TelecomChurnCapstone

Nirmal Patel

November 5, 2018

Introduction

The telecommunication industry has come a long way since its beginning of being just a phone service industry. Once the telephone became mobile over 45 years ago, technological advances have skyrocketed. This has forced the major companies to accommodate and increase its client base by adding more services to make their store a one-stop for all their technological needs. Over time, telephone companies have had to increase many more products and services. Products induce tablets, watches, smartphones, flip-phones, home-security monitors, and even voice controlled speakers. While the services include Phone, Internet, Online Security, Online Backup, Device Protection, Tech Support, and even Streaming TV/ Movies. The data from these companies can be useful in customer retention in order to minimize the number of customers leaving the company.

Disclaimer

The data provided does not contain any personal information of customers such as name, address, phone number and location. To keep this anonymity a Customer ID number was provided by IBM.

Dataset: https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/

Data

The dataset consists of 147,924 entries with 7044 rows and 21 columns. The column variables and their descriptions are:

Variable	Description
customerID	Customer ID
genderCustomer	Gender (female, male)
SeniorCitizen	Whether the customer is a senior citizen or not (1, 0)
PartnerWhether	The customer has a partner or not (Yes, No)
DependentsWhether	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 or 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 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)
PaperlessBilling	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 monthlyl
TotalCharges	The total amount charged to the customer
ChurnWhether	The customer churned or not (Yes or No)

Data Wrangling

I loaded the dataset as a CSV file and renamed it telecom and added necessary libraries to it. In this section, I looked for outliers, missing values, and if the variable was crucial for the customer churn analysis.

Structure

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..:
5376 3963 2565 5536 6512 6552 1003 4771 5605 4535 ...
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1
2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1
...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2
...
```

```
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2
...
##  $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1
2 1 3 3 2 3 1 ...
##  $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2
2 2 1 2 1 ...
##  $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3
3 3 1 1 1 3 1 3 ...
##  $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1
3 1 1 1 3 1 1 3 ...
##  $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3
1 3 1 3 1 1 3 1 ...
##  $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1
1 3 1 1 1 1 3 1 ...
##  $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1
1 1 1 3 3 1 3 1 ...
##  $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1
1 1 1 3 1 1 3 1 ...
##  $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1
1 1 2 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1
...
##  $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3
4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1
...
```

The structure of the telecom company depicts four variables that are integers or numerical type. It would be beneficial if we look into these and see if there are any NA or blank variables. We can also see that most of the data is well organized and has 2 to 4 factors (options for the data). This data looks clean with well defined variable names.

There are a few variables with factorial variables such as Yes and No

Summary of Telecom Churn Data

```
##       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
```

The summary function gives us a further break down of the variables including the mean(average), minimum(smallest value), median (middle value), maximum(largest value), and if the variable includes a blank value (NA) this function will let us know. If the variable is a factor, the summary method will give us the total of each factor.

Data Exploration

Summary of Telecom Tenure

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00    9.00   29.00   32.37   55.00   72.00
```

We check this for any outliers or NA entries. Since there are no outliers nor NA points, we can move on to the next numerical variable

Summary of Telecom Monthly Charges

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.25   35.50   70.35   64.76   89.85  118.75
```

The monthly charges range from $18.25 to $118.75. There are no blanks (NA’s).

Summary of Telecom Total Charges

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
##    18.8   401.4  1397.5  2283.3  3794.7  8684.8      11
```

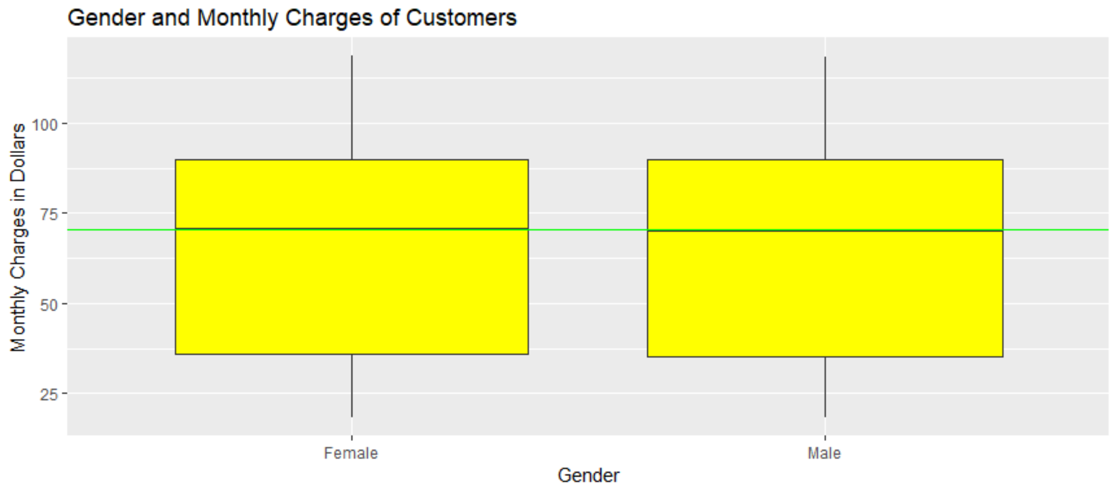
The total charges range from $18.8 to $8684.8, the max seems very high it could possibly be an outlier. This variable also has 11 blank points. Therefore we will have to determine if removing the rows with blanks would be better than keeping them. I decided to keep them in my data for now. but would use the omit function if needed to remove them. We have 11 NA out of 7043 points.

Compare Total and Monthly Charges

```
##    MonthlyCharges TotalCharges TotalDivide12
## 1           29.85        29.85       2.48750
## 2           56.95      1889.50     157.45833
## 3           53.85       108.15       9.01250
## 4           42.30      1840.75     153.39583
## 5           70.70       151.65      12.63750
## 6           99.65       820.50      68.37500
## 7           89.10      1949.40     162.45000
## 8           29.75       301.90      25.15833
## 9          104.80      3046.05     253.83750
## 10          56.15      3487.95     290.66250
```

We need to perform a check to see if the monthly charges are equal to the total charges divided by 12. The number 12 is used because there are 12 months in a year. From the looks of the the total charges column is not uniform and has some charges that might be monthly and some that might be yearly and of various time frames, therefore it makes more sense to omit this column for two reasons: missing values and unstructured method of calculating the total.

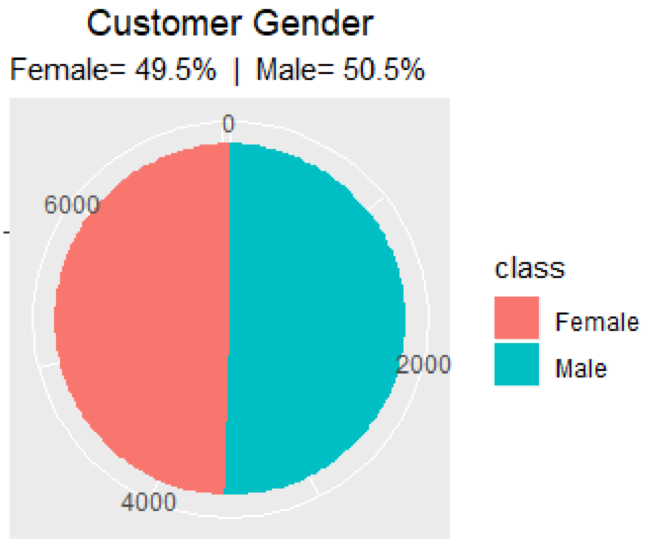
Gender and Monthly Charges



The data shows that both male and female monthly costs were about the same with a median of $70.35 represented by the gtreen line. This seems fair and unbais toward a particular sex getting an immense discount.

gender

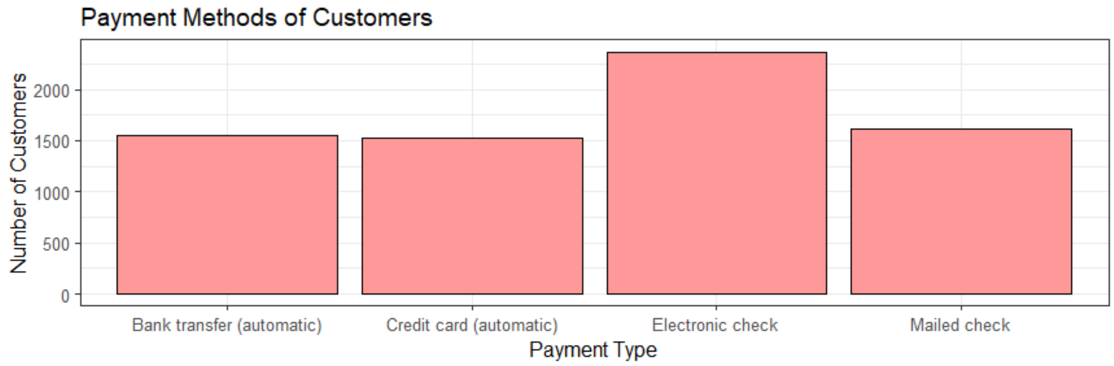
Is this data bias toward men? Are men more likely to have have coverage?



We have 48.5% Female and 50.5% Male in our data.Therefore the data seems normally distributed and large enough to be unbias.

Payment

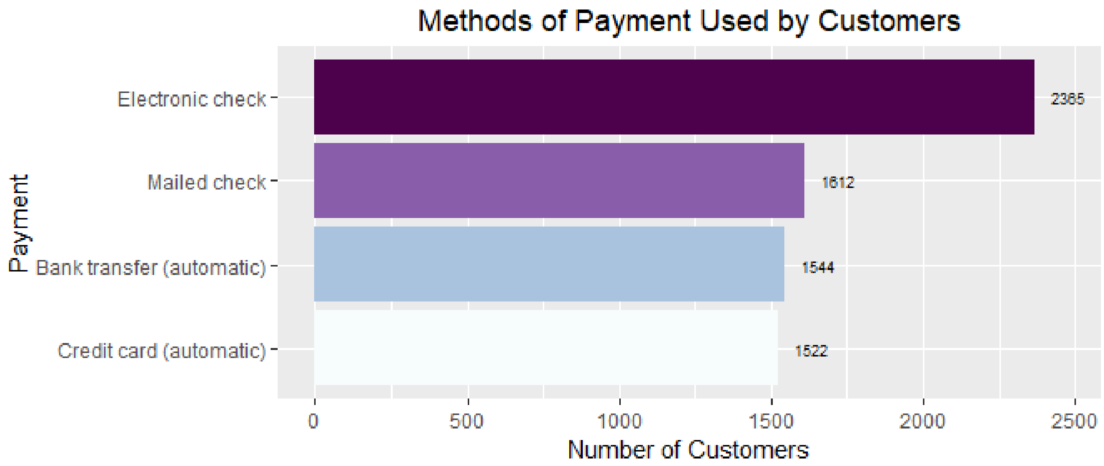
What is the most common form of payment?



```
##
## Bank transfer (automatic)   Credit card (automatic)
##                 0.2192248                 0.2161011
##          Electronic check              Mailed check
##                 0.3357944                 0.2288797
```

The Most common form of payment was Electronic check accounting for 33.5%. Second most common was Mailed check at 22.9%. Third was Bank transfer (automatic) at 22%. Least common was Credit card (automatic) at 21.6%.

Summary of Telecom Payment Methods



Monthly charges

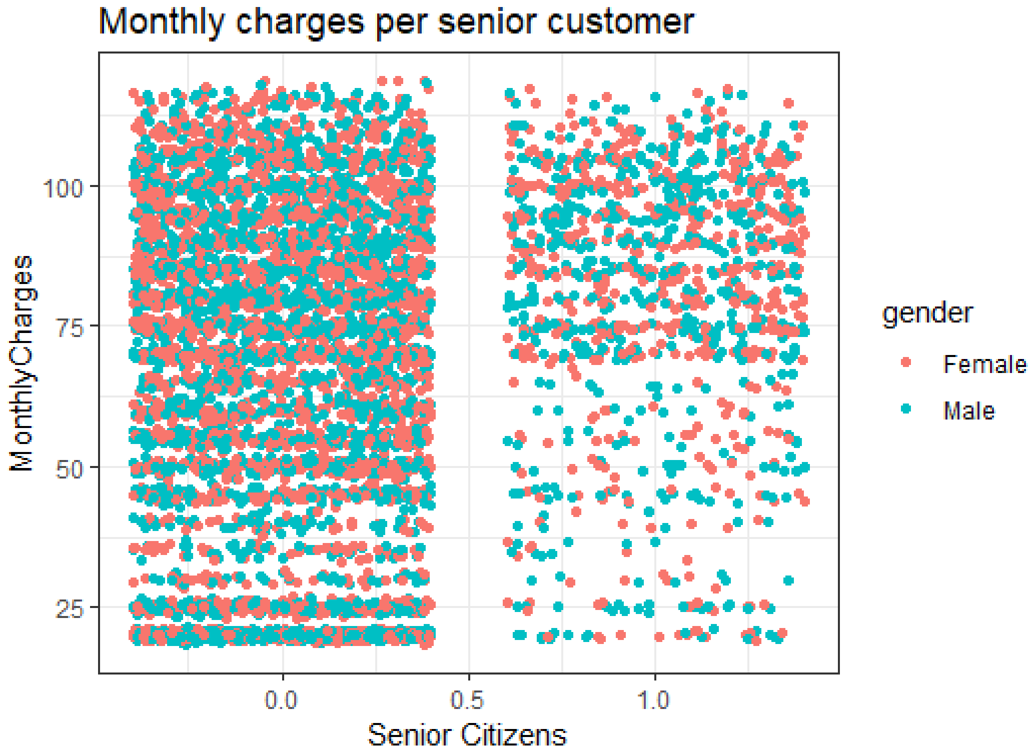
What is the range of monthly charges?

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.25   35.50   70.35   64.76   89.85  118.75
```

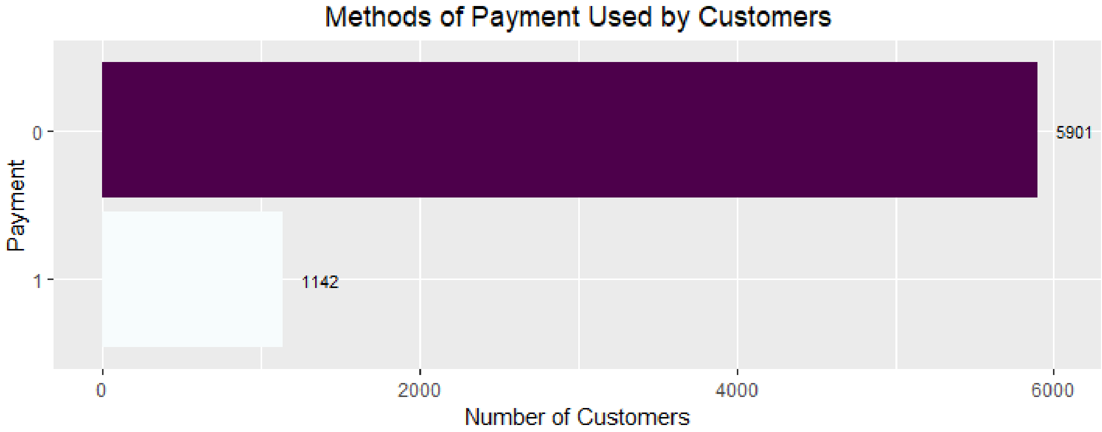
The monthly charges range from $18.25 to $118.75 per month. The mean/average monthly bill is $64.76. While the median bill is $70.35.

Do Senior Citizen's pay less in comparison to everyone else?

Do senior citizens get a discount? Are there more senior customers?



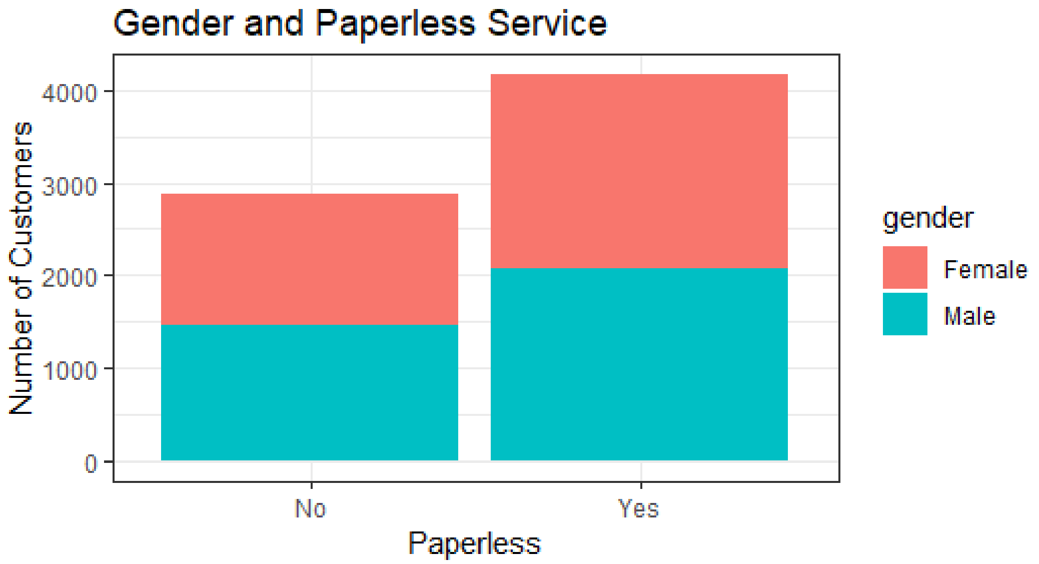
The telecom company seems to have about the same minimum and mazimum as for the Senior citizens as the regular non senior customers; However, the majority of Senior Citizens seem to be paying above ~$70. Therfore it seems that being a senior does not give a bonus to all customers. We would have to do a bar plot to see the difference between number of senior and non senior customers.



There are very few Senior Customers in comparison to non senior customers.

Ecofriendly initative by going paperless

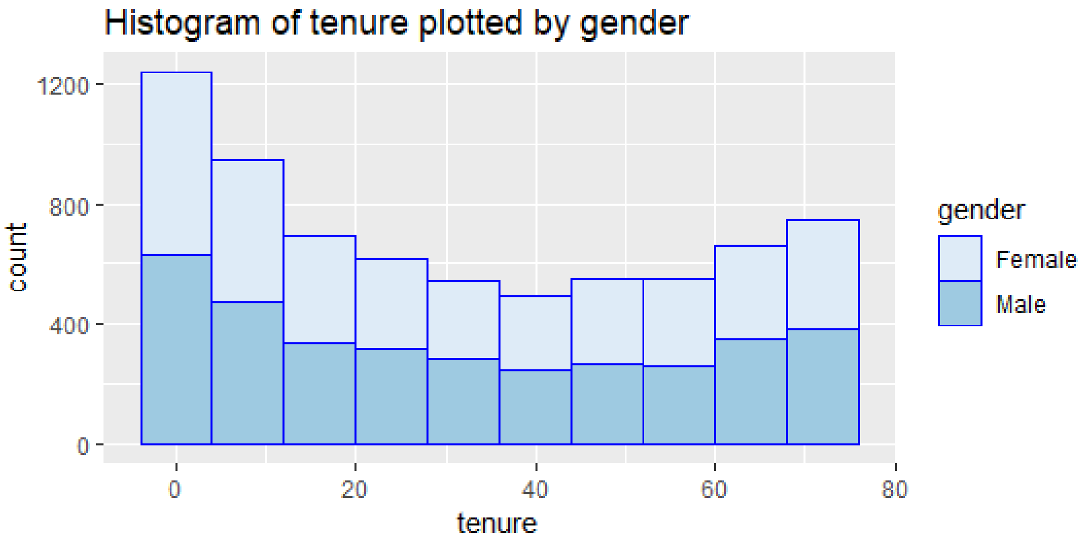
How eco friendly is the brand? Does it have higher percentage of paperless billing?



```
##
##        No       Yes
## 0.4077808 0.5922192
```

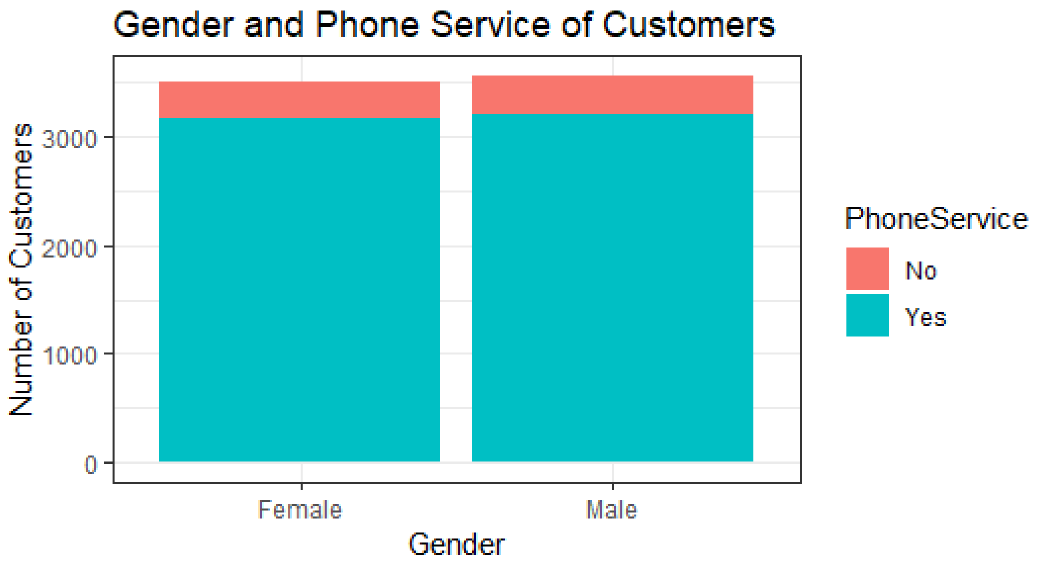
About 59% of customers choose the paperless route, while 41% still want a paper copy of the bill. Though this can be inproved by giving an incentive to go paperless, which would save the company on stationary supplies.

CustomerTenure



There seems to be an equal amount of tenure retention between male and female customers. It also seems that after 40 months the customer is more likely to stay. However from 0 to 40 months it seems that the customer is likely to churn and the comapny should focus on retaining their customers during this period.

Gender and Phoneservice

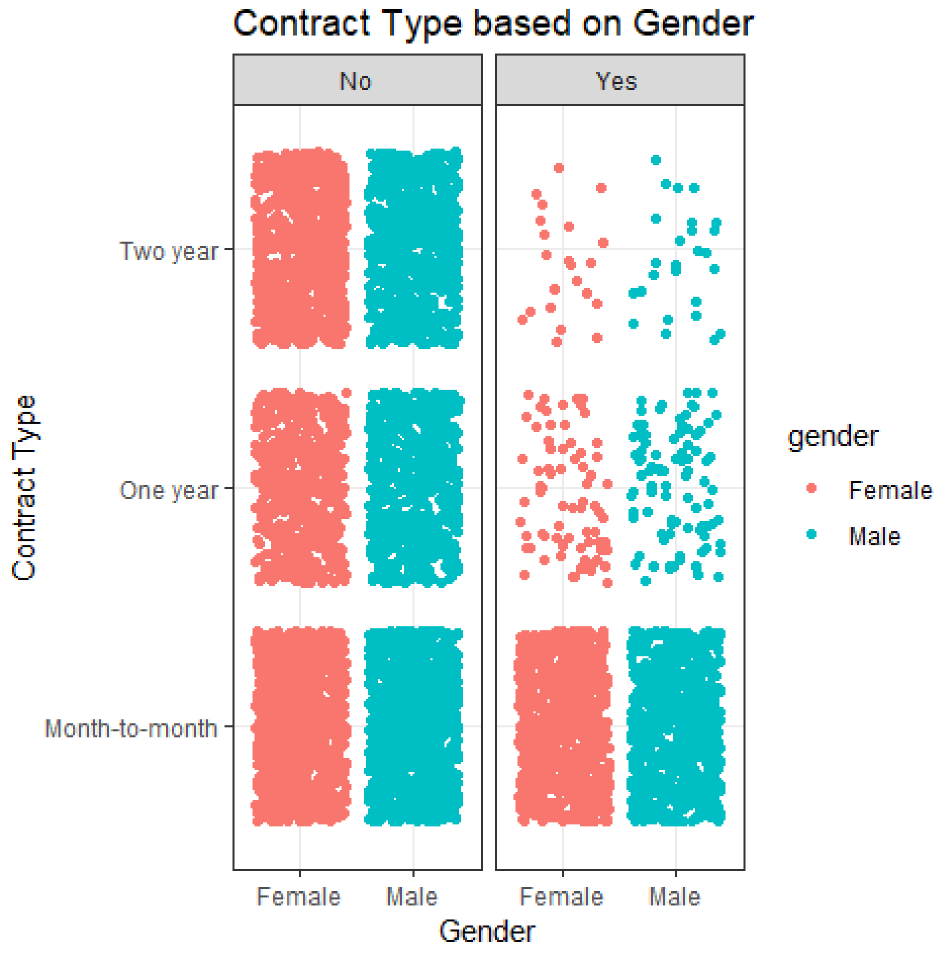


```
##
##                   No        Yes
##   Female 0.04699702 0.44824649
##   Male   0.04983672 0.45491978
```

44% women have a phone service with our company while 4% women do not. 45% male customers have the phone service, while 5% male do not have the phone service.

Gender and Preferred Type of Payment

Do more men or women prefer each type of contract? Month-to-month, One year, Two year?



There seems to be an equal number of male and women per Contract type. However by adding a Churn statistics to this data we see that the month to month customers where most likely to churn while, One year Contractees were less likely to churn and Two year contractees were least likely to churn.

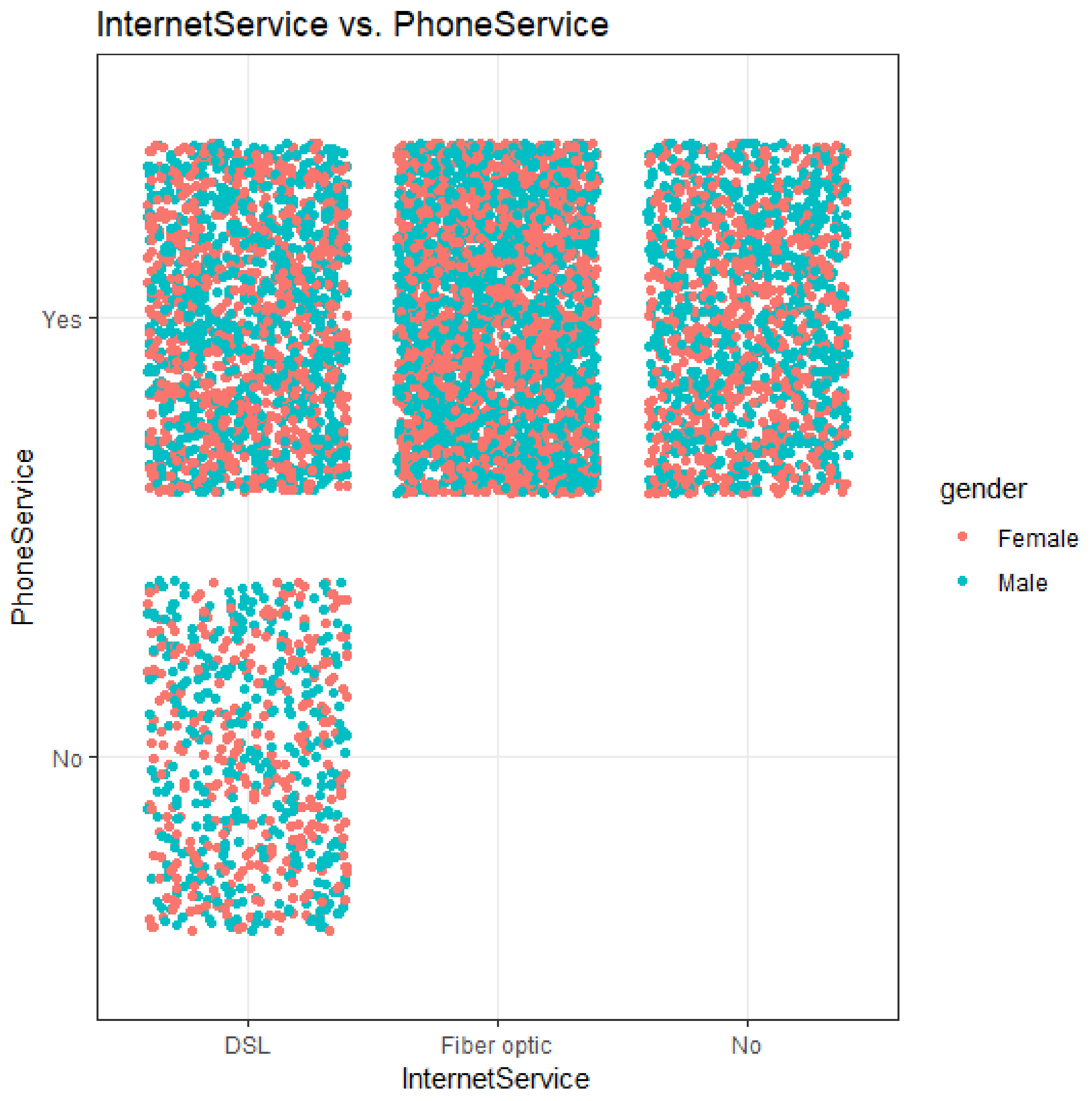
```
## 
##          Month-to-month  One year  Two year
##   Female      0.2733210 0.1019452 0.1199773
##   Male        0.2768707 0.1071986 0.1206872
```

The data shows that about 27% male and 27% female have a month to month contract. While 10% female and 11% male have a One year contract. And 12% females and 12% males have a two year contract.

```
## 
## Month-to-month       One year       Two year 
##      0.5501917      0.2091438      0.2406645
```

The most common contract type is month to month at 55% total, followed by Two year contract at 24%, and least common was one year contract at 21% of total contracts.
##Internet or Phone sells more?

What is more common phone service or internet service?



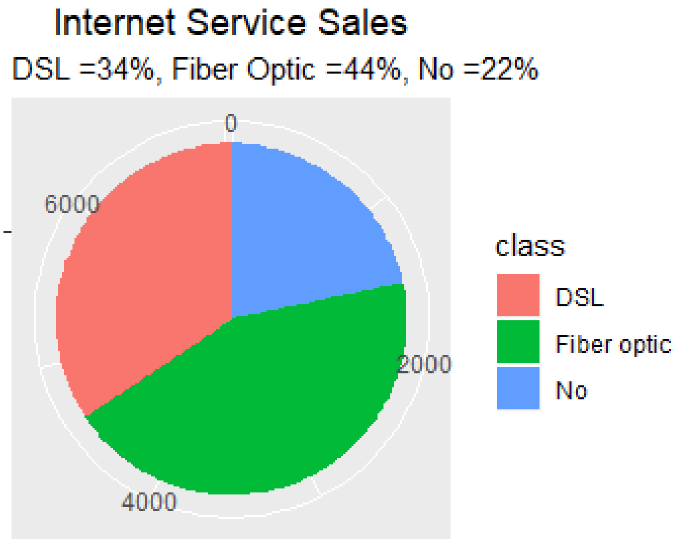
```
##
##                        No        Yes
##   DSL         0.09683374 0.24691183
##   Fiber optic 0.00000000 0.43958540
##   No          0.00000000 0.21666903

##
##         DSL Fiber optic          No
##   0.3437456   0.4395854   0.2166690
```

```
## 
##         No        Yes 
## 0.09683374 0.90316626
```

78.3% of the customers have Internet service while 90% have phone service. Therefore Phone serive is more common. This may lead to the company havingto work more on marketing a better way to increase internet sales.

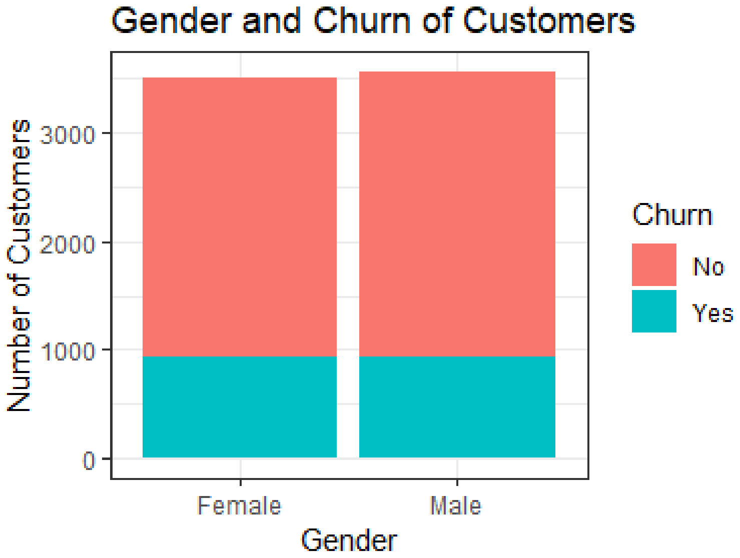
Internet Sales



The internet Service sales indicate that customers chose DSL =34%, Fiber Optic =44%,and No internetv service =22% of the times.

gender and Churn

What percentage of customers stay?

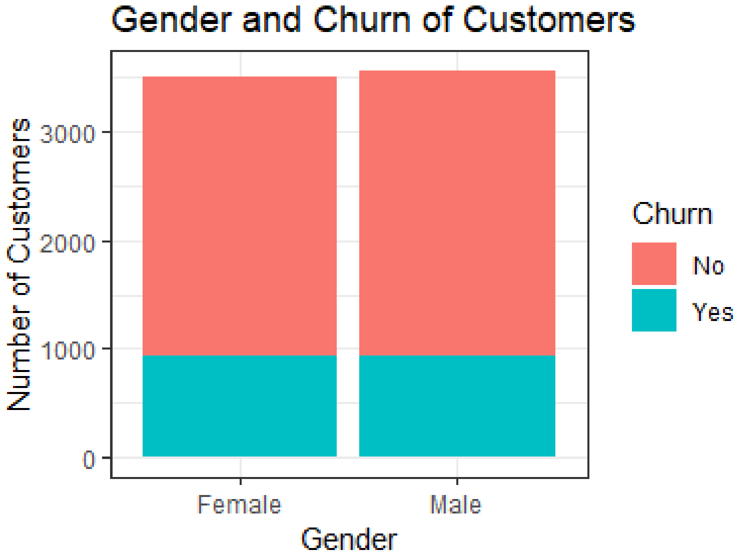


```
##
##                No       Yes
##   Female 0.3619196 0.1333239
##   Male   0.3727105 0.1320460
```

About 26.5% of customers churned of these 13.3% where Female while 13.2% were male. While 73.5% stayed with out telecom company. Of these 36.2% where female while 37.3% where male.

Predicting Churning on Gender

Is it possible to accurately predict which customer would churn based on sex?



```
## 
##                  No       Yes
##   Female 0.3619196 0.1333239
##   Male   0.3727105 0.1320460

## 
##        No       Yes 
## 0.7346301 0.2653699
```

The churn Rate of both men and women seem equal at 13% each. While 36% of female and 37% of male customers will remain with the company. This totals to 73.5% of customers would stay with the company while 26.5% would churn away from the company to another company.

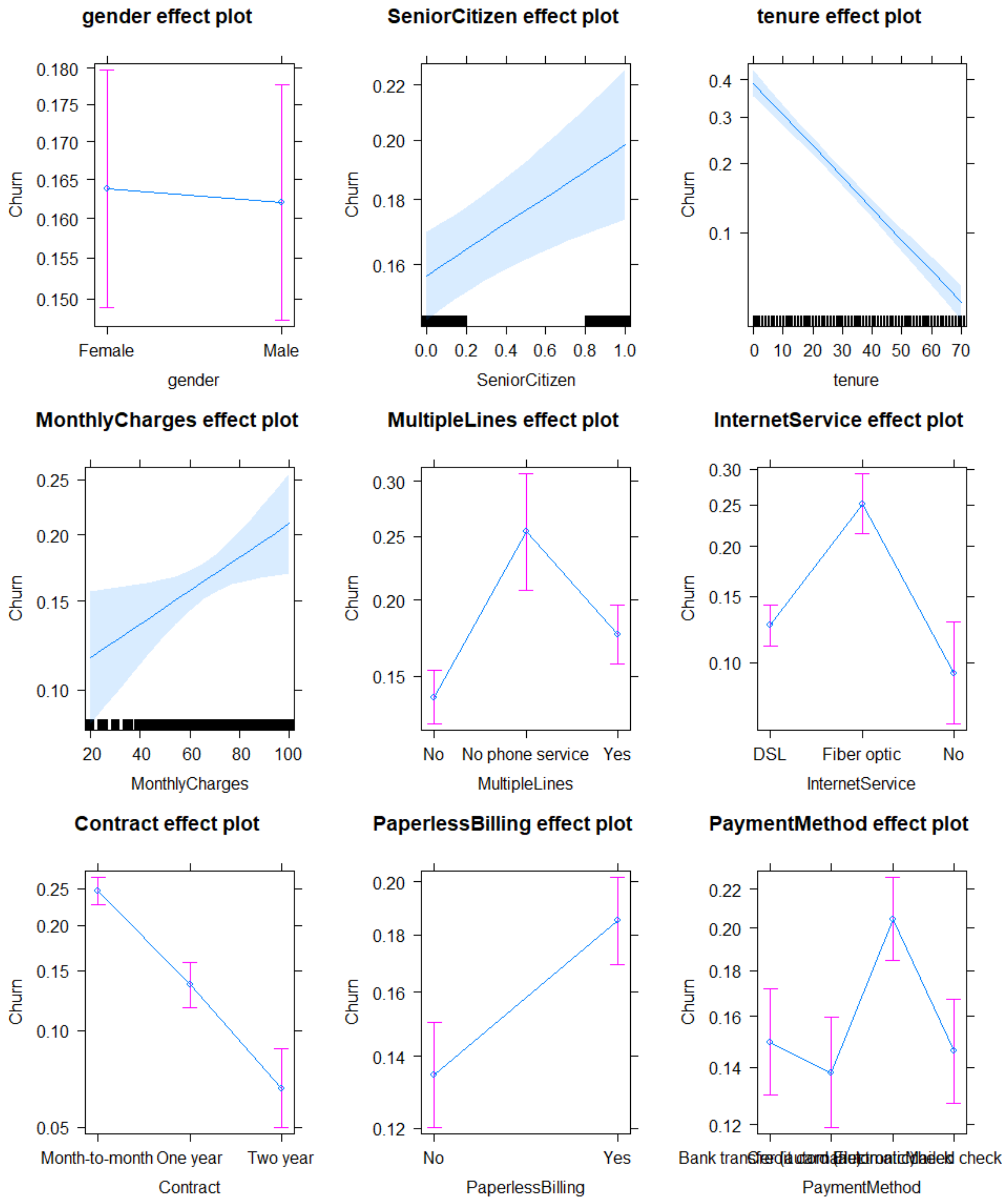
#Machine Learning ##Logistic Regression

```
## 
## Call:
## glm(formula = Churn ~ gender + SeniorCitizen + Partner + Dependents + 
##     tenure + PhoneService + MultipleLines + InternetService + 
##     OnlineSecurity + OnlineBackup + DeviceProtection + TechSupport + 
##     StreamingTV + StreamingMovies + Contract + PaperlessBilling + 
##     PaymentMethod + MonthlyCharges, family = "binomial", data = telecom)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.9780  -0.6707  -0.2946   0.6918   3.1454  
## 
## Coefficients: (7 not defined because of singularities)
##                                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)                          0.612080   0.811986   0.754  0.45097
## genderMale                          -0.020514   0.064885  -0.316  0.75189
## SeniorCitizen                        0.217015   0.084920   2.556  0.01060
## PartnerYes                          -0.002440   0.077741  -0.031  0.97496
## DependentsYes                       -0.167071   0.089678  -1.863  0.06246
## tenure                              -0.034172   0.002366 -14.443  < 2e-16
## PhoneServiceYes                      0.165499   0.652460   0.254  0.79976
## MultipleLinesNo phone service              NA         NA      NA       NA
## MultipleLinesYes                     0.462796   0.178054   2.599  0.00934
## InternetServiceFiber optic           1.720069   0.803709   2.140  0.03234
## InternetServiceNo                   -1.622325   0.811846  -1.998  0.04568
## OnlineSecurityNo internet service          NA         NA      NA       NA
## OnlineSecurityYes                   -0.199497   0.179719  -1.110  0.26698
## OnlineBackupNo internet service            NA         NA      NA       NA
## OnlineBackupYes                      0.049975   0.176251   0.284  0.77676
## DeviceProtectionNo internet service        NA         NA      NA       NA
## DeviceProtectionYes                  0.162576   0.177303   0.917  0.35918
## TechSupportNo internet service             NA         NA      NA       NA
## TechSupportYes                      -0.168836   0.181586  -0.930  0.35248
## StreamingTVNo internet service             NA         NA      NA       NA
## StreamingTVYes                          0.593806   0.328488   1.808  0.07065
## StreamingMoviesNo internet service           NA         NA      NA       NA
## StreamingMoviesYes                      0.608397   0.328840   1.850  0.06429
## ContractOne year                       -0.666321   0.106644  -6.248 4.15e-10
## ContractTwo year                       -1.356836   0.173956  -7.800 6.20e-15
## PaperlessBillingYes                     0.335906   0.074277   4.522 6.12e-06
## PaymentMethodCredit card (automatic) -0.086598   0.114085  -0.759  0.44782
## PaymentMethodElectronic check           0.314319   0.094582   3.323  0.00089
## PaymentMethodMailed check            -0.005299   0.113719  -0.047  0.96283
## MonthlyCharges                       -0.032716   0.031940  -1.024  0.30570
##
## (Intercept)
## genderMale
## SeniorCitizen                        *
## PartnerYes
## DependentsYes                        .
## tenure                               ***
## PhoneServiceYes
## MultipleLinesNo phone service
## MultipleLinesYes                     **
## InternetServiceFiber optic           *
## InternetServiceNo                    *
## OnlineSecurityNo internet service
## OnlineSecurityYes
## OnlineBackupNo internet service
## OnlineBackupYes
## DeviceProtectionNo internet service
## DeviceProtectionYes
## TechSupportNo internet service
## TechSupportYes
## StreamingTVNo internet service
## StreamingTVYes                       .
## StreamingMoviesNo internet service
## StreamingMoviesYes                   .
## ContractOne year                     ***
## ContractTwo year                     ***
## PaperlessBillingYes                  ***
## PaymentMethodCredit card (automatic)
## PaymentMethodElectronic check        ***
## PaymentMethodMailed check
## MonthlyCharges
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 8150.1  on 7042  degrees of freedom
## Residual deviance: 5851.0  on 7020  degrees of freedom
## AIC: 5897
```

```
##
## Number of Fisher Scoring iterations: 6
```

The most significant variables were SeniorCitizen, tenure, MultipleLines, InternetService, Contract, PaperlessBilling, and PaymentMethod.

Predicting Churning on Multiple Variables



Conclusion

The telecom customer churn analysis depicts various interesting results some of which include: 1. Females are slightly more likely to churn
2. Senior Citizens are more likely to churn
3. Higher tenure less likely to churn
4. Higher monthly charges to customers are more likely to churn
5. People with multiple lines but no phone service are more likley to churn in caomparision to people with/without multiple lines
6. Customers with Fiber Optics are more likely to churn in comparison to customers with DSL and no internet service
7. Month to Month customers are most liekly to churn
8. Customers with paperless billing are more likely to churn
9. Customers paying with Electronic Check are most likely to churn in comparison to any other payment method

Though more research would need to be done to see if these trend are specific to this data set or can be used to speak of other telecom data sets as well.