Data Mining and Analytics II

I: Tool Selection:

Data extraction and imported into R by the following code:

```
# Load the Telco Churn Data
Telco_Churn_2 <- read.csv("~/Desktop/r_intro/Telco_Churn.csv")
```

Here it is showing that it was executed successfully:

```
> Telco_Churn_2 <- read.csv("~/Desktop/r_intro/Telco_Churn.csv")
```

```
Data
Telco_Churn_2 7043 obs. of 21 variables
```

There are 7,043 observations and 21 variables in the data.

A. Though there are benefits to each tool for this project, R was chosen for 2 different reasons. (1) R is an open source free program that runs on multiple different types of platforms. Though SAS is a very viable option for this project, the cost outweighs the benefits as SAS is extremely pricey. (2) Unlike other free statistical programs, R is extremely powerful and programable. Also, it has many packages that can be installed and used to analyze the data set. Having access to add on packages and an extensive library of statistical tools not only saves time but allows for accurate calculations and analysis.
B. The following project will focus on the following goals: (1) Describe the data (2) Find trends or patterns that may be present in the data, (3) Identify why customers are leaving and potential indicators to explain why those customers are leaving and mitigate further customer loss.
C. When trying to understand your data it is important to look at it from many different viewpoints. For this reason, a descriptive method and a non-descriptive method will both be used. The main descriptive statistical method that will be used is hierarchical clustering using the k-modes algorithm. Also, as part of the descriptive method the summary(), str(), and visual representations of the descriptive data will be used to give a quick and brief overview of the data set. Both the hierarchal clustering method and the functions give a quick glance at each variable. Through this quick glance we can discover possible meaning or possible abnormalities that can help us analyze the data further and gain a deeper understanding. Hierarchal clustering by the mode is appropriate because, most of the data is categorical and has such responses as "Yes", "No", "Male", "Female" that cannot be handled in other methods that require only numerical data. Also, the hierarchal clustering by the mode can handle mixed data (i.e. handles both categorical and numerical data). Through hierarchal clustering, summary, structure, and visual representations of the data, the project goals will begin to be completed. In the

descriptive method we might start to identify potential indicators of why customers are leaving. However, it is in the non-descriptive method that we will get an even better understanding of the trends and key indicators that are affecting customer churn. The non-descriptive method that will be used is logistic regression by using the glm() function in R. Like the hierarchal clustering method with k-modes, logistic regression can handle categorical data very well while other methods cannot or are not as great at doing so. Which makes logistic regression a perfect fit for our data set that is riddled with categorical data.

II: Data Exploration and Preparation

D. The target variable is "Churn" and is a Nominal categorical binary variable stated as a "yes" or a "no" response. What does that mean? It means that you can group the observations or data by a qualitative measure, a "yes" or "no" response for example. Also, it is nominal because it is unordered. Or in other words there is no value assigned to either yes or no in terms of which is greater or worth more. Both "yes" and "no" are on equal footing. Lastly, Churn is a binary variable because there are only two responses, "yes" and "no". As shown in the summary below, it was imported into the data as a factor with two levels. The summary function also provides the count of each customer per an answer. We can see that out of all the customers present in the data set 1,869 of them churned (left the company).

```
> ## Churn
> str(Telco_Churn_2$Churn)
 Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
> summary(Telco_Churn_2$Churn)
  No  Yes 
5174 1869 
```

E. One of the independent predictor variables is "Contract". Contract is a categorical ordinal variable that is not binary. As with the target variable, Churn, Contract is categorical because of its ability to be grouped. There are three possible levels or answers that a customer can fall into, "Month-to-Month", "One year", and "Two year". However, unlike Churn, Contract has some value to its values which is time. One can clearly state that contract lengths have an identifiable order to them. Contract has also been converted into a factor. The summary function also provides the count of each customer per an answer.

a.

```
> ## Contract
> str(Telco_Churn_2$Contract)
 Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
> summary(Telco_Churn_2$Contract)
Month-to-month       One year       Two year 
          3875           1473           1695 
```

F. See section G and H.

G. In this project it will be attempted to disseminate what kind of customer in a telecommunications company is most likely to churn. The aim in analyzing the telecommunications company data is to (1) discover what type of customer would be most likely to churn, (2) mitigate further customer loss, and (3) to identify potential indicators to why customers are leaving the telecommunications company. One phenomenon that I would like to see if it is present or not is if contract length and/or tenure with the company affects whether a customer is more likely to churn or not.

As for the data set, the majority of the data is qualitative with a few of the variables being quantitative. As stated above the dependent variable in this scenario is binary and categorical. The independent or predictor variables on the other hand, are a combination between binary, continuous and categorical. This is important to know as we go through the data analysis process and will ensure correct methods are used and to ensure that R will produce accurate and desired results.

H. Goals of Data Cleaning are to [1] find and remove missing values and [2] and address any anomalies in the data. Missing values in the data were found in with the following code:

```r
## Find missing values
sapply(Telco_Churn_2, function(x) sum(is.na(x)))
```

Results:

```
> sapply(Telco_Churn_2, function(x) sum(is.na(x))) #TotalCharges Variable has 11 missing values
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService 
               0                0                0                0                0                0                0 
   MultipleLines  InternetService   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV 
               0                0                0                0                0                0                0 
 StreamingMovies         Contract PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn 
               0                0                0                0                0               11                0 
```

The results show that there are 11 missing values in total charges. As part of the cleaning of the data all of the missing values in the TotalCharges variable are removed by applying the following code:

```r
## Remove the 11 missing values in the TotalCharges Variable
Telco_Churn_2 <- Telco_Churn_2[complete.cases(Telco_Churn_2), ]
```

The sapply function is run again and we can now see that the 11 missing values in TotalCharges is no longer present:

```
> sapply(Telco_Churn_2, function(x) sum(is.na(x))) #TotalCharges Variable has 11 missing values
      customerID           gender    SeniorCitizen          Partner       Dependents           tenure     PhoneService 
               0                0                0                0                0                0                0 
   MultipleLines  InternetService   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport      StreamingTV 
               0                0                0                0                0                0                0 
 StreamingMovies         Contract PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges            Churn 
               0                0                0                0                0                0                0 
```

From the str() command we can see a brief overview of each variable and how it has been imported into R:

```
> str(Telco_Churn_2)
'data.frame':	7032 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 ...
```

To analyze the data further, summary() is run on the data set Telco_Churn_2:

```
> summary(Telco_Churn_2)
      customerID      gender     SeniorCitizen     Partner    Dependents     tenure      PhoneService          MultipleLines
 0002-ORFBO:   1   Female:3483   Min.   :0.0000   No :3639   No :4933   Min.   : 1.00   No : 680     No              :3385
 0003-MKNFE:   1   Male  :3549   1st Qu.:0.0000   Yes:3393   Yes:2099   1st Qu.: 9.00   Yes:6352     No phone service: 680
 0004-TLHLJ:   1                 Median :0.0000                         Median :29.00                Yes             :2967
 0011-IGKFF:   1                 Mean   :0.1624                         Mean   :32.42
 0013-EXCHZ:   1                 3rd Qu.:0.0000                         3rd Qu.:55.00
 0013-MHZWF:   1                 Max.   :1.0000                         Max.   :72.00
 (Other)   :7026
    InternetService             OnlineSecurity              OnlineBackup             DeviceProtection
 DSL        :2416   No                 :3497   No                 :3087   No                 :3094
 Fiber optic:3096   No internet service:1520   No internet service:1520   No internet service:1520
 No         :1520   Yes                :2015   Yes                :2425   Yes                :2418



             TechSupport                StreamingTV             StreamingMovies             Contract      PaperlessBilling
 No                 :3472   No                 :2809   No                 :2781   Month-to-month:3875   No :2864
 No internet service:1520   No internet service:1520   No internet service:1520   One year      :1472   Yes:4168
 Yes                :2040   Yes                :2703   Yes                :2731   Two year      :1685



                   PaymentMethod  MonthlyCharges    TotalCharges    Churn
 Bank transfer (automatic):1542   Min.   : 18.25   Min.   :  18.8   No :5163
 Credit card (automatic)  :1521   1st Qu.: 35.59   1st Qu.: 401.4   Yes:1869
 Electronic check         :2365   Median : 70.35   Median :1397.5
 Mailed check             :1604   Mean   : 64.80   Mean   :2283.3
                                  3rd Qu.: 89.86   3rd Qu.:3794.7
                                  Max.   :118.75   Max.   :8684.8
```

A brief look at the summary shows that six variables have the values “Yes, No, & No internet service”. The variable “No internet service” is already present in the variable “InternetService” and does not need to be repeated and should be removed from the other six variables (OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, streamingTV, streamingMovies) as follows:

```
recode_Columns <- c(10:15)
for(i in 1:ncol(Telco_Churn_2[,recode_Columns])) {
  Telco_Churn_2[,recode_Columns][,i] <- as.factor(mapvalues
                                                  (Telco_Churn_2[,recode_Columns][,i], from =c("No internet service"),to=c("No")))
}
```

Also, the variable “Multiplelines” has three values (No phone service, No, Yes). The value “No phone service” is repetitive and is not needed since the value “No” would also include a value of “No Phone Service” for all intents and purposes. Therefore, the value of “No phone service” will be changed to “No” in the variable “Multiplelines” as follows:

```
Telco_Churn_2$MultipleLines <- as.factor(mapvalues(Telco_Churn_2$MultipleLines,
                                                   from=c("No phone service"),
                                                   to=c("No")))
```

To verify that the change took place we will run the following commands and view the data as follows:

Summary(Telco_Churn_2)

```
> summary(Telco_Churn_2)
     customerID       gender      SeniorCitizen      Partner    Dependents      tenure       PhoneService          MultipleLines
 0002-ORFBO:   1   Female:3483   Min.   :0.0000   No :3639   No :4933   Min.   : 1.00   No : 680     No              :3385
 0003-MKNFE:   1   Male  :3549   1st Qu.:0.0000   Yes:3393   Yes:2099   1st Qu.: 9.00   Yes:6352     No phone service: 680
 0004-TLHLJ:   1                 Median :0.0000                         Median :29.00                Yes             :2967
 0011-IGKFF:   1                 Mean   :0.1624                         Mean   :32.42
 0013-EXCHZ:   1                 3rd Qu.:0.0000                         3rd Qu.:55.00
 0013-MHZWF:   1                 Max.   :1.0000                         Max.   :72.00
 (Other)   :7026
    InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies           Contract
 DSL        :2416   No :5017       No :4607     No :4614         No :4992    No :4329    No :4301        Month-to-month:3875
 Fiber optic:3096   Yes:2015       Yes:2425     Yes:2418         Yes:2040    Yes:2703    Yes:2731        One year      :1472
 No         :1520                                                                                        Two year      :1685


 PaperlessBilling                   PaymentMethod  MonthlyCharges    TotalCharges     Churn
 No :2864         Bank transfer (automatic):1542   Min.   : 18.25   Min.   :  18.8   No :5163
 Yes:4168         Credit card (automatic)  :1521   1st Qu.: 35.59   1st Qu.: 401.4   Yes:1869
                  Electronic check         :2365   Median : 70.35   Median :1397.5
                  Mailed check             :1604   Mean   : 64.80   Mean   :2283.3
                                                   3rd Qu.: 89.86   3rd Qu.:3794.7
                                                   Max.   :118.75   Max.   :8684.8
```

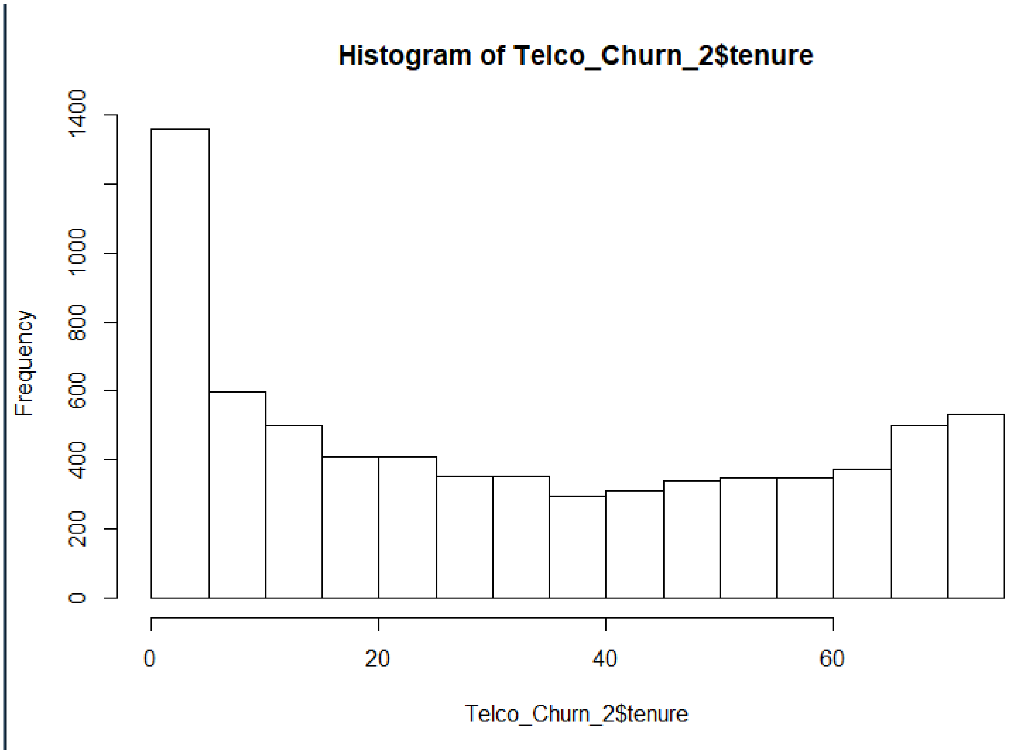
str(Telco_Churn_2)

```
> str(Telco_Churn_2)
'data.frame':	7032 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/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

Next, we will look at the variable “tenure”. Tenure was imported as an integer and indicates the number of months a customer stayed or has been with the telecommunications company. The minimum tenure is 1 month and the maximum is 72.

```
> min(Telco_Churn$tenure); max(Telco_Churn$tenure)
[1] 1
[1] 72
```

A histogram is used to see the frequencies' distribution better:

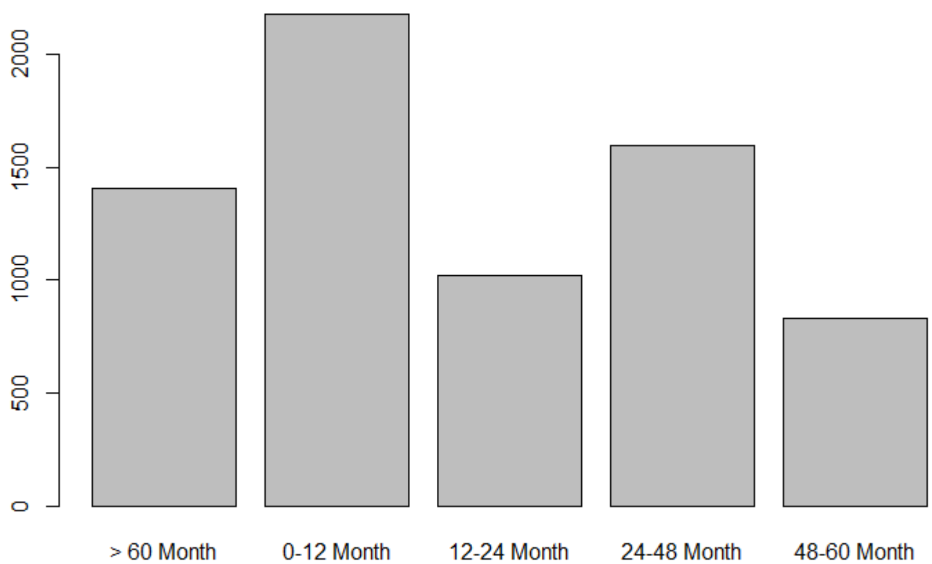


The histogram is revealing on the two tails showing that the majority of the customers have a tenure between 1 – 12 months and 60 + months. However, the middle of the tenure data is still vague and can be visualized better by grouping. The variable "tenure" will be grouped into five groups or bins and a new variable called "tenure_group" will be created which will be used in the data analysis as follows:

```
## Group tenure into 5 groups
group_tenure <- function(tenure){
  if (tenure >= 0 & tenure <= 12){
    return('0-12 Month')
  }else if(tenure > 12 & tenure <= 24){
    return('12-24 Month')
  }else if (tenure > 24 & tenure <= 48){
    return('24-48 Month')
  }else if (tenure > 48 & tenure <=60){
    return('48-60 Month')
  }else if (tenure > 60){
    return('> 60 Month')
  }
}
Telco_Churn_2$tenure_group <- sapply(Telco_Churn_2$tenure,group_tenure)
Telco_Churn_2$tenure_group <- as.factor(Telco_Churn_2$tenure_group)
```

Now if we plot the grouping we get the following result:

```
plot(Telco_Churn_2$tenure_group)
```



With the groupings we are able to see that the greatest number of customers are in the 0-12-month and the 24-48-month categories. With groupings (transforming the data into categorical data), we are able to find deeper meaning in the data.

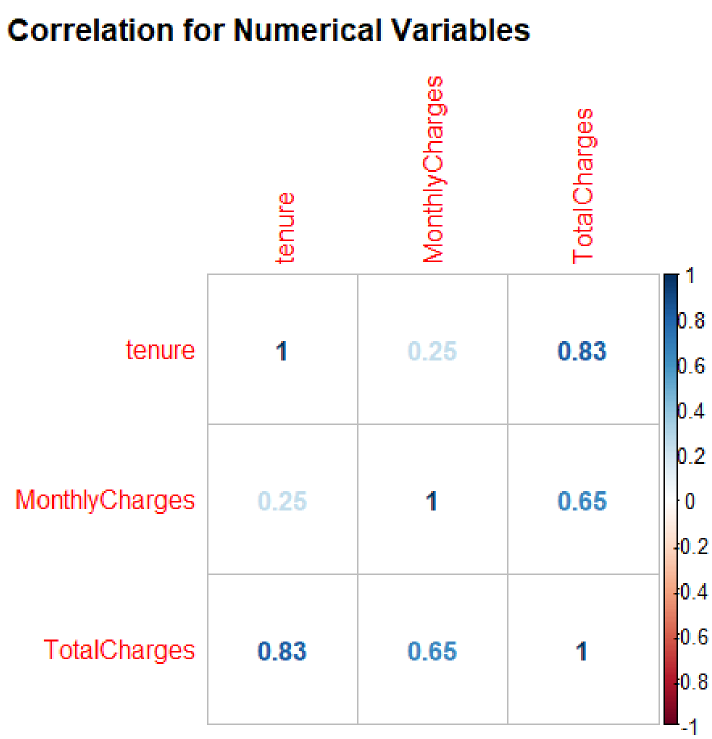
Next, to create uniformity the values in the variable “SeniorCitizen” will be changed from 0 or 1 to No or Yes respectively as follows:

```
# -----Senior Citizen-----
# Senior Citizen: Change Senior Citizen identifier from '0 or 1' to 'No or Yes'
is.numeric(Telco_Churn_2$SeniorCitizen)
is.factor(Telco_Churn_2$SeniorCitizen)
Telco_Churn_2$SeniorCitizen <- as.factor(mapvalues(Telco_Churn_2$SeniorCitizen,
                                                   from=c("0","1"),
                                                   to=c("No", "Yes")))
```

To discover if the numeric variables have correlation or if they are independent the following code is run:

```
# -----Correlation-----
# Discover Correlation between Numneric Variables
numeric_variables <- sapply(Telco_Churn_2, is.numeric)
matrix <- cor(Telco_Churn_2[,numeric_variables])
corrplot(matrix, main="\n\nCorrelation for Numerical Variables", method="number")
```

A correlation matrix is produced showing that a positive correlation between the numeric variables “MonthlyCharges and “TotalCharges” does exist which is shown by the correlation coefficient being at 0.65 (i.e. 1 or -1 indicates a perfectly correlated and 0 would indicate no correlation between the variables is present):



Due to a high correlation that Total charges has with the variables “tenure” and “Monthly Charges”, Total Charges is removed from the data as follows:

```r
# Remove Total Charges due to strong correlation between variables
Telco_Churn_2$TotalCharges <- NULL
```

To finish off our data cleaning and to make the data set a little cleaner and remove unused or unneeded data, the variables “customerID” and “tenure” (The variable “tenure” is removed in favor or the variable “tenure_group”) will be removed as follows:

```r
# Remove columns not needed for analysis
Telco_Churn$customerID <- NULL
Telco_Churn$tenure <- NULL
```

```r
# View Data to ensure that all changes have been successful
View(Telco_Churn_2)
str(Telco_Churn_2)
summary(Telco_Churn_2)
```

III: Data Analysis

I. The following clips will attempt to show through univariate analysis what the distribution of each variable is. To start off, the functions str() and summary() are run to give an overall feel of the cleaned data. There are now 19 variables and 7,032 observations opposed to the 21 variables and 7,043 observations that were in the uncleaned data set (Remember that the ID, Tenure and Total Charges variables were removed and tenure_group was created to replace tenure.

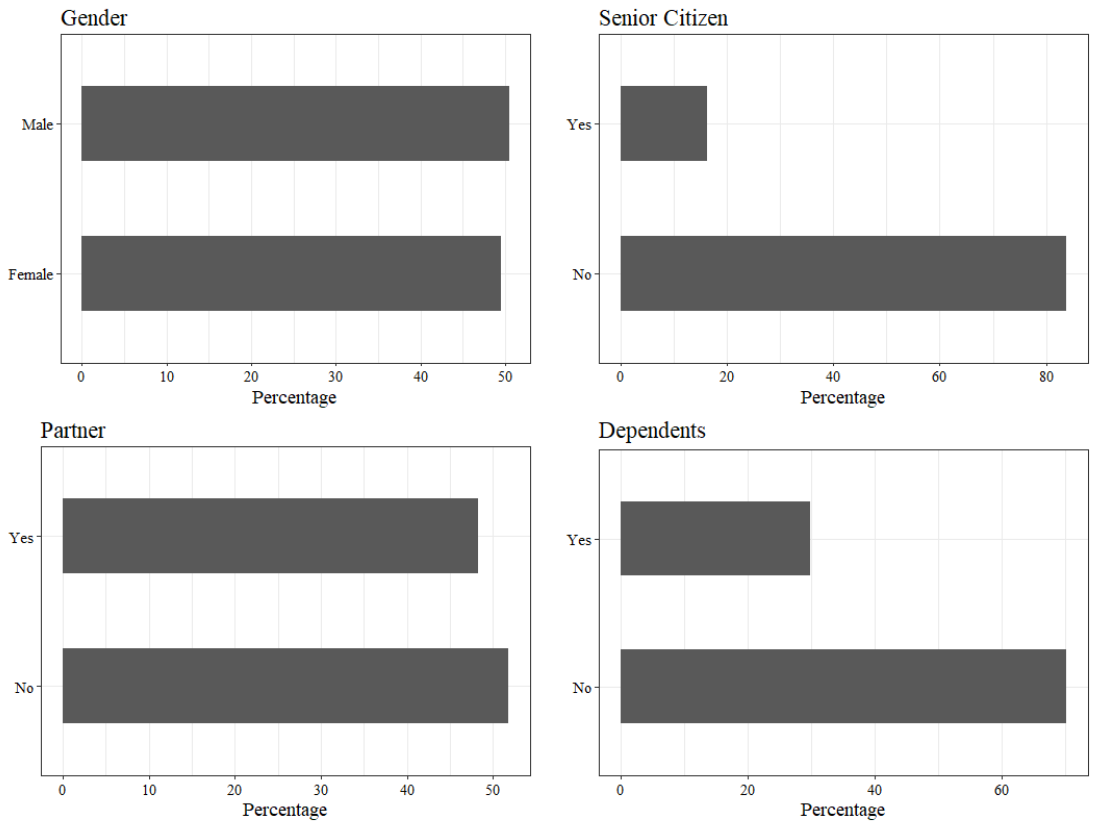
```
> str(Telco_Churn_2)
'data.frame':	7032 obs. of  19 variables:
 $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
 $ SeniorCitizen   : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 1 1 1 1 ...
 $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
 $ 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/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
 $ OnlineSecurity  : Factor w/ 2 levels "No","Yes": 1 2 2 2 1 1 1 2 1 2 ...
 $ OnlineBackup    : Factor w/ 2 levels "No","Yes": 2 1 2 1 1 1 2 1 1 2 ...
 $ DeviceProtection: Factor w/ 2 levels "No","Yes": 1 2 1 2 1 2 1 1 2 1 ...
 $ TechSupport     : Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 2 1 ...
 $ StreamingTV     : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 2 1 2 1 ...
 $ StreamingMovies : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 2 1 1 2 1 ...
 $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
 $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
 $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
 $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
 $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
 $ tenure_group    : Factor w/ 5 levels "> 60 Month","0-12 Month",..: 2 4 2 4 2 2 3 2 4 1 ...
```

```
> summary(Telco_Churn_2)
    gender     SeniorCitizen Partner    Dependents PhoneService MultipleLines    InternetService OnlineSecurity OnlineBackup
 Female:3483   No :5890      No :3639   No :4933   No : 680     No :4065      DSL        :2416   No :5017       No :4607
 Male  :3549   Yes:1142      Yes:3393   Yes:2099   Yes:6352     Yes:2967      Fiber optic:3096   Yes:2015       Yes:2425
                                                                              No         :1520


 DeviceProtection TechSupport StreamingTV StreamingMovies           Contract     PaperlessBilling                   PaymentMethod
 No :4614         No :4992    No :4329    No :4301        Month-to-month:3875   No :2864         Bank transfer (automatic):1542
 Yes:2418         Yes:2040    Yes:2703    Yes:2731        One year      :1472   Yes:4168         Credit card (automatic)  :1521
                                                          Two year      :1685                    Electronic check         :2365
                                                                                                 Mailed check             :1604


 MonthlyCharges    Churn          tenure_group
 Min.   : 18.25   No :5163   > 60 Month :1407
 1st Qu.: 35.59   Yes:1869   0-12 Month :2175
 Median : 70.35              12-24 Month:1024
 Mean   : 64.80              24-48 Month:1594
 3rd Qu.: 89.86              48-60 Month: 832
 Max.   :118.75
```

The majority of the variables are categorical. Bar plots were created for 18 variables (categorical) to identify the distribution by percentage (Li, 2017).

```
# Bar plot Theme
windowsFonts()
theme_new <- theme_bw() +
  theme(plot.background = element_rect(size = 1, color = "white", fill = "white"),
        text=element_text(size = 14, family = "serif", color = "black"),
        axis.text.y = element_text(colour = "black"),
        axis.text.x = element_text(colour = "black"),
        panel.background = element_rect(fill = "white"),
        strip.background = element_rect(fill = ("gray")))
```

```
# Bar PLots 1, Variables: Gender, Senior Citizen, Partner, Dependants
"Shows percentages of each response per a variable. Trying to Determine distribution of variables"
p1 <- ggplot(Telco_Churn_2, aes(x=gender)) + ggtitle("Gender") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p2 <- ggplot(Telco_Churn_2, aes(x=SeniorCitizen)) + ggtitle("Senior Citizen") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() +theme_new
p3 <- ggplot(Telco_Churn_2, aes(x=Partner)) + ggtitle("Partner") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p4 <- ggplot(Telco_Churn_2, aes(x=Dependents)) + ggtitle("Dependents") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p1, p2, p3, p4, ncol=2)
```



<u>Gender</u> (*Frequency & Percentage*):

- Nearly a 50 percent split.

```
> table(Telco_Churn_2$gender)

Female   Male
  3483   3549
> table(Telco_Churn_2$gender)/length(Telco_Churn_2$gender)

   Female      Male
0.4953072 0.5046928
```

<u>Senior Citizen</u> (*Frequency & Percentage*):

- Significantly more Non-Senior Citizens

```
> table(Telco_Churn_2$SeniorCitizen)

  No  Yes
5890 1142
> table(Telco_Churn_2$SeniorCitizen)/length(Telco_Churn_2$SeniorCitizen)

       No       Yes
0.8375995 0.1624005
```

<u>Partner</u> (*Frequency & Percentage*):

- Nearly a 50 percent split.

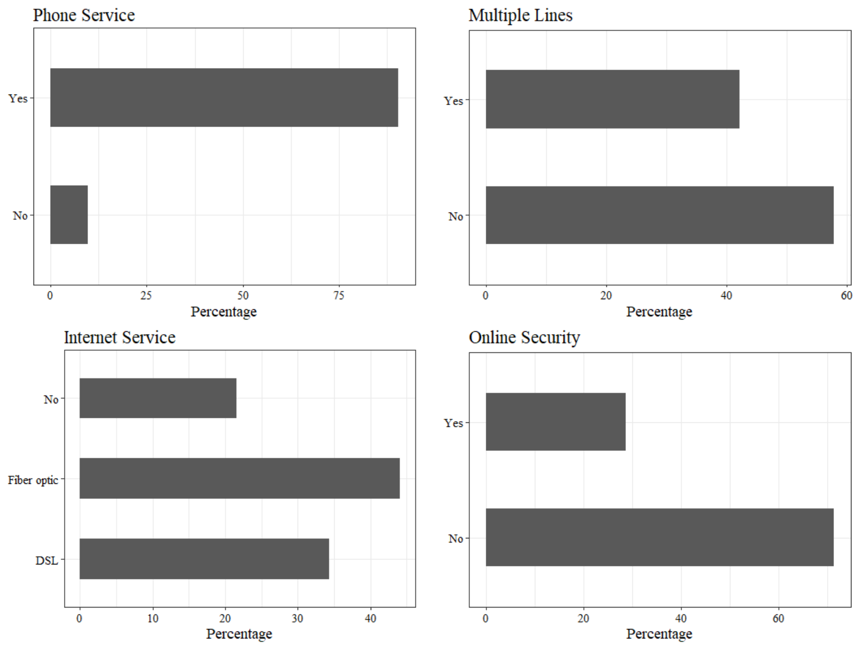
```
> table(Telco_Churn_2$Partner)

  No  Yes
3639 3393
> table(Telco_Churn_2$Partner)/length(Telco_Churn_2$Partner

       No       Yes
0.5174915 0.4825085
```

<u>Dependents</u> (*Frequency & Percentage*):

- 70 percent of customers do not have dependents

```
> table(Telco_Churn_2$Dependents)

  No  Yes
4933 2099
> table(Telco_Churn_2$Dependents)/length(Telco_Churn_2$Dependents)

       No       Yes
0.7015074 0.2984926
```

```
# Bar Plot 2, Variables: PHone Serivce, Mulptiple Lines, Internet Servic, Online Security.
p5 <- ggplot(Telco_Churn_2, aes(x=PhoneService)) + ggtitle("Phone Service") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p6 <- ggplot(Telco_Churn_2, aes(x=MultipleLines)) + ggtitle("Multiple Lines") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p7 <- ggplot(Telco_Churn_2, aes(x=InternetService)) + ggtitle("Internet Service") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p8 <- ggplot(Telco_Churn_2, aes(x=OnlineSecurity)) + ggtitle("Online Security") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p5, p6, p7, p8, ncol=2)
```



<u>Phone Service</u> (*Frequency & Percentage*):

- 96.7 percent of customers have phone service

```
> table(Telco_Churn_2$PhoneService)

  No  Yes
 680 6352
> table(Telco_Churn_2$PhoneService)/length(Telco_Churr

       No       Yes
0.0967008 0.9032992
```

<u>Multiple Lines</u> (*Frequency & Percentage*):

- 57 percent of customer do not have multiple lines.

```
> table(Telco_Churn_2$MultipleLines)

  No  Yes
4065 2967
> table(Telco_Churn_2$MultipleLines)/length(Telco_Churn_2$MultipleLin

       No       Yes
0.5780717 0.4219283
```

<u>Internet Service</u> (*Frequency & Percentage*):

- 44 percent of customers have Fiber optic

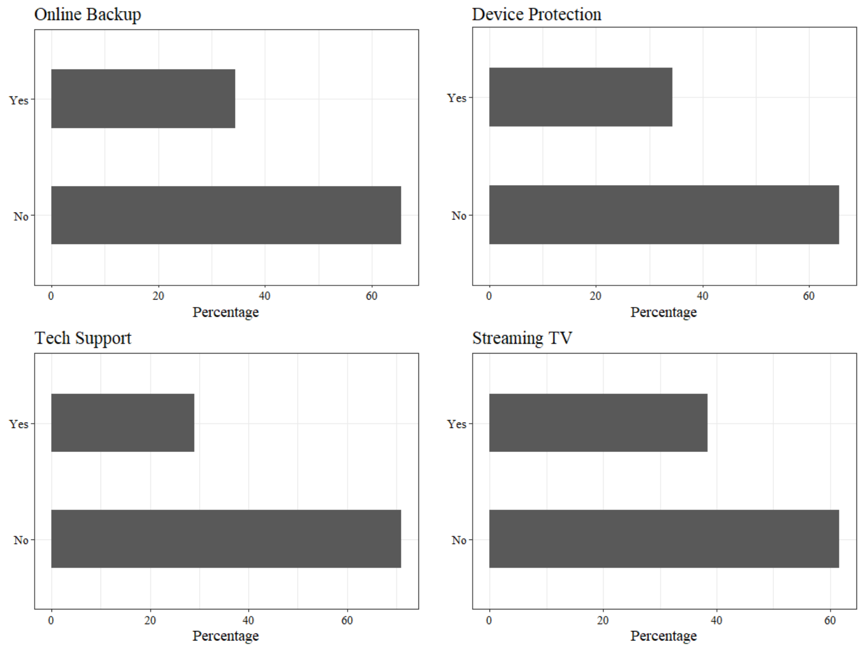
```
> table(Telco_Churn_2$InternetService)

        DSL Fiber optic          No
       2416        3096        1520
> table(Telco_Churn_2$InternetService)/length(Telco

        DSL Fiber optic          No
  0.3435722   0.4402730   0.2161547
```

<u>Online Security</u> (*Frequency & Percentage*):

- 71.3 percent of customers do not have online security.

```
> table(Telco_Churn_2$OnlineSecurity)

  No  Yes
5017 2015
> table(Telco_Churn_2$OnlineSecurity)/length(Telco_Churn_2$Online

       No       Yes
0.7134528 0.2865472
```

```
# Bar PLots 3, Variables: Online Backup, Device Protection, Tech Support, Streaming TV
p9 <- ggplot(Telco_Churn_2, aes(x=OnlineBackup)) + ggtitle("Online Backup") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p10 <- ggplot(Telco_Churn_2, aes(x=DeviceProtection)) + ggtitle("Device Protection") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p11 <- ggplot(Telco_Churn_2, aes(x=TechSupport)) + ggtitle("Tech Support") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p12 <- ggplot(Telco_Churn_2, aes(x=StreamingTV)) + ggtitle("Streaming TV") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p9, p10, p11, p12, ncol=2)
```



Online Backup (*Frequency & Percentage*):

- 65.5 percent do not have online backup

```
> table(Telco_Churn_2$OnlineBackup)

  No  Yes
4607 2425
> table(Telco_Churn_2$OnlineBackup)/length(Telco_C

       No       Yes
0.6551479 0.3448521
```

Device Protection (*Frequency & Percentage*):

- 65.6 percent do not have device protection

```
> table(Telco_Churn_2$DeviceProtection)

  No  Yes
4614 2418
> table(Telco_Churn_2$DeviceProtection)/length(Telco_Churn_2$

       No       Yes
0.6561433 0.3438567
```

Tech Support (*Frequency & Percentage*):

- 71 percent do not have tech support.

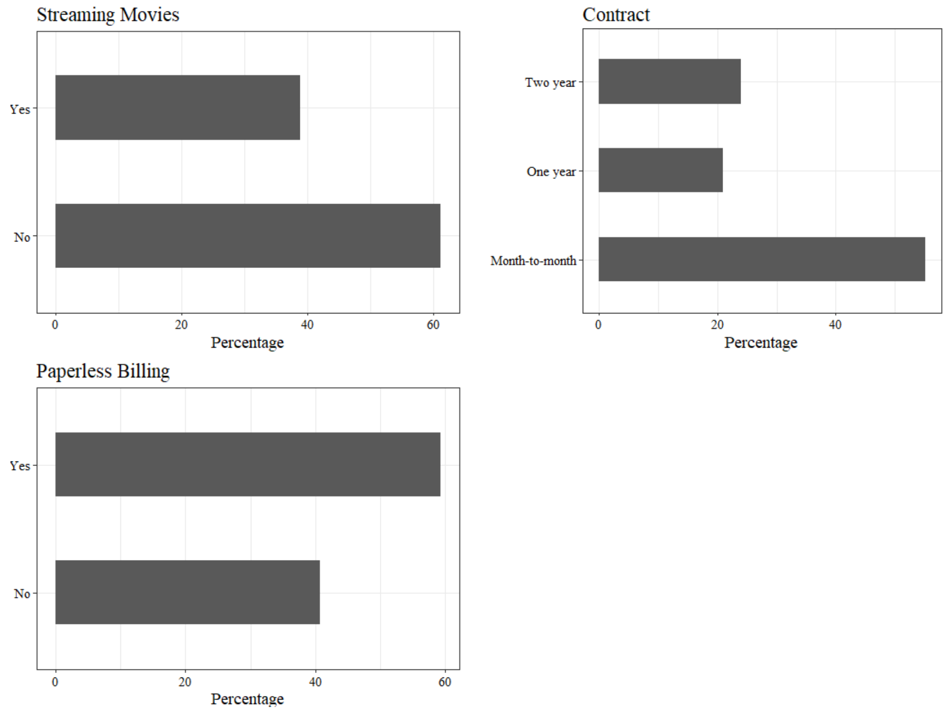
```
> table(Telco_Churn_2$TechSupport)

  No  Yes
4992 2040
> table(Telco_Churn_2$TechSupport)/length(Telco_Churn_

       No       Yes
0.7098976 0.2901024
```

Streaming TV (*Frequency & Percentage*):

- 61.6 percent do not have Streaming TV

```
> table(Telco_Churn_2$StreamingTV)

  No  Yes
4329 2703
> table(Telco_Churn_2$StreamingTV)/length(Telco_Churn_2$StreamingTV)

       No       Yes
0.6156143 0.3843857
```

```
#Bar Plots 4, Variables: Streaming Moveis, Contact, Paperless Billing, Payment Method, Tenure Group
p13 <- ggplot(Telco_Churn_2, aes(x=StreamingMovies)) + ggtitle("Streaming Movies") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p14 <- ggplot(Telco_Churn_2, aes(x=Contract)) + ggtitle("Contract") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
p15 <- ggplot(Telco_Churn_2, aes(x=PaperlessBilling)) + ggtitle("Paperless Billing") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p13, p14, p15, ncol=2)
```



<u>Streaming Movies</u> (*Frequency & Percentage*):

- 61.2 percent do not have Streaming movies

```
> table(Telco_Churn_2$StreamingMovies)

  No  Yes
4301 2731
> table(Telco_Churn_2$StreamingMovies)/length(

       No       Yes
0.6116325 0.3883675
```

<u>Paperless Billing</u> (*Frequency & Percentage*):

- 59.3 percent have paperless billing.

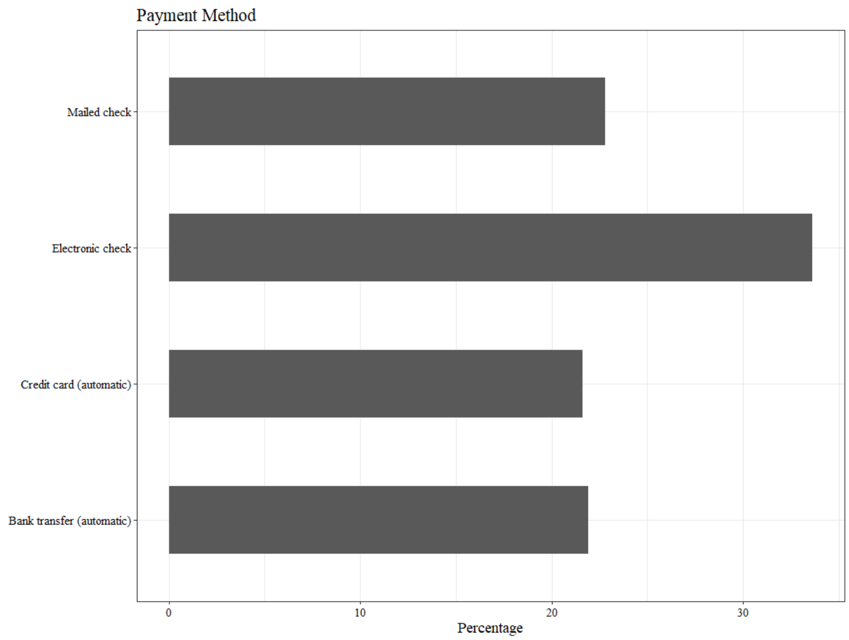
```
> table(Telco_Churn_2$PaperlessBilling)

  No  Yes
2864 4168
> table(Telco_Churn_2$PaperlessBilling)/length(Te

      No      Yes
0.407281 0.592719
```

<u>Contract</u> (*Frequency & Percentage*):

- 55.1 percent have a month-to-month contract.

```
> table(Telco_Churn_2$Contract)

Month-to-month       One year       Two year
          3875           1472           1685
> table(Telco_Churn_2$Contract)/length(Telco_Churn_2$Cont

Month-to-month       One year       Two year
     0.5510523      0.2093288      0.2396189
```

```
#Bar Plots 6, Variables: Tenure Group
p17 <- ggplot(Telco_Churn_2, aes(x=tenure_group)) + ggtitle("Tenure Group") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p17, ncol=1)
```

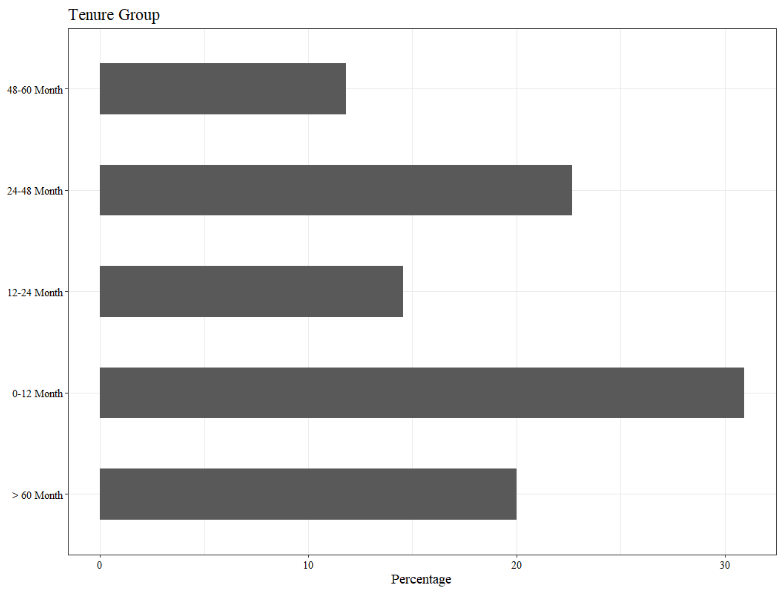


<u>Payment Method</u> (*Frequency & Percentage*):

- Most payment methods seem to be somewhat equal. However, electronic check is nearly 11 percent higher then the rest of the observations sitting at 33.6 percent.

```
> table(Telco_Churn_2$PaymentMethod)

Bank transfer (automatic)   Credit card (automatic)          Electronic check              Mailed check
                     1542                      1521                      2365                      1604
> table(Telco_Churn_2$PaymentMethod)/length(Telco_Churn_2$PaymentMethod)

Bank transfer (automatic)   Credit card (automatic)          Electronic check              Mailed check
                0.2192833                 0.2162969                 0.3363197                 0.2281001
```

```
#Bar Plots 6, Variables: Tenure Group
p17 <- ggplot(Telco_Churn_2, aes(x=tenure_group)) + ggtitle("Tenure Group") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p17, ncol=1)
```

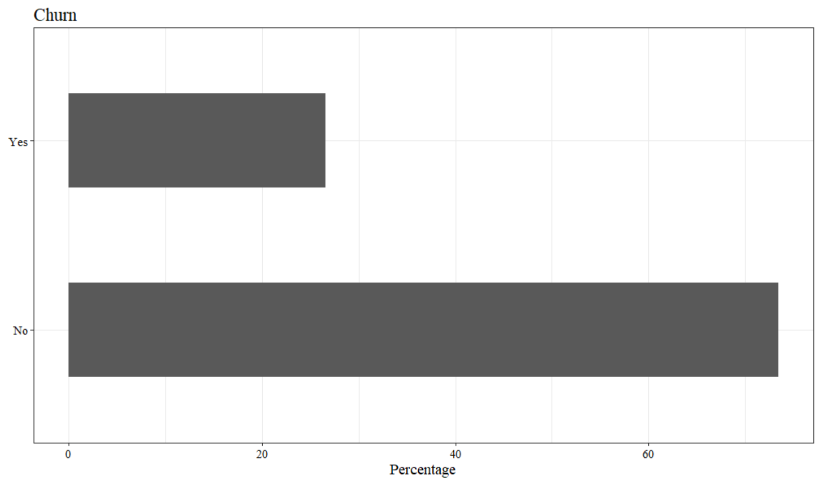


<u>**Tenure Group**</u> (*Frequency & Percentage*):

- The top three tenure groups as far as number of customers are (1) 0-12 month at 30.9 percent, (2) 24-48 month at 22.7 percent and, (3) > 60 month at 20 percent.

```
> # Tenure Group
> table(Telco_Churn_2$tenure_group)

 > 60 Month  0-12 Month 12-24 Month 24-48 Month 48-60 Month
      1407        2175        1024        1594         832
> table(Telco_Churn_2$tenure_group)/length(Telco_Churn_2$tenure_group)

 > 60 Month  0-12 Month 12-24 Month 24-48 Month 48-60 Month
 0.2000853   0.3093003   0.1456200   0.2266780   0.1183163
```

```
#Bar Plots 6, Variables: Churn
p18 <- ggplot(Telco_Churn_2, aes(x=Churn)) + ggtitle("Churn") + xlab("") +
  geom_bar(aes(y = 100*(..count..)/sum(..count..)), width = 0.5) + ylab("Percentage") + coord_flip() + theme_new
grid.arrange(p18, ncol=1)
hist(Telco_Churn_2$MonthlyCharges)
```



<u>Churn</u> (*Frequency & Percentage*):

- Out of the 7,032 customers that we are analyzing 26.6 percent of them left (churned) the telecommunications company.

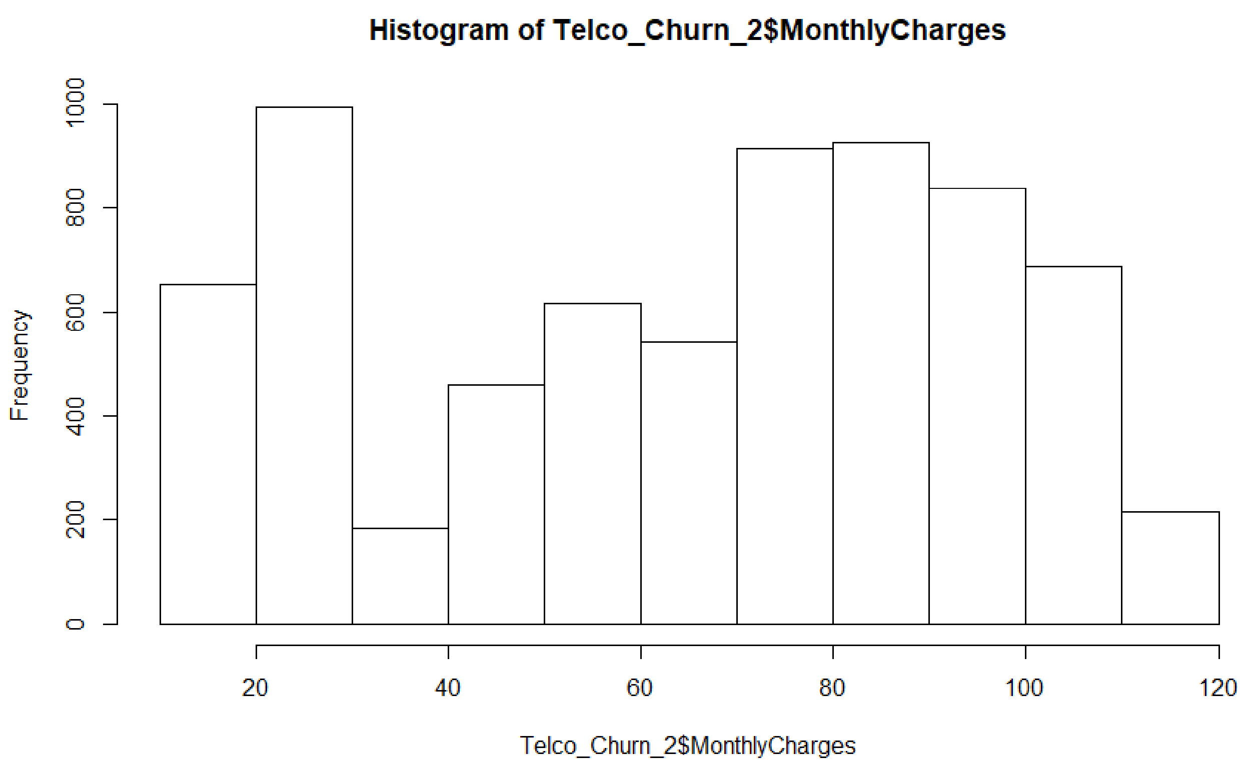
```
> # Churn
> table(Telco_Churn_2$Churn)

  No  Yes 
5163 1869 
> table(Telco_Churn_2$Churn)/length(Telco_Churn_2$Churn)

      No      Yes 
0.734215 0.265785 
```

Monthly Charges:

- The variable Monthly Charges is continuous. Therefore, the distribution will be shown differently than the categorical variables. Below, the head function was run to call the first 6 variables. They are quite diverse and seem random and continuous.

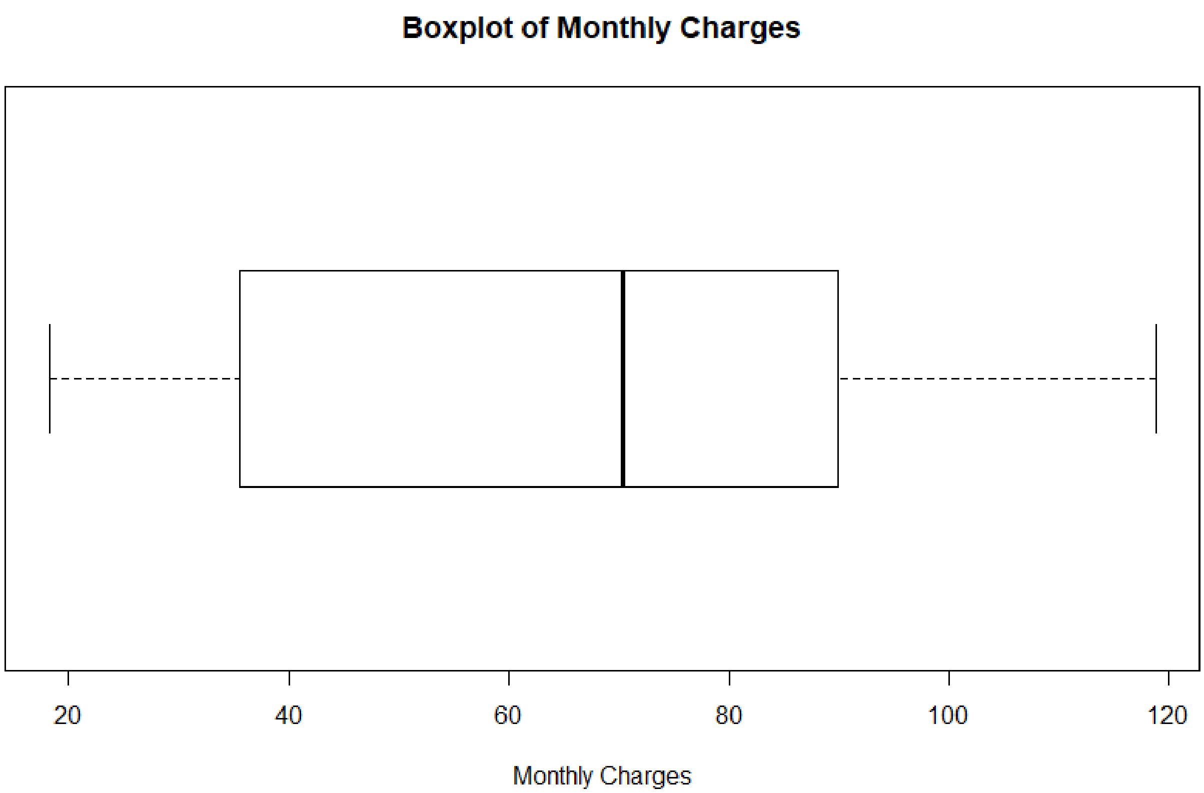
```
> head(Telco_Churn_2$MonthlyCharges)
[1] 29.85 56.95 53.85 42.30 70.70 99.65
> mean(Telco_Churn_2$MonthlyCharges)
[1] 64.79821
> median(Telco_Churn_2$MonthlyCharges)
[1] 70.35
> var(Telco_Churn_2$MonthlyCharges)
[1] 905.1658
> sd(Telco_Churn_2$MonthlyCharges)
[1] 30.08597
> range(Telco_Churn_2$MonthlyCharges)
[1]  18.25 118.75
```

Next, a histogram of Monthly Charges is run. From the histogram, we can start to see that Monthly Charges is slightly left skewed. The left skewedness is also confirmed by the mean being slightly less than the median.



To show the left skewness further a boxplot of Monthly charges is run.

```
boxplot(Telco_Churn_2$MonthlyCharges, horizontal = TRUE,
        main = "Boxplot of Monthly Charges", xlab = "Monthly Charges")
```



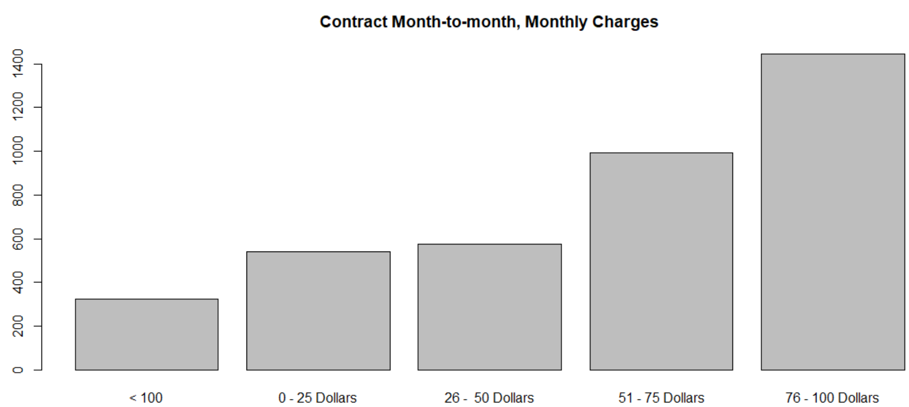

```
> quantile(Telco_Churn_2$MonthlyCharges)
      0%      25%      50%      75%     100%
 18.2500  35.5875  70.3500  89.8625 118.7500
```

```
 MonthlyCharges
 Min.   : 18.25
 1st Qu.: 35.59
 Median : 70.35
 Mean   : 64.80
 3rd Qu.: 89.86
 Max.   :118.75
```

J. **<u>Contract Length vs Monthly Charges Analysis</u>**

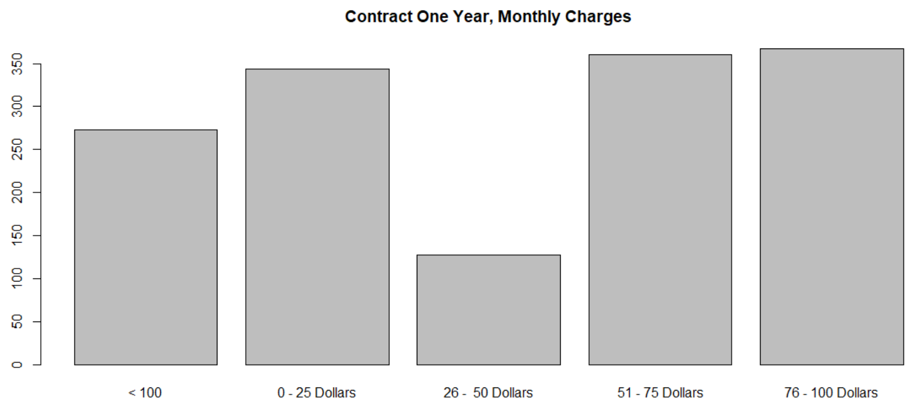
We will now use bivariate statistics to analyze the data further and compare two variables at a time. We will start with the variables "Contract" and "Monthly Charges". The variable contract in the next three bar plots is broken into its three subgroups "Month-to-Month", "One-year", and "Two-year". The code used is below:

```
## PLots Contract Length Vs Monthly Charges
MonthtoMonth_Contract <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "Month-to-month")
plot(MonthtoMonth_Contract$MonthlyCharges_group, main = "Contract Month-to-month, Monthly Charges")

OneYear_Contract <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "One year")
plot(OneYear_Contract$MonthlyCharges_group, main = "Contract One Year, Monthly Charges")

TwoYear_Contract <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "Two year")
plot(TwoYear_Contract$MonthlyCharges_group, main = "Contract Two Year, Monthly Charges")
```

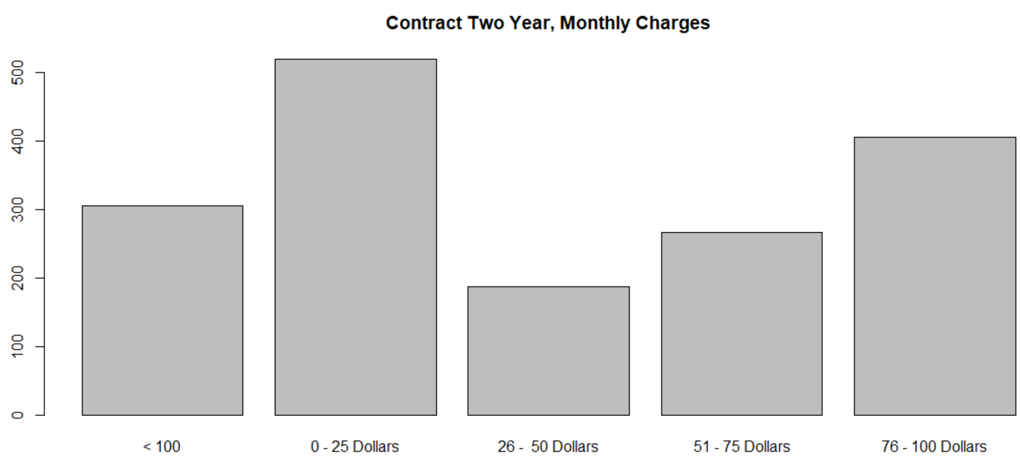


Majority of Monthly charges are between 76 and 100 dollars in a Month-to-month contract. If you go back to the summary() function used in the in univariate statistics you will also see that "Month-to-month" is the largest subgroup of the variable "Contract". Below is a table for easy reference:

```
> table(Telco_Churn_2$Contract)

Month-to-month       One year       Two year
          3875           1472           1685
```



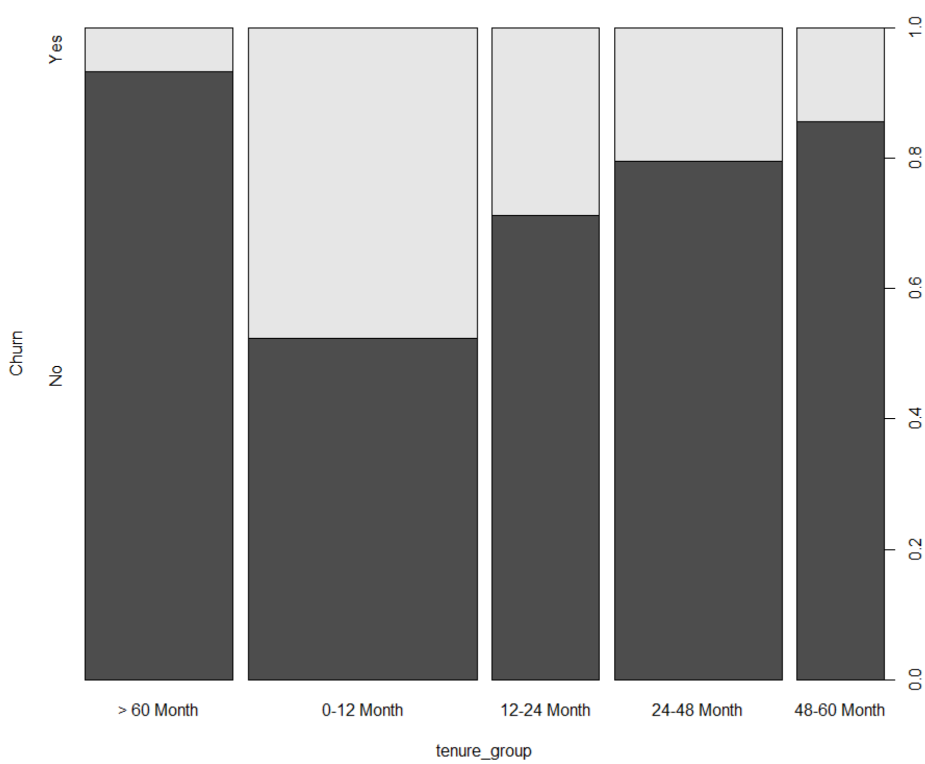
With a One-year contract monthly charges are more evenly distributed except in the 25 – 50 dollar range.



With a two-year contract, customers seem to fall into two major groups of monthly charges, a 0–25 dollar group and a 76–100 dollar group.

Tenure Group vs Churn

```
plot(Churn ~ tenure_group, data = Telco_Churn_2)
```

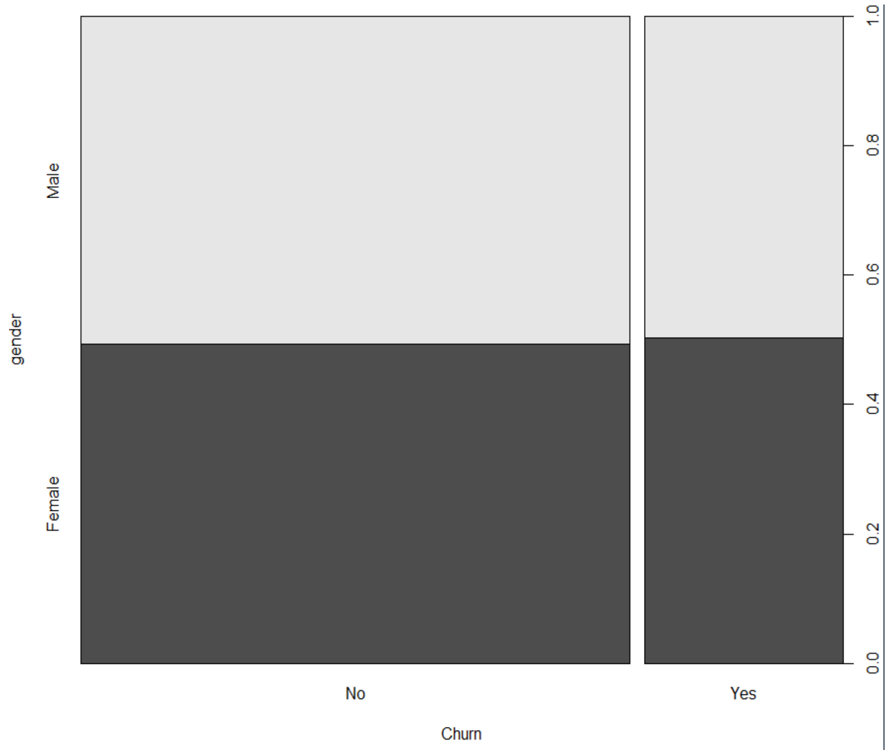


The above plot indicates that if you are in the 0-12 Month tenure group you are approximately 50 percent likely to churn. The next most likely to churn is someone in the 12-24 Month tenure group.

Gender vs Churn

```
> table(Telco_Churn_2$gender, Telco_Churn_2$Churn)/length(Telco_Churn_2$gender)

                No       Yes
  Female 0.3617747 0.1335324
  Male   0.3724403 0.1322526
```

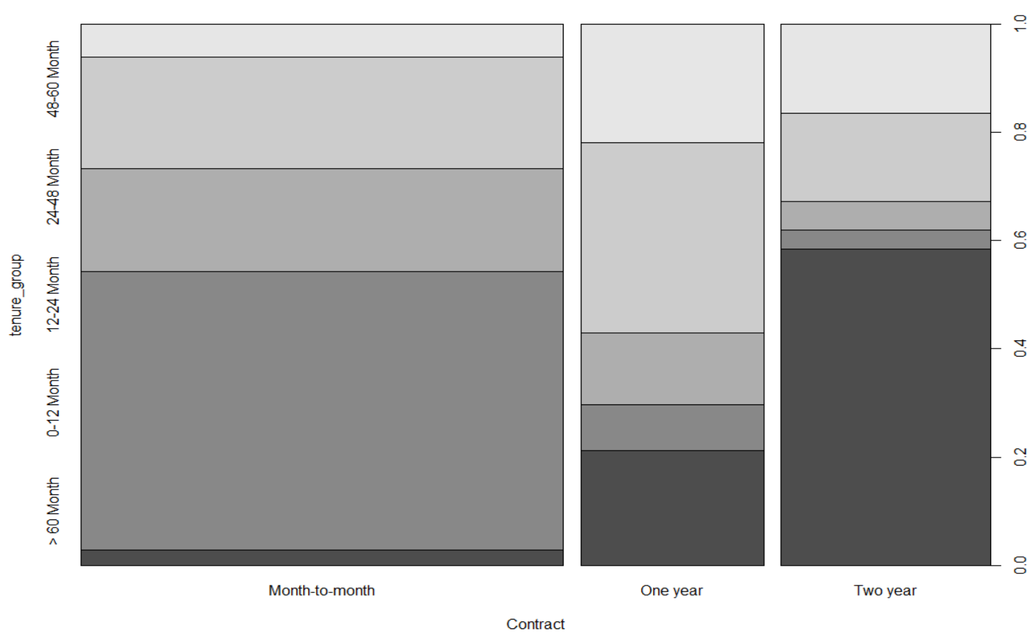
```
plot(gender ~ Churn, data = Telco_Churn_2)
```



Gender does not appear to have significant influence at this point over whether a customer would churn or not.

Tenure Group vs Contract Length

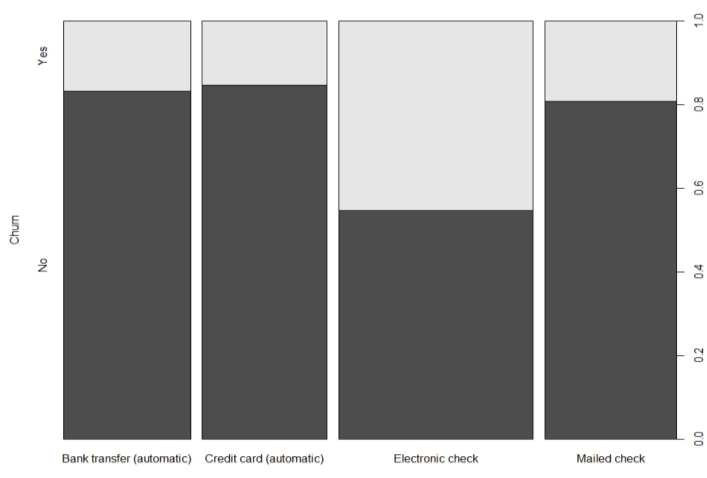
```
plot(tenure_group ~ Contract, data = Telco_Churn_2)
```



There seems to be some significance to the tenure of a customer and what type of contract they have with the telecommunications company. It is especially apparent in the greater than 60-month tenure group. With a two-year contract, nearly 60 percent of the customers stay more than 60 months. On the other hand, if the customer has a month-to-month contract they are less than 5 percent likely to stay with the company greater than 60 months.

Payment Method vs Churn

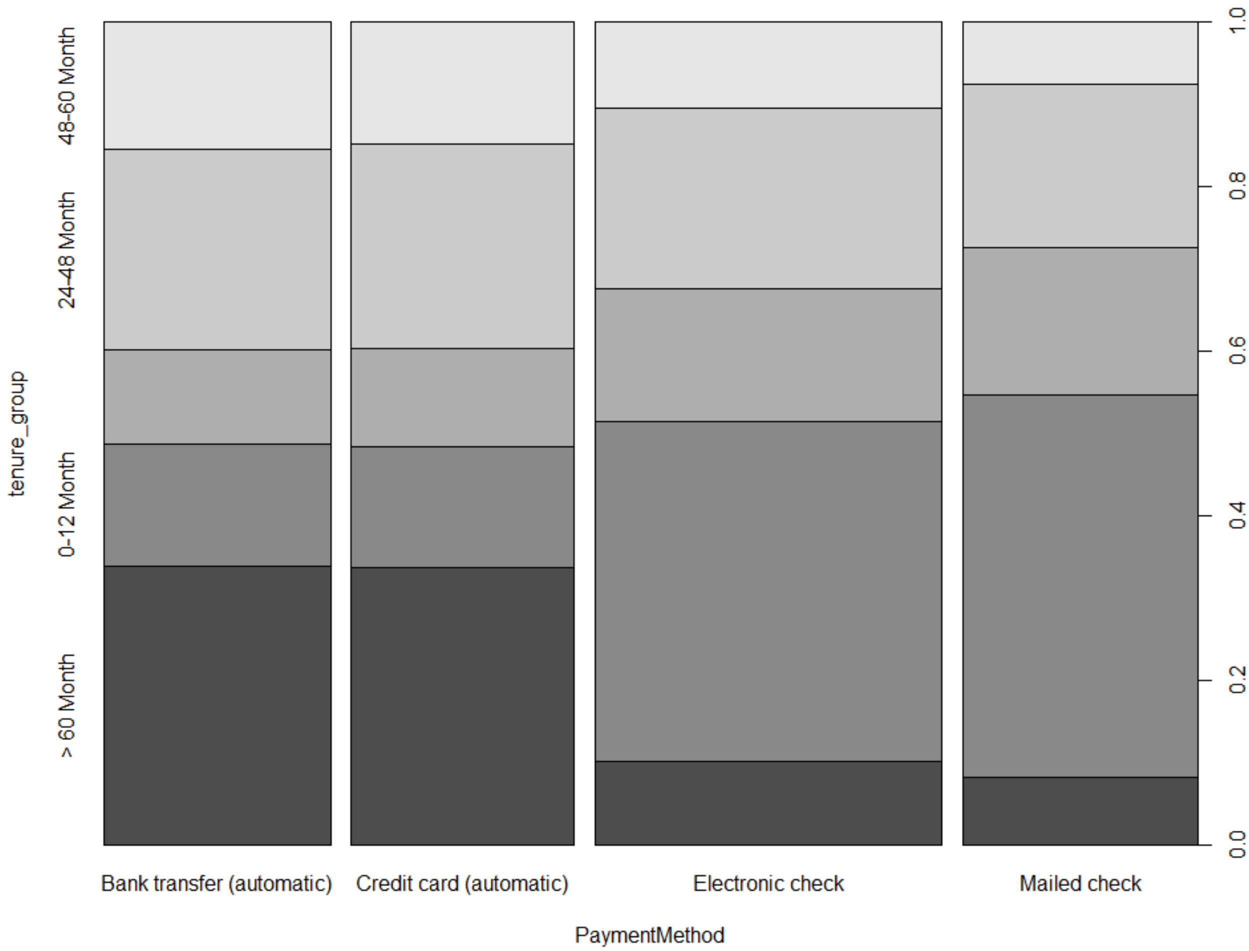
```
plot(Churn ~ PaymentMethod, data = Telco_Churn_2)
```



If the customer utilizes electronic checks, it appears that they are more likely to churn.

Payment Method vs Tenure Group

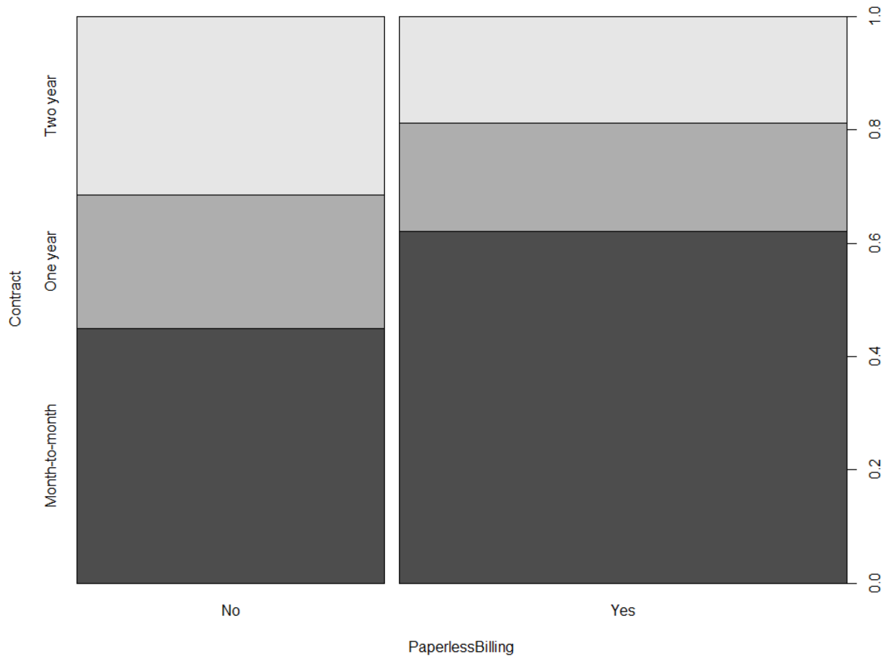
```
plot(tenure_group ~ PaymentMethod, data = Telco_Churn_2)
```



As with payment method vs churn, we can see through a comparison of the variables "tenure group" and "payment method" that a customer is more likely to be with the company greater than 60 months if they do an automatic bank transfer or credit card payment versus using an electronic check or paper mailed check.

Contract vs Paperless Billing

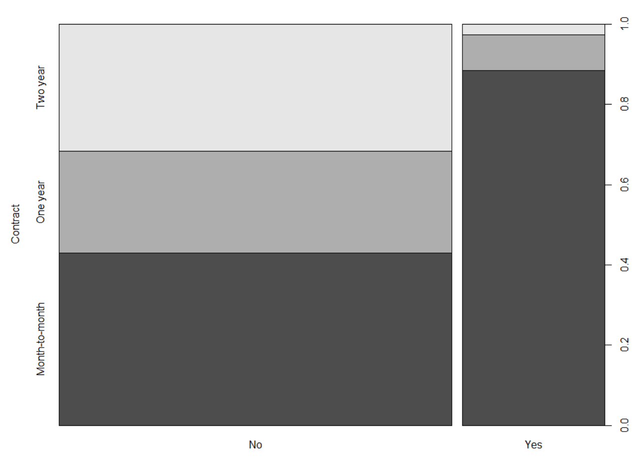
```
plot(Contract ~ PaperlessBilling, data = Telco_Churn_2)
```



Approximately 60 percent of the customers who utilize paperless billing also have a month-to-month contract.

Contract vs Churn

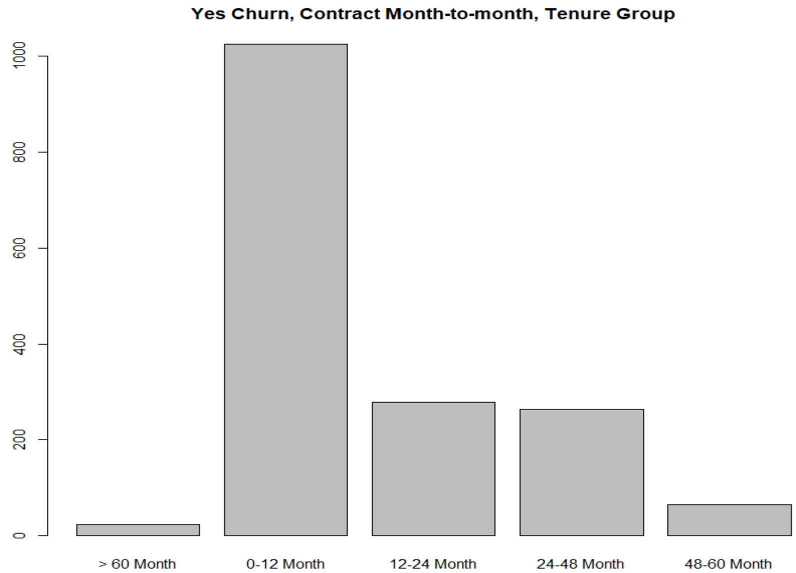
```
plot(Contract ~ Churn, data = Telco_Churn_2)
```



Month-to-month contract is the variable that sees the highest amount of Churn. A two-year contract seems to reduce the probability for a customer to Churn.

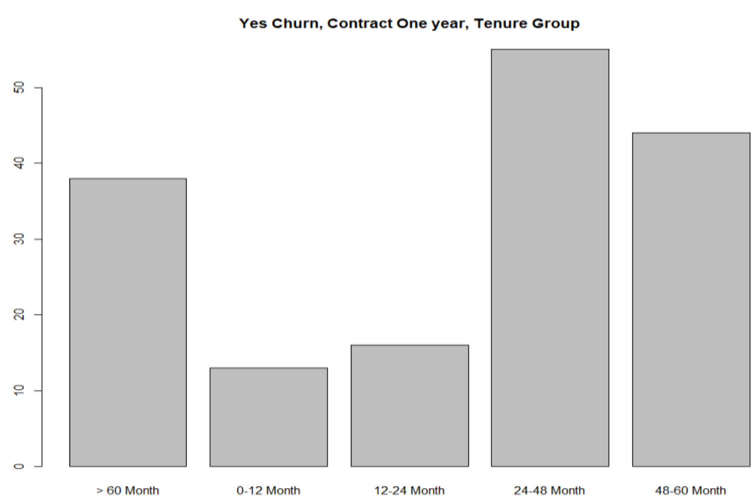
Bonus Analysis – Churn vs Contact length vs Tenure Group

```
# Month-to-month Contract vs Churn vs Tenure Group
MonthtoMonth_Contract_Churn_Yes_tenure_group <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "Month-to-month",
                                                        Telco_Churn_MonthlyCharges_Groups$Churn == "Yes")
plot(MonthtoMonth_Contract_Churn_Yes_tenure_group$tenure_group, main = "Yes Churn, Contract Month-to-month, Tenure Group")
```



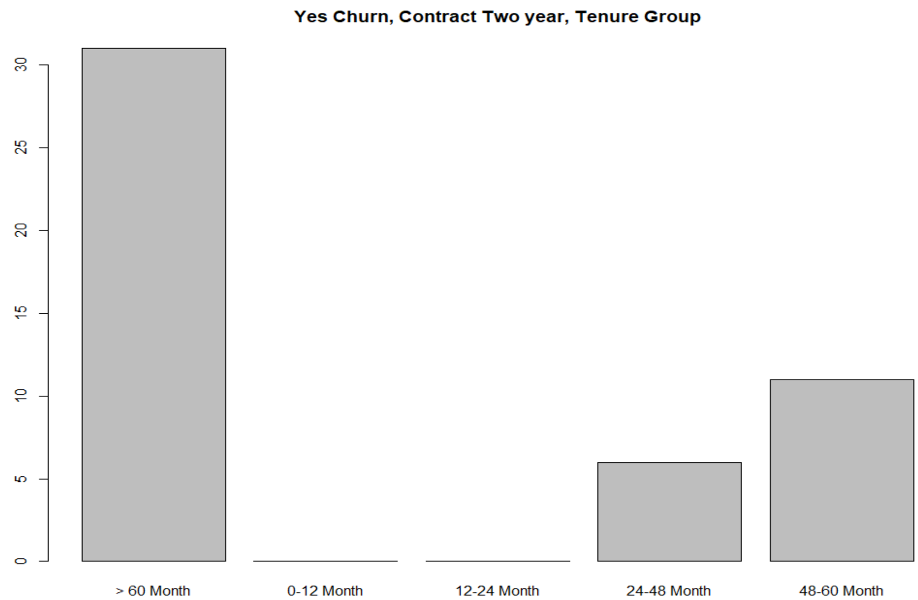
When a customer has a month-to-month contract, the most significant amount of Churn is found in the tenure group a 0–12 months.

```
# One Year Contract vs Churn vs Tenure Group
Oneyear_Contract_Churn_Yes_tenure_group <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "One year",
                                                   Telco_Churn_MonthlyCharges_Groups$Churn == "Yes")
plot(Oneyear_Contract_Churn_Yes_tenure_group$tenure_group, main = "Yes Churn, Contract One year, Tenure Group")
```



When a customer has a one-year contract, we can see the churn amount decreases significantly for the 0–12 months vs the customers with a month-to-month contract.

```
# Two Year Contract vs Churn vs Tenure Group
Twoyear_Contract_Churn_Yes_tenure_group <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "Two year",
                                                  Telco_Churn_MonthlyCharges_Groups$Churn == "Yes")
plot(Twoyear_Contract_Churn_Yes_tenure_group$tenure_group, main = "Yes Churn, Contract Two year, Tenure Group")
```

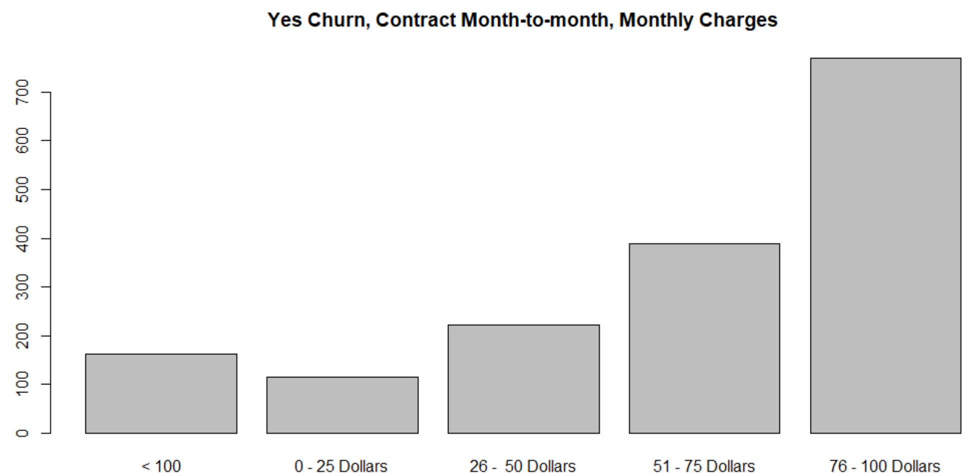


With a two-year contract, we see an almost expected result that there is no churn during the term of the contract.

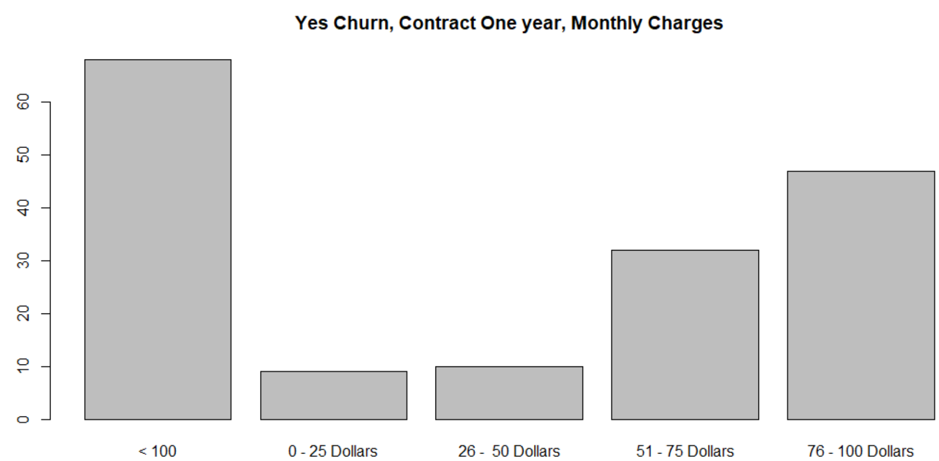
Bonus Analysis - Monthly Charges/Contract Length/Churn

```
## PLots Contract Length Vs Monthly Charges vs Churn
MonthtoMonth_Contract_Churn_Yes <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract
                                          == "Month-to-month", Telco_Churn_MonthlyCharges_Groups$Churn == "Yes")
plot(MonthtoMonth_Contract_Churn_Yes$MonthlyCharges_group, main = "Yes Churn, Contract Month-to-month, Monthly Charges")

# MonthtoMonth_Contract_Churn_No <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract
# plot(MonthtoMonth_Contract_Churn_No$MonthlyCharges_group, main = "No Churn, Contract Month-to-month, Monthly Charges")

OneYear_Contract_Churn_Yes <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract ==
                                     "One year", Telco_Churn_MonthlyCharges_Groups$Churn == "Yes")
plot(OneYear_Contract_Churn_Yes$MonthlyCharges_group, main = "Yes Churn, Contract One year, Monthly Charges")

# OneYear_Contract_Churn_No <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "O
# plot(OneYear_Contract_Churn_No$MonthlyCharges_group, main = "No Churn, Contract One year, Monthly Charges")

TwoYear_Contract_Churn_Yes <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract ==
                                     "Two year", Telco_Churn_MonthlyCharges_Groups$Churn == "Yes")
plot(TwoYear_Contract_Churn_Yes$MonthlyCharges_group, main = "Yes Churn, Contract Two Year, Monthly Charges")

# TwoYear_Contract_Churn_No <- filter(Telco_Churn_MonthlyCharges_Groups, Telco_Churn_MonthlyCharges_Groups$Contract == "T
# plot(TwoYear_Contract_Churn_No$MonthlyCharges_group, main = "No Churn, Contract Two Year, Monthly Charges")
```

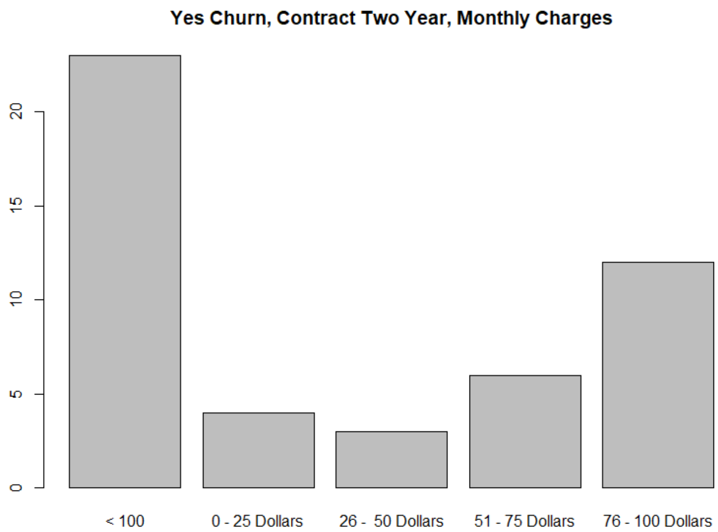
Next, as an extra, a multivariate analysis will be looked at with the variables Contact, Monthly Charges, and Churn. Again, the subgroups of the variable Contract will be used to gain further possible meaning. Also, Churn is separated into its two subgroups of “Yes” or “No” response. The code used is below:



With a month-to-month contract the majority of customers who churned fall into the 76–100 dollar monthly charges range.



With a one-year contract, the majority very few customers churned if their monthly bill was between $0 and $50.



Customers with a two-year contract saw the most amount of churn if their bill was greater than $100.

K. The analytic method that will be applied is hierarchal clustering with k-modes. The evaluative method that will be applied to the data is logistic regression:

<u>Hierarchal Clustering K – Modes:</u>

```
######################### K-Modes Algorithm #####################################
# Example of kmodes
data.to.cluster <- read.csv(dataset.csveader = TRUE, sep = ',')
cluster.results <- kmodes(data.to.cluster[,2:5], 3, iter.max = 10, weighted = FALSE )
setwd("C:/Users/Adam/CatCluster/kmodes")

# kmodes my way
data_to_cluster <- Telco_Churn_2
cluster_results <- kmodes(data_to_cluster, 3, iter.max = 10, weighted = FALSE)
cluster_results
summary(cluster_results)

# Output found clusters to csv file
cluster_output <- cbind(data_to_cluster,cluster_results$cluster)
write.csv(cluster_output, file = "kmodes clusters.csv", row.names = TRUE)
```

```
> data_to_cluster <- Telco_Churn_2
> cluster_results <- kmodes(data_to_cluster, 3, iter.max = 10, weighted = FALSE)
> cluster_results
K-modes clustering with 3 clusters of sizes 2854, 2550, 1628

Cluster modes:
  gender SeniorCitizen Partner Dependents PhoneService MultipleLines InternetService
1   Male            No      No         No          Yes           Yes     Fiber optic
2 Female            No      No         No          Yes            No             DSL
3   Male            No     Yes         No          Yes           Yes             DSL
  OnlineSecurity OnlineBackup DeviceProtection TechSupport StreamingTV StreamingMovies
1             No           No               No          No          No              No
2             No           No               No          No          No              No
3            Yes          Yes              Yes         Yes         Yes             Yes
        Contract PaperlessBilling             PaymentMethod MonthlyCharges Churn
1 Month-to-month              Yes          Electronic check          19.95    No
2 Month-to-month              Yes          Electronic check          20.05    No
3       Two year              Yes Bank transfer (automatic)          79.20    No
  tenure_group
1   0-12 Month
2   0-12 Month
3   > 60 Month
```

Feature Analysis:

Paperless Billing: “Yes” is the most reoccurring answer in all three clusters.

Contract: In two of the three clusters, month-to-month is the most reoccurring.

Tenure_group: In two of the three clusters 0-12 month is the most reoccurring.

PaymentMethod: In two of the three clusters, electronic check is the most reoccurring.

```
Clustering vector:
   [1] 2 2 2 2 2 1 1 2 1 2 2 1 1 1 1 3 2 3 2 2 2 1 1 2 2 2 1 2 3 2 3 1 3 1 2 3 1 2 1 1 2
  [42] 3 2 3 3 1 2 2 3 3 1 2 2 1 3 1 3 3 1 3 1 1 3 2 1 2 1 3 2 1 2 2 3 1 2 3 2 2 2 1 2 2
  [83] 2 3 2 1 2 2 2 2 1 1 3 3 3 1 3 1 1 2 1 2 1 3 3 2 3 2 1 1 1 2 2 1 3 1 2 1 2 1 1 2 1
 [124] 2 2 1 1 1 1 3 1 3 2 2 2 1 2 2 1 2 3 2 3 1 1 3 1 2 2 3 2 1 3 3 2 2 1 1 3 1 2 1 2 3
 [165] 2 1 2 3 2 2 3 1 3 2 2 1 1 2 3 1 2 2 1 3 2 2 2 2 2 2 3 1 2 3 1 2 1 1 3 1 3 2 3 1 1
 [206] 2 3 1 2 3 2 2 1 2 1 3 3 3 3 2 2 1 1 2 2 2 3 2 1 3 3 1 1 2 1 1 2 1 2 1 3 3 3 2 2 1
 [247] 1 1 2 1 2 2 2 3 1 3 3 3 2 2 2 1 1 3 3 3 1 2 1 1 2 3 2 1 1 1 1 3 2 2 2 2 1 2 1 1 2
 [288] 1 1 2 1 2 1 2 2 1 2 1 1 2 2 1 1 3 1 1 1 3 1 3 2 2 3 2 1 3 1 3 1 1 3 3 2 2 3 3 2 2
 [329] 2 1 2 2 1 1 1 3 1 3 2 3 2 3 3 1 2 1 2 2 2 3 3 2 2 1 2 2 2 3 2 1 1 3 2 2 1 1 2 2 3
 [370] 3 1 1 2 3 1 3 2 3 3 1 2 2 3 1 3 1 1 2 1 2 1 3 2 2 1 3 3 2 2 2 2 2 1 2 1 1 3 1 2 1
 [411] 1 1 1 1 2 2 3 1 2 2 1 2 1 1 2 2 3 1 1 3 2 3 2 2 1 2 1 3 1 2 3 1 3 2 2 1 2 1 1 3 1
 [452] 2 3 2 1 1 3 3 1 1 1 2 1 1 3 2 3 1 2 2 2 2 2 2 2 1 3 2 1 2 1 2 2 3 2 1 1 3 1 3 2 2
 [493] 3 1 2 2 2 2 1 3 1 3 2 3 1 1 1 1 2 3 1 1 1 1 1 2 2 2 1 3 1 1 1 1 1 3 1 1 3 2 1 1 3
 [534] 1 1 1 2 2 3 3 2 2 1 2 1 1 3 3 1 3 3 1 1 1 3 2 3 1 2 3 1 2 3 2 3 3 3 1 2 2 1 2 1 3
 [575] 3 3 1 1 1 2 2 1 2 2 1 2 1 3 3 3 1 3 1 2 2 3 1 1 1 1 1 1 3 3 3 3 2 3 3 2 2 1 3 3 2
 [616] 2 3 1 1 2 1 1 2 2 2 1 3 3 1 3 2 3 1 2 3 3 3 3 1 1 2 1 2 1 1 2 3 1 1 2 2 1 1 2 1 3
 [657] 3 1 1 2 2 1 3 2 2 2 2 2 3 2 1 1 1 2 3 1 2 1 1 1 3 2 1 2 2 1 2 2 1 1 3 1 1 2 2 3 2
 [698] 1 3 3 1 1 2 2 2 2 3 1 3 1 2 1 1 1 1 3 1 2 1 1 1 1 2 1 1 1 2 1 3 1 1 1 1 1 2 3 3 1
 [739] 3 2 2 1 1 2 3 3 2 1 2 2 3 3 2 1 1 2 2 3 1 2 1 3 1 2 1 1 1 3 1 1 2 2 1 3 2 1 3 2 3
 [780] 2 2 2 3 3 3 2 3 3 3 3 2 2 2 2 1 1 2 1 1 1 1 2 1 3 2 3 1 1 3 3 1 2 1 1 2 3 3 3 3 3
 [821] 1 2 1 2 1 1 1 2 3 1 3 1 2 1 3 2 3 1 1 3 2 2 1 2 3 2 3 3 3 3 1 2 2 1 3 2 1 3 1 3 1
 [862] 2 1 1 3 2 1 1 3 3 1 2 1 2 2 3 2 1 3 1 2 3 2 1 2 3 1 3 1 3 1 2 1 1 3 1 2 1 2 1 1 3
 [903] 3 1 2 2 2 3 1 1 3 1 1 1 2 3 1 3 1 1 1 3 2 3 1 2 3 2 2 1 1 1 2 1 1 3 2 2 2 1 3 3 1
 [944] 1 1 1 1 2 2 2 2 1 2 1 2 3 2 2 1 2 3 2 1 1 1 3 1 2 1 3 2 3 3 1 1 3 2 2 1 2 1 1 1 2
 [985] 1 1 1 2 1 1 1 1 1 2 1 1 2 2 1 1
 [ reached getOption("max.print") -- omitted 6032 entries ]

Within cluster simple-matching distance by cluster:
[1] 18533 15901 10780

Available components:
[1] "cluster"    "size"       "modes"      "withindiff" "iterations" "weighted"
> summary(cluster_results)
           Length Class      Mode
cluster    7032   -none-     numeric
size          3   table      numeric
modes        19   data.frame list
withindiff    3   -none-     numeric
iterations    1   -none-     numeric
weighted      1   -none-     logical
```

Logistic Regression:

```
> # ----------------------------------LOGISITIC REGRESSION-----------------
---------
> nrow(Telco_Churn_2)
[1] 7032
> train <- createDataPartition(Telco_Churn_2$Churn,p=0.7,list=FALSE)
> set.seed(2017)
> training <- Telco_Churn_2[train,]
> testing <- Telco_Churn_2[-train,]
> # Check Spliting Results
> dim(training); dim(testing)
[1] 4924   19
[1] 2108   19
> # Fitting the LOg Regresssion Model
> mod_fit <- glm(Churn ~ .,family=binomial(link="logit"),data=training)
> mod_fit
```

```
> summary(mod_fit)

Call:
glm(formula = Churn ~ ., family = binomial(link = "logit"), data = training)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0123  -0.6750  -0.2959   0.6764   3.1239

Coefficients:
                                      Estimate Std. Error z value Pr(>|z|)
(Intercept)                          -0.757885   0.982041  -0.772 0.440266
genderMale                            0.020537   0.077454   0.265 0.790898
SeniorCitizenYes                      0.217599   0.100763   2.160 0.030811 *
PartnerYes                           -0.082088   0.093355  -0.879 0.379235
DependentsYes                        -0.179643   0.106980  -1.679 0.093109 .
PhoneServiceYes                       0.714876   0.772446   0.925 0.354721
MultipleLinesYes                      0.558000   0.210999   2.645 0.008180 **
InternetServiceFiber optic            2.159851   0.949798   2.274 0.022965 *
InternetServiceNo                    -2.135237   0.960333  -2.223 0.026187 *
OnlineSecurityYes                    -0.029481   0.213149  -0.138 0.889993
OnlineBackupYes                       0.091448   0.208882   0.438 0.661531
DeviceProtectionYes                   0.292064   0.209834   1.392 0.163958
TechSupportYes                       -0.035915   0.213906  -0.168 0.866662
StreamingTVYes                        0.808742   0.388045   2.084 0.037147 *
StreamingMoviesYes                    0.763061   0.389398   1.960 0.050044 .
ContractOne year                     -0.723807   0.126481  -5.723 1.05e-08 ***
ContractTwo year                     -1.707730   0.219988  -7.763 8.31e-15 ***
PaperlessBillingYes                   0.333745   0.089021   3.749 0.000177 ***
PaymentMethodCredit card (automatic) -0.008611   0.133598  -0.064 0.948607
PaymentMethodElectronic check         0.376810   0.113633   3.316 0.000913 ***
PaymentMethodMailed check            -0.071253   0.137325  -0.519 0.603856
MonthlyCharges                       -0.053017   0.037762  -1.404 0.160330
tenure_group0-12 Month                1.716441   0.203310   8.442  < 2e-16 ***
tenure_group12-24 Month               0.843996   0.200251   4.215 2.50e-05 ***
tenure_group24-48 Month               0.544207   0.181722   2.995 0.002747 **
tenure_group48-60 Month               0.145186   0.197988   0.733 0.463372
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5702.8  on 4923  degrees of freedom
Residual deviance: 4116.0  on 4898  degrees of freedom
AIC: 4168

Number of Fisher Scoring iterations: 6
```

```
> qchisq(0.95, 4898)
[1] 5061.928
```

Feature Analysis:

The critical value at 95 percent confidence and 4898 degrees of freedom is 5,061.928. Since the residual deviance of 4,116.0 is less than the critical value the null model is not rejected. In other words, we have a reliable model at 95 percent confidence level. Also, we can see that the four most significant variables are “Contract, PaperlessBilling, PaymentMethod, tenure_group”.

Deviance Analysis Table:

```
> # ANOVA of model_log
> anova(mod_fit, test="Chisq")
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
NULL                              4923     5702.8
gender            1     0.02      4922     5702.7  0.877032
SeniorCitizen     1   105.60      4921     5597.1 < 2.2e-16 ***
Partner           1   133.32      4920     5463.8 < 2.2e-16 ***
Dependents        1    36.25      4919     5427.6 1.732e-09 ***
PhoneService      1     2.68      4918     5424.9  0.101455
MultipleLines     1     9.04      4917     5415.8  0.002646 **
InternetService   2   465.12      4915     4950.7 < 2.2e-16 ***
OnlineSecurity    1   151.13      4914     4799.6 < 2.2e-16 ***
OnlineBackup      1    70.92      4913     4728.7 < 2.2e-16 ***
DeviceProtection  1    31.77      4912     4696.9 1.739e-08 ***
TechSupport       1    71.00      4911     4625.9 < 2.2e-16 ***
StreamingTV       1     2.11      4910     4623.8  0.146679
StreamingMovies   1     0.00      4909     4623.8  0.965628
Contract          2   304.22      4907     4319.6 < 2.2e-16 ***
PaperlessBilling  1    15.57      4906     4304.0 7.940e-05 ***
PaymentMethod     3    35.87      4903     4268.1 7.987e-08 ***
MonthlyCharges    1     2.25      4902     4265.9  0.133386
tenure_group      4   149.86      4898     4116.0 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The deviance table shows that as you add a variable there is a reduction in the deviance. However, some variables have a greater impact on reduction of deviance than others. The variables “InternetService”, “Contract”, “OnlineSecurity” and “tenure_group” have some the greatest reduction impact and all have low p-values. On the other hand, the variables “Dependents”, “DeviceProtection”, “PaymentMethod”, and “PaperlessBilling” have low low p-values be have a much smaller impact on the reduction of the residual deviance.

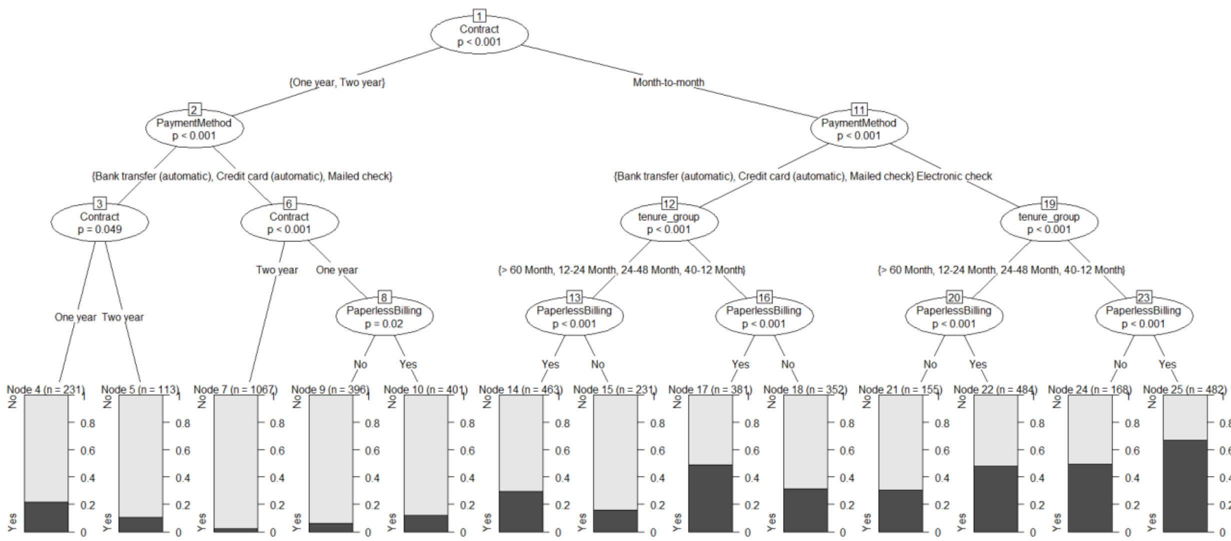
Odds Ratio

```
> exp(cbind(OR=coef(mod_fit), confint(mod_fit)))
Waiting for profiling to be done...
                                          OR      2.5 %     97.5 %
(Intercept)                        0.4686567 0.06828594  3.2112898
genderMale                         1.0207489 0.87699598  1.1881811
SeniorCitizenYes                   1.2430885 1.02008586  1.5143358
PartnerYes                         0.9211913 0.76715700  1.1062459
DependentsYes                      0.8355682 0.67699334  1.0298494
PhoneServiceYes                    2.0439341 0.45009824  9.3049256
MultipleLinesYes                   1.7471746 1.15590379  2.6438383
InternetServiceFiber optic         8.6698475 1.35095870 55.9850647
InternetServiceNo                  0.1182166 0.01795584  0.7754751
OnlineSecurityYes                  0.9709490 0.63918728  1.4743630
OnlineBackupYes                    1.0957603 0.72765874  1.6505776
DeviceProtectionYes                1.3391890 0.88779323  2.0213647
TechSupportYes                     0.9647224 0.63410474  1.4669892
StreamingTVYes                     2.2450815 1.05030222  4.8097565
StreamingMoviesYes                 2.1448309 1.00074330  4.6071689
ContractOne year                   0.4849028 0.37739954  0.6198106
ContractTwo year                   0.1812768 0.11613159  0.2755869
PaperlessBillingYes                1.3961868 1.17300158  1.6629929
PaymentMethodCredit card (automatic) 0.9914259 0.76298373  1.2884442
PaymentMethodElectronic check      1.4576278 1.16762091  1.8232158
PaymentMethodMailed check          0.9312264 0.71171525  1.2195032
MonthlyCharges                     0.9483639 0.88060876  1.0211497
tenure_group0-12 Month             5.5646881 3.74951104  8.3237835
tenure_group12-24 Month            2.3256413 1.57502407  3.4550103
tenure_group24-48 Month            1.7232410 1.21059144  2.4698274
tenure_group48-60 Month            1.1562546 0.78453007  1.7062755
```

The odds that a certain incident is going to transpire is fascinating to explore when carrying out logistic regression. But I will not go into too much detail here about the odds ratio as it is not in the scope of the assignment. However, it is something neat to look at with the data.

Bonus - Decision Tree:

```
# Decision Tree
tree <- ctree(Churn~Contract+PaperlessBilling+PaymentMethod+tenure_group, training)
plot(tree)
```




```
> # Decision Tree Accuracy
> p1 <- predict(tree, training)
> tab1 <- table(Predicted = p1, Actual = training$Churn)
> tab2 <- table(Predicted = pred_tree, Actual = testing$Churn)
> print(paste('Decision Tree Accuracy',sum(diag(tab2))/sum(tab2)))
[1] "Decision Tree Accuracy 0.793643263757116"
```

```
> # Decision Tree Confusion Matrix
> pred_tree <- predict(tree, testing)
> print("Confusion Matrix for Decision Tree"); table(Predicted = pred_tree, Actual = testing$Churn)
[1] "Confusion Matrix for Decision Tree"
         Actual
Predicted    0    1
      No  1481  408
      Yes   67  152
```

Feature Analysis - Decision Tree:

The four most significant variables were used to create the above decision tree. We can see that the accuracy is slightly lower than the logistic regression. Of the four variables, the most significant is contract. Meaning the variable contract and the contract length a customer has is the best variable to indicate whether or not a customer will churn. Three other things we can pull from the data are (1) A customer in month-to-month contract is more likely to churn than a customer with one or two-year contract. (2) If a customer has paperless billing, they are more likely to Churn if they have a month-to-month or one-year contract. (3) If a customer utilizes the "Electronic Check" payment method, he or she is more likely to churn. (4) Tenure appears to have an influence on the potential to Churn. A customer that has been with the company for 0-12 months is more likely to churn than a customer that has been with the company greater than 60 months.

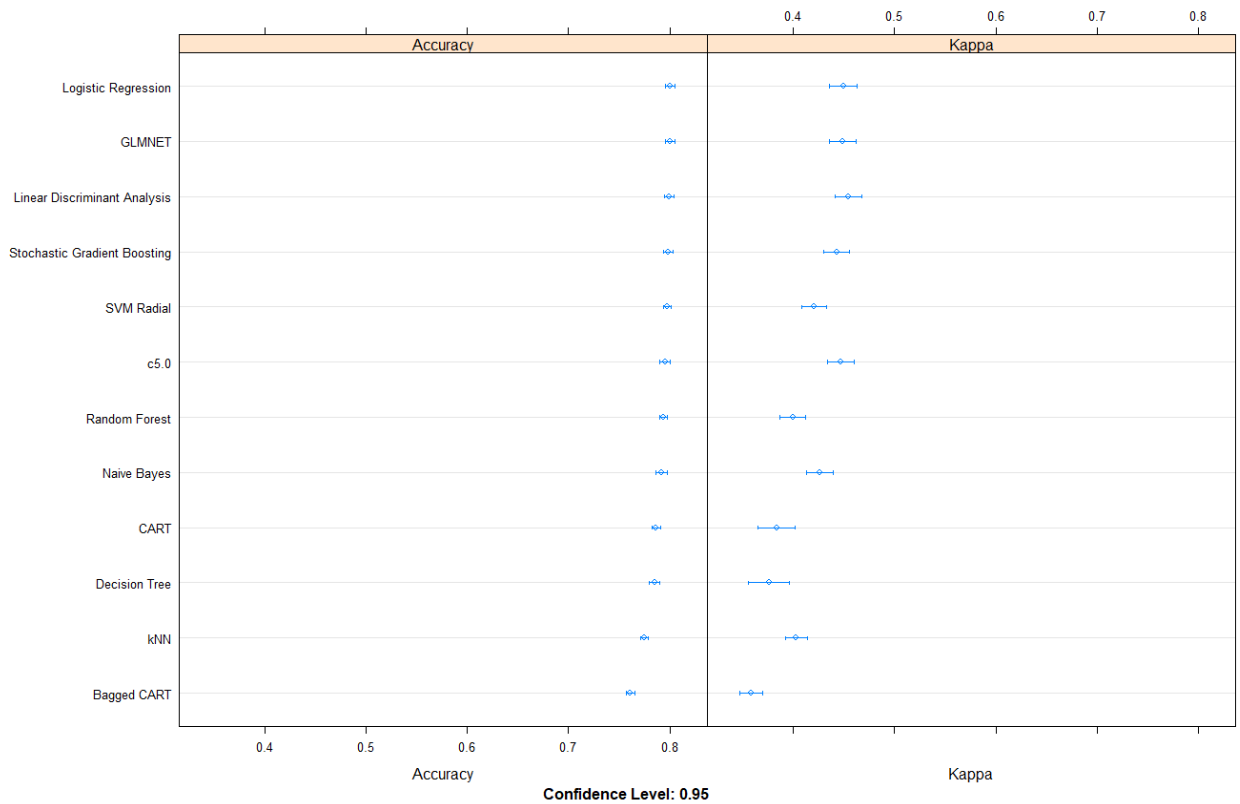
L. When choosing a method to evaluate a data set, there can be an ample amount of options. To quickly choose a potential method, the boxplot below was created to compare multiple model types (Brownlee, 2016). See code and box plot below.

```
# rename dataset to keep code below generic
dataset_test <- Telco_Churn_2

control <- trainControl(method="repeatedcv", number=10, repeats=3)
seed <- 7

metric <- "Accuracy"
preProcess=c("center", "scale")
```

```
# Linear Discriminant Analysis
set.seed(seed)
fit.lda <- train(Churn~., data=dataset_test, method="lda", metric=metric, preProc=c("center", "scale"), trControl=control)
# Logistic Regression
set.seed(seed)
fit.glm <- train(Churn~., data=dataset_test, method="glm", metric=metric, trControl=control)
# GLMNET
set.seed(seed)
fit.glmnet <- train(Churn~., data=dataset_test, method="glmnet", metric=metric, preProc=c("center", "scale"), trControl=control)
# SVM Radial
set.seed(seed)
fit.svmRadial <- train(Churn~., data=dataset_test, method="svmRadial", metric=metric, preProc=c("center", "scale"), trControl=control, fit=FALSE)
# kNN
set.seed(seed)
fit.knn <- train(Churn~., data=dataset_test, method="knn", metric=metric, preProc=c("center", "scale"), trControl=control)
# Naive Bayes
set.seed(seed)
fit.nb <- train(Churn~., data=dataset_test, method="nb", metric=metric, trControl=control)
# CART
set.seed(seed)
fit.cart <- train(Churn~., data=dataset_test, method="rpart", metric=metric, trControl=control)
# C5.0
set.seed(seed)
fit.c50 <- train(Churn~., data=dataset_test, method="C5.0", metric=metric, trControl=control)
# Bagged CART
set.seed(seed)
fit.treebag <- train(Churn~., data=dataset_test, method="treebag", metric=metric, trControl=control)
# Random Forest
set.seed(seed)
fit.rf <- train(Churn~., data=dataset_test, method="rf", metric=metric, trControl=control)
# Stochastic Gradient Boosting (Generalized Boosted Modeling)
set.seed(seed)
fit.gbm <- train(Churn~., data=dataset_test, method="gbm", metric=metric, trControl=control, verbose=FALSE)
# Decision Tree
set.seed(seed)
fit.dt <- train(Churn~., data=dataset_test, method="rpart", metric=metric, trControl=control)
```

```
results <- resamples(list("Linear Discriminant Analysis"=fit.lda, "Logistic Regression"=fit.glm, GLMNET=fit.glmnet,
                          "SVM Radial"=fit.svmRadial, kNN=fit.knn, "Naive Bayes"=fit.nb, CART=fit.cart, "c5.0"=fit.c50,
                          "Bagged CART"=fit.treebag, "Random Forest"=fit.rf, "Stochastic Gradient Boosting"=fit.gbm, "Decision Tree" =fit.dt ))
# Table comparison
summary(results)
```

```
# boxplot comparison
bwplot(results)
# Dot-plot comparison
dotplot(results)
```



As you can see, out of all the options for an evaluative method, logistic regression was the most accurate of each of the models. Therefore, it was chosen to analyze the data set. Likewise, logistic regression is able to handle all of the categorical variables with ease. As mentioned above in feature analysis of logistic regression in section k, the critical value at a 95 percent confidence level compared to the residual deviance indicates a good fit for this particular data set. The bonus method, decision tree method, was chosen for its visualization easiness to represent the data but was slightly less accurate than logistic regression. See below for actual number comparison:

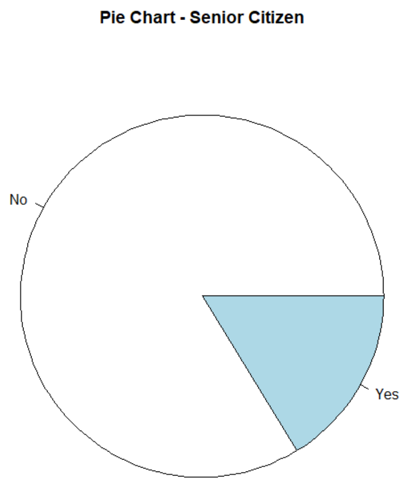
```
# Logistic Regression Accuracy or the predictive ability of the model_log
testing$Churn <- as.character(testing$Churn)
testing$Churn[testing$Churn=="No"] <- "0"
testing$Churn[testing$Churn=="Yes"] <- "1"
fitted.results <- predict(mod_fit,newdata=testing,type='response')
fitted.results <- ifelse(fitted.results > 0.5,1,0)
misClasificError <- mean(fitted.results != testing$Churn)
print(paste('Logistic Regression Accuracy',1-misClasificError))
```

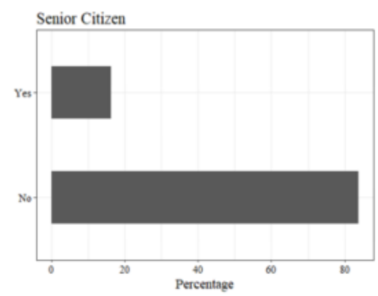
```
[1] "Logistic Regression Accuracy 0.808823529411765"
```

```
[1] "Decision Tree Accuracy 0.793643263757116"
```

Hierarchal Clustering K – Modes was chosen for the analytic method because of its ability to handle categorical data and represent which values occur most in the data set. PCA and other options lacked the ability to cope with the categorical data the data set contained.

M. The methods that were chosen to visually present the data are the best because they easily and accurately tell the story. For example, I could have done a pie chart in the univariate and bivariate analysis. However, a pie chart is vague and does not give much detail. Take the pie chart of the variable Senior Citizen for example. We can visually see that there are more non-senior citizens than senior citizens in the data but we are unsure or the exact difference. On the other hand, a boxplot and a table show much more detail and accurately represent the data. For similar reasons, the analytic and evaluative visualization methods were chosen. They give a quick and accurate interpretation of the data. However, since the accuracy of the decision tree method was close to the accuracy of the logistic regression method, I included it since it is visually more appealing and easier to interpret than the logistic regression output.





Senior Citizen (*Frequency & Percentage*):

- Significantly more Non-Senior Citizens

```
> table(Telco_Churn_2$SeniorCitizen)

  No  Yes
5890 1142
> table(Telco_Churn_2$SeniorCitizen)/length(Telco_Churn_2$SeniorCitizen)

       No       Yes
0.8375995 0.1624005
```

IV: Data Summary

N. **Discriminate Analysis**

Two ways were used to test that the data was discriminating. First, the Chi-squared was calculated in section K with the logistic regression. We saw that with a 95 percent confidence interval, the critical value was larger than the residual deviance of the logistic regression. This indicated a good model fit and that the data was discriminating. See below:

```
> qchisq(0.95, 4898)
[1] 5061.928
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5702.8  on 4923  degrees of freedom
Residual deviance: 4116.0  on 4898  degrees of freedom
AIC: 4168

Number of Fisher Scoring iterations: 6
```

The second way to show that the data was discriminating is to produce a ROC Curve and calculate the area under that curve. In a ROC Curve the value under the curve has a range of 0.50 to 1.00. A calculated area of .80 or greater demonstrates that the model does a fantastic work in discriminating (Analytics, n.d.). The area under the ROC Curve for the logistic regression model came to 0.8434287. See below:

ROC Curve

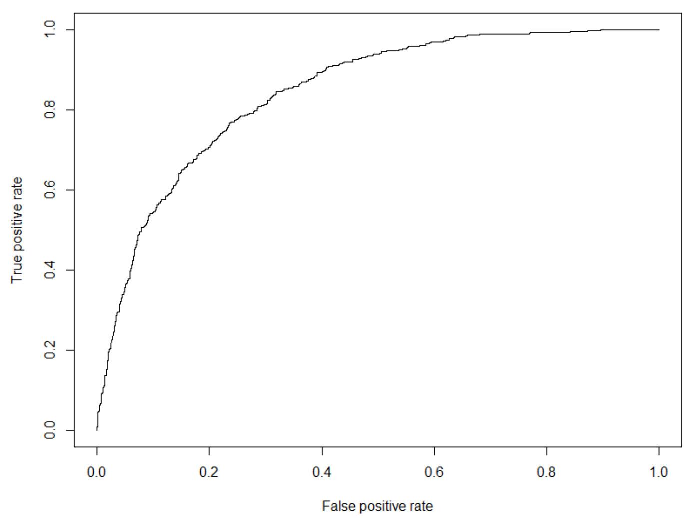
We calculate the ROC Curve by using the following lines of code:

```
# Compute AUC for predicting Class with the model
prob <- predict(mod_fit_one, newdata=testing, type="response")
pred <- prediction(prob, testing$Churn)
perf <- performance(pred, measure = "tpr", x.measure = "fpr")
plot(perf)

auc <- performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

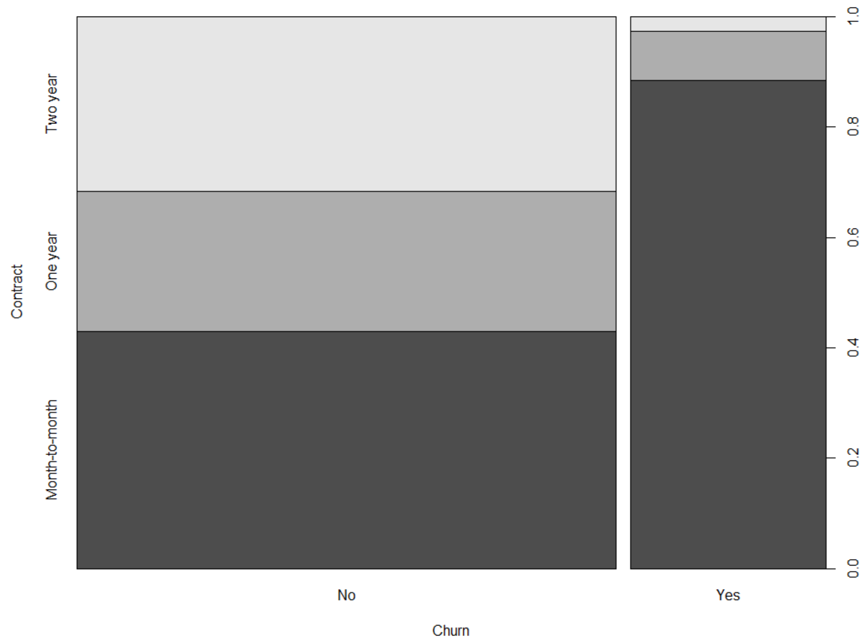
What we are most concerned about is the area under the ROC Curve which comes out to 0.8434287 which shows that the model does a good job in discriminating.

```
> auc <- performance(pred, measure = "auc")
> auc <- auc@y.values[[1]]
> auc
[1] 0.8434287
```

Below is a visual representation of the ROC Curve:

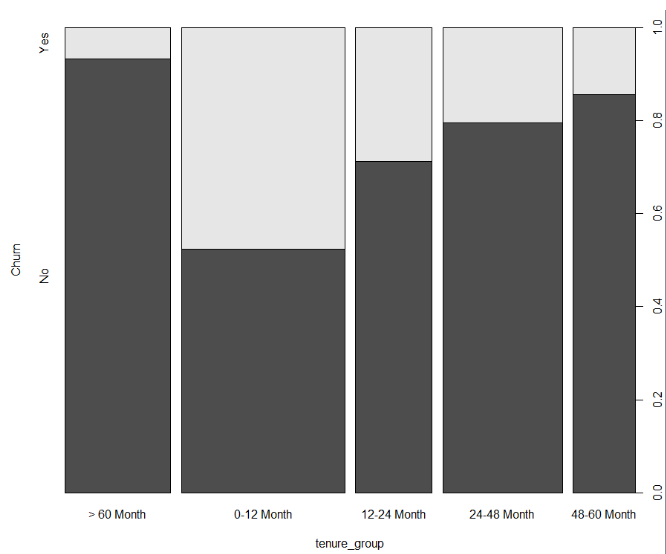


The phenomenon that we wanted to detect was if contract length and/or tenure with the company affects whether a customer is more likely to churn or not. Throughout the analysis, it became apparent that the type of contract and the length of time a customer has been with the company surely does affect whether a customer will churn. As you will see in the next section, the top four most significant or important variables are “tenure_group0-12 Month”, “ContractTwo year”, “ContractOne year”, and “tenure_group12-24 Month”. For instance, as a customer’s contract length increases, that customer is less likely to churn. See below:



Similar to contract length, the likelihood of a customer churning decreases as the tenure of a customer increases. See below:

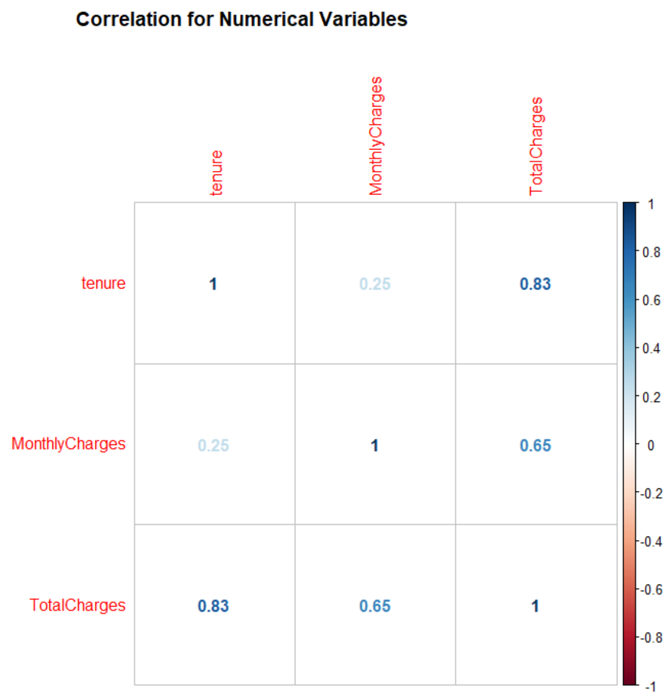
<u>Tenure Group vs Churn</u>



Other possible indicators that a customer might churn include: they utilize the electronic check payment method or they have paperless billing and a month-to-month or one year contract. Over all, the length of time that a customer has been with the company and whether they have a contract or not seem to be the most significant influencers on whether a customer will churn or not.

O. One of the ways that was used to detect interactions between numeric variables was to check for correlations with a correlation matrix. See below:

```
# -----Correlation-----
# Discover Correlation between Numneric Variables
numeric_variables <- sapply(Telco_Churn_2, is.numeric)
matrix <- cor(Telco_Churn_2[,numeric_variables])
corrplot(matrix, main="\n\nCorrelation for Numerical Variables", method="number")
```



As explained in the data cleaning process, correlation is high if it approaches 1 or -1. In this case we see that TotalCharges is highly related to tenure and MonthyCharges. Therefore, TotalCharges was removed from the data set in favor of MonthyCharges. Also, tenrue was grouped to provide a deeper understanding of the data and likelihood of a Churn.

As for selecting the most important predictor variables this was a process. To start the process a logistic regression model was performed with all of the variables from the cleaned data set. Also, a deviance analysis table was produced to see how the model was affected by adding one variable at a time. As we saw from above in section K, the variables “InternetService”, “Contract”, “OnlineSecurity” and “tenure_group” have some the greatest reduction impact and all have low p-values. On the other hand, the variables “Dependents”, “DeviceProtection”, “PaymentMethod”, and “PaperlessBilling” have low p-values and have a much smaller impact on the reduction of the residual deviance.

With this in mind, the accuracy of the first logistic model was run. It came out to 80.9 percent which is quite respectable. However, I wanted to confirm that the best variables in the cleaned data set were being selected. Therefore I created a logistic model with only the most significant variables which were tenure_group, Contract, MultipleLines, SeniorCitizen, PaperlessBilling, PaymentMethod, InternetService, StreamingTV, MonthlyCharges, DeviceProtection, PhoneService, Partner, and Dependents. However, when the accuracy was run it was ever so slightly lower than including all the variables in the cleaned data set at 80.8 percent only 0.1 difference.

```
[1] "Logistic Regression Accuracy 0.808349146110057"
```

But, because of the difference and the slightly better accuracy the first logistic model with all variables was used. The below output shows the most significant/important variables in the data set off the logistic regression model.

```
> mod_fit_3 <- train(Churn ~ ., data=training, method="glm", family="binomial")
> varImp(mod_fit_3)
glm variable importance

  only 20 most important variables shown (out of 25)

                                       Overall
`tenure_group0-12 Month`               100.000
`ContractTwo year`                      96.205
`ContractOne year`                      66.614
`tenure_group12-24 Month`               47.883
MultipleLinesYes                        36.118
SeniorCitizenYes                        35.293
PaperlessBillingYes                     30.808
StreamingMoviesYes                      30.320
`PaymentMethodElectronic check`         29.309
`InternetServiceFiber optic`            29.190
InternetServiceNo                       28.761
`tenure_group24-48 Month`               27.722
StreamingTVYes                          24.685
MonthlyCharges                          19.093
DeviceProtectionYes                     15.020
`PaymentMethodCredit card (automatic)`  11.329
PhoneServiceYes                          9.784
PartnerYes                               9.579
DependentsYes                            7.524
`tenure_group48-60 Month`                6.636
```

I. References

Analytics, M. (n.d.). *Logistic Regression in R – Part Two*. Retrieved from https://mathewanalytics.com/2015/09/02/logistic-regression-in-r-part-two/

Brownlee, J. (2016, February 1). *How to Evaluate Machine Learning Algorithms with R*. Retrieved from Machine Learning Mastery: https://machinelearningmastery.com/evaluate-machine-learning-algorithms-with-r/

Li, S. (2017, November 16). *Predict Customer Churn with R*. Retrieved from Towards Data Science: https://towardsdatascience.com/predict-customer-churn-with-r-9e62357d47b4