesis tests, we were able to validate that there are statistically significant relationships between Churn and the Contract Type a customer holds, their Monthly Charges and their Tenure.

More specifically, we found that Churn is not independent of contract type, the mean tenure of customers who churned is less than the mean tenure of customers who did not churn and the mean monthly charges incurred by the customers who churned is higher than the mean monthly charges incurred by the customers who churned.

This means that customers with different levels commitment to a company have different probabilties of churning, customers who have longer tenures with a company are less likely to churn, on average, and customers who have larger monthly bills are more likely to churn, on average.

We are interested in seeing how well the three aforementioned variables would be able predict churn in a logistic regression model.