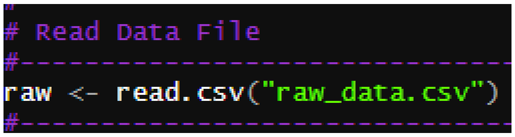
C744 Project Write Up

James Shea

Section 1:

Data Extraction:

Data was extracted from the provided file using the following code:



```
# Read Data File
#--------------------------------
raw <- read.csv("raw_data.csv")
#--------------------------------
```

Part A:

I have chosen to use R over SAS and Python due to its widespread usage, its large variety of possible packages to use, and because it is free. While R may have issues with larger data sets, this data set is relatively small and well within the capability of R. Thus, I can still run my analyses quickly, freely, and get the benefits of using an open source software with many available packages that are updated frequently.

Part B:

The goal of this data analysis is to attempt to predict whether a customer will churn. This analysis will be using the provided data set and thus will be limited to the variables, levels, and observations provided within. The data provided will be used by using relevant non-churn variables to predict churn. This analysis could provide insight into the behaviors of the customers of this company by providing information on what customers who churn have in common.

Part C:

The descriptive method that will be used is Multiple Correspondence Analysis. FactoMineR, a package available in R, provides this capability. This method will be used to detect patterns of relationships between churn and the rest of the variables. This will allow for the detection of which variables have the greatest impact on churn. This method is appropriate to this analysis due to the large quantity of nominal variables provided in the data.

The predictive method that will be used is logistic regression. This method is used to predict a binary outcome variable. Logistic regression is used when predicting whether an outcome variable is either 0 or 1 (false or true) and there are no other values possible. This method is appropriate as churn is a binary variable.

Section 2:

Part D:

The target variable in the data is churn. Churn in this data is binary. The two values possible for churn in the data are Yes and No. No other values are present, nor are any values missing. A Yes value means that the customer did churn (stopped being a customer). A No value means that the customer did not churn (is still a customer).

Part E:

An independent predictor variable in the data is gender. Gender is nominal and binary. The two possible values for this variable are Male and Female. There are no missing values or any other values present. Male and Female have no order, thus the variable is nominal.

Part F:

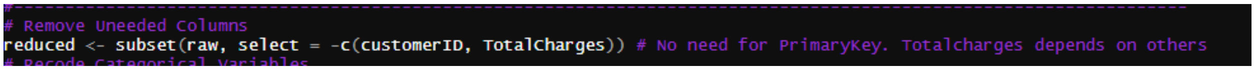
There are several goals when manipulating and preparing data. First, columns that are not needed for the analysis should be removed. Second, to identify missing and duplicated values/observations and deal with them appropriately. Third, to recode and/or impute variables when needed and appropriate in order to make analysis possible. Finally, I also rename some columns to make column name capitalization formats consistent.

Part G:

The target population is customers, particularly the customers who churn. The statistical entity studied is individual persons. The phenomenon to be predicted is churn. CustomerID is not needed for this analysis as there is no need to identify particular observations. TotalCharges is not needed for this analysis as it is a function of other variables in the data, in particular MonthlyCharges and Tenure, but possibly others as well. The rest of the criteria are essential as each potentially provides unique information about customer churn. These variables will all be turned into binary variables in order to conduct the analysis later.

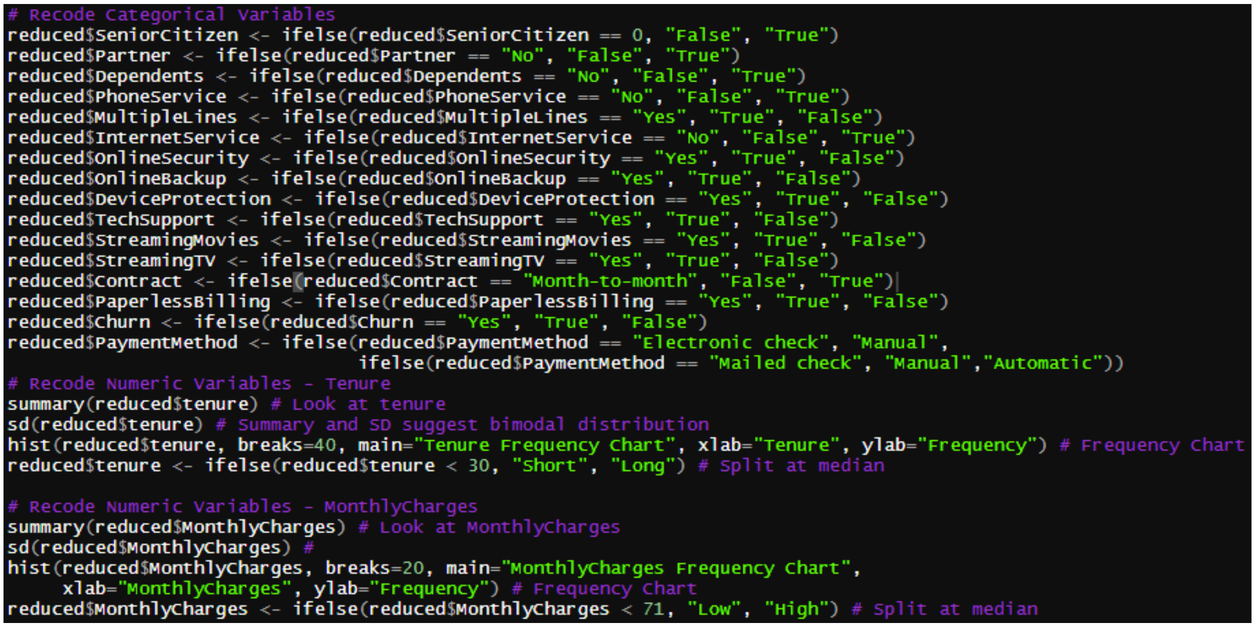
Part H:

1) Removed columns not needed for analysis.



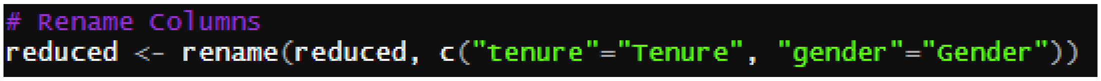
```
#----------------------------------------------------------------------------------------------------
# Remove Uneeded Columns
reduced <- subset(raw, select = -c(customerID, TotalCharges)) # No need for PrimaryKey. Totalcharges depends on others
# Recode Categorical Variables
```

2) Recoded variables.



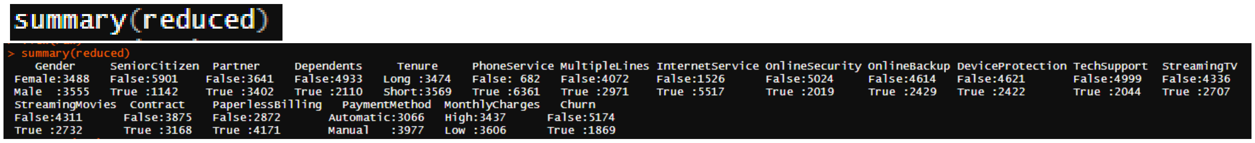
```
# Recode Categorical Variables
reduced$SeniorCitizen <- ifelse(reduced$SeniorCitizen == 0, "False", "True")
reduced$Partner <- ifelse(reduced$Partner == "No", "False", "True")
reduced$Dependents <- ifelse(reduced$Dependents == "No", "False", "True")
reduced$PhoneService <- ifelse(reduced$PhoneService == "No", "False", "True")
reduced$MultipleLines <- ifelse(reduced$MultipleLines == "Yes", "True", "False")
reduced$InternetService <- ifelse(reduced$InternetService == "No", "False", "True")
reduced$OnlineSecurity <- ifelse(reduced$OnlineSecurity == "Yes", "True", "False")
reduced$OnlineBackup <- ifelse(reduced$OnlineBackup == "Yes", "True", "False")
reduced$DeviceProtection <- ifelse(reduced$DeviceProtection == "Yes", "True", "False")
reduced$TechSupport <- ifelse(reduced$TechSupport == "Yes", "True", "False")
reduced$StreamingMovies <- ifelse(reduced$StreamingMovies == "Yes", "True", "False")
reduced$StreamingTV <- ifelse(reduced$StreamingTV == "Yes", "True", "False")
reduced$Contract <- ifelse(reduced$Contract == "Month-to-month", "False", "True")
reduced$PaperlessBilling <- ifelse(reduced$PaperlessBilling == "Yes", "True", "False")
reduced$Churn <- ifelse(reduced$Churn == "Yes", "True", "False")
reduced$PaymentMethod <- ifelse(reduced$PaymentMethod == "Electronic check", "Manual",
                                ifelse(reduced$PaymentMethod == "Mailed check", "Manual","Automatic"))
# Recode Numeric Variables - Tenure
summary(reduced$tenure) # Look at tenure
sd(reduced$tenure) # Summary and SD suggest bimodal distribution
hist(reduced$tenure, breaks=40, main="Tenure Frequency Chart", xlab="Tenure", ylab="Frequency") # Frequency Chart
reduced$tenure <- ifelse(reduced$tenure < 30, "Short", "Long") # Split at median

# Recode Numeric Variables - MonthlyCharges
summary(reduced$MonthlyCharges) # Look at MonthlyCharges
sd(reduced$MonthlyCharges) #
hist(reduced$MonthlyCharges, breaks=20, main="MonthlyCharges Frequency Chart",
     xlab="MonthlyCharges", ylab="Frequency") # Frequency Chart
reduced$MonthlyCharges <- ifelse(reduced$MonthlyCharges < 71, "Low", "High") # Split at median
```

3) Renamed columns. This was done so all column names began with a capital letter for consistency.



```
# Rename Columns
reduced <- rename(reduced, c("tenure"="Tenure", "gender"="Gender"))
```

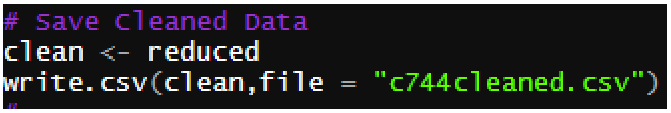
4) Checked for missing values. There were no missing values in the variables of interest and thus no action taken.



```
summary(reduced)
```

```
> summary(reduced)
   Gender      SeniorCitizen  Partner      Dependents     Tenure      PhoneService MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport  StreamingTV
 Female:3488   False:5901    False:3641   False:4933   Long :3474   False: 682   False:4072    False:1526      False:5024     False:4614   False:4621       False:4999   False:4336
 Male  :3555   True :1142    True :3402   True :2110   Short:3569   True :6361   True :2971    True :5517      True :2019     True :2429   True :2422       True :2044   True :2707
 StreamingMovies  Contract    PaperlessBilling  PaymentMethod  MonthlyCharges   Churn
 False:4311      False:3875   False:2872       Automatic:3066  High:3437       False:5174
 True :2732      True :3168   True :4171       Manual   :3977  Low :3606       True :1869
```

5) Saved the clean data.

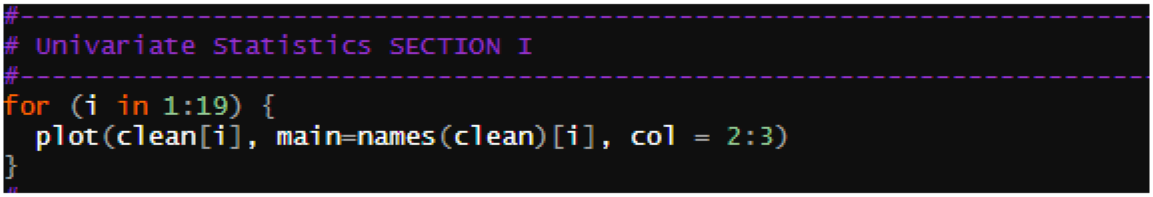


```
# Save Cleaned Data
clean <- reduced
write.csv(clean,file = "c744cleaned.csv")
```

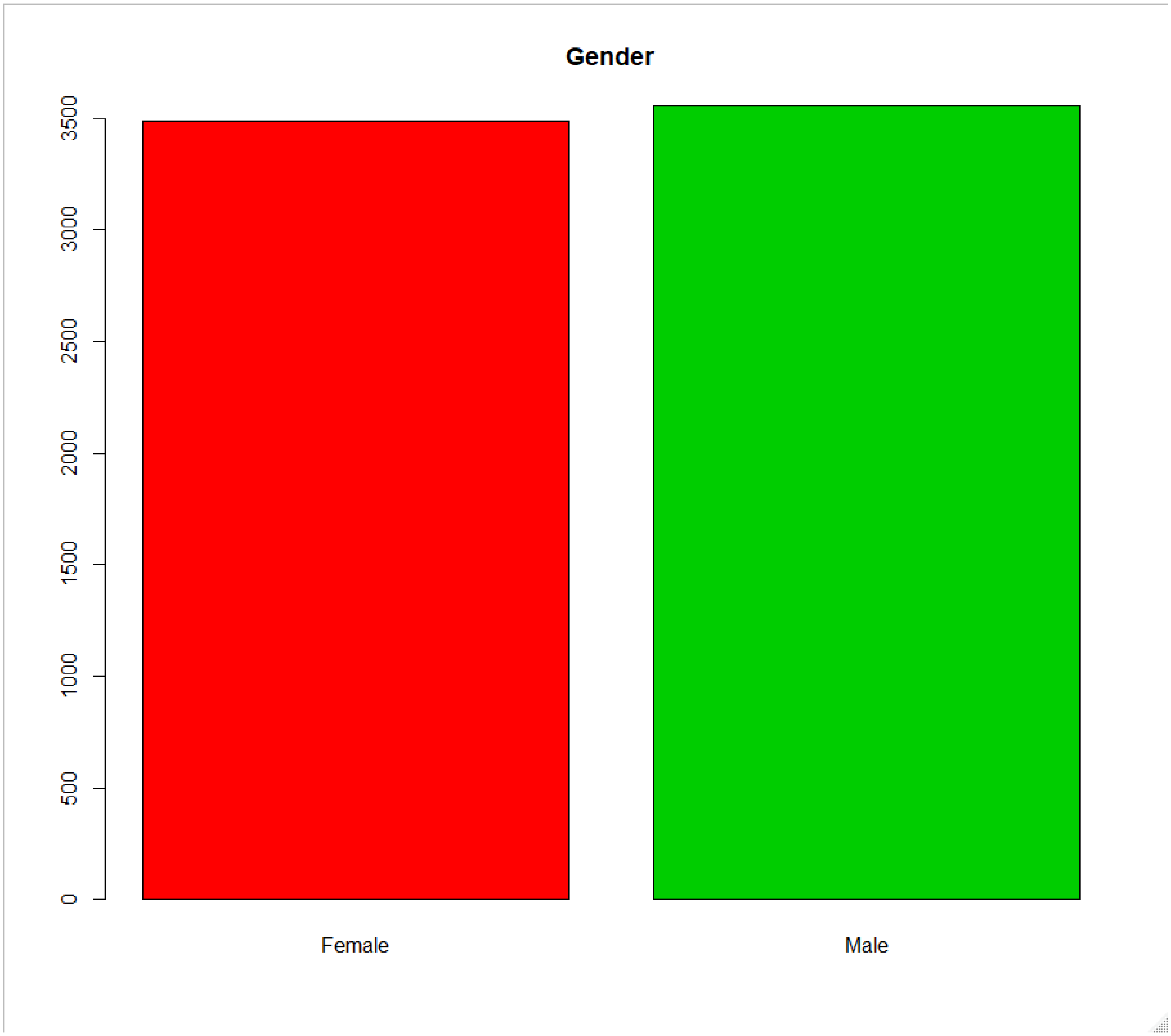
Section 3:

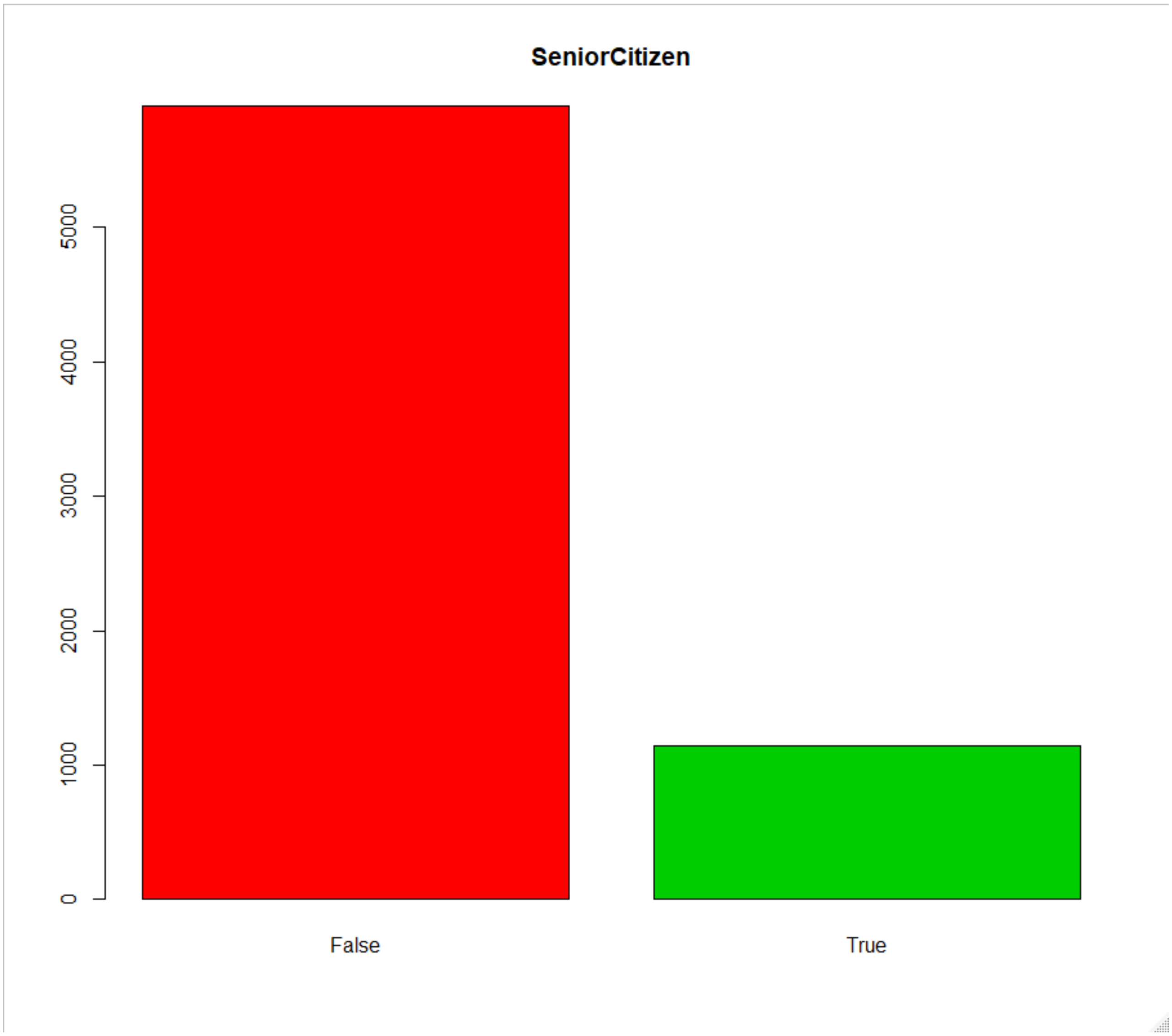
Part I:

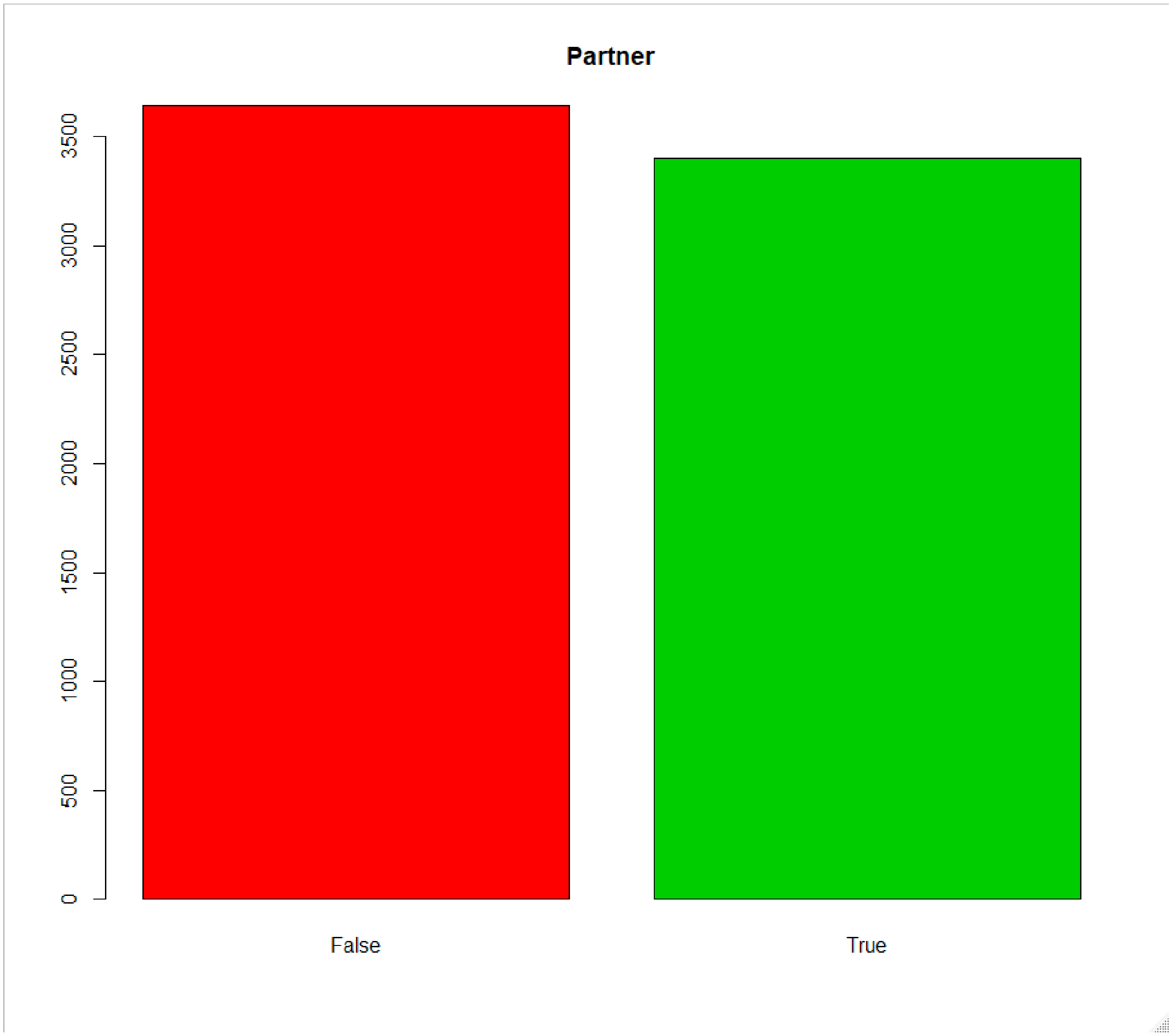
For the univariate statistics, I generated a bar graph for all variables. This first screen shot is a for loop which generated all the graphs in the rest of the screen shots for this part.

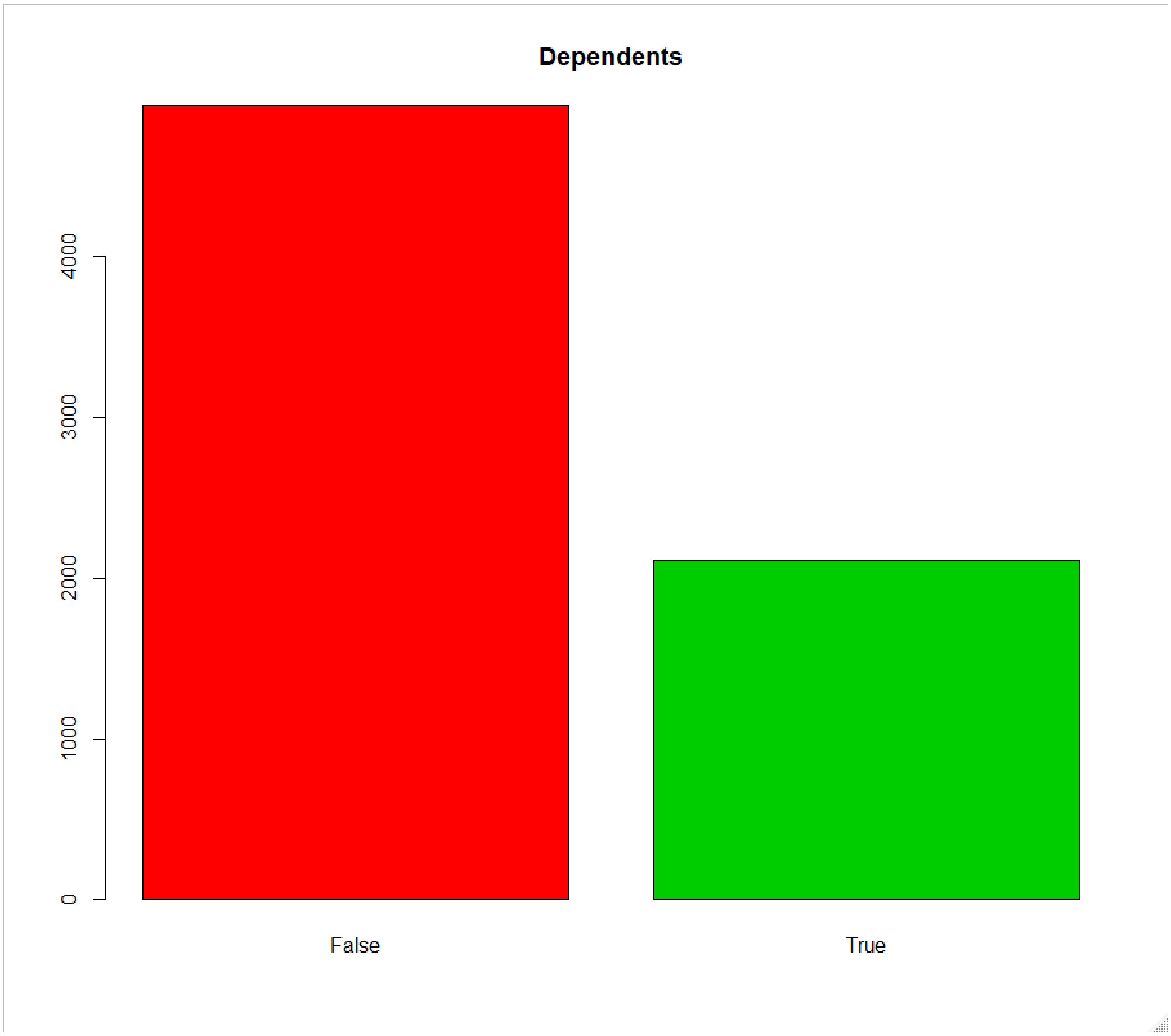


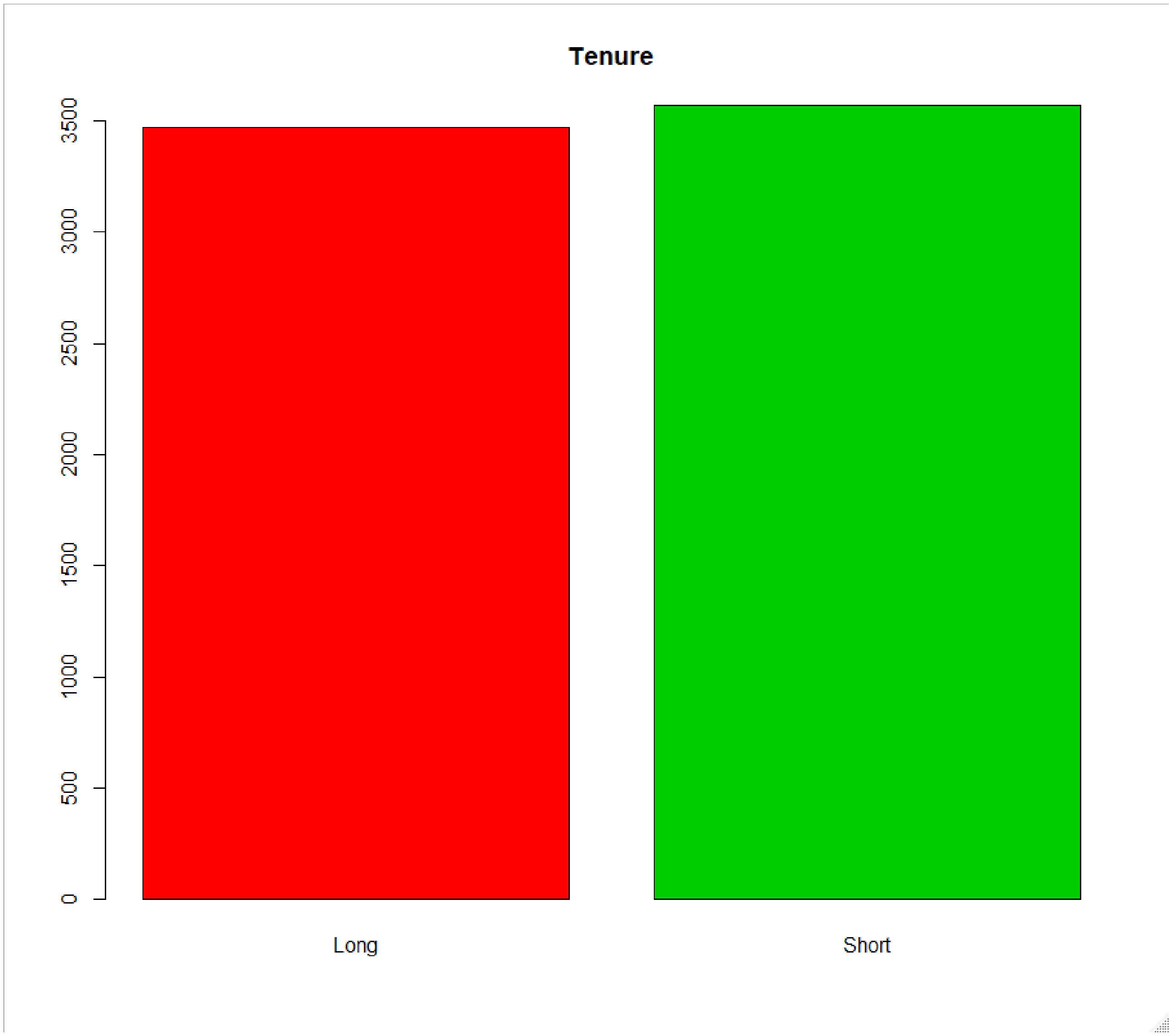
```
#-----------------------------------------------------------------------
# Univariate Statistics SECTION I
#-----------------------------------------------------------------------
for (i in 1:19) {
  plot(clean[i], main=names(clean)[i], col = 2:3)
}
```

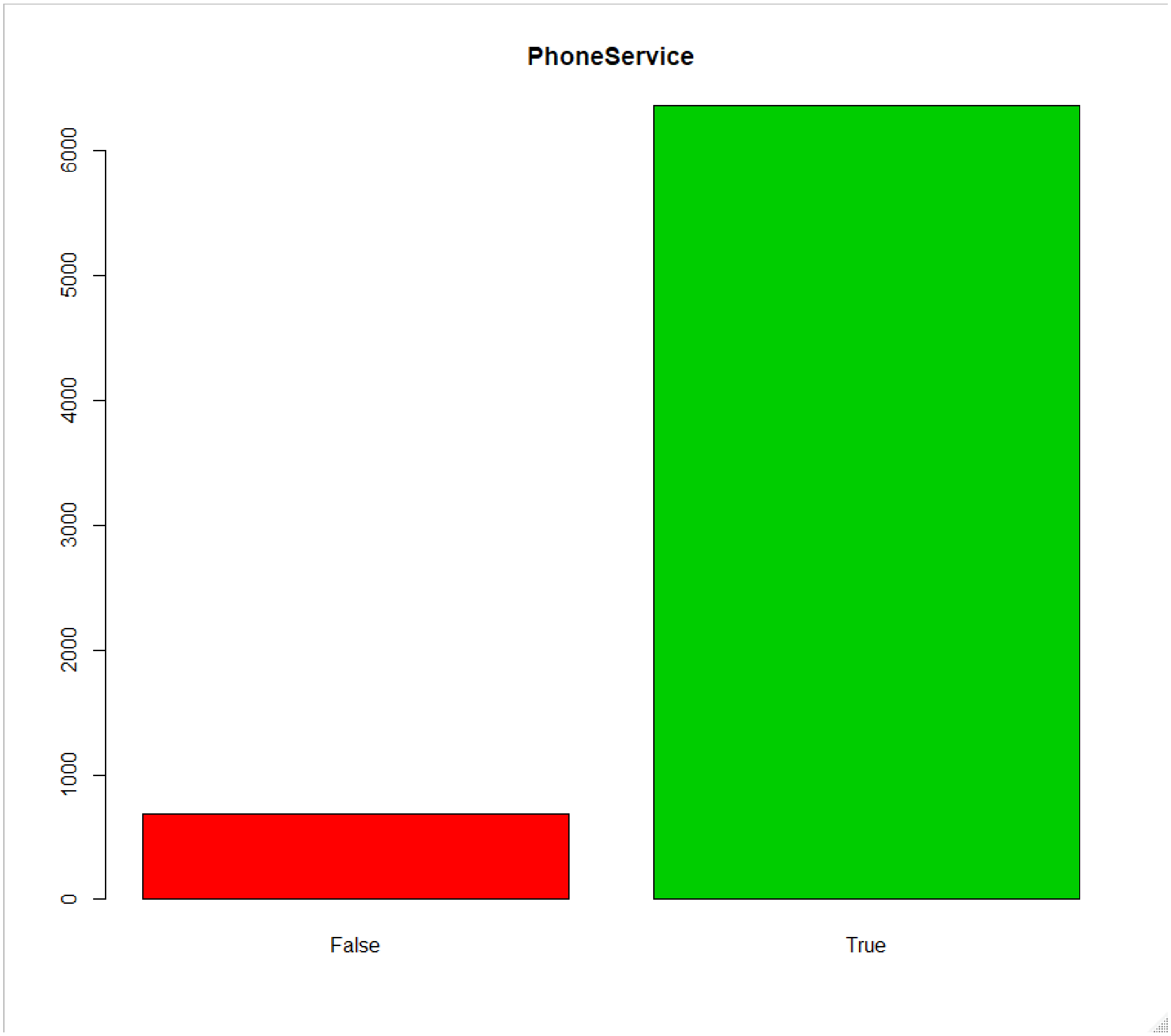

Gender
0
500
1000
1500
2000
2500
3000
3500
Female
Male

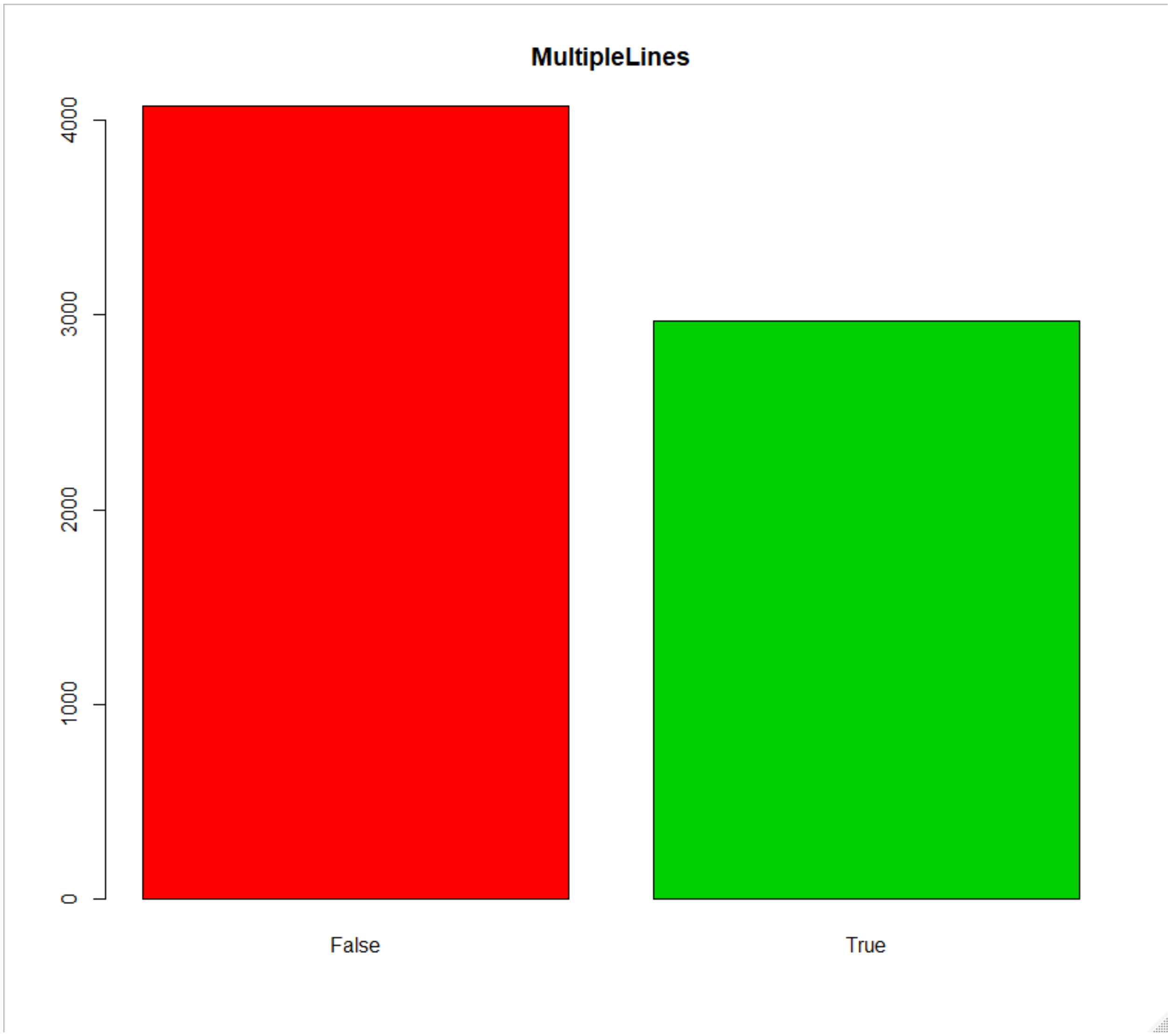

SeniorCitizen
0
1000
2000
3000
4000
5000
False
True

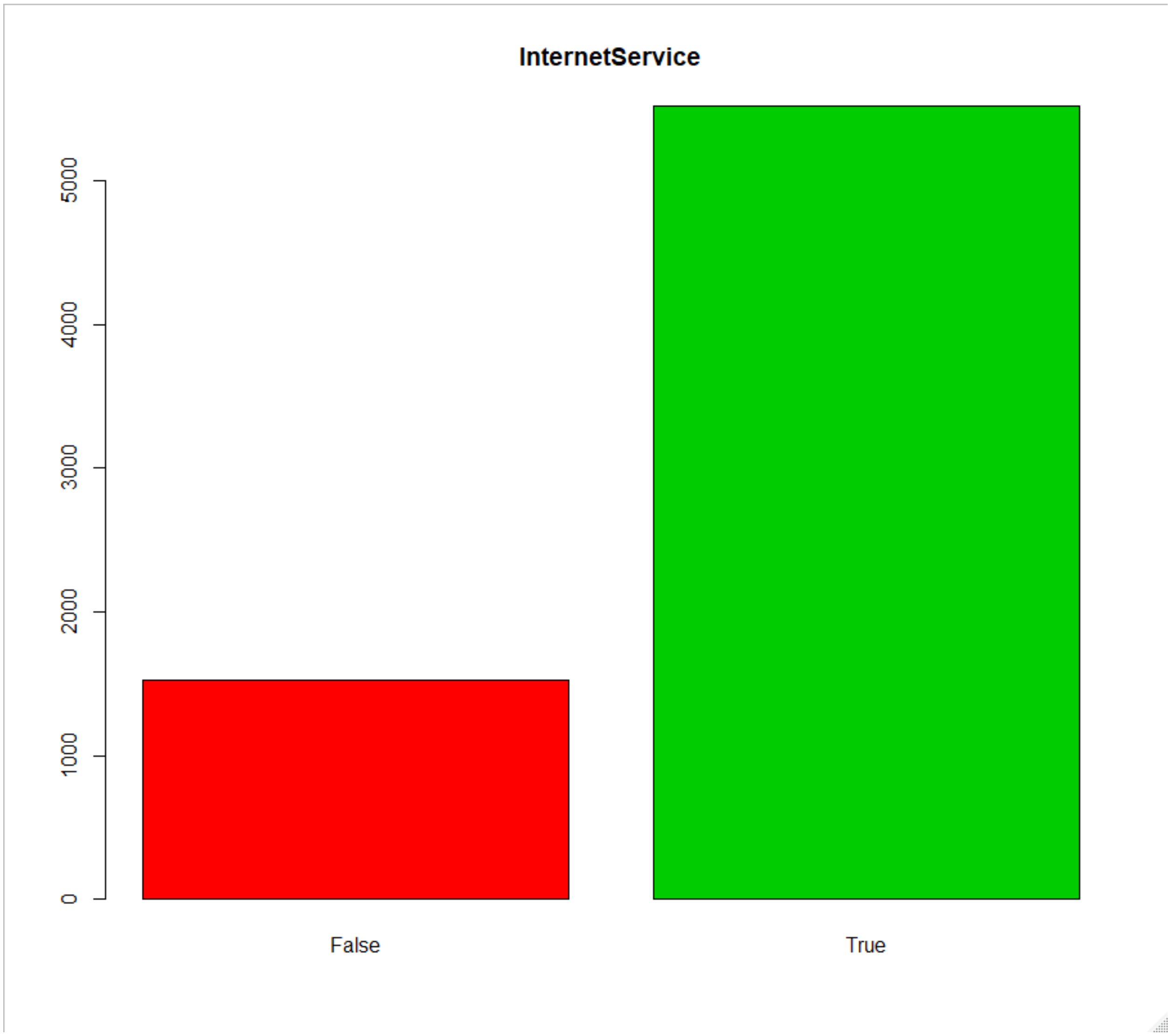

Partner
0
500
1000
1500
2000
2500
3000
3500
False
True

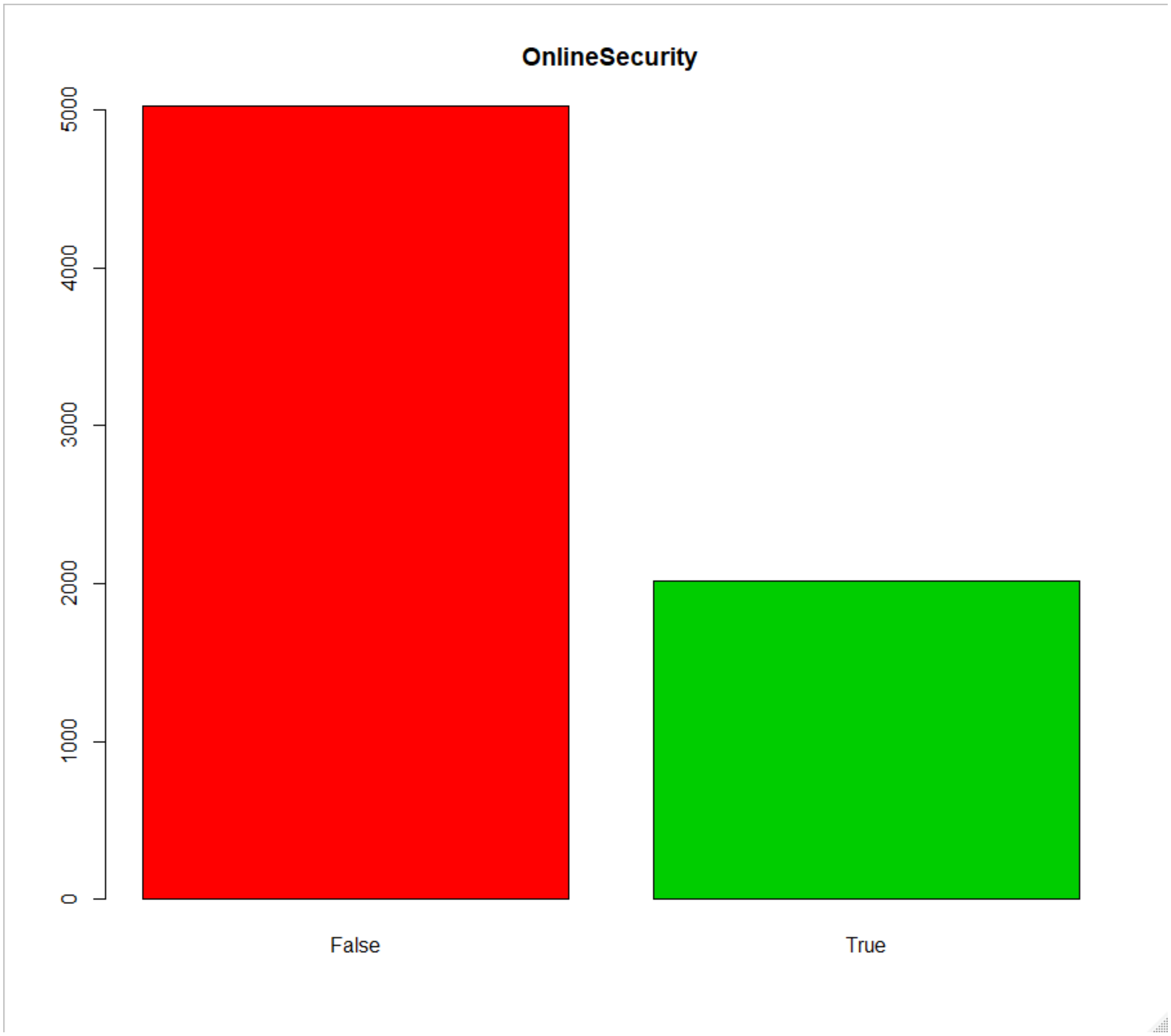

Dependents
0
1000
2000
3000
4000
False
True

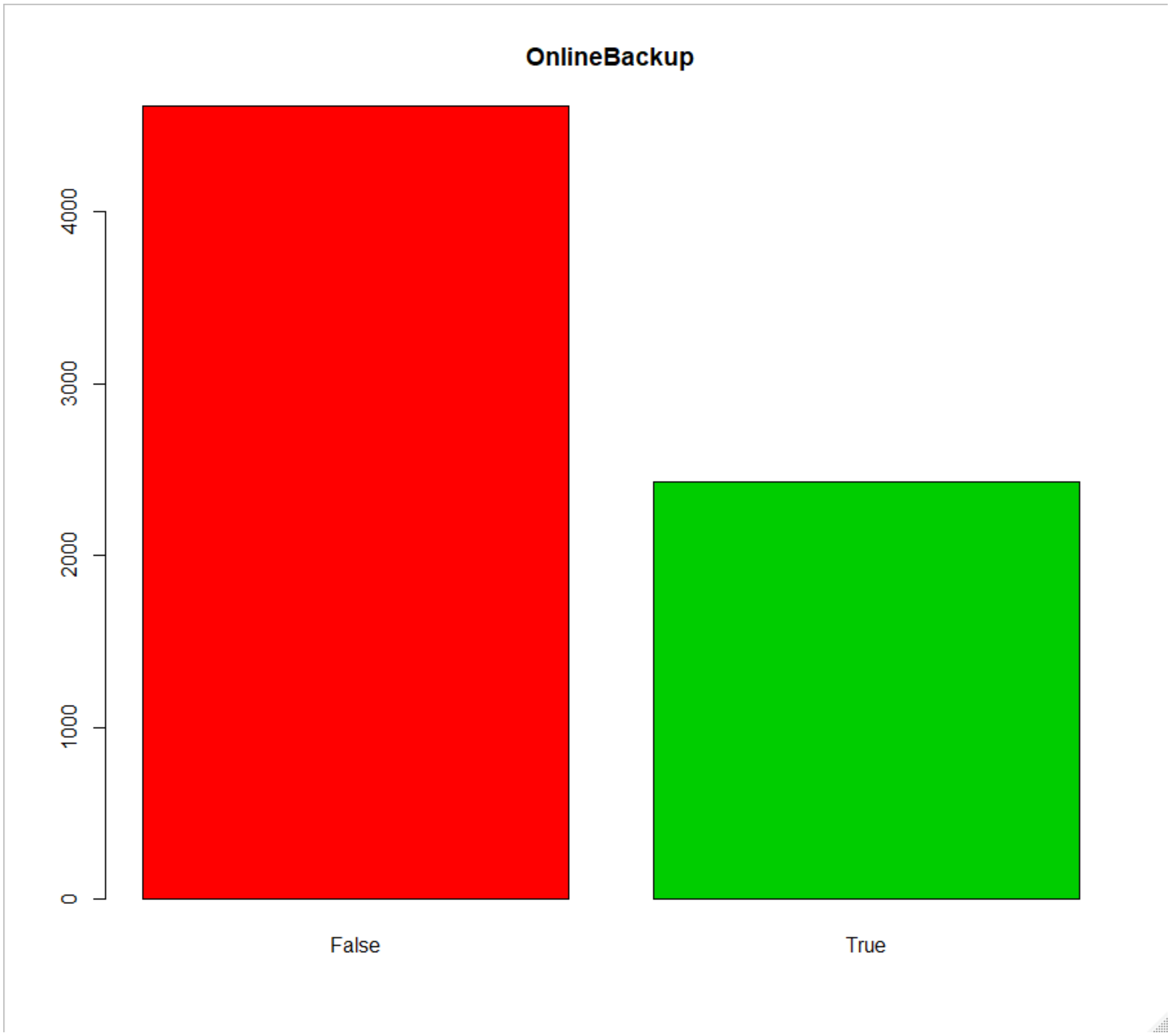

Tenure
0
500
1000
1500
2000
2500
3000
3500
Long
Short

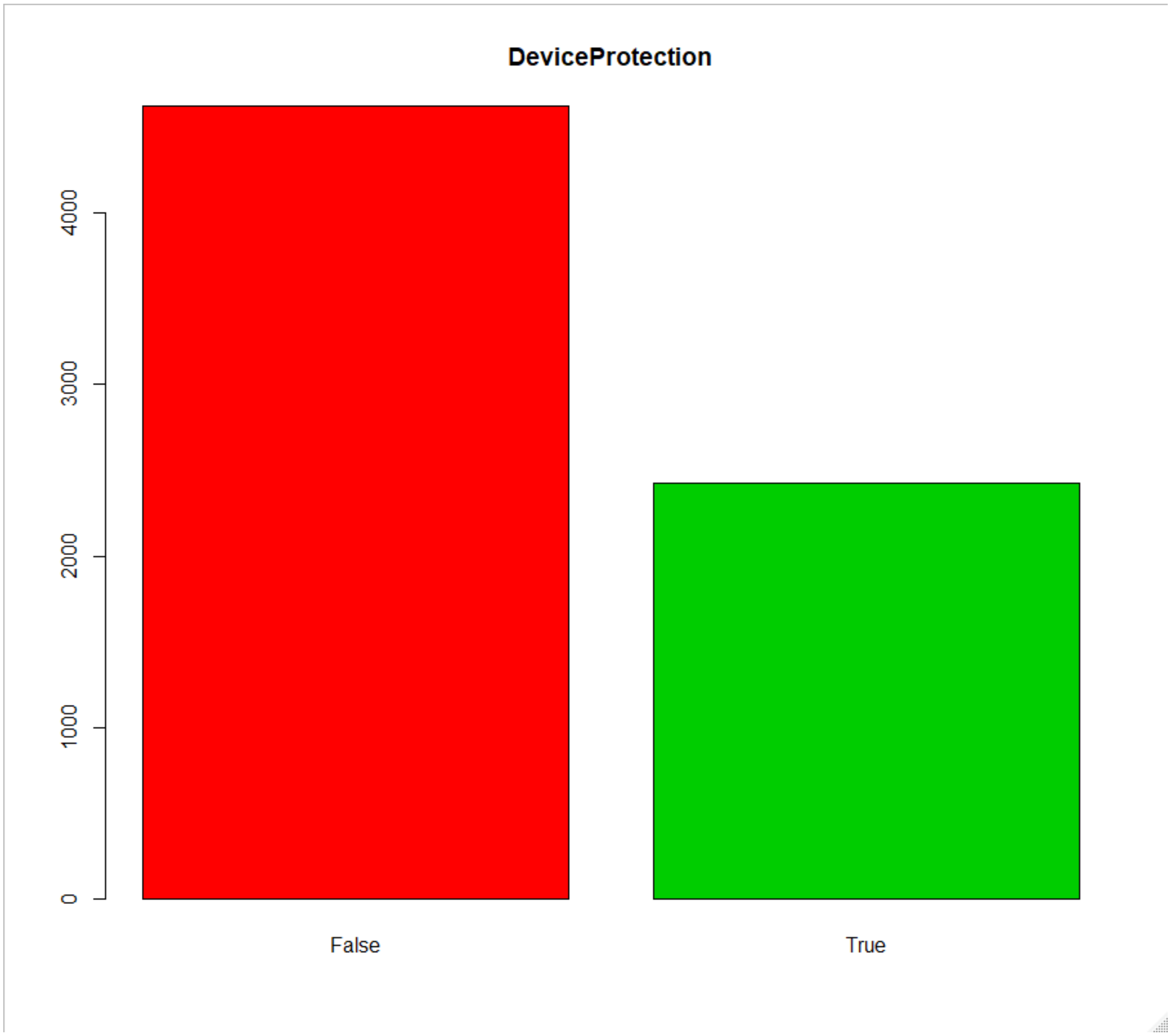

PhoneService
6000
5000
4000
3000
2000
1000
0
False
True

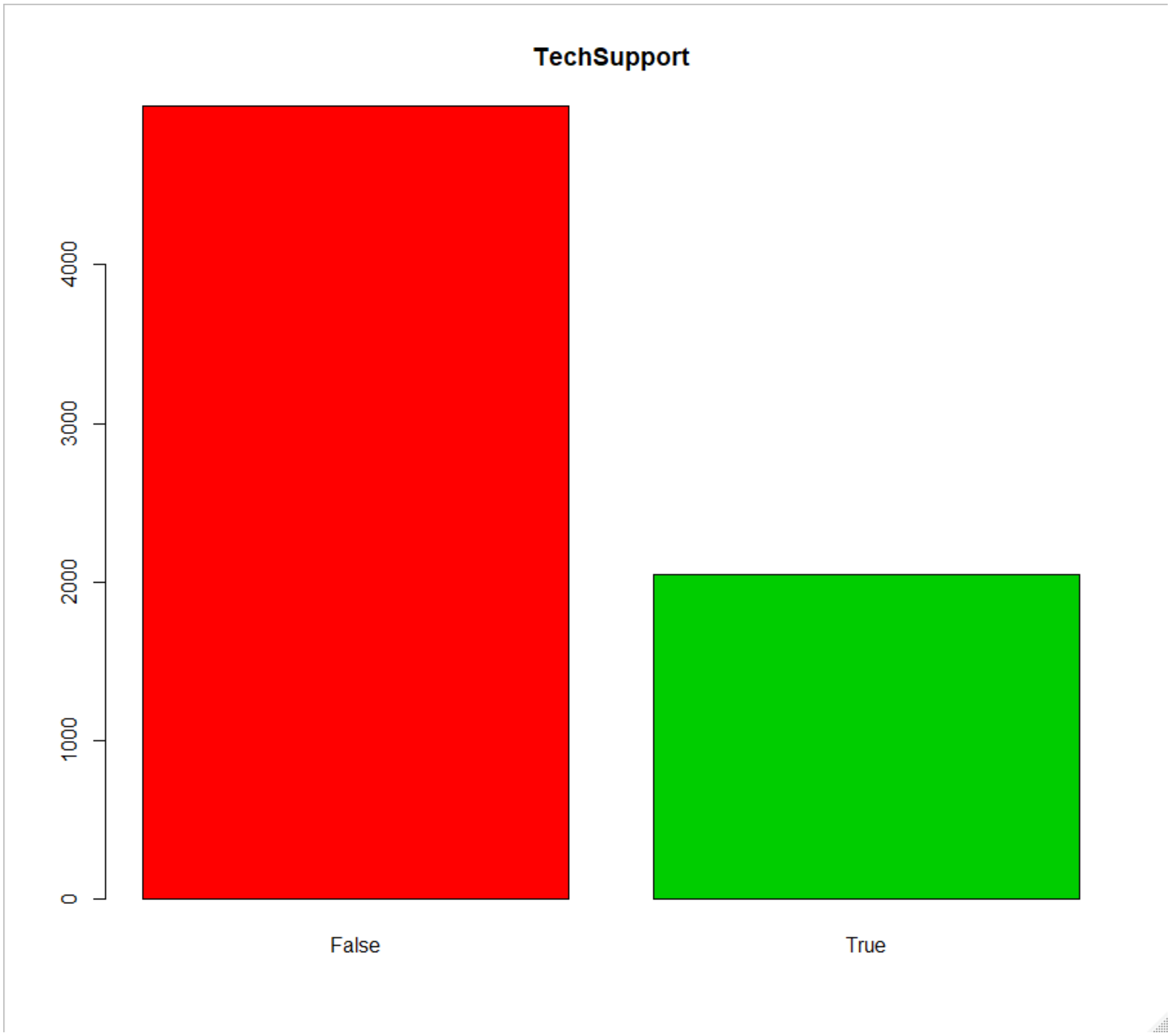

MultipleLines
4000
3000
2000
1000
0
False
True

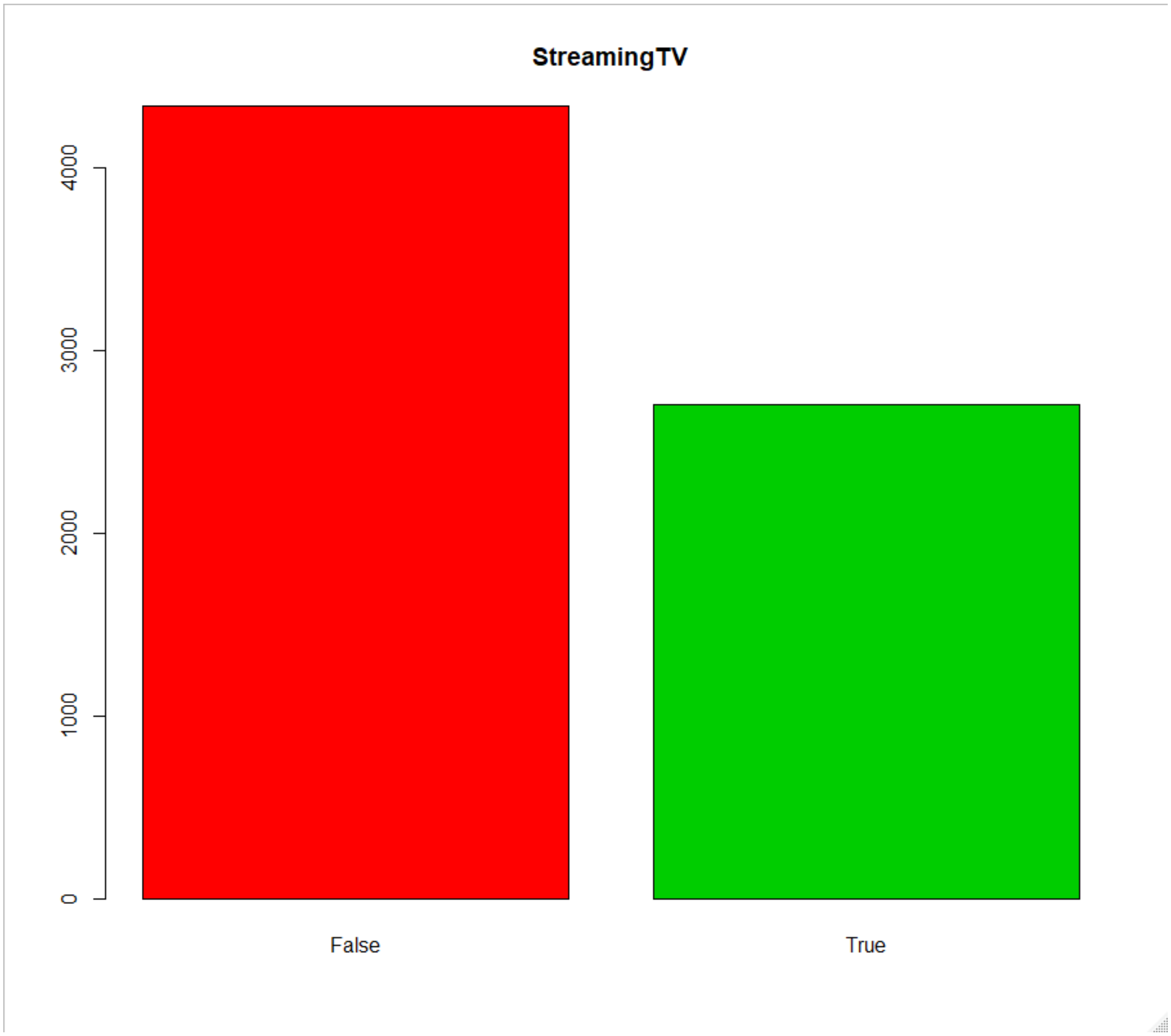

InternetService
0
1000
2000
3000
4000
5000
False
True

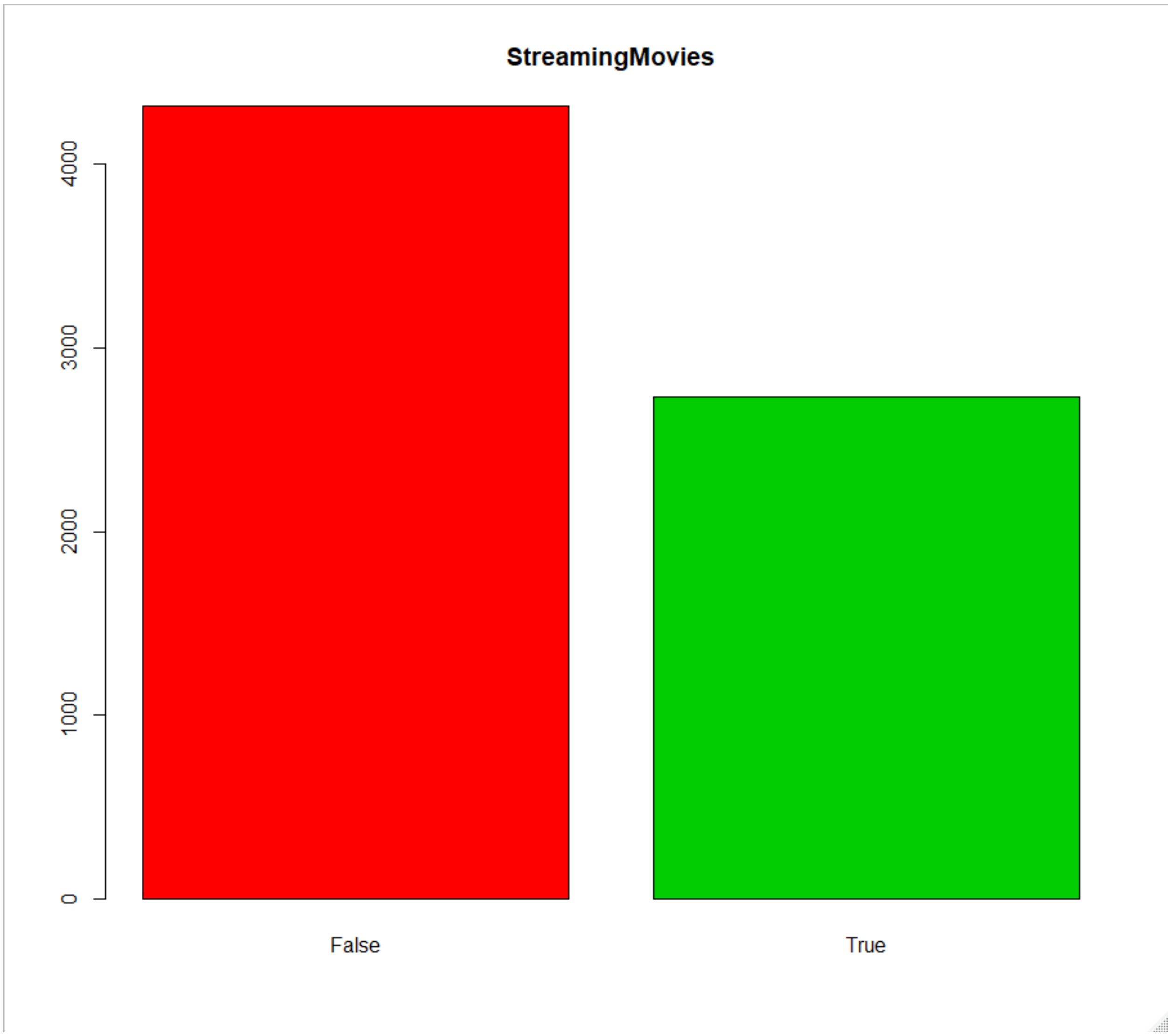

OnlineSecurity
0
1000
2000
3000
4000
5000
False
True

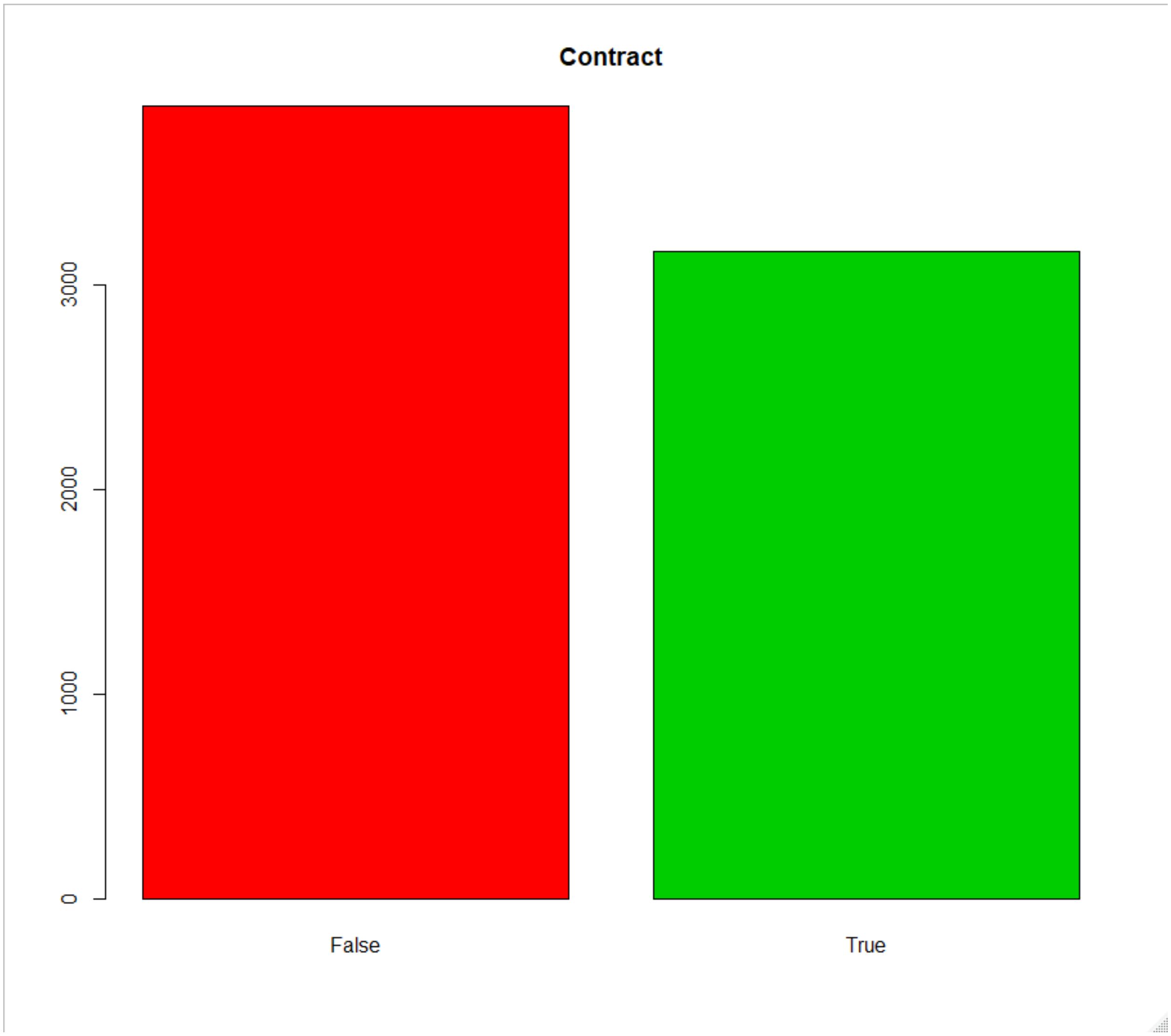

OnlineBackup
0
1000
2000
3000
4000
False
True

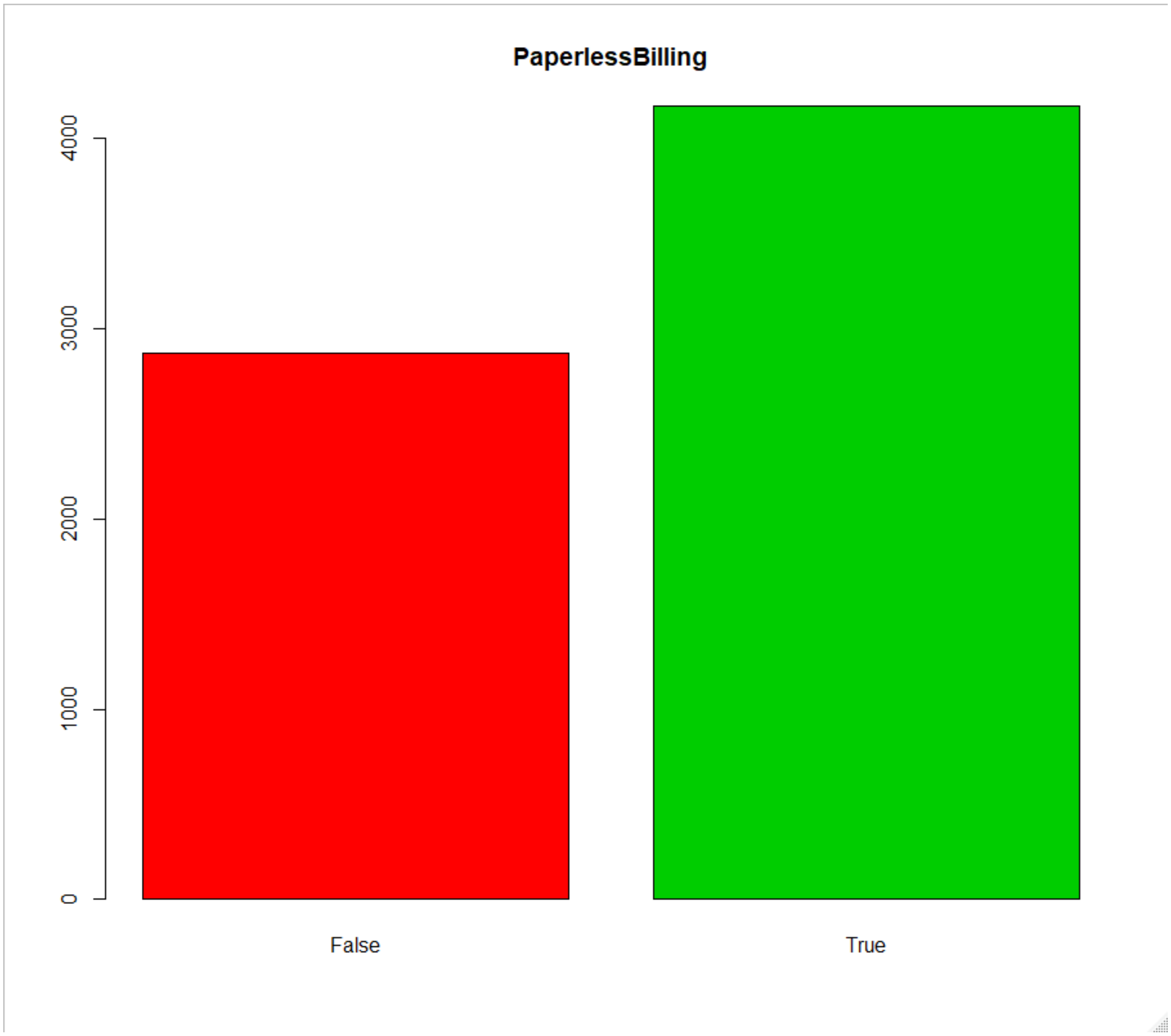

DeviceProtection
0
1000
2000
3000
4000
False
True

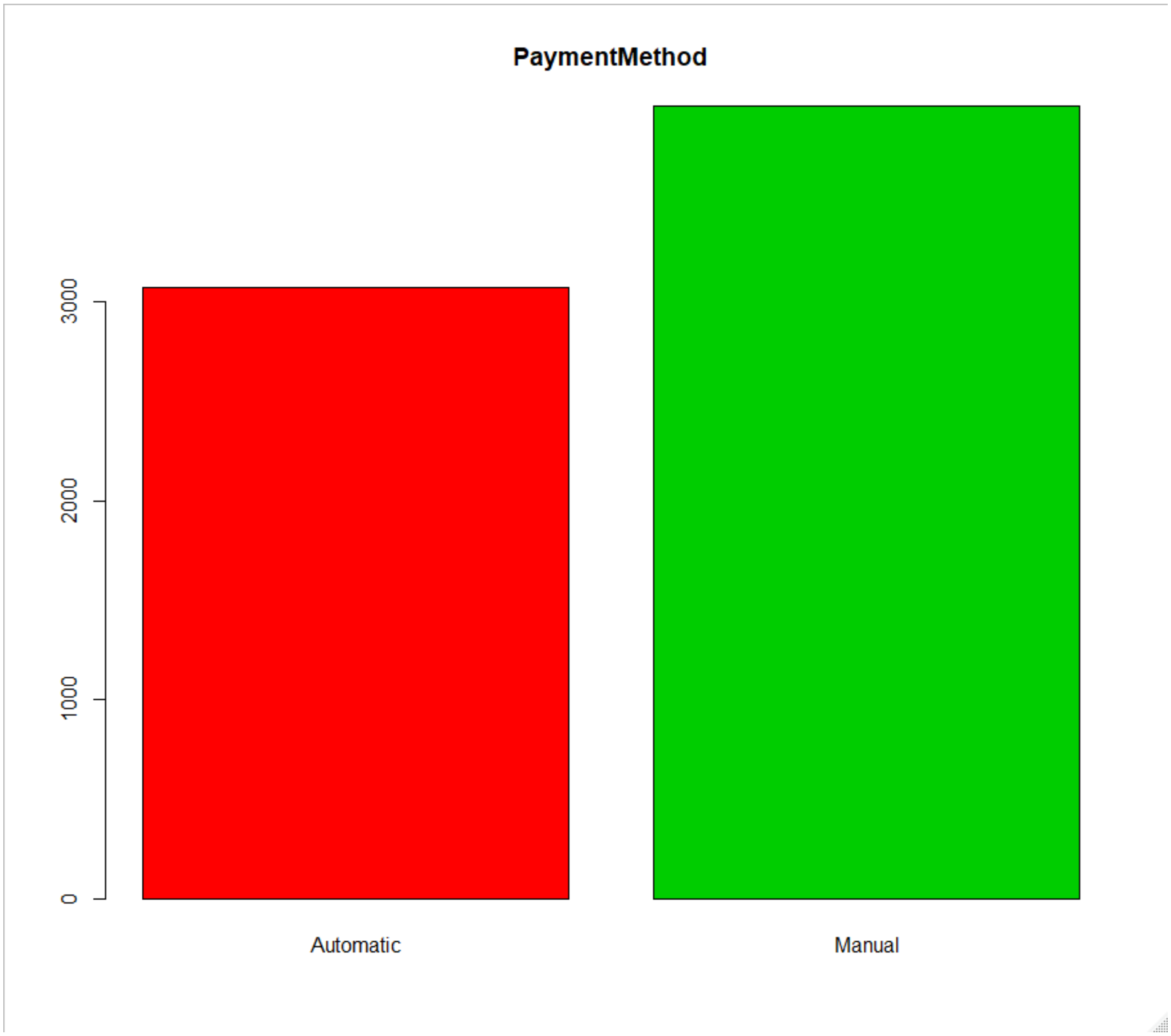

TechSupport
0
1000
2000
3000
4000
False
True

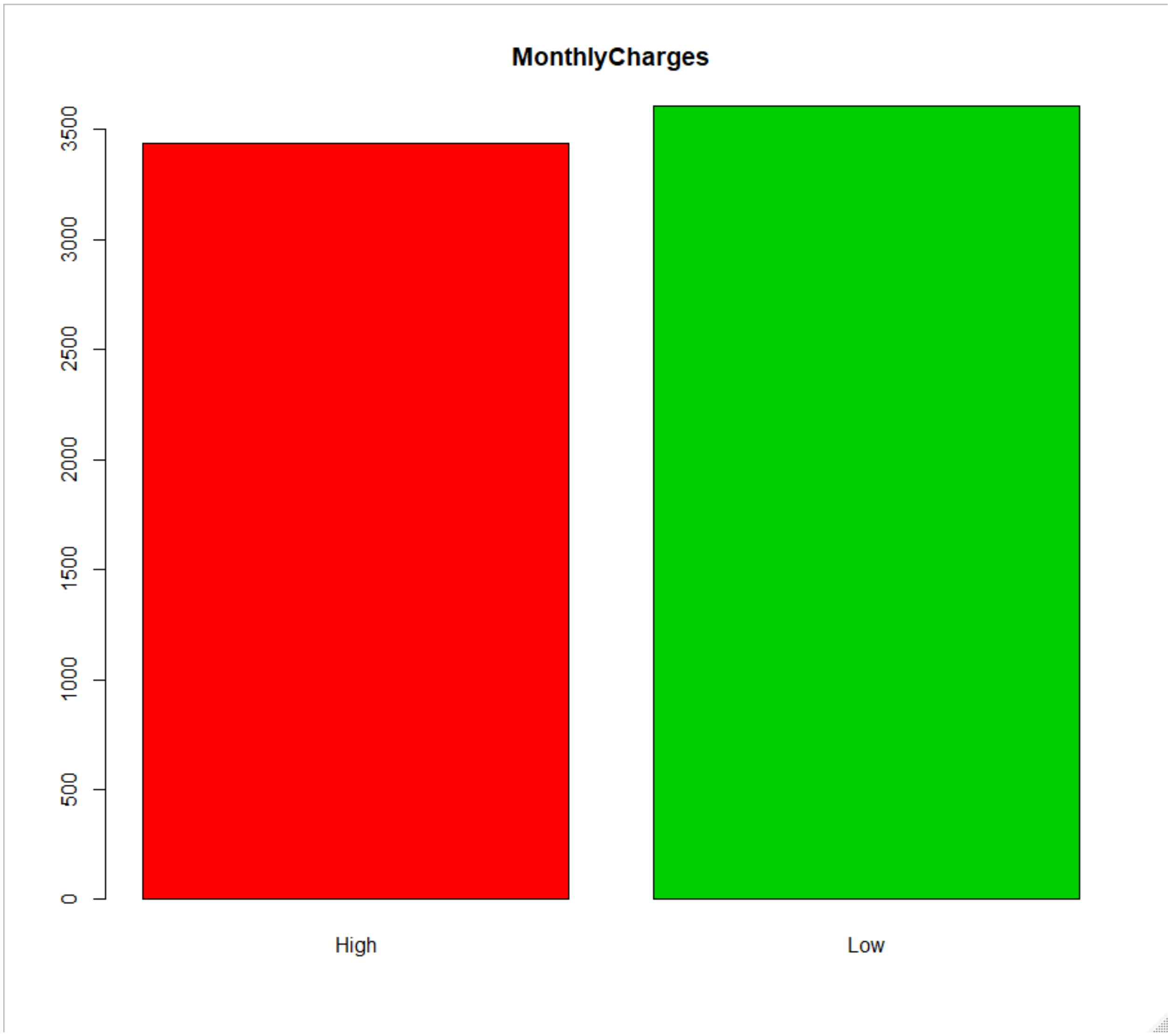

StreamingTV
0
1000
2000
3000
4000
False
True

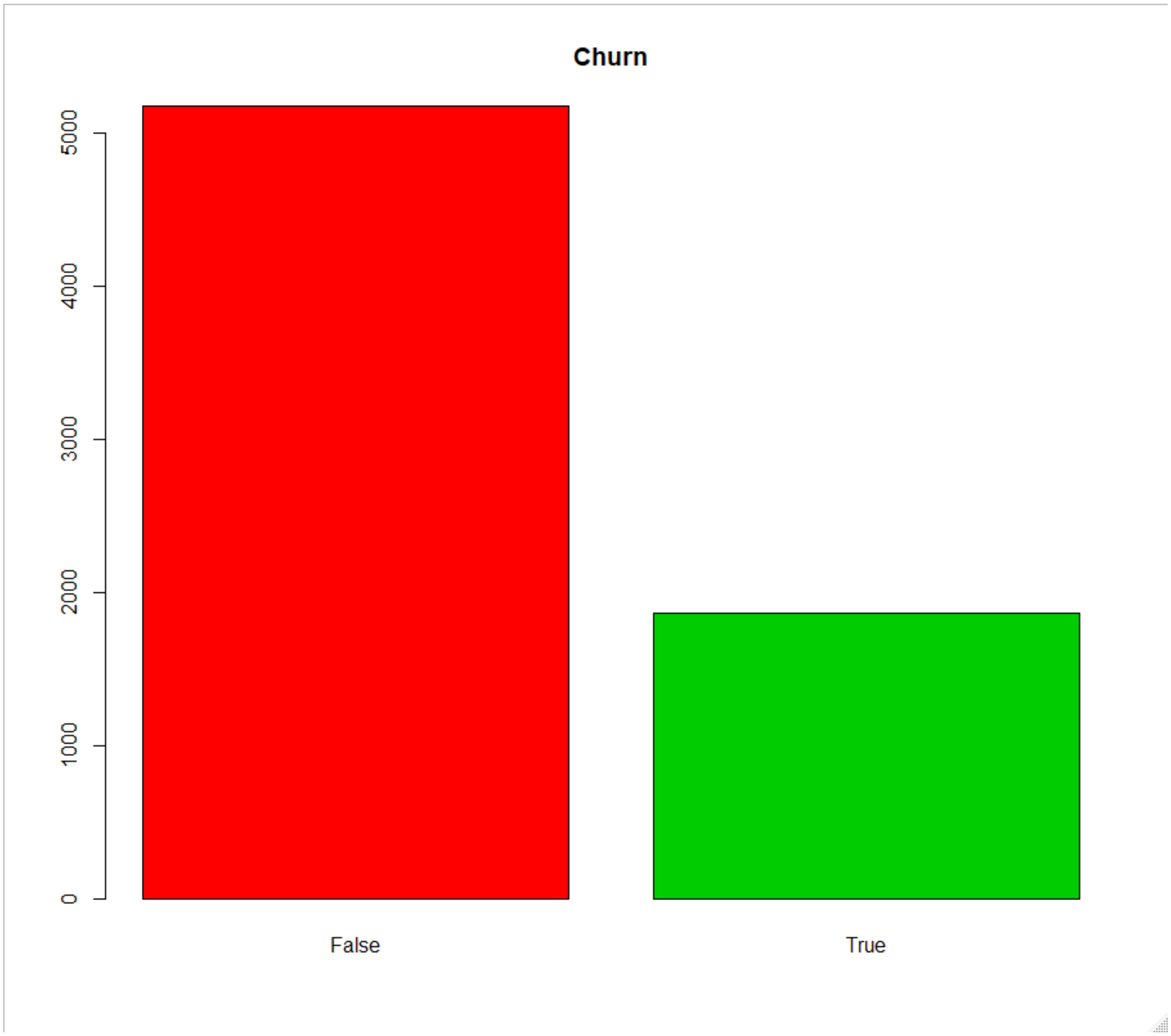

StreamingMovies
0
1000
2000
3000
4000
False
True


Contract
0
1000
2000
3000
False
True


PaperlessBilling
0
1000
2000
3000
4000
False
True

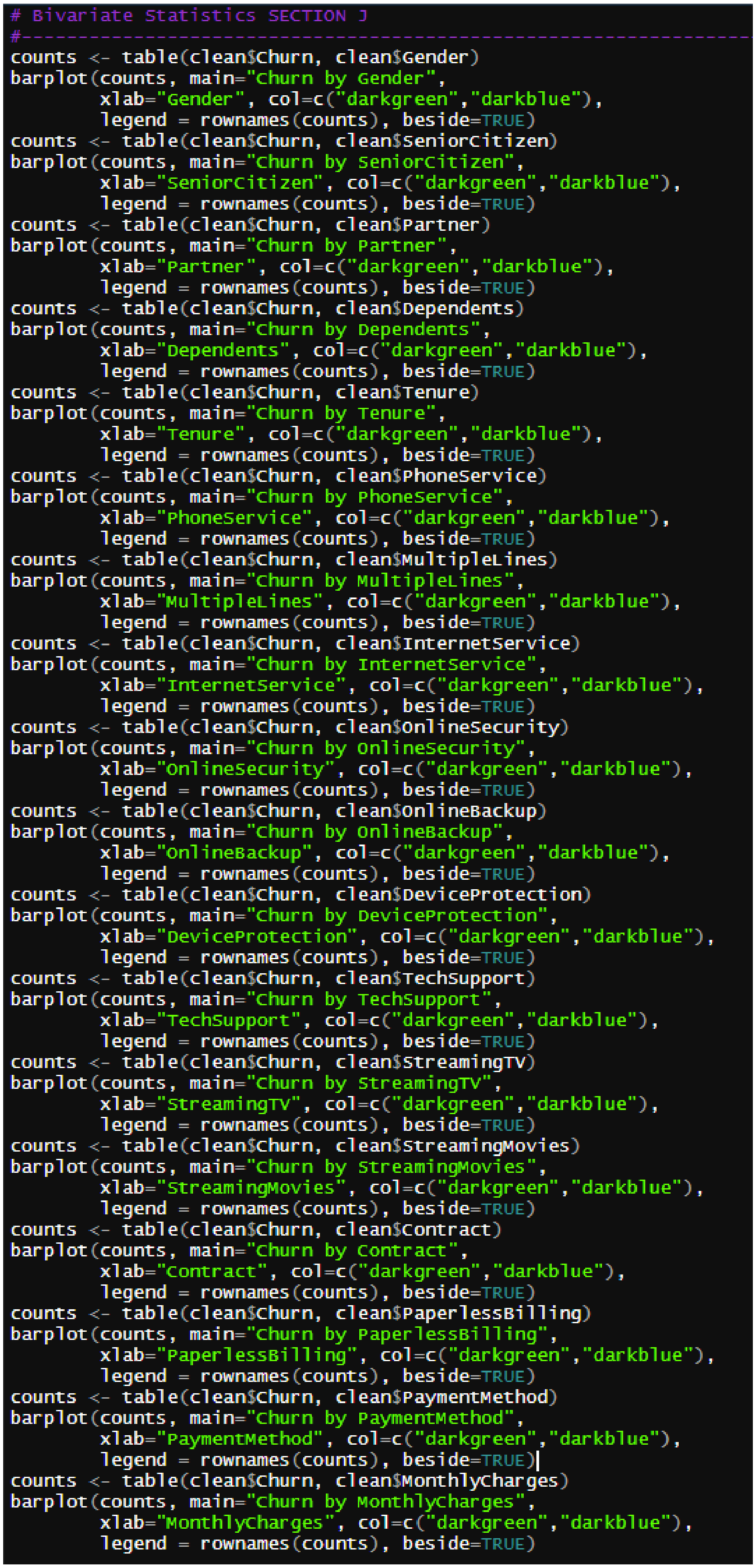

PaymentMethod
3000
2000
1000
0
Automatic
Manual


MonthlyCharges
0
500
1000
1500
2000
2500
3000
3500
High
Low

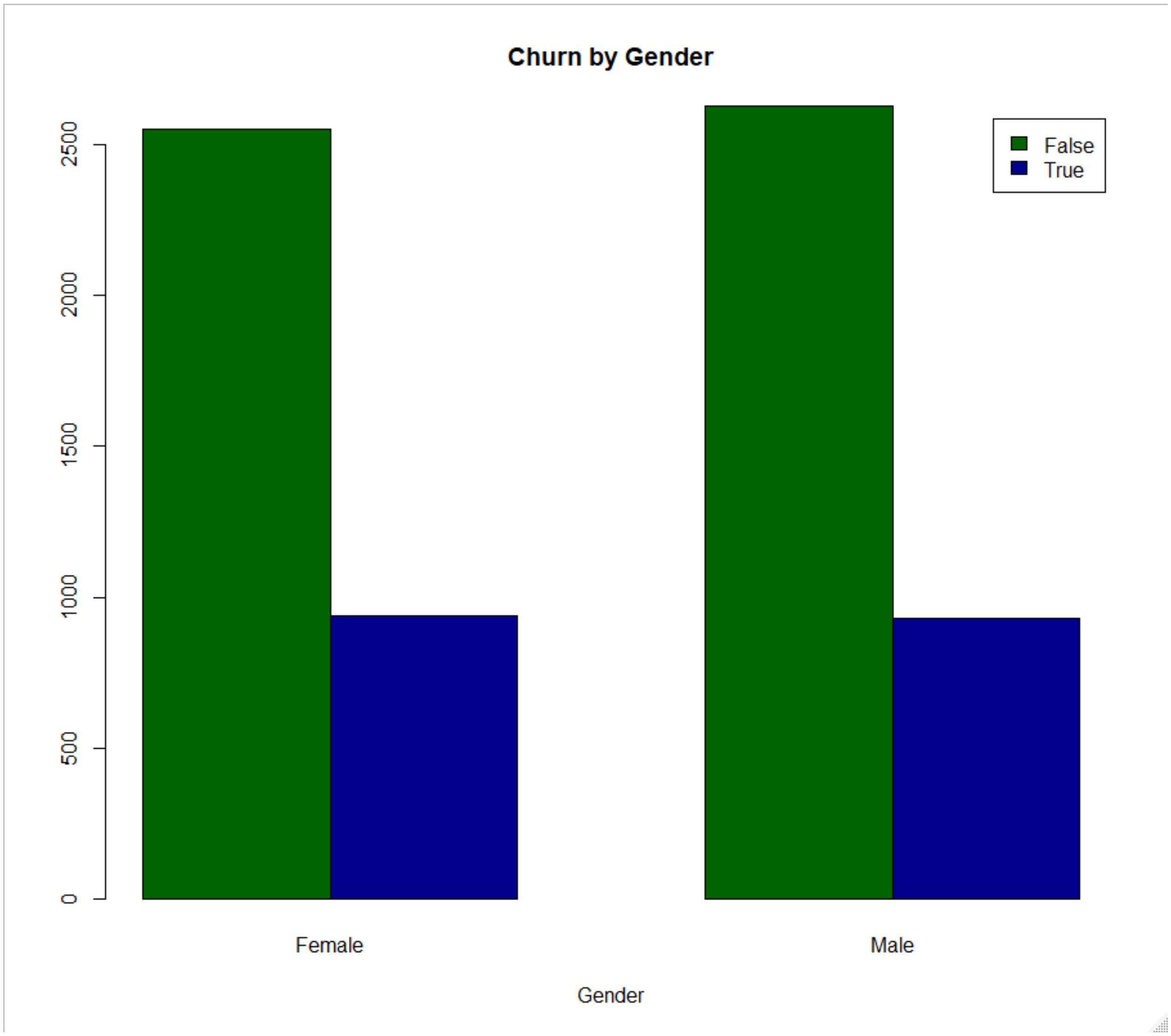


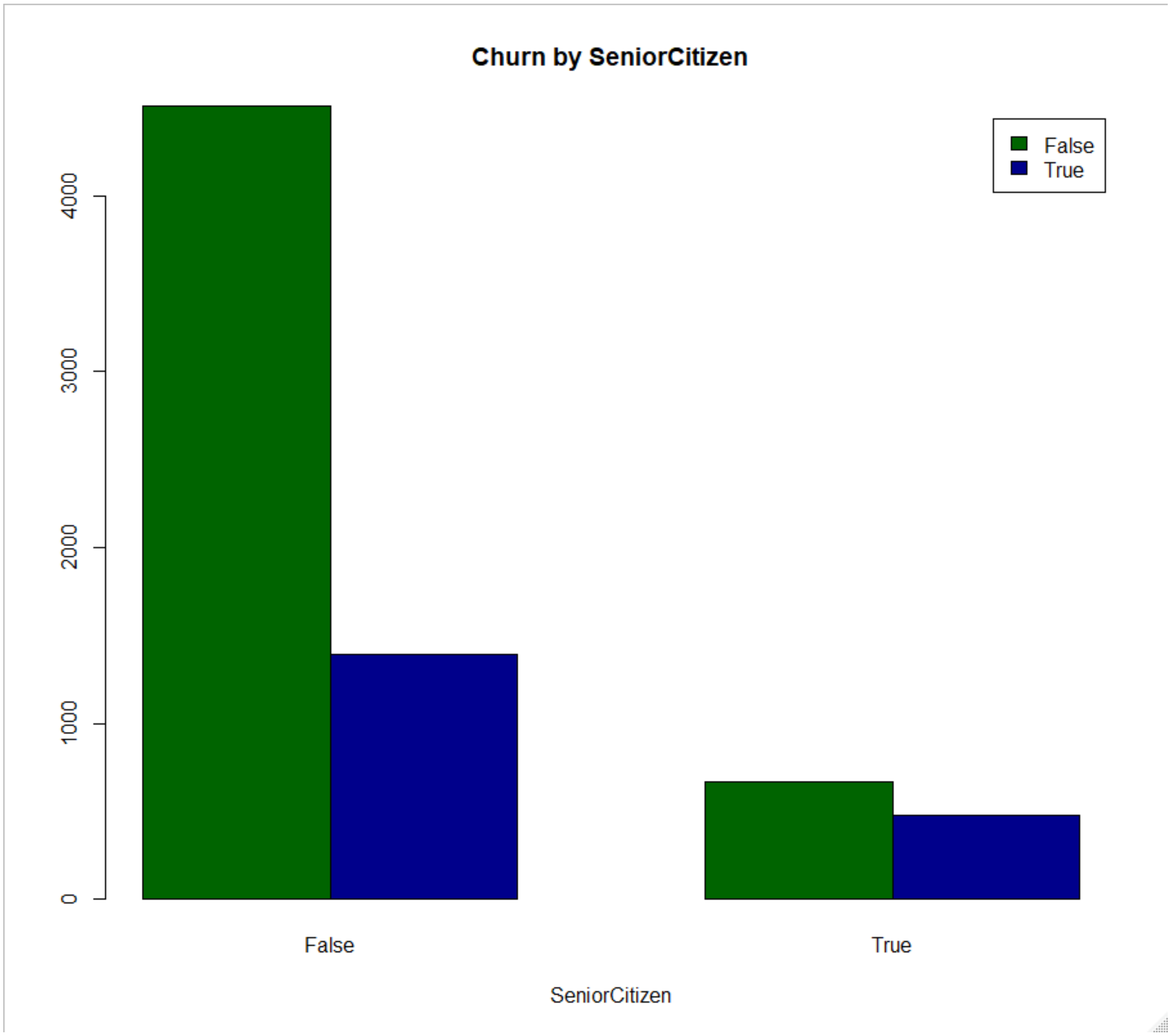
Part J:

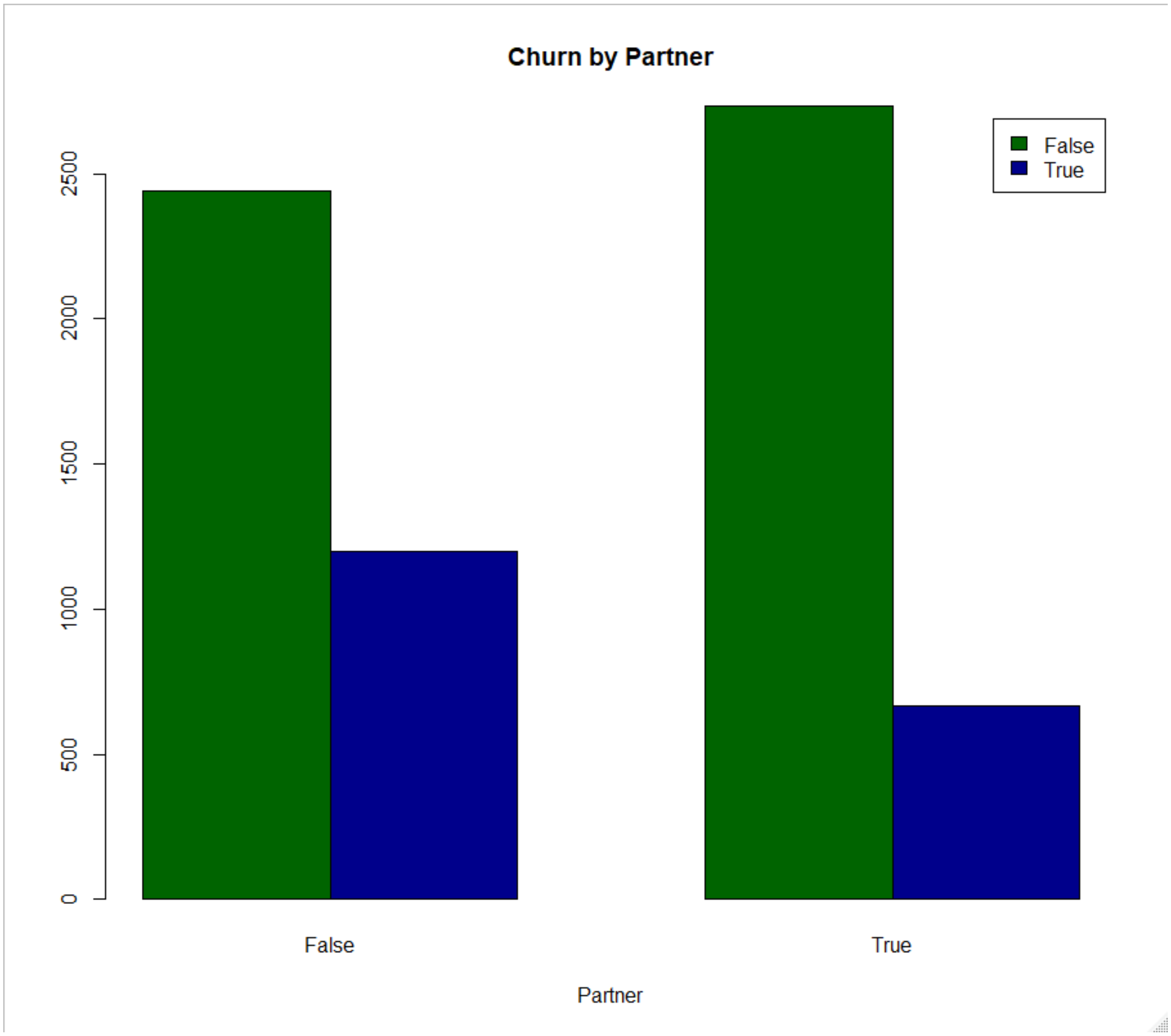
For the bivariate bar graphs, the code used to generate all graphs is provided in the first screen shot. The graphs generated then follow in the rest of the screen shots for this section. This was done for simplicity as the code for each graph is basically the with only names and variables changing.

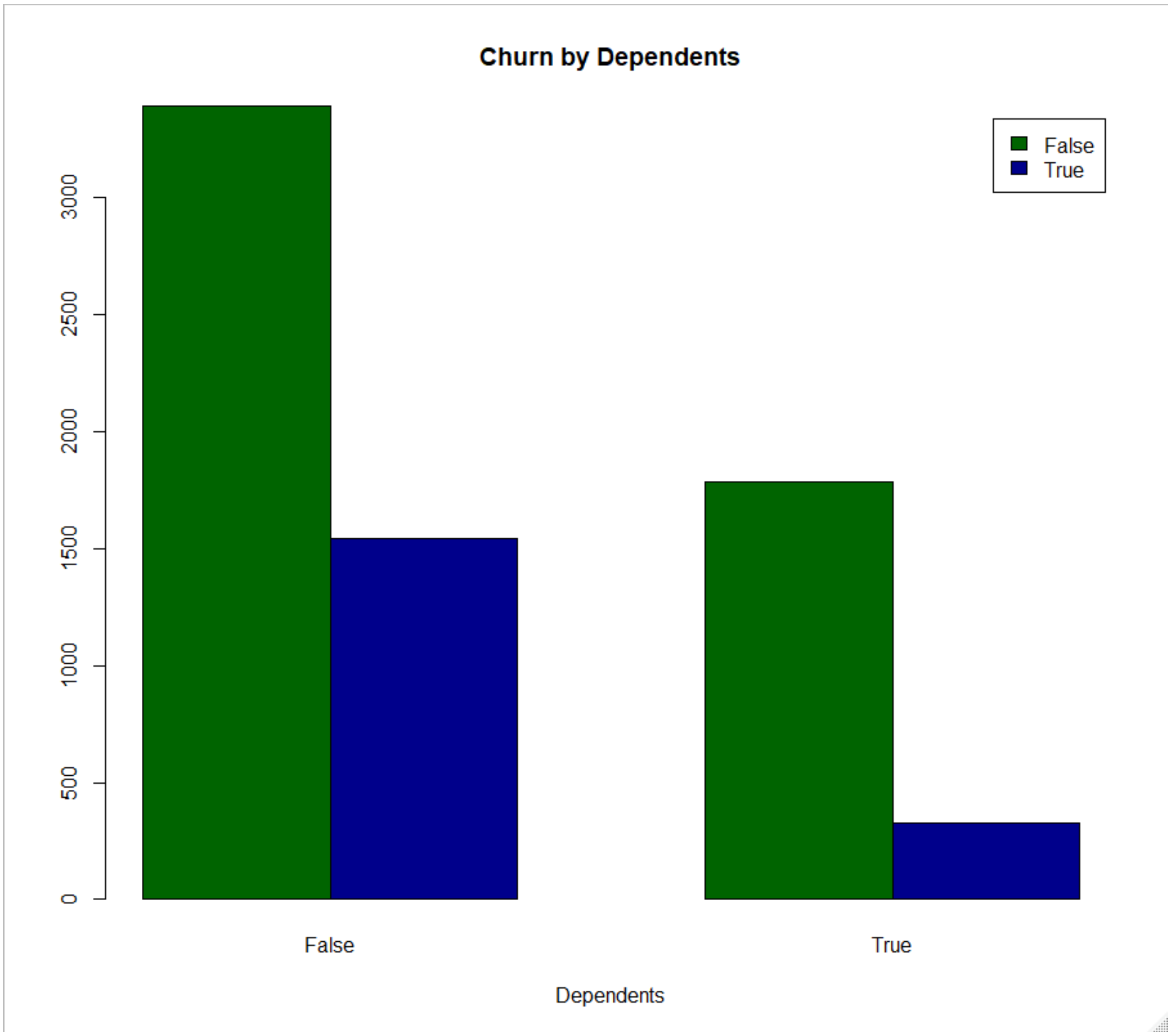


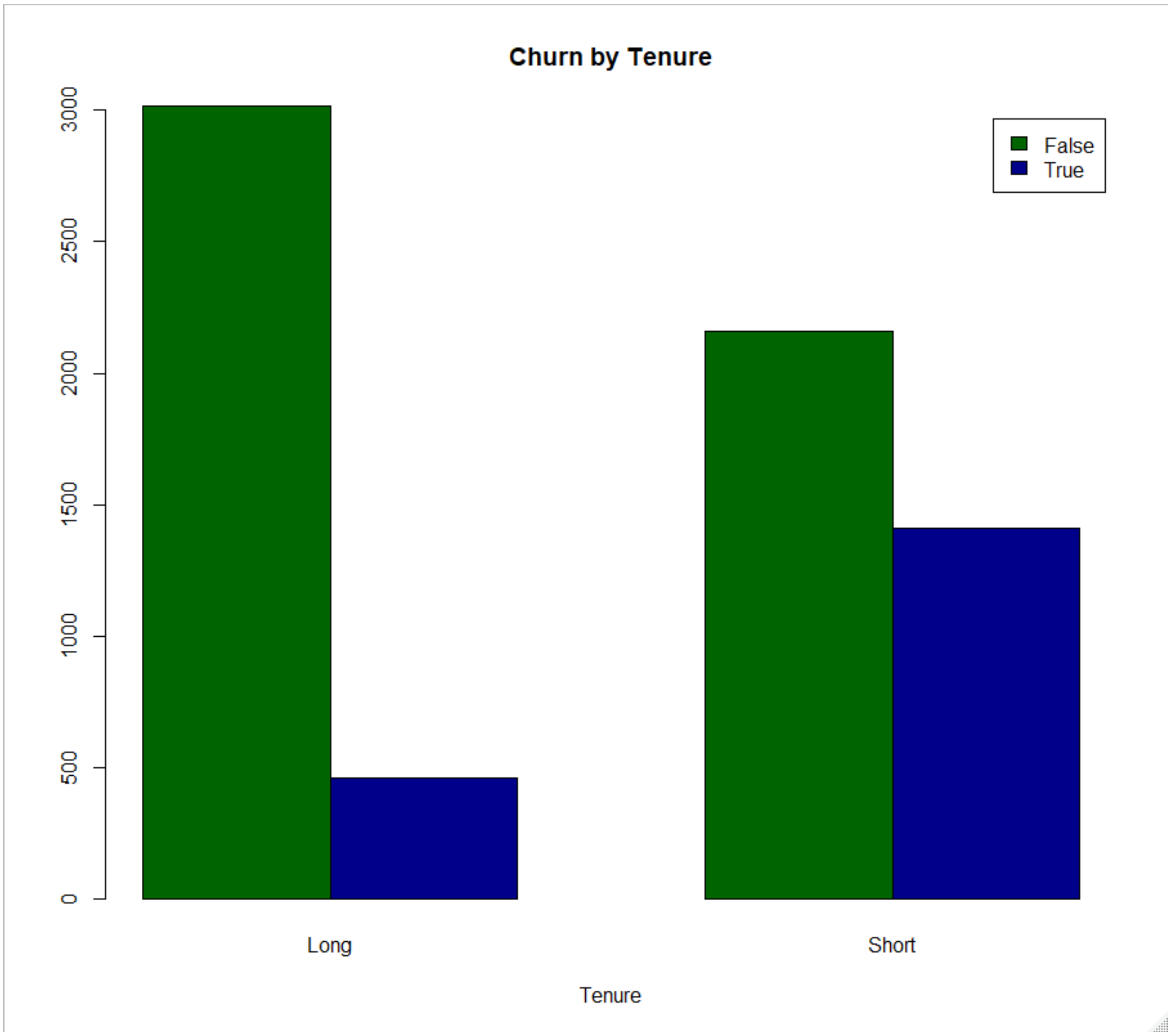
```
# Bivariate Statistics SECTION J
#-------------------------------------------------------------------
counts <- table(clean$Churn, clean$Gender)
barplot(counts, main="Churn by Gender",
        xlab="Gender", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$SeniorCitizen)
barplot(counts, main="Churn by SeniorCitizen",
        xlab="SeniorCitizen", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$Partner)
barplot(counts, main="Churn by Partner",
        xlab="Partner", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$Dependents)
barplot(counts, main="Churn by Dependents",
        xlab="Dependents", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$Tenure)
barplot(counts, main="Churn by Tenure",
        xlab="Tenure", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$PhoneService)
barplot(counts, main="Churn by PhoneService",
        xlab="PhoneService", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$MultipleLines)
barplot(counts, main="Churn by MultipleLines",
        xlab="MultipleLines", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$InternetService)
barplot(counts, main="Churn by InternetService",
        xlab="InternetService", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$OnlineSecurity)
barplot(counts, main="Churn by OnlineSecurity",
        xlab="OnlineSecurity", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$OnlineBackup)
barplot(counts, main="Churn by OnlineBackup",
        xlab="OnlineBackup", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$DeviceProtection)
barplot(counts, main="Churn by DeviceProtection",
        xlab="DeviceProtection", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$TechSupport)
barplot(counts, main="Churn by TechSupport",
        xlab="TechSupport", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$StreamingTV)
barplot(counts, main="Churn by StreamingTV",
        xlab="StreamingTV", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$StreamingMovies)
barplot(counts, main="Churn by StreamingMovies",
        xlab="StreamingMovies", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$Contract)
barplot(counts, main="Churn by Contract",
        xlab="Contract", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$PaperlessBilling)
barplot(counts, main="Churn by PaperlessBilling",
        xlab="PaperlessBilling", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$PaymentMethod)
barplot(counts, main="Churn by PaymentMethod",
        xlab="PaymentMethod", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
counts <- table(clean$Churn, clean$MonthlyCharges)
barplot(counts, main="Churn by MonthlyCharges",
        xlab="MonthlyCharges", col=c("darkgreen","darkblue"),
        legend = rownames(counts), beside=TRUE)
```

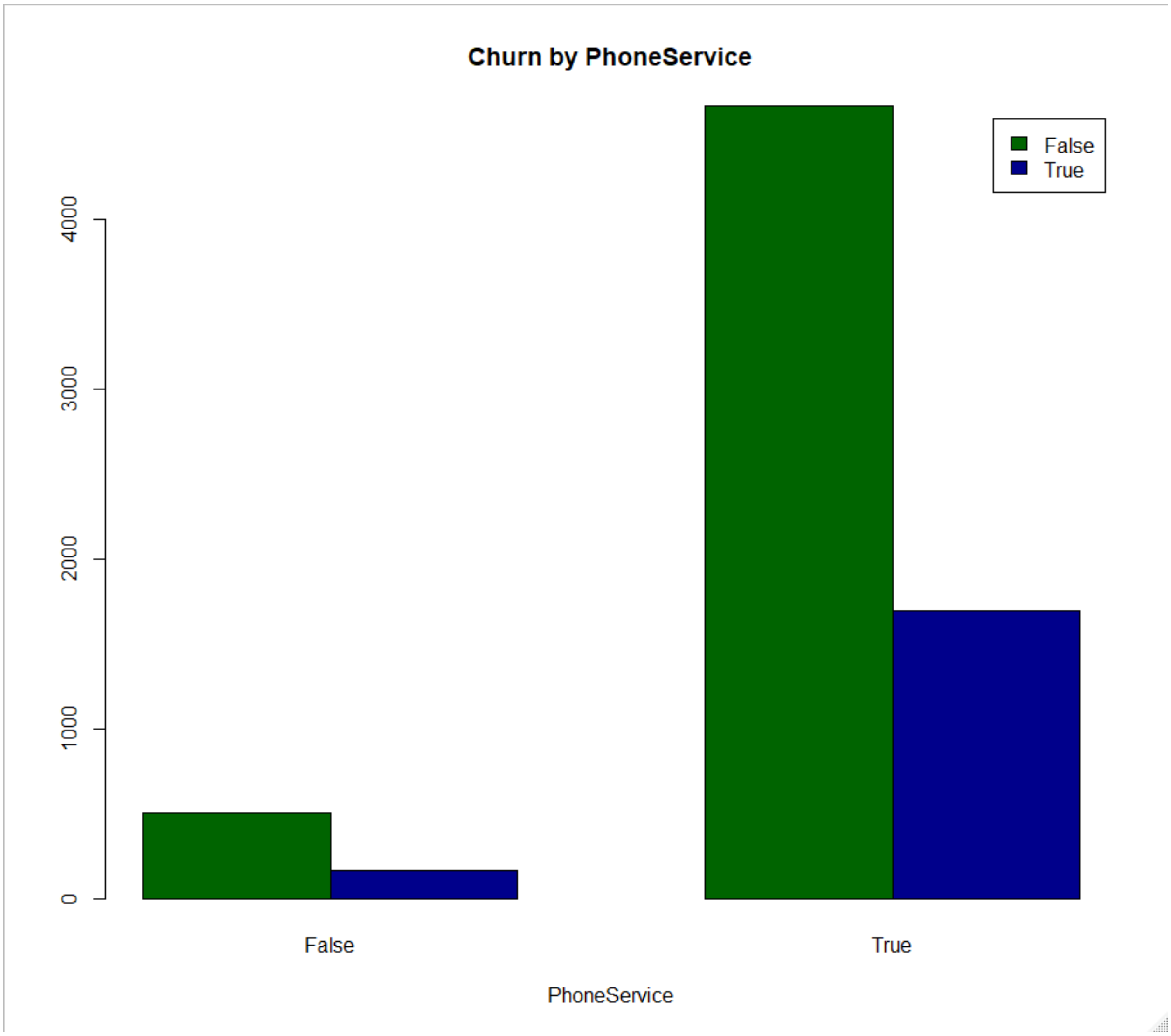

Churn by Gender
False
True
2500
2000
1500
1000
500
0
Female
Male
Gender

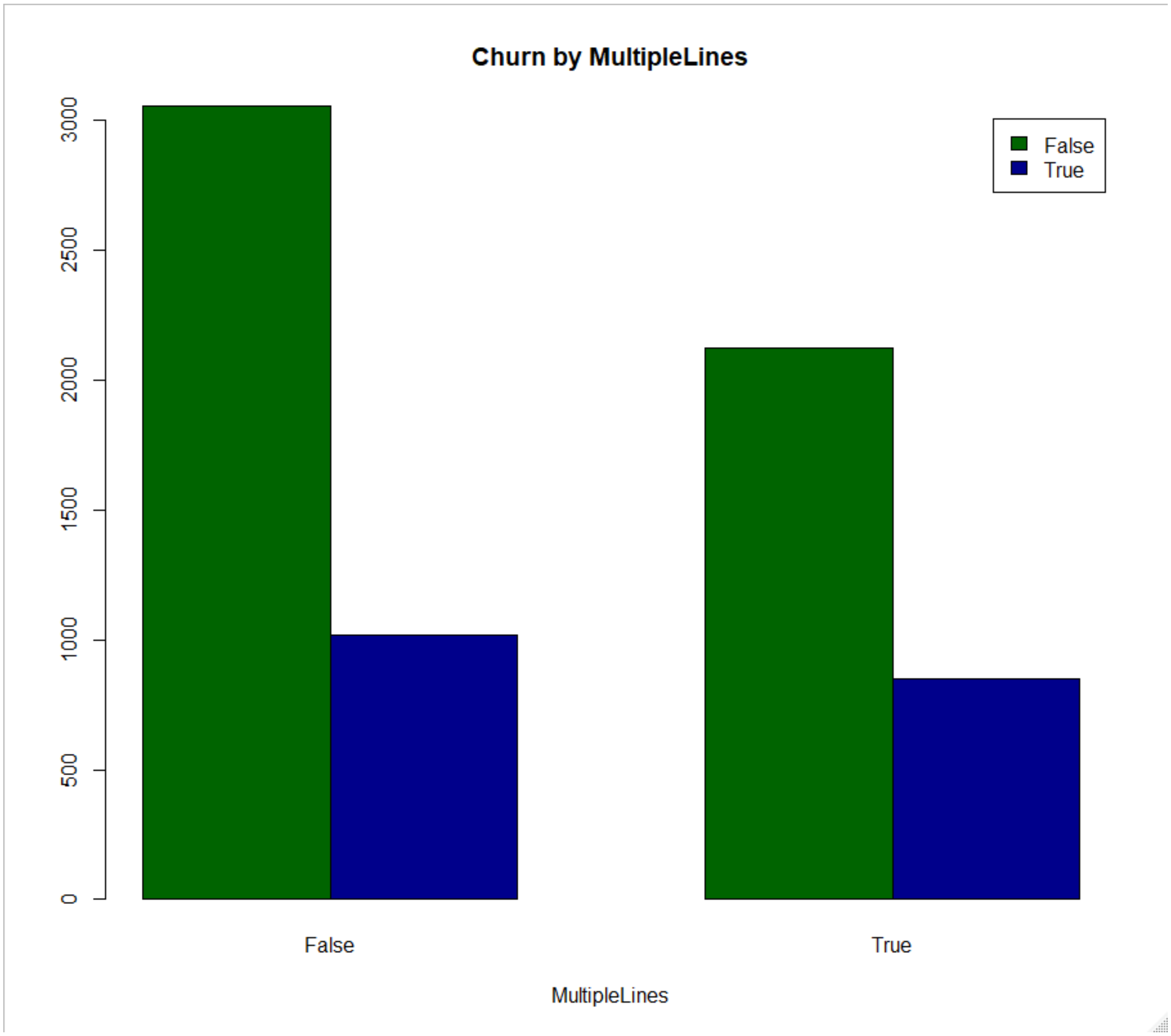

Churn by SeniorCitizen
🟩 False
🟦 True
0
1000
2000
3000
4000
False
True
SeniorCitizen

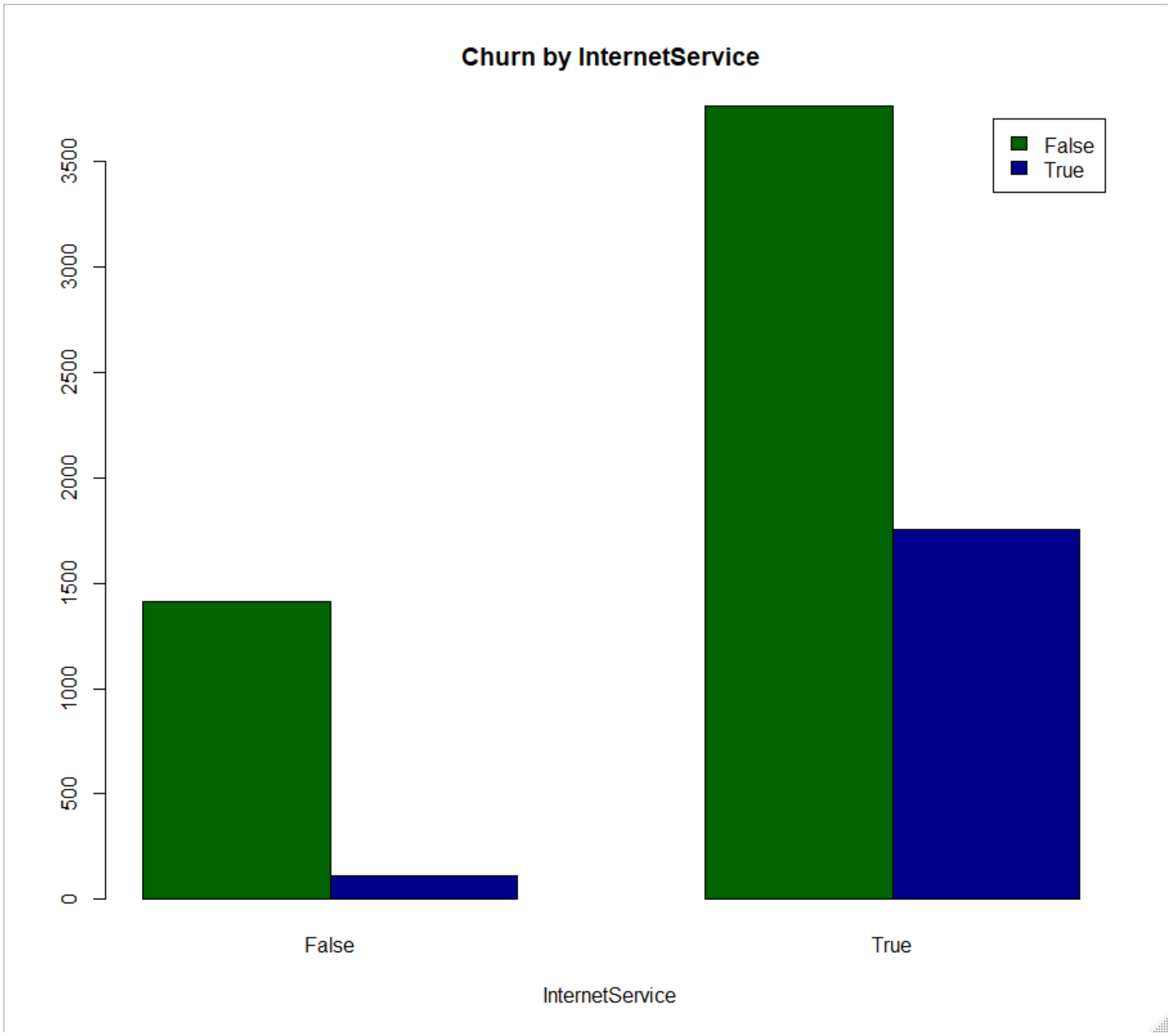

Churn by Partner
2500
2000
1500
1000
500
0
False
True
Partner
🟩 False
🟦 True

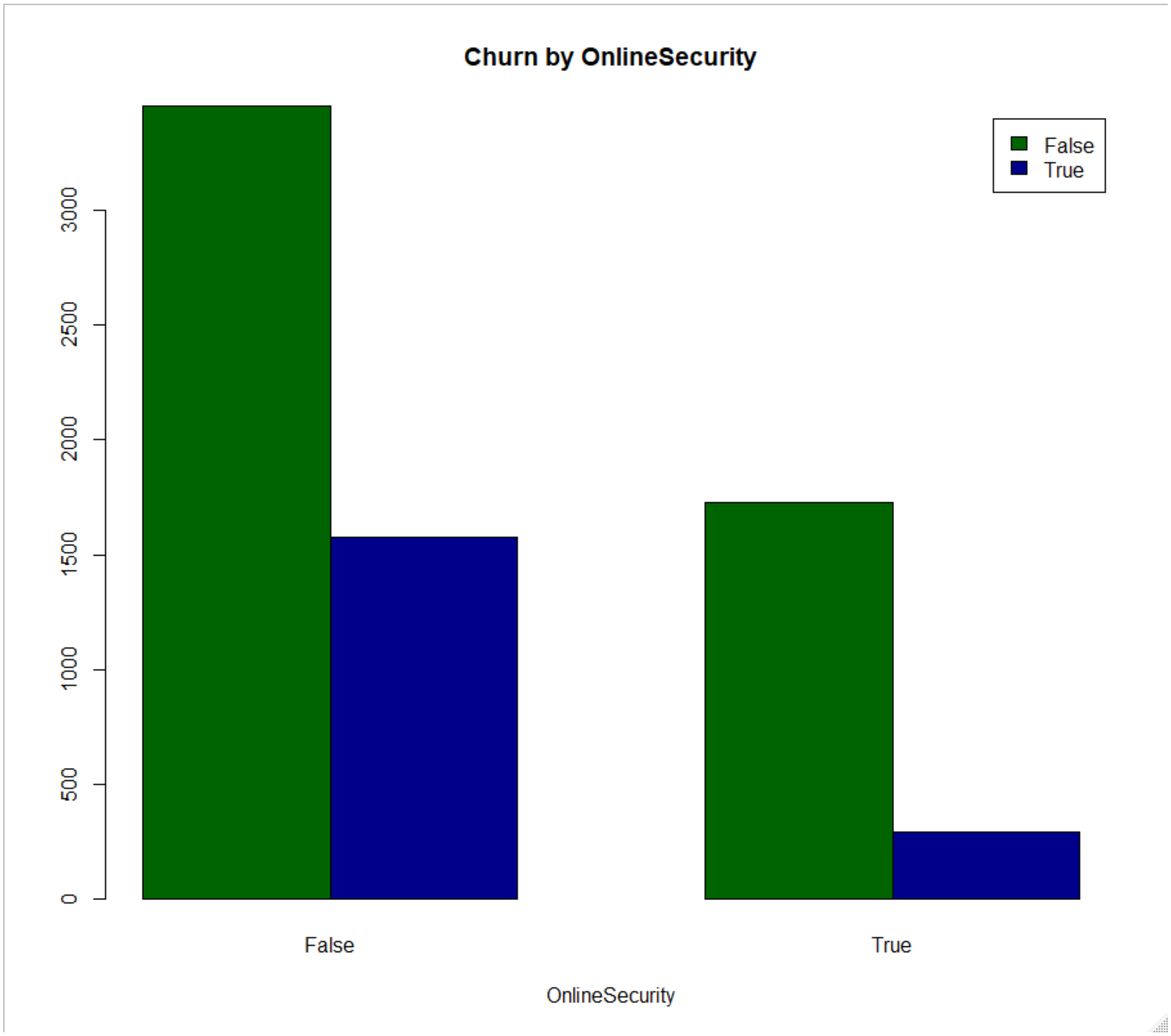

Churn by Dependents
🟩 False
🟦 True
0
500
1000
1500
2000
2500
3000
False
True
Dependents

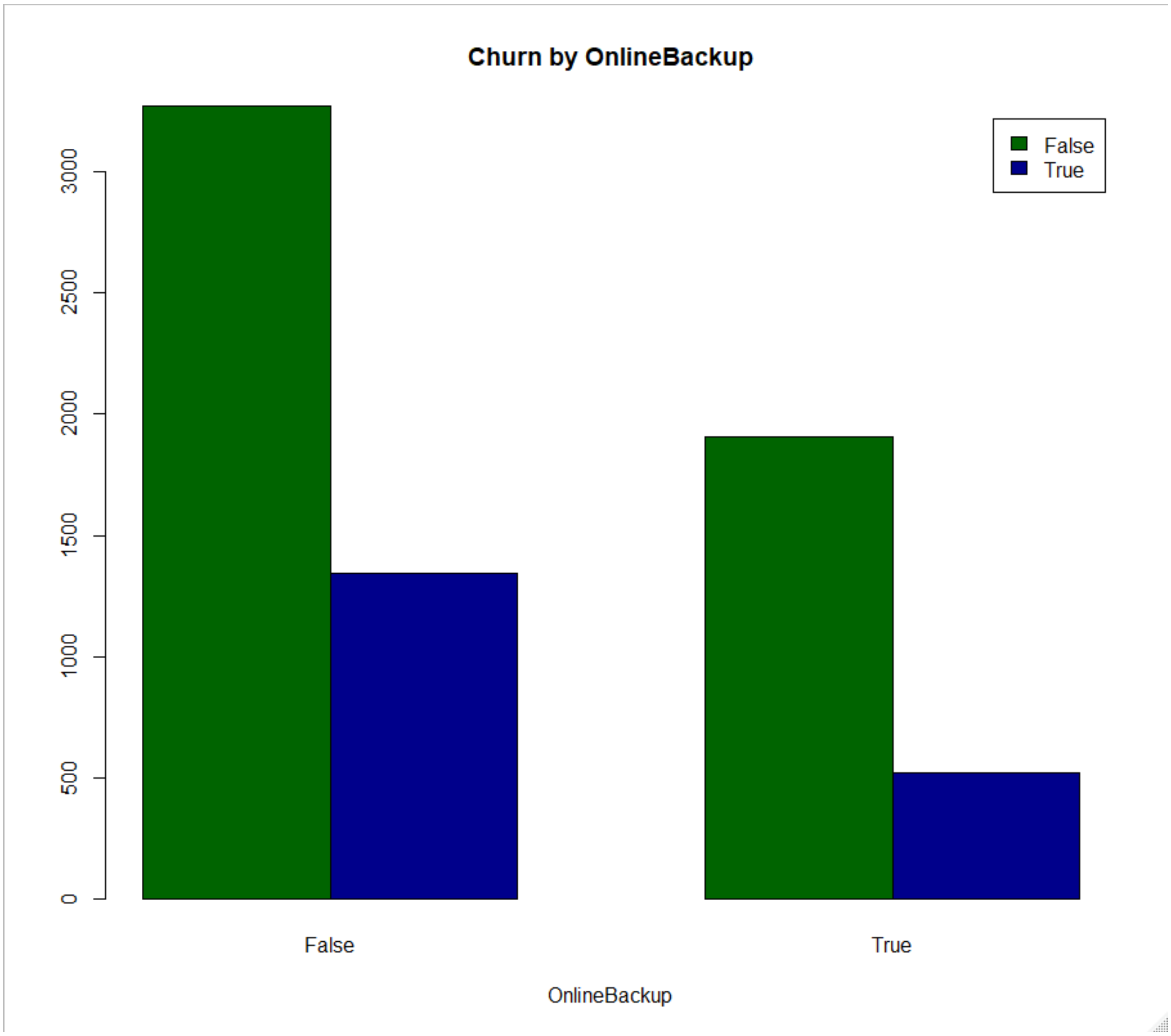

Churn by Tenure
3000
2500
2000
1500
1000
500
0
🟩 False
🟦 True
Long
Short
Tenure

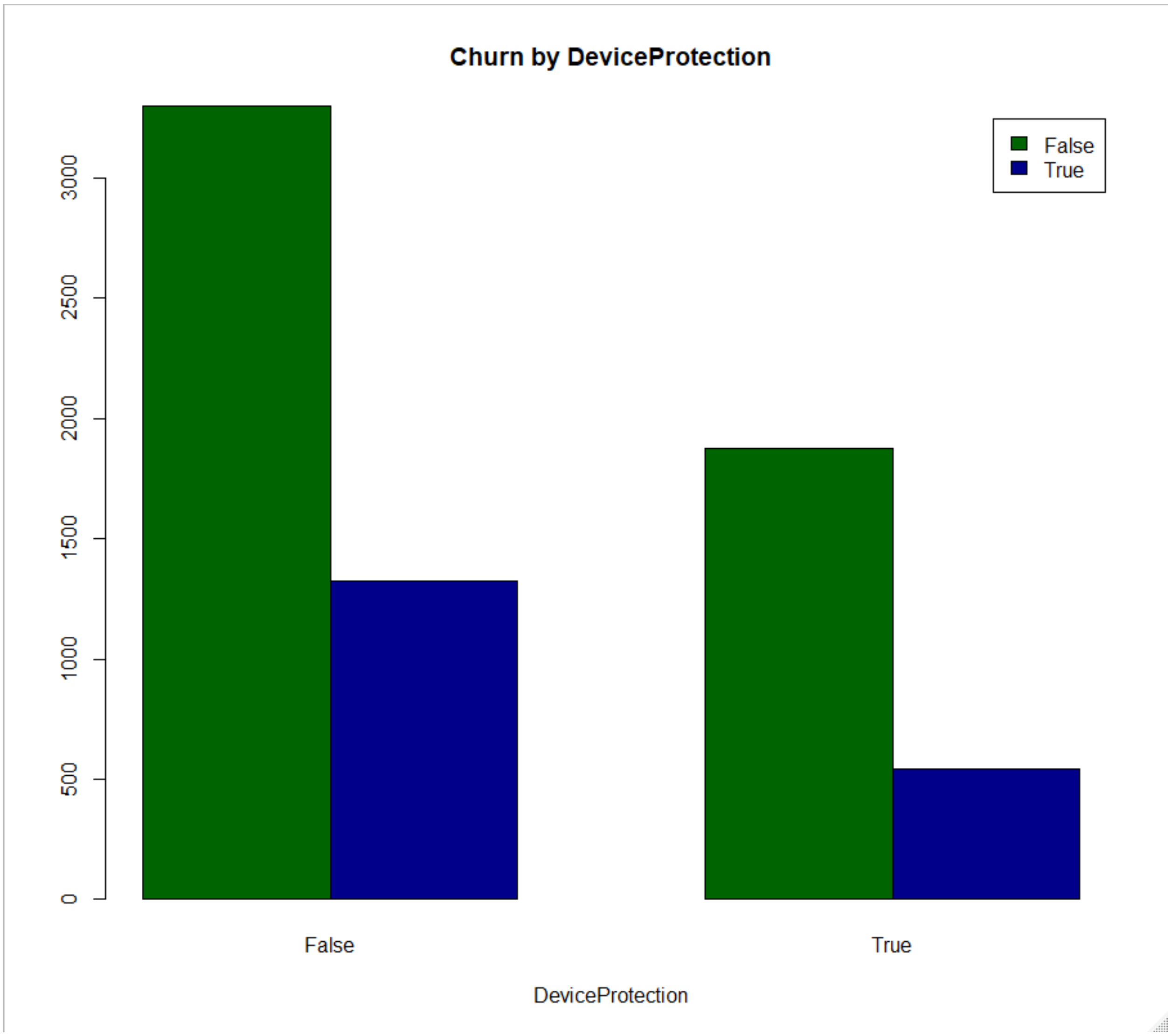

Churn by PhoneService
False
True
0
1000
2000
3000
4000
False
True
PhoneService

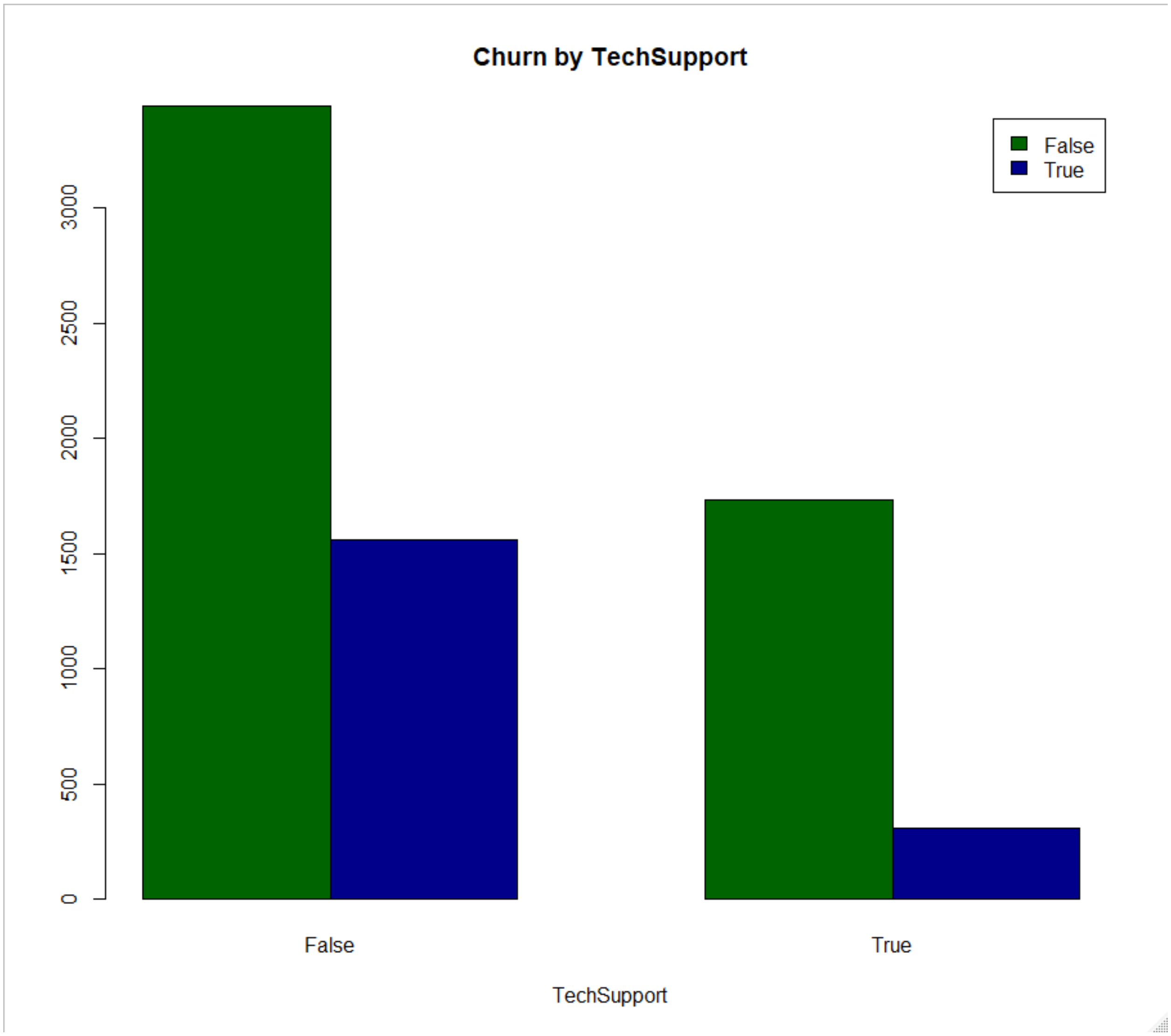

Churn by MultipleLines
3000
2500
2000
1500
1000
500
0
🟩 False
🟦 True
False
True
MultipleLines

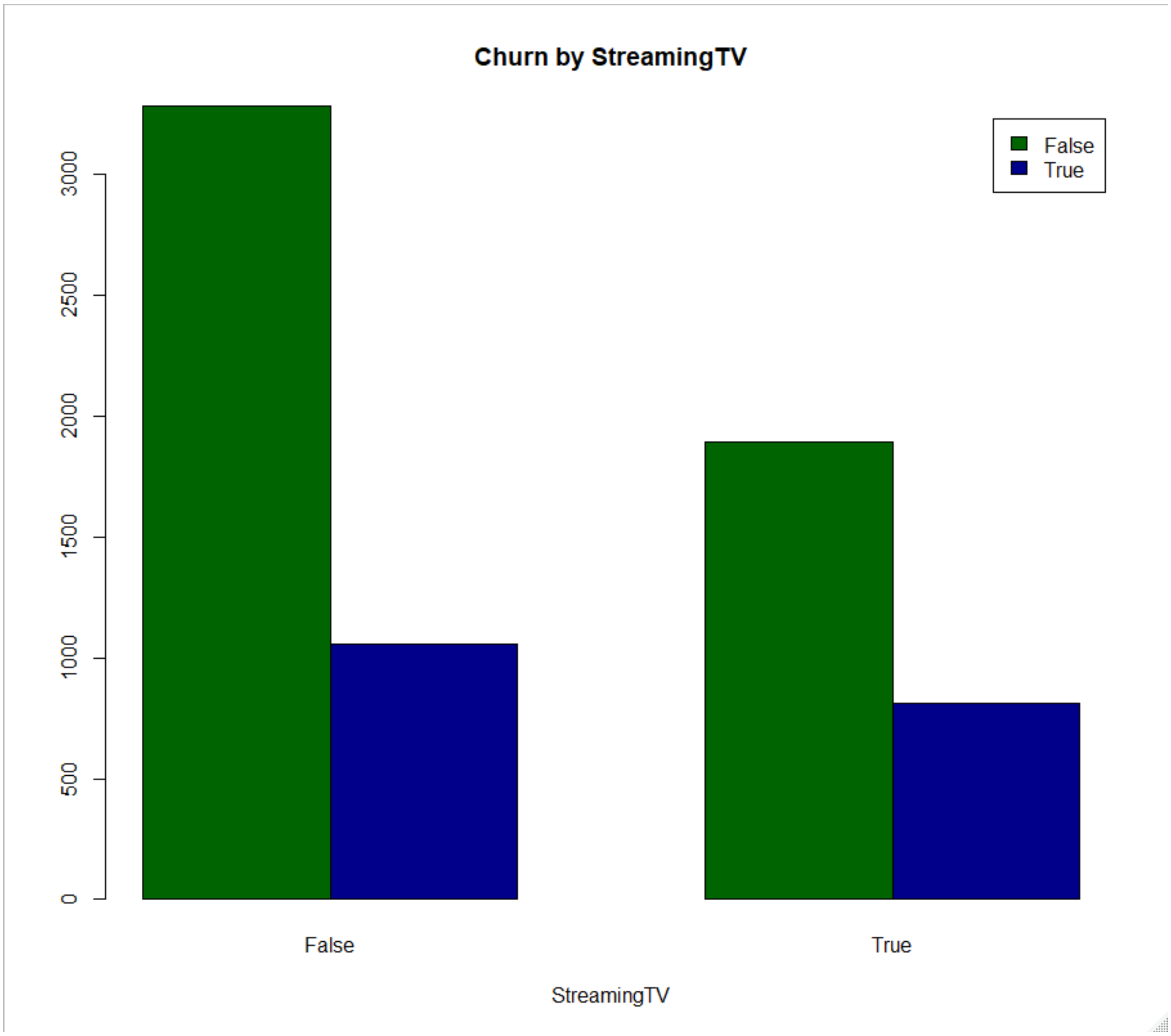

Churn by InternetService
3500
3000
2500
2000
1500
1000
500
0
🟩 False
🟦 True
False
True
InternetService

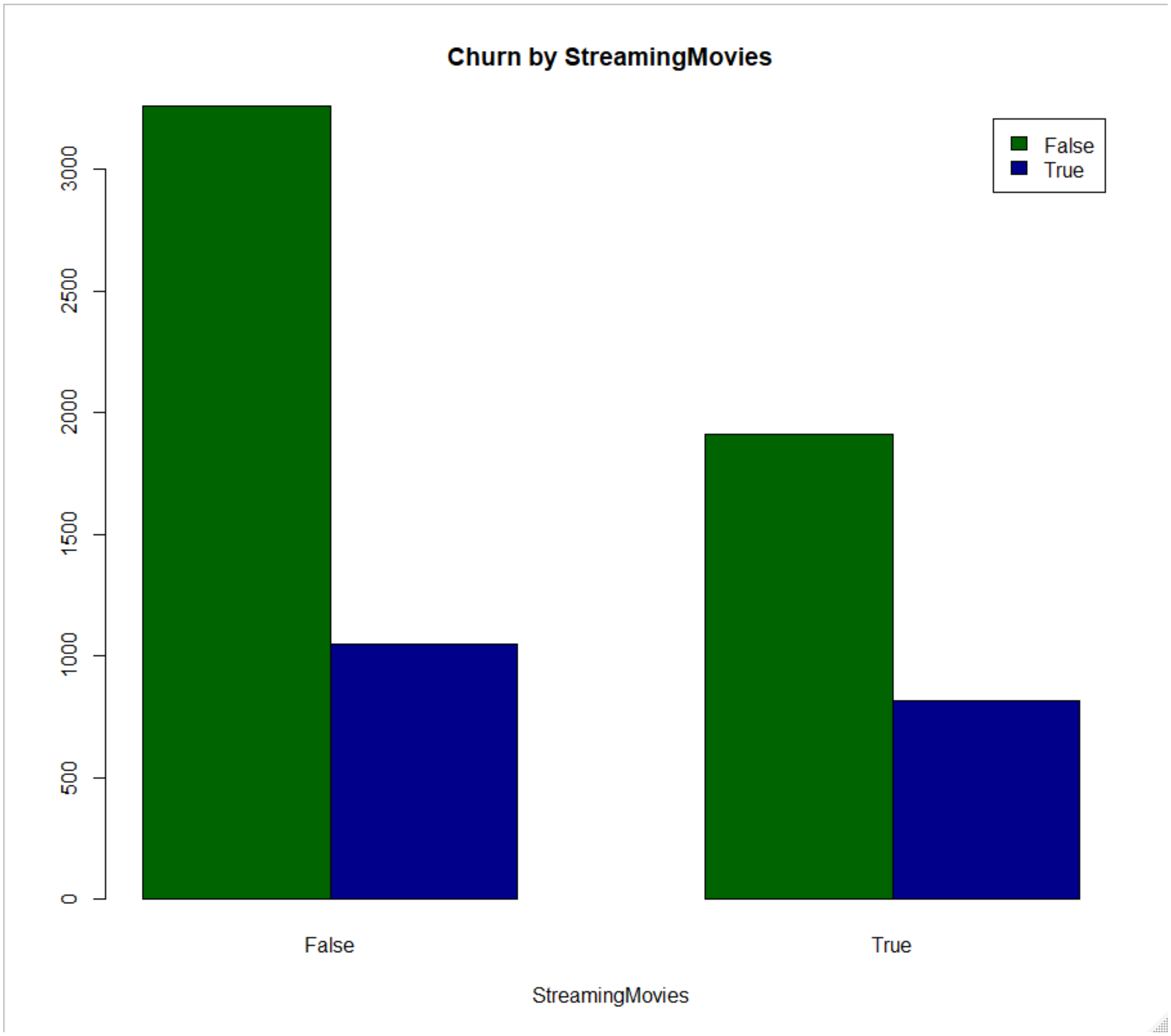

Churn by OnlineSecurity
🟩 False
🟦 True
0
500
1000
1500
2000
2500
3000
False
True
OnlineSecurity

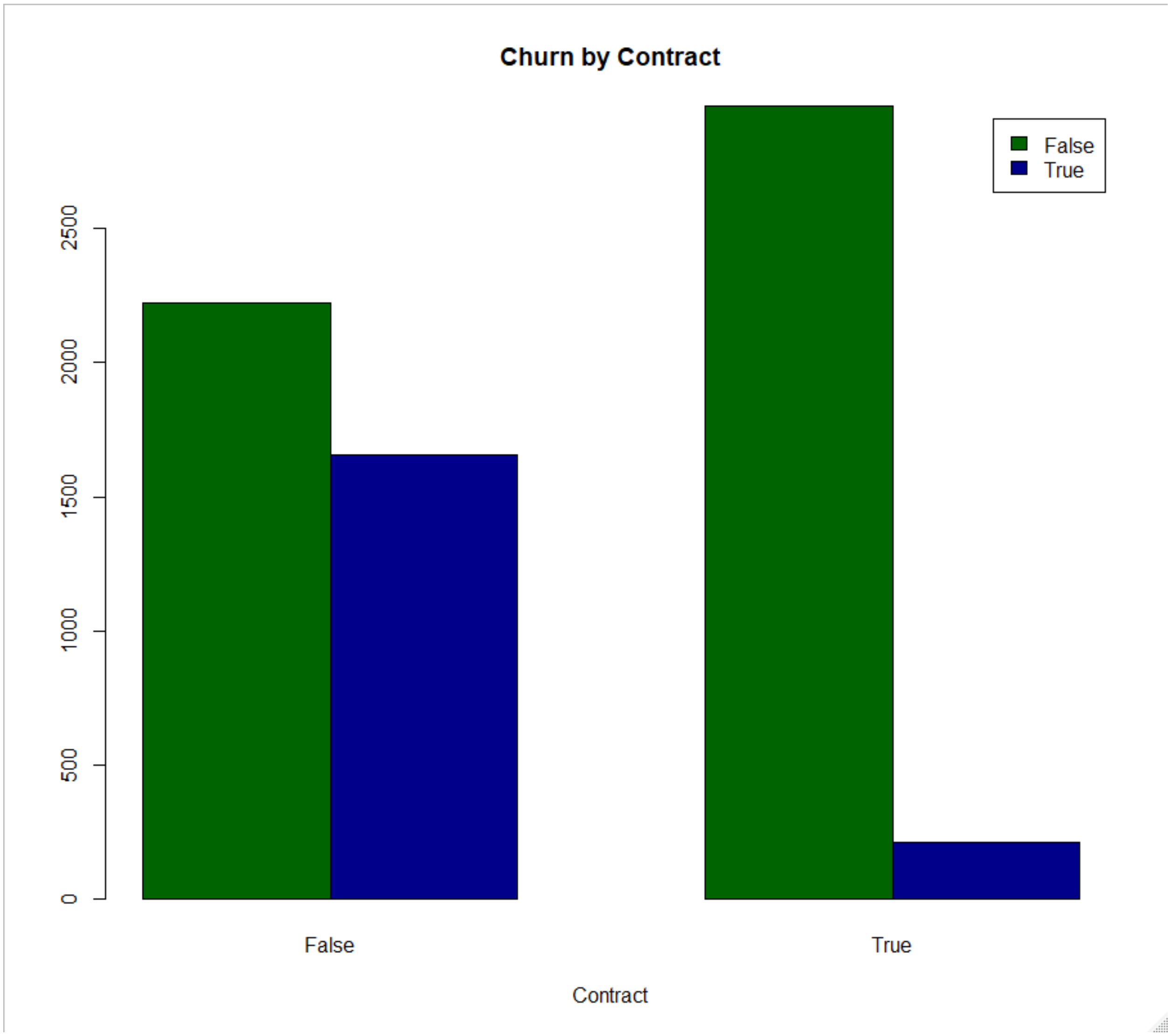

Churn by OnlineBackup
🟩 False
🟦 True
0
500
1000
1500
2000
2500
3000
False
True
OnlineBackup

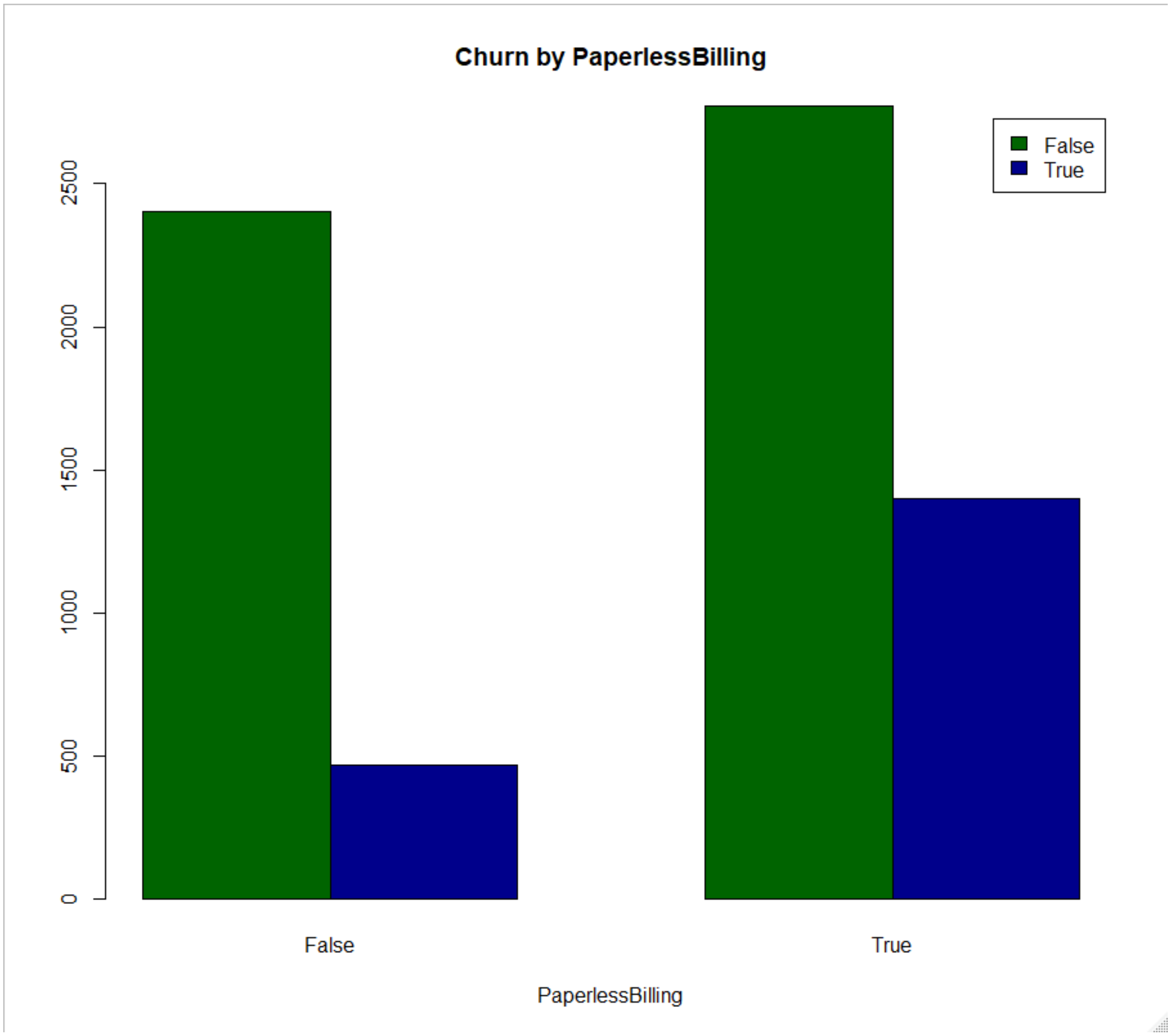

Churn by DeviceProtection
🟩 False
🟦 True
0
500
1000
1500
2000
2500
3000
False
True
DeviceProtection

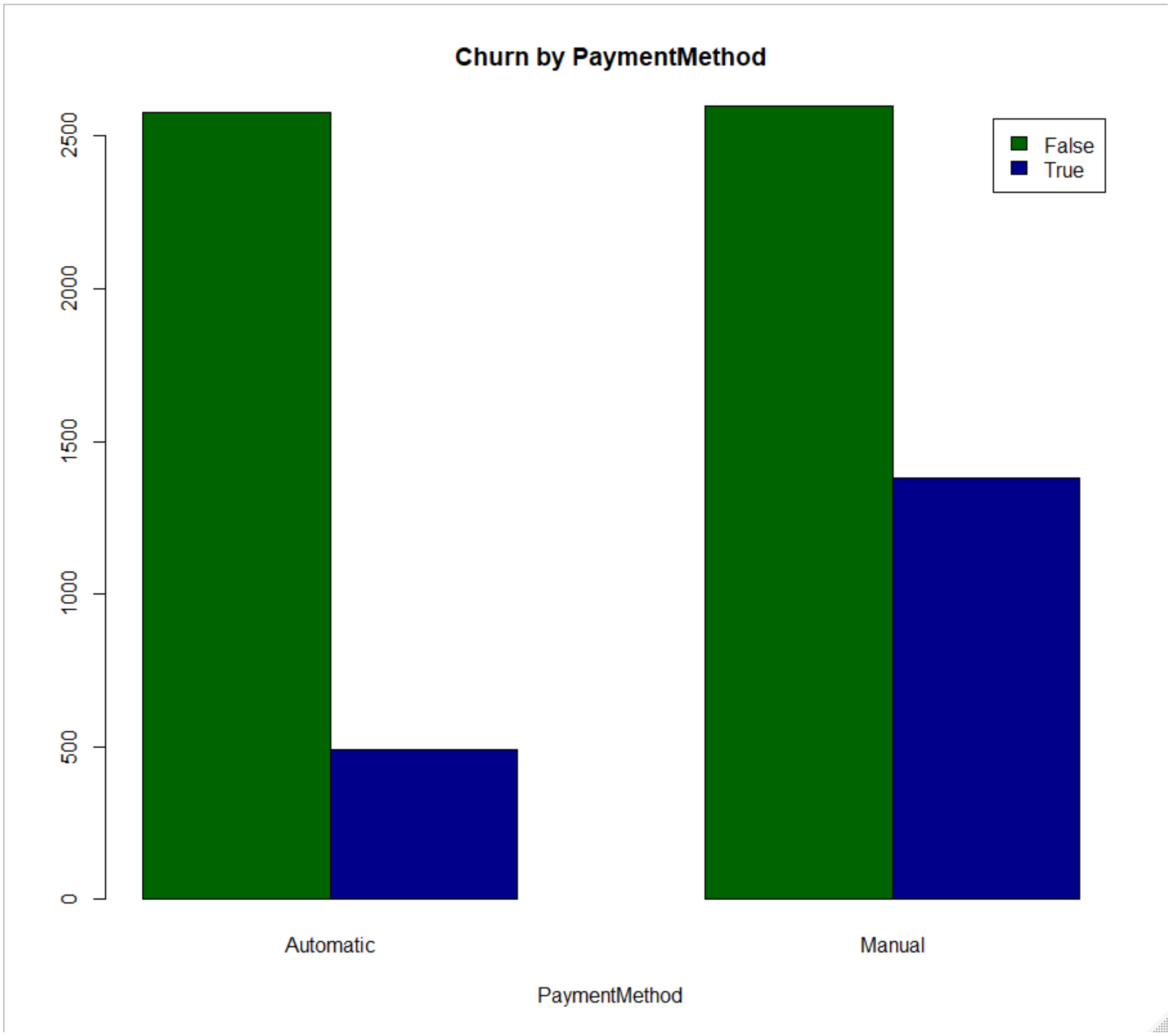

Churn by TechSupport
🟩 False
🟦 True
0
500
1000
1500
2000
2500
3000
False
True
TechSupport

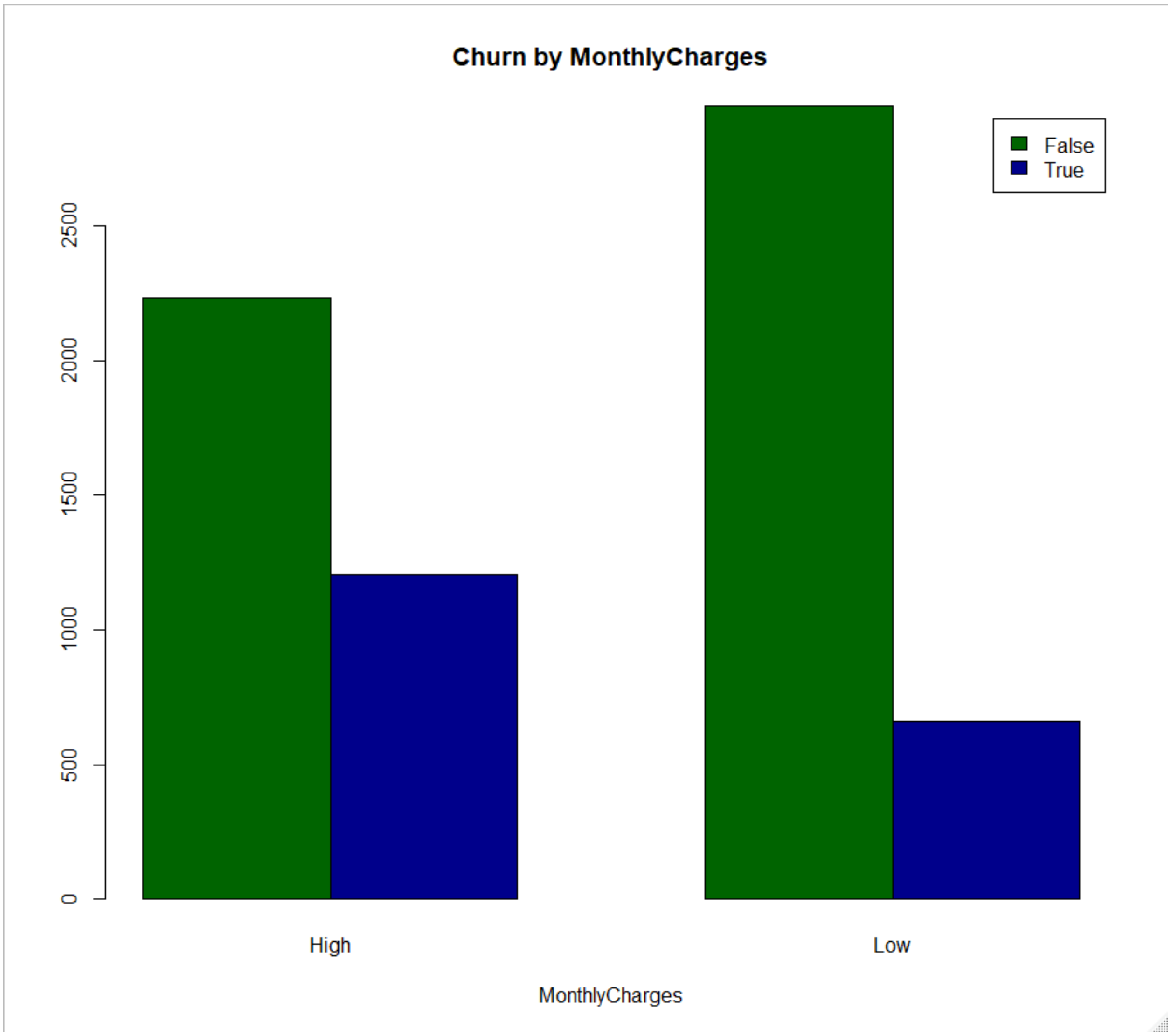

Churn by StreamingTV
🟩 False
🟦 True
0
500
1000
1500
2000
2500
3000
False
True
StreamingTV


Churn by StreamingMovies
0
500
1000
1500
2000
2500
3000
🟩 False
🟦 True
False
True
StreamingMovies


Churn by Contract
🟩 False
🟦 True
0
500
1000
1500
2000
2500
False
True
Contract

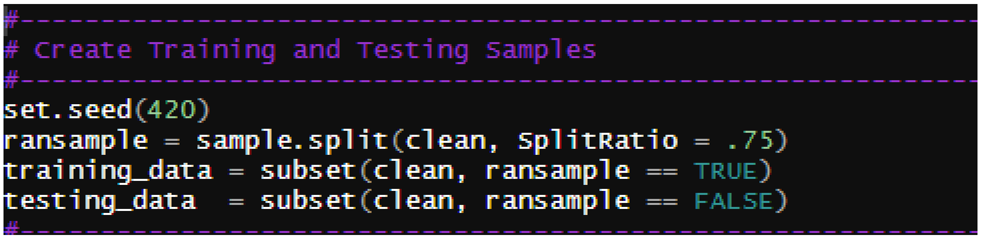

Churn by PaperlessBilling
False
True
0
500
1000
1500
2000
2500
False
True
PaperlessBilling


Churn by PaymentMethod
False
True
0
500
1000
1500
2000
2500
Automatic
Manual
PaymentMethod



Part K:

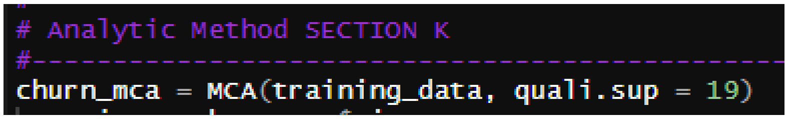
First, the data was split into two samples. The first sample is the training data while the second sample is the testing data. This allows for the logistic regression model created later to be made using the training data and then be applied to the testing data to help determine the efficacy of the model.



```
#-------------------------------------------------------------
# Create Training and Testing Samples
#-------------------------------------------------------------
set.seed(420)
ransample = sample.split(clean, SplitRatio = .75)
training_data = subset(clean, ransample == TRUE)
testing_data  = subset(clean, ransample == FALSE)
#-------------------------------------------------------------
```

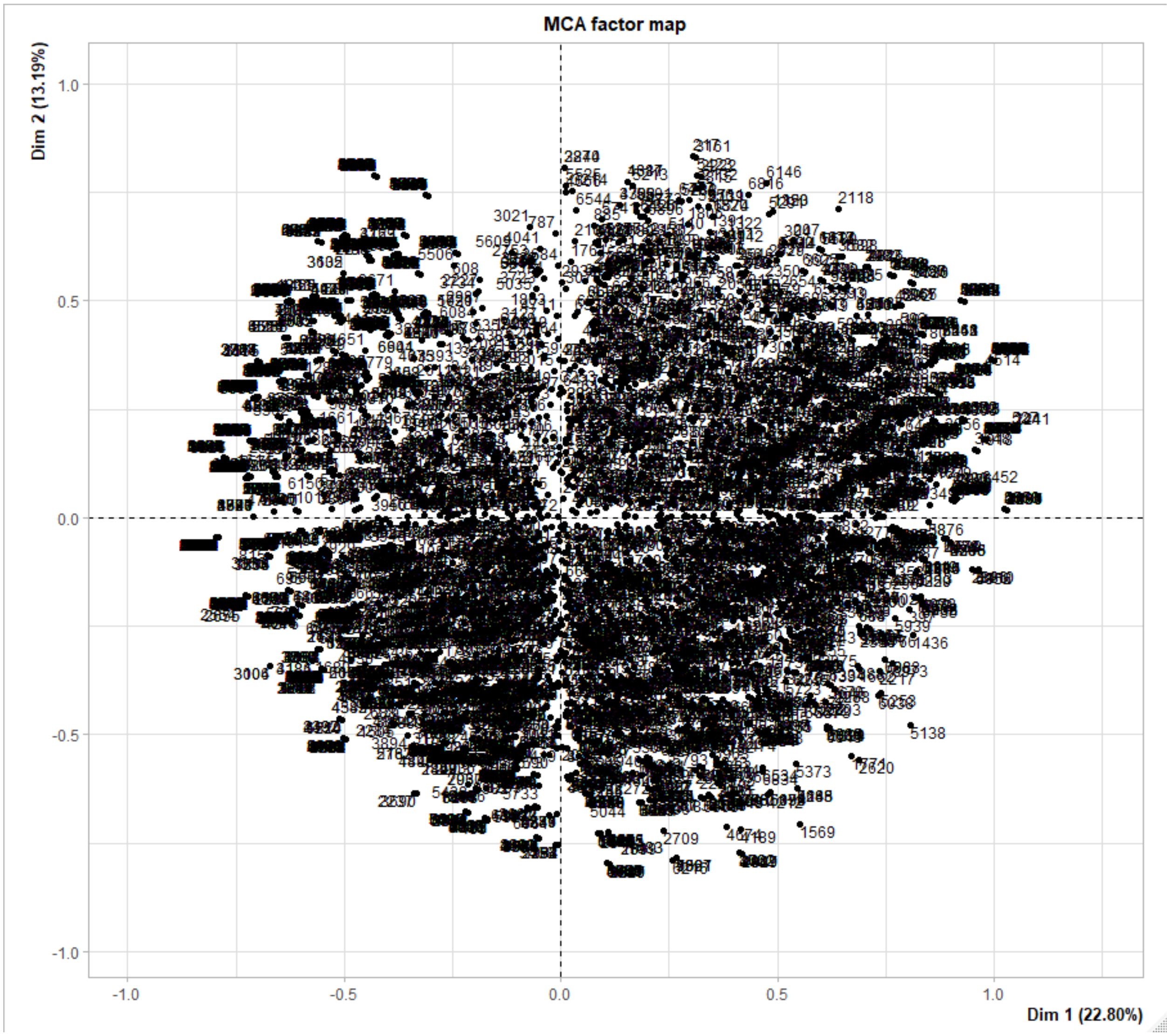
Analytic Method:

The analytic method used in this analysis is Multiple Correspondence Analysis (MCA). This method was done by using the R package FactoMineR. This first line of code both applies the MCA method to our data and generates some useful plots.

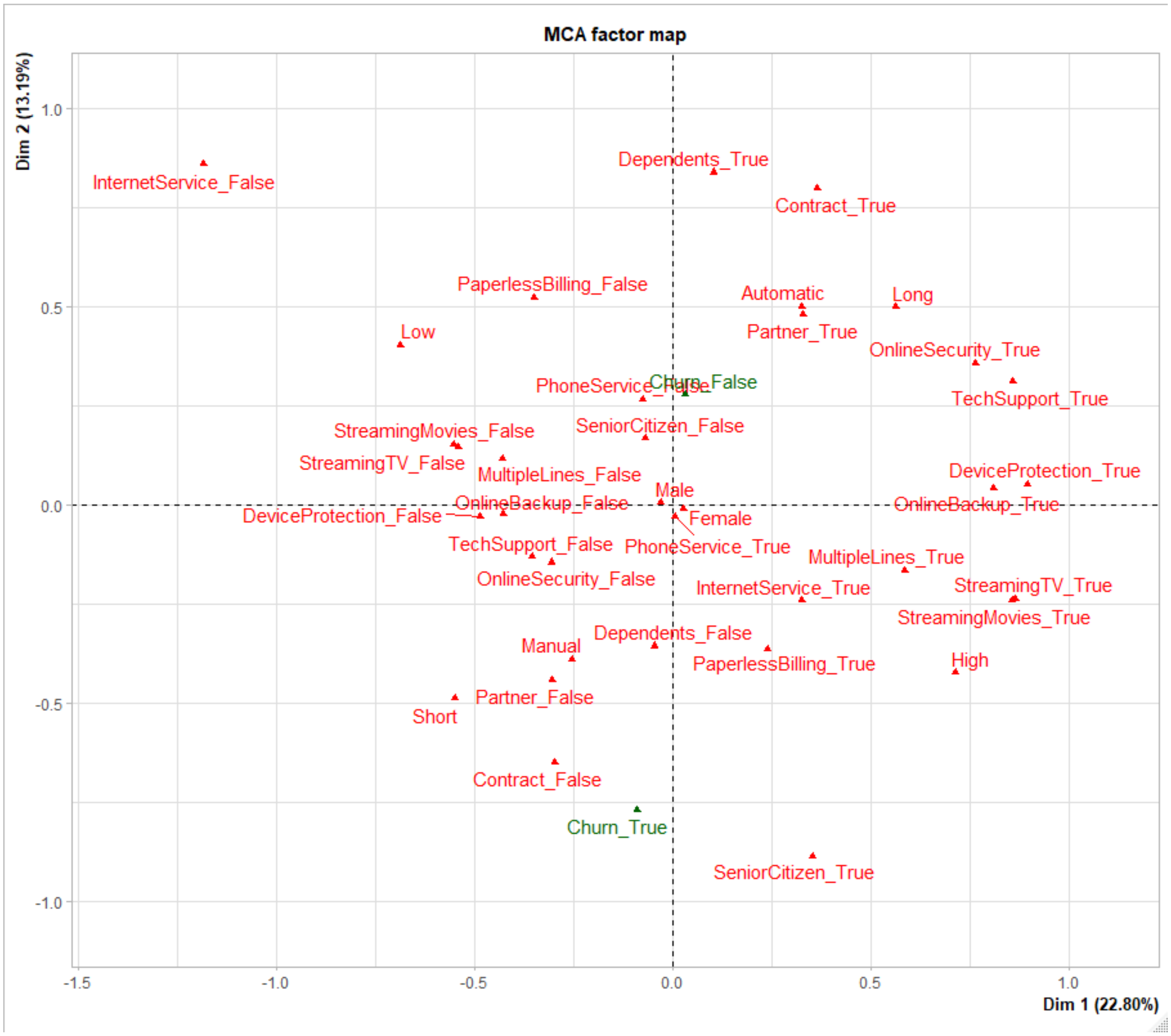


```
# Analytic Method SECTION K
#--------------------------------------------------
churn_mca = MCA(training_data, quali.sup = 19)
```

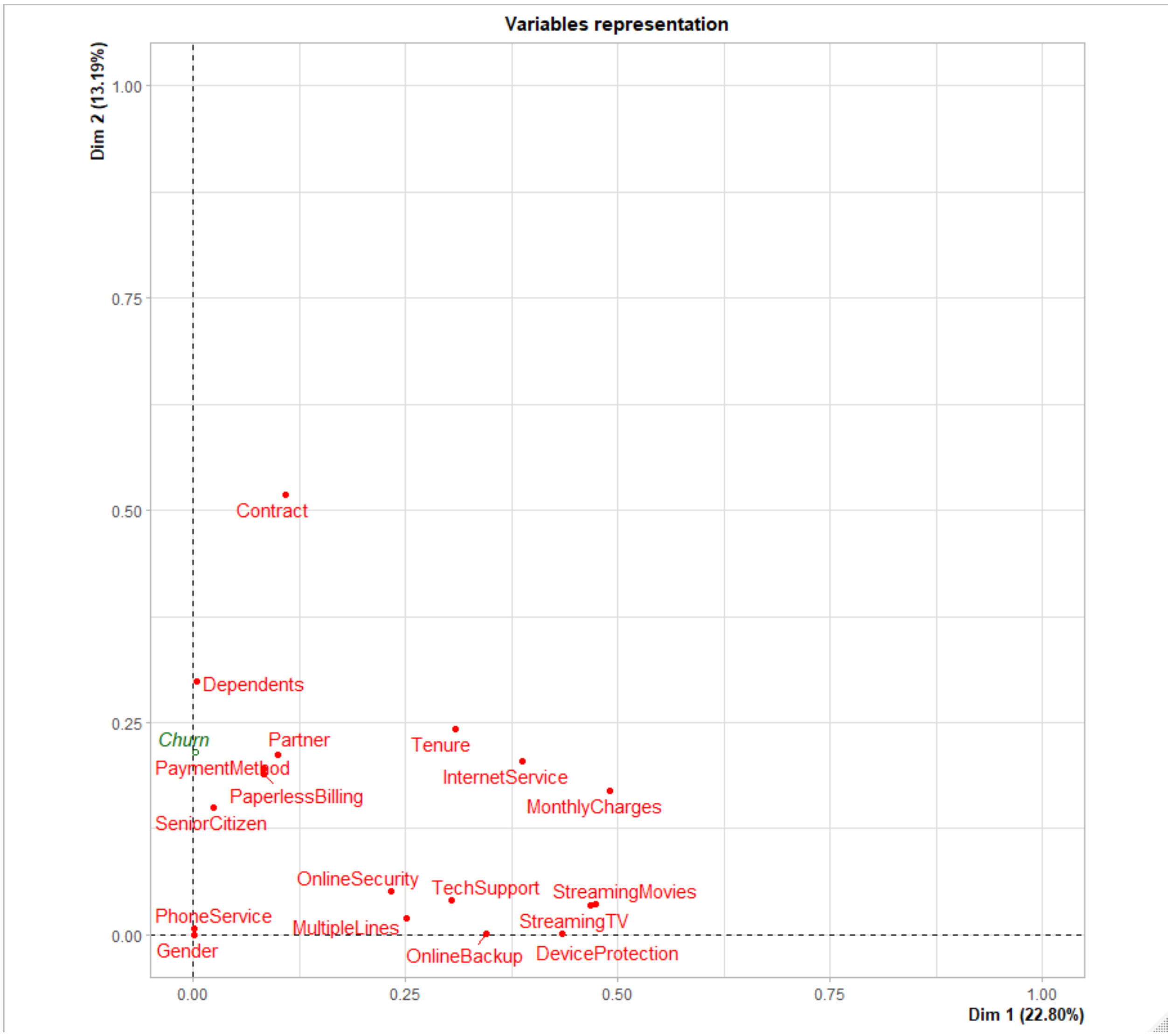
Scatterplot of Individuals. As this projects all individuals on the graph, it is hard to make out much information from this graph. There are over 7000 observations graphed, which makes this messy. Below are graphs which better depict how the MCA can help analyse the data.



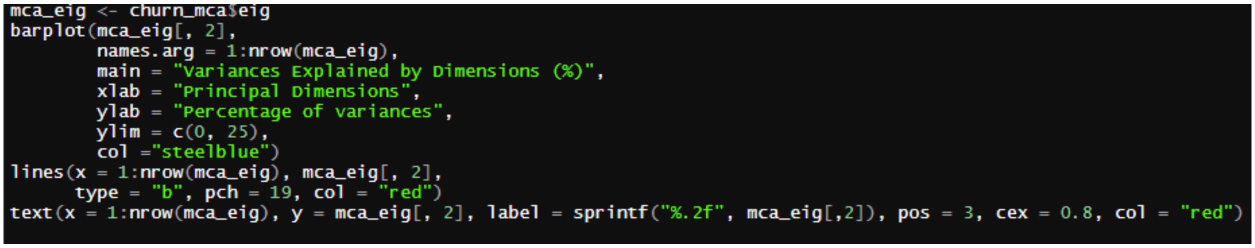
MCA Factor Map for Variables. The variables are projected on the graph and shows the relationship between variables and dimensions. The next graph better demonstrates how variables effect dimensions.



Variable contributions to each dimension. Dimension 1 is along the X axis while dimension 2 is along the Y axis. The location of a variable on the graph indicates the strength of its effect on each dimension. As the variables move right, they have a greater effect on dimension 1. As the variables move up, they have a greater influence on dimension 2. If a variable moves at a 45 degree angle, it suggests that it is having a greater effect on both dimensions equally.

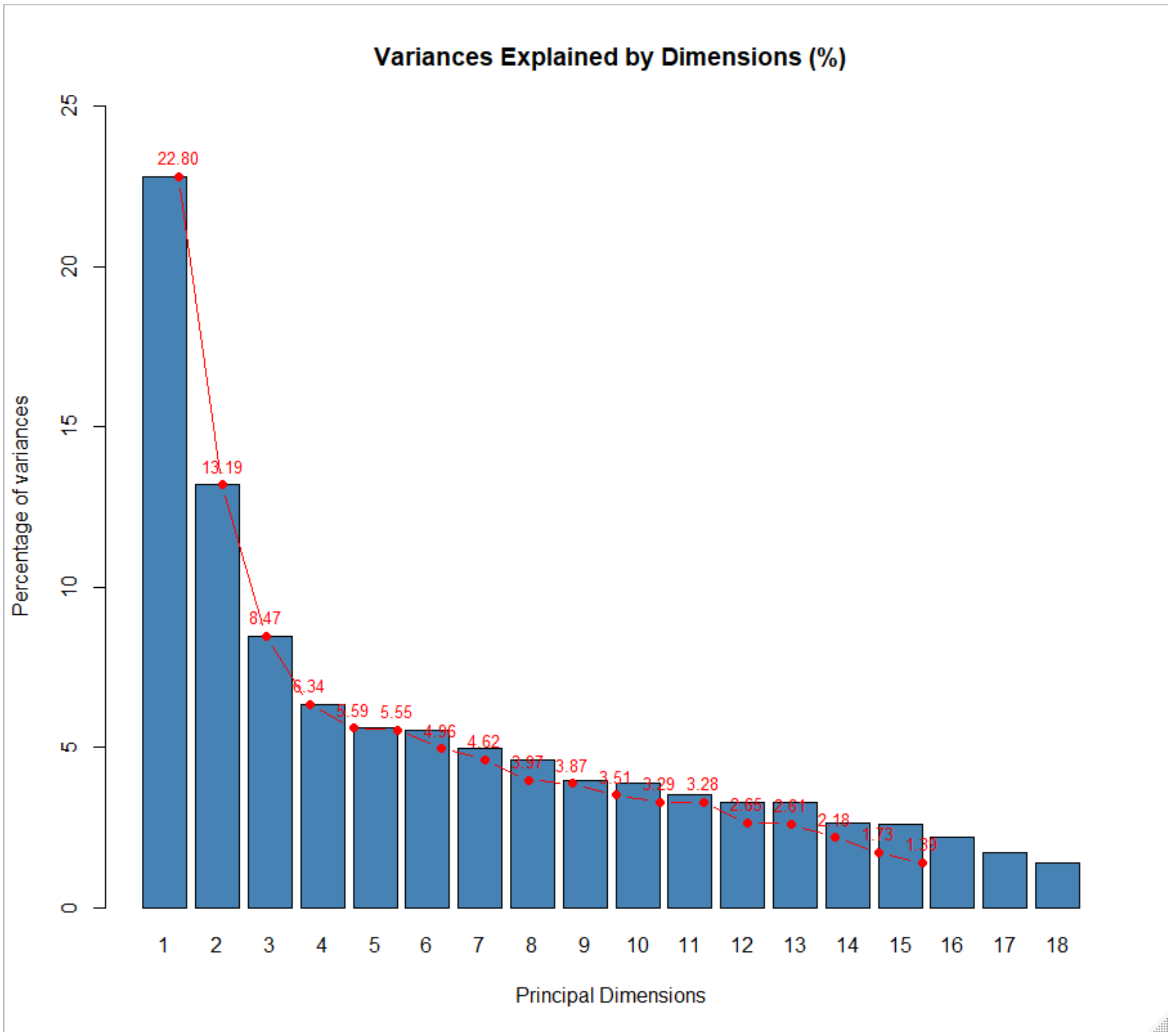


Here, we extract the eigen values for the dimensions created by the MCA analysis.



```
mca_eig <- churn_mca$eig
barplot(mca_eig[, 2],
        names.arg = 1:nrow(mca_eig),
        main = "Variances Explained by Dimensions (%)",
        xlab = "Principal Dimensions",
        ylab = "Percentage of variances",
        ylim = c(0, 25),
        col ="steelblue")
lines(x = 1:nrow(mca_eig), mca_eig[, 2],
      type = "b", pch = 19, col = "red")
text(x = 1:nrow(mca_eig), y = mca_eig[, 2], label = sprintf("%.2f", mca_eig[,2]), pos = 3, cex = 0.8, col = "red")
```

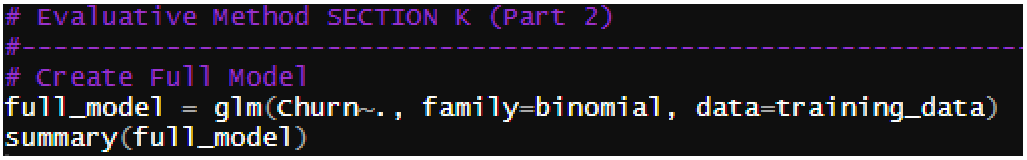
Eigen values for the dimensions created by the MCA analysis graph. The first thing we can see in this visualization of the eigen values is that the variance explained by each dimension decreases rapidly. The first two dimensions have a much greater drop between them than the next two, which again have a much greater drop than the next two. Due to these large drops, this chart suggests that two dimensions would likely be appropriate for this data, although, three dimensions could also be considered.



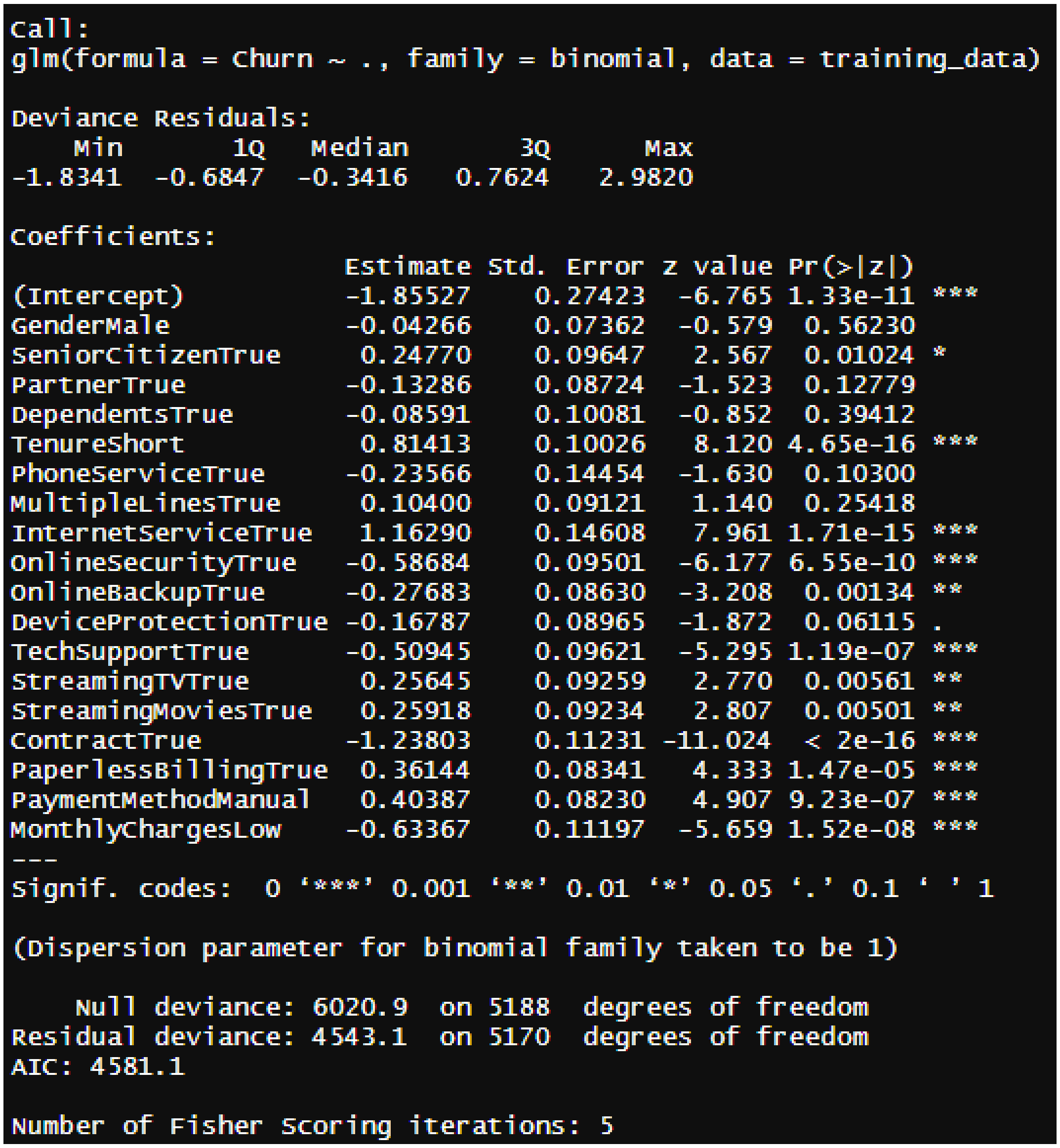
Evaluative Method:

The evaluative method used in this analysis is Logistic Regression.

First the full model is created and summarized.

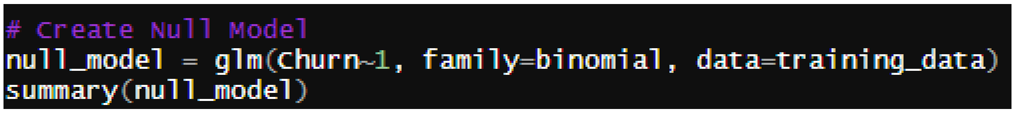


```
# Evaluative Method SECTION K (Part 2)
#-------------------------------------------------------------
# Create Full Model
full_model = glm(Churn~., family=binomial, data=training_data)
summary(full_model)
```

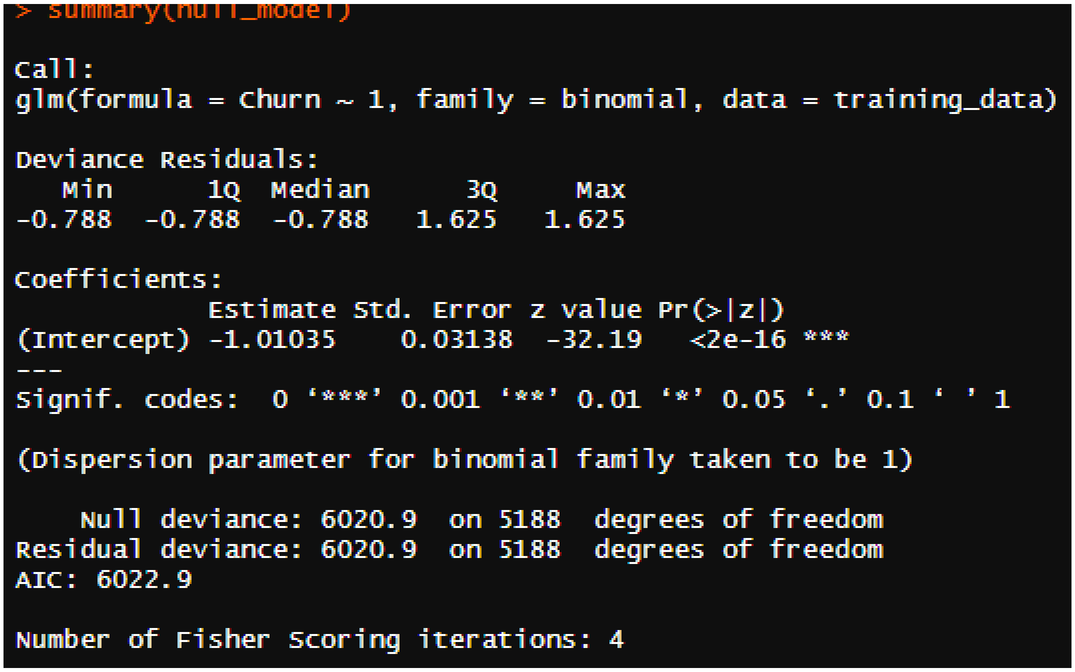


```
Call:
glm(formula = Churn ~ ., family = binomial, data = training_data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8341  -0.6847  -0.3416   0.7624   2.9820

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -1.85527    0.27423  -6.765 1.33e-11 ***
GenderMale           -0.04266    0.07362  -0.579  0.56230
SeniorCitizenTrue     0.24770    0.09647   2.567  0.01024 *
PartnerTrue          -0.13286    0.08724  -1.523  0.12779
DependentsTrue       -0.08591    0.10081  -0.852  0.39412
TenureShort           0.81413    0.10026   8.120 4.65e-16 ***
PhoneServiceTrue     -0.23566    0.14454  -1.630  0.10300
MultipleLinesTrue     0.10400    0.09121   1.140  0.25418
InternetServiceTrue   1.16290    0.14608   7.961 1.71e-15 ***
OnlineSecurityTrue   -0.58684    0.09501  -6.177 6.55e-10 ***
OnlineBackupTrue     -0.27683    0.08630  -3.208  0.00134 **
DeviceProtectionTrue -0.16787    0.08965  -1.872  0.06115 .
TechSupportTrue      -0.50945    0.09621  -5.295 1.19e-07 ***
StreamingTVTrue       0.25645    0.09259   2.770  0.00561 **
StreamingMoviesTrue   0.25918    0.09234   2.807  0.00501 **
ContractTrue         -1.23803    0.11231 -11.024  < 2e-16 ***
PaperlessBillingTrue  0.36144    0.08341   4.333 1.47e-05 ***
PaymentMethodManual   0.40387    0.08230   4.907 9.23e-07 ***
MonthlyChargesLow    -0.63367    0.11197  -5.659 1.52e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6020.9  on 5188  degrees of freedom
Residual deviance: 4543.1  on 5170  degrees of freedom
AIC: 4581.1

Number of Fisher Scoring iterations: 5
```

Next the null model is created and summarized.



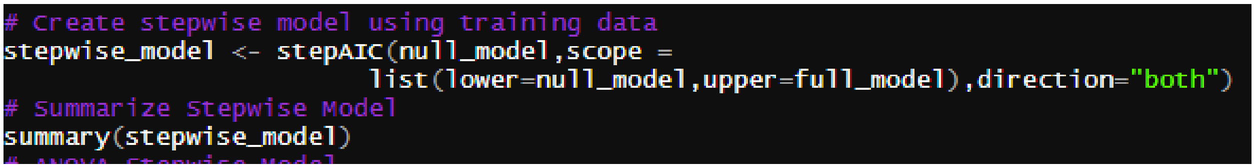
```
# Create Null Model
null_model = glm(Churn~1, family=binomial, data=training_data)
summary(null_model)
```



```
> summary(null_model)

Call:
glm(formula = Churn ~ 1, family = binomial, data = training_data)

Deviance Residuals:
   Min      1Q  Median      3Q     Max  
-0.788  -0.788  -0.788   1.625   1.625  

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.01035    0.03138  -32.19   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6020.9  on 5188  degrees of freedom
Residual deviance: 6020.9  on 5188  degrees of freedom
AIC: 6022.9

Number of Fisher Scoring iterations: 4
```

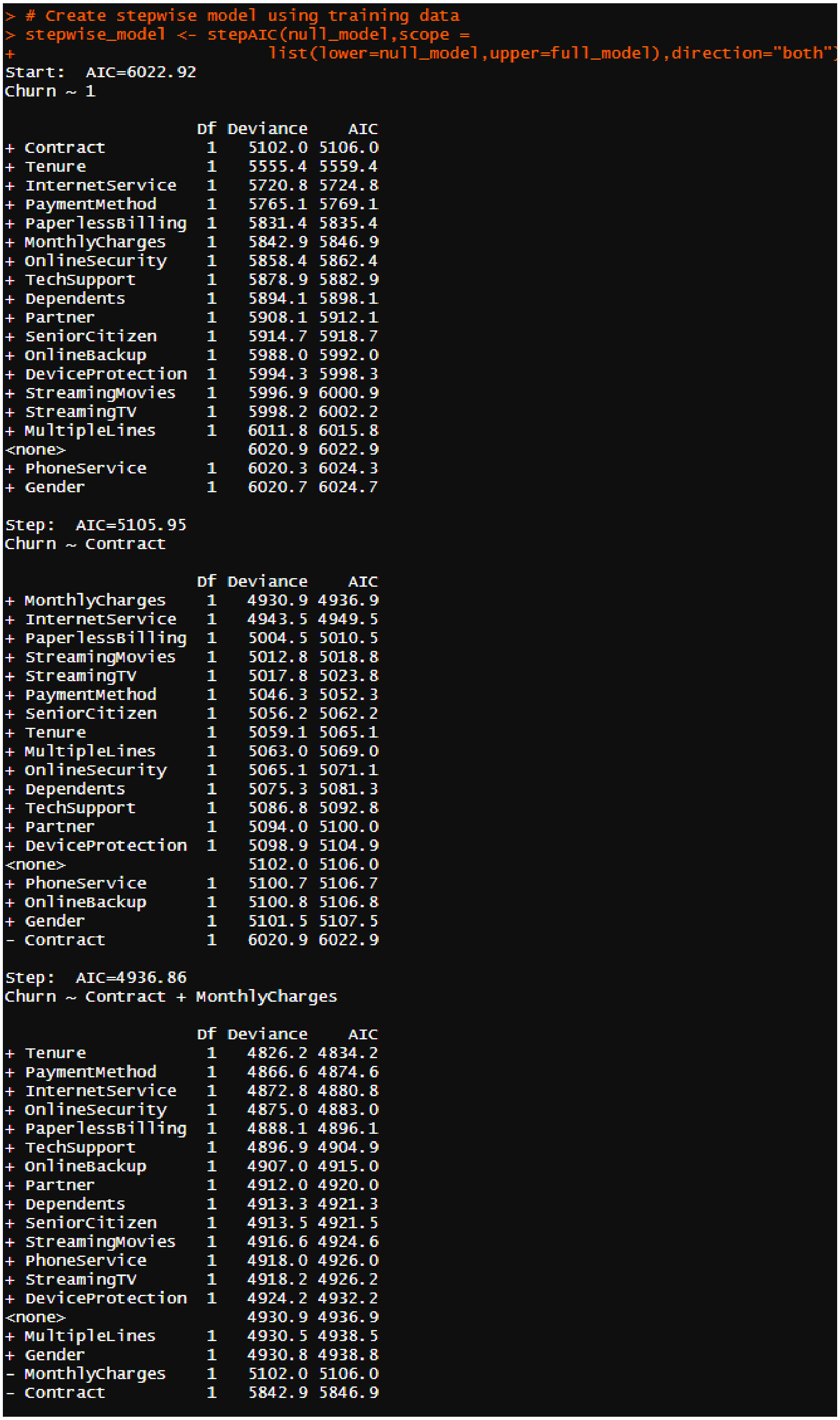
By simply looking at the AIC value, we can see that the full model is better than the null model.

Next we use stepwise logistic regression by AIC to build the model. Stepwise means that predictor variables are automatically selected in order to improve the model. By AIC means that predictor variables are selected or rejected in order to lower the AIC value. A lower AIC value indicates a better model fit. For this model, the direction both has been selected, which is a combination of both forward and backward selection, testing to select or reject variables at each step of the process.

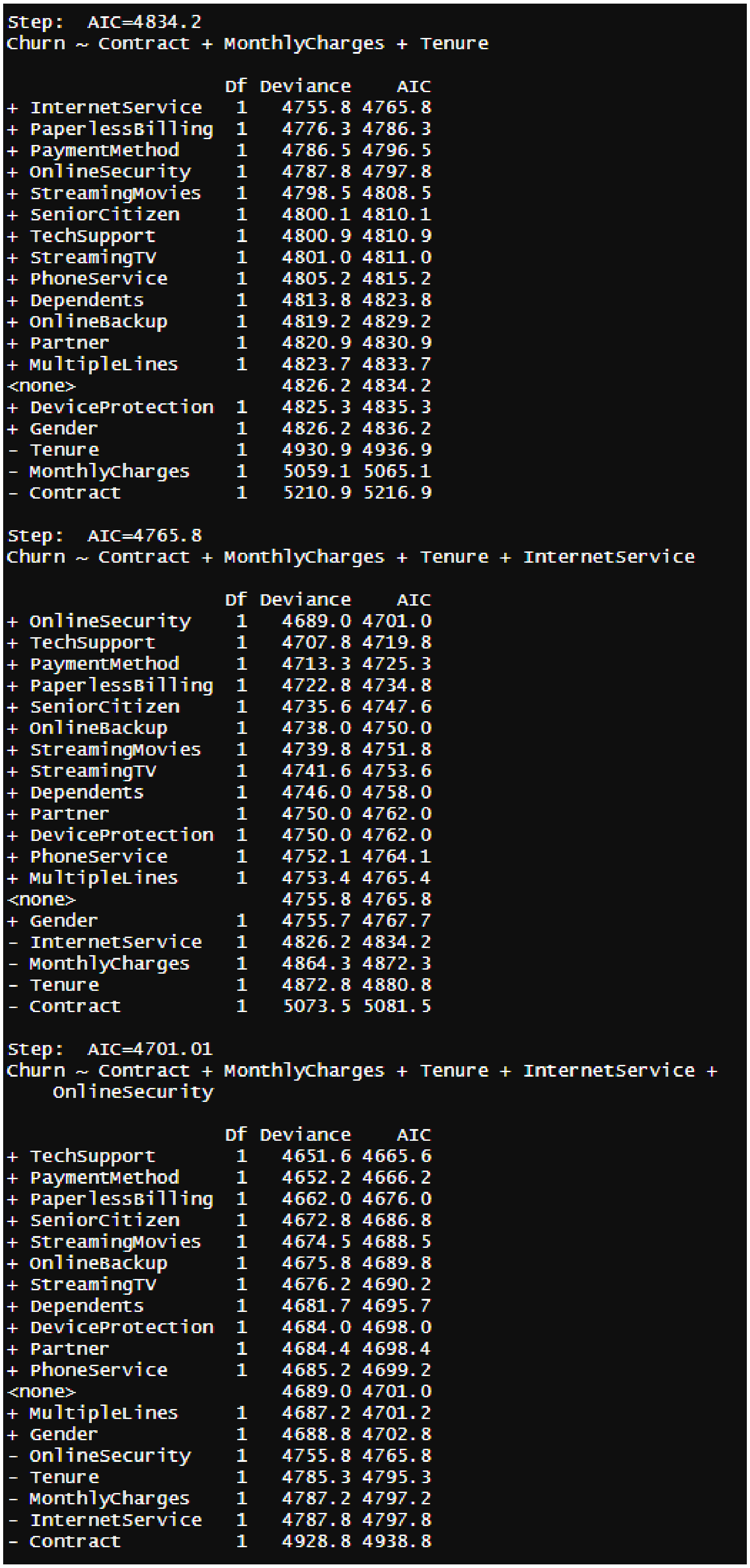


```
# Create stepwise model using training data
stepwise_model <- stepAIC(null_model,scope =
                       list(lower=null_model,upper=full_model),direction="both")
# Summarize Stepwise Model
summary(stepwise_model)
# ANOVA Stepwise Model
```

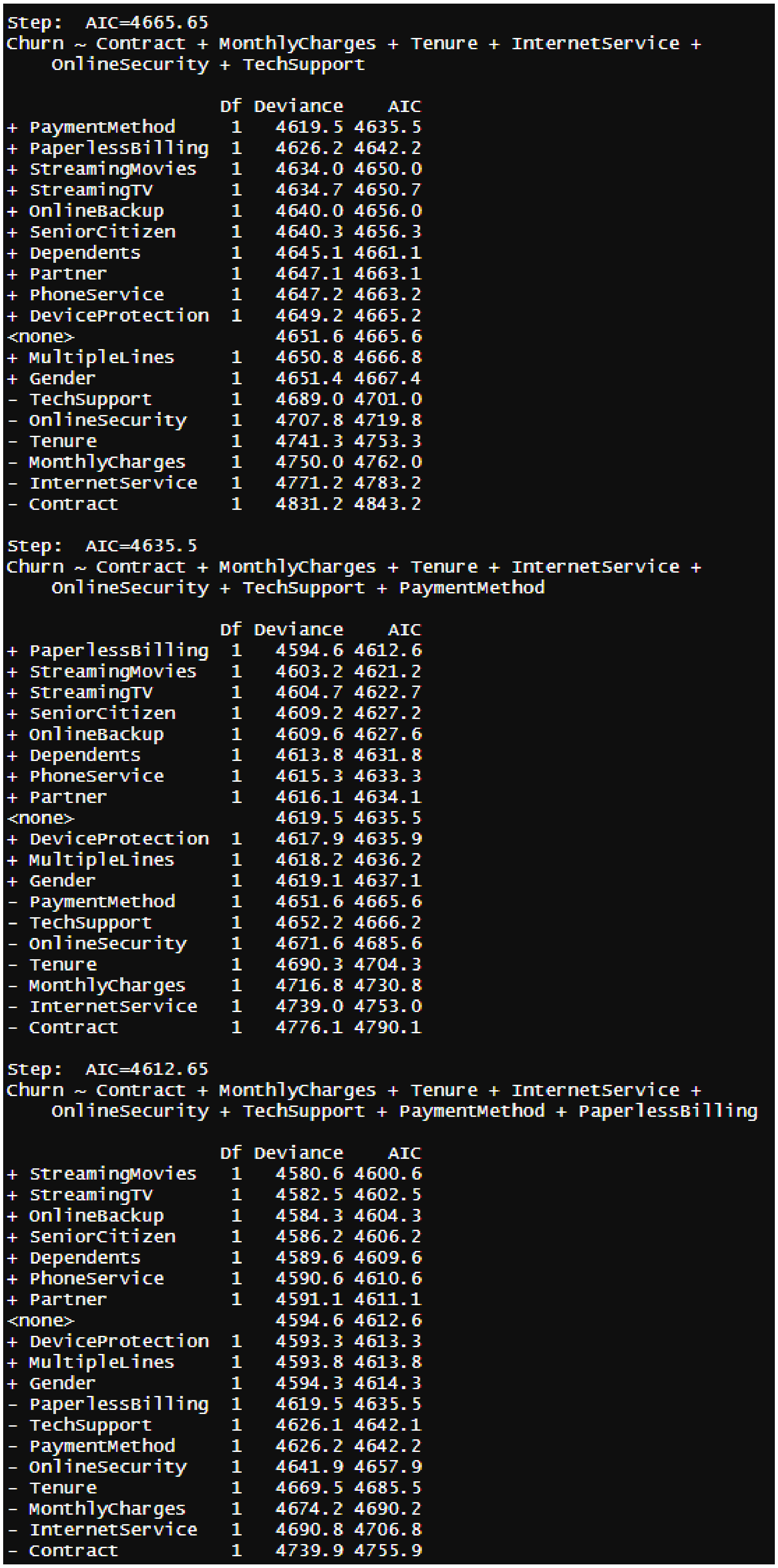
Stepwise model creation (note-this took several screenshots to capture and were pasted in order)



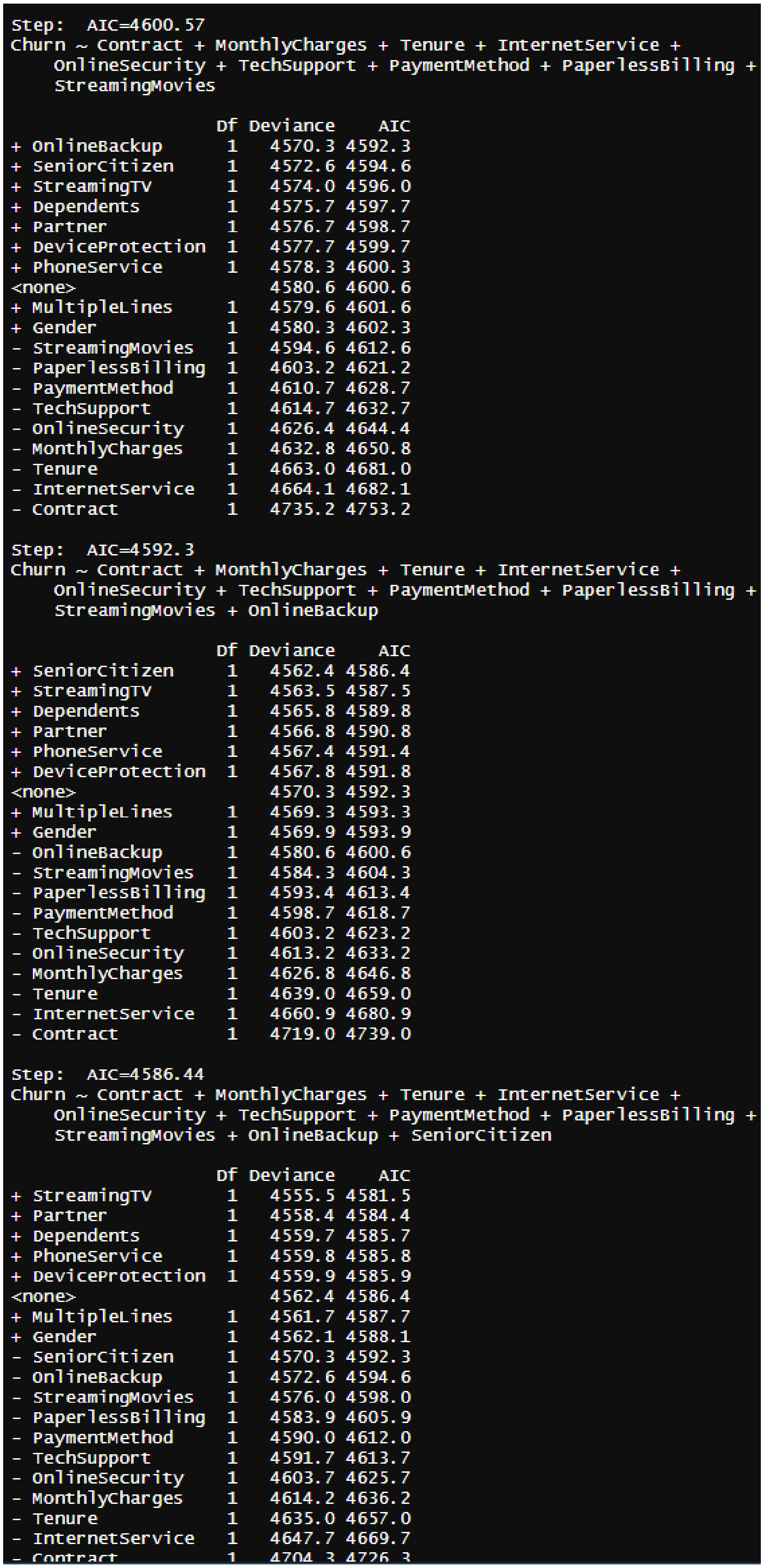
```
> # Create stepwise model using training data
> stepwise_model <- stepAIC(null_model,scope =
+                           list(lower=null_model,upper=full_model),direction="both")
Start:  AIC=6022.92
Churn ~ 1

                   Df Deviance    AIC
+ Contract          1   5102.0 5106.0
+ Tenure            1   5555.4 5559.4
+ InternetService   1   5720.8 5724.8
+ PaymentMethod     1   5765.1 5769.1
+ PaperlessBilling  1   5831.4 5835.4
+ MonthlyCharges    1   5842.9 5846.9
+ OnlineSecurity    1   5858.4 5862.4
+ TechSupport       1   5878.9 5882.9
+ Dependents        1   5894.1 5898.1
+ Partner           1   5908.1 5912.1
+ SeniorCitizen     1   5914.7 5918.7
+ OnlineBackup      1   5988.0 5992.0
+ DeviceProtection  1   5994.3 5998.3
+ StreamingMovies   1   5996.9 6000.9
+ StreamingTV       1   5998.2 6002.2
+ MultipleLines     1   6011.8 6015.8
<none>                  6020.9 6022.9
+ PhoneService      1   6020.3 6024.3
+ Gender            1   6020.7 6024.7

Step:  AIC=5105.95
Churn ~ Contract

                   Df Deviance    AIC
+ MonthlyCharges    1   4930.9 4936.9
+ InternetService   1   4943.5 4949.5
+ PaperlessBilling  1   5004.5 5010.5
+ StreamingMovies   1   5012.8 5018.8
+ StreamingTV       1   5017.8 5023.8
+ PaymentMethod     1   5046.3 5052.3
+ SeniorCitizen     1   5056.2 5062.2
+ Tenure            1   5059.1 5065.1
+ MultipleLines     1   5063.0 5069.0
+ OnlineSecurity    1   5065.1 5071.1
+ Dependents        1   5075.3 5081.3
+ TechSupport       1   5086.8 5092.8
+ Partner           1   5094.0 5100.0
+ DeviceProtection  1   5098.9 5104.9
<none>                  5102.0 5106.0
+ PhoneService      1   5100.7 5106.7
+ OnlineBackup      1   5100.8 5106.8
+ Gender            1   5101.5 5107.5
- Contract          1   6020.9 6022.9

Step:  AIC=4936.86
Churn ~ Contract + MonthlyCharges

                   Df Deviance    AIC
+ Tenure            1   4826.2 4834.2
+ PaymentMethod     1   4866.6 4874.6
+ InternetService   1   4872.8 4880.8
+ OnlineSecurity    1   4875.0 4883.0
+ PaperlessBilling  1   4888.1 4896.1
+ TechSupport       1   4896.9 4904.9
+ OnlineBackup      1   4907.0 4915.0
+ Partner           1   4912.0 4920.0
+ Dependents        1   4913.3 4921.3
+ SeniorCitizen     1   4913.5 4921.5
+ StreamingMovies   1   4916.6 4924.6
+ PhoneService      1   4918.0 4926.0
+ StreamingTV       1   4918.2 4926.2
+ DeviceProtection  1   4924.2 4932.2
<none>                  4930.9 4936.9
+ MultipleLines     1   4930.5 4938.5
+ Gender            1   4930.8 4938.8
- MonthlyCharges    1   5102.0 5106.0
- Contract          1   5842.9 5846.9
```



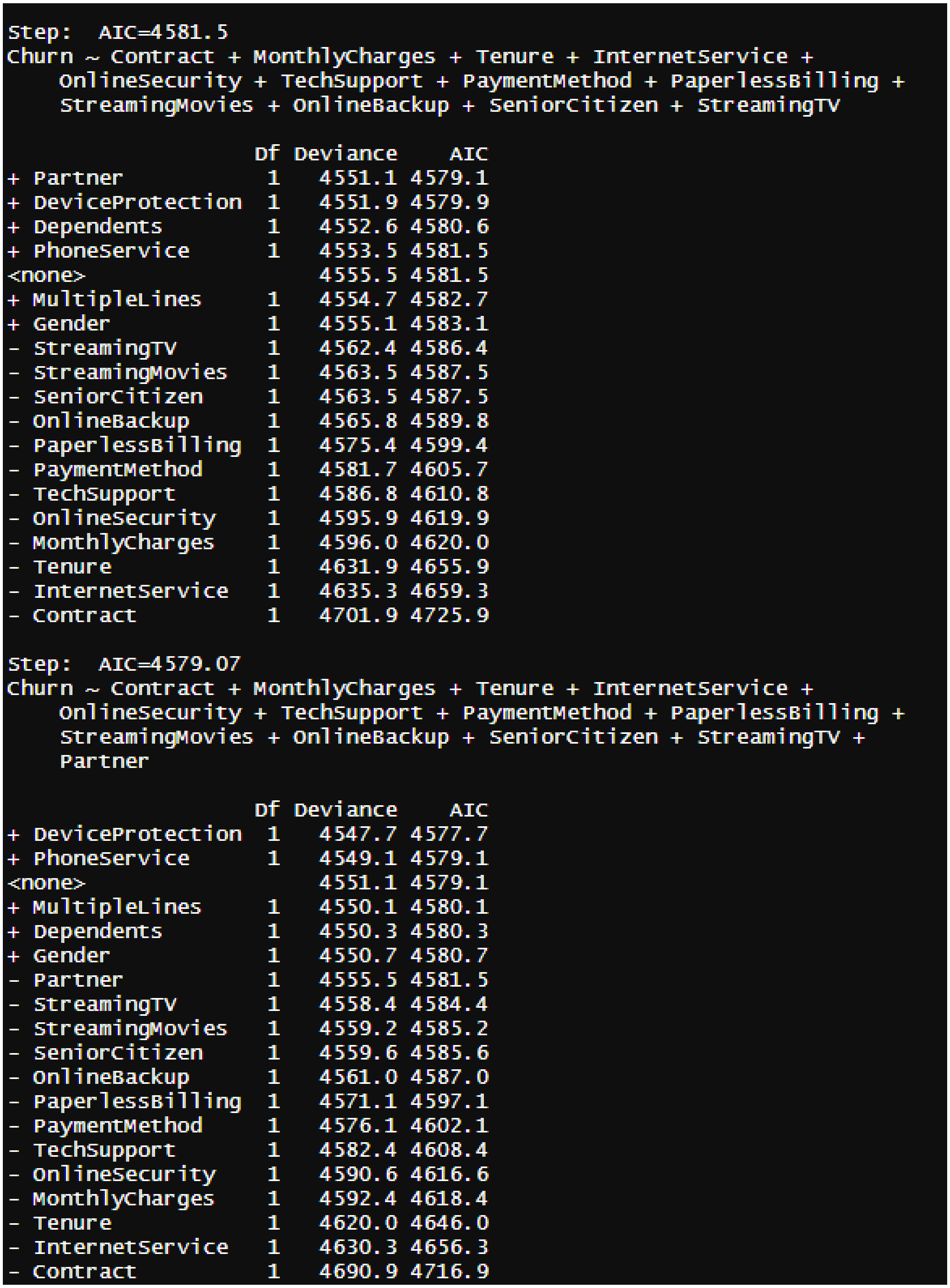
```
Step:  AIC=4834.2
Churn ~ Contract + MonthlyCharges + Tenure

                  Df Deviance    AIC
+ InternetService  1   4755.8 4765.8
+ PaperlessBilling 1   4776.3 4786.3
+ PaymentMethod    1   4786.5 4796.5
+ OnlineSecurity   1   4787.8 4797.8
+ StreamingMovies  1   4798.5 4808.5
+ SeniorCitizen    1   4800.1 4810.1
+ TechSupport      1   4800.9 4810.9
+ StreamingTV      1   4801.0 4811.0
+ PhoneService     1   4805.2 4815.2
+ Dependents       1   4813.8 4823.8
+ OnlineBackup     1   4819.2 4829.2
+ Partner          1   4820.9 4830.9
+ MultipleLines    1   4823.7 4833.7
<none>                 4826.2 4834.2
+ DeviceProtection 1   4825.3 4835.3
+ Gender           1   4826.2 4836.2
- Tenure           1   4930.9 4936.9
- MonthlyCharges   1   5059.1 5065.1
- Contract         1   5210.9 5216.9

Step:  AIC=4765.8
Churn ~ Contract + MonthlyCharges + Tenure + InternetService

                  Df Deviance    AIC
+ OnlineSecurity   1   4689.0 4701.0
+ TechSupport      1   4707.8 4719.8
+ PaymentMethod    1   4713.3 4725.3
+ PaperlessBilling 1   4722.8 4734.8
+ SeniorCitizen    1   4735.6 4747.6
+ OnlineBackup     1   4738.0 4750.0
+ StreamingMovies  1   4739.8 4751.8
+ StreamingTV      1   4741.6 4753.6
+ Dependents       1   4746.0 4758.0
+ Partner          1   4750.0 4762.0
+ DeviceProtection 1   4750.0 4762.0
+ PhoneService     1   4752.1 4764.1
+ MultipleLines    1   4753.4 4765.4
<none>                 4755.8 4765.8
+ Gender           1   4755.7 4767.7
- InternetService  1   4826.2 4834.2
- MonthlyCharges   1   4864.3 4872.3
- Tenure           1   4872.8 4880.8
- Contract         1   5073.5 5081.5

Step:  AIC=4701.01
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity

                  Df Deviance    AIC
+ TechSupport      1   4651.6 4665.6
+ PaymentMethod    1   4652.2 4666.2
+ PaperlessBilling 1   4662.0 4676.0
+ SeniorCitizen    1   4672.8 4686.8
+ StreamingMovies  1   4674.5 4688.5
+ OnlineBackup     1   4675.8 4689.8
+ StreamingTV      1   4676.2 4690.2
+ Dependents       1   4681.7 4695.7
+ DeviceProtection 1   4684.0 4698.0
+ Partner          1   4684.4 4698.4
+ PhoneService     1   4685.2 4699.2
<none>                 4689.0 4701.0
+ MultipleLines    1   4687.2 4701.2
+ Gender           1   4688.8 4702.8
- OnlineSecurity   1   4755.8 4765.8
- Tenure           1   4785.3 4795.3
- MonthlyCharges   1   4787.2 4797.2
- InternetService  1   4787.8 4797.8
- Contract         1   4928.8 4938.8
```



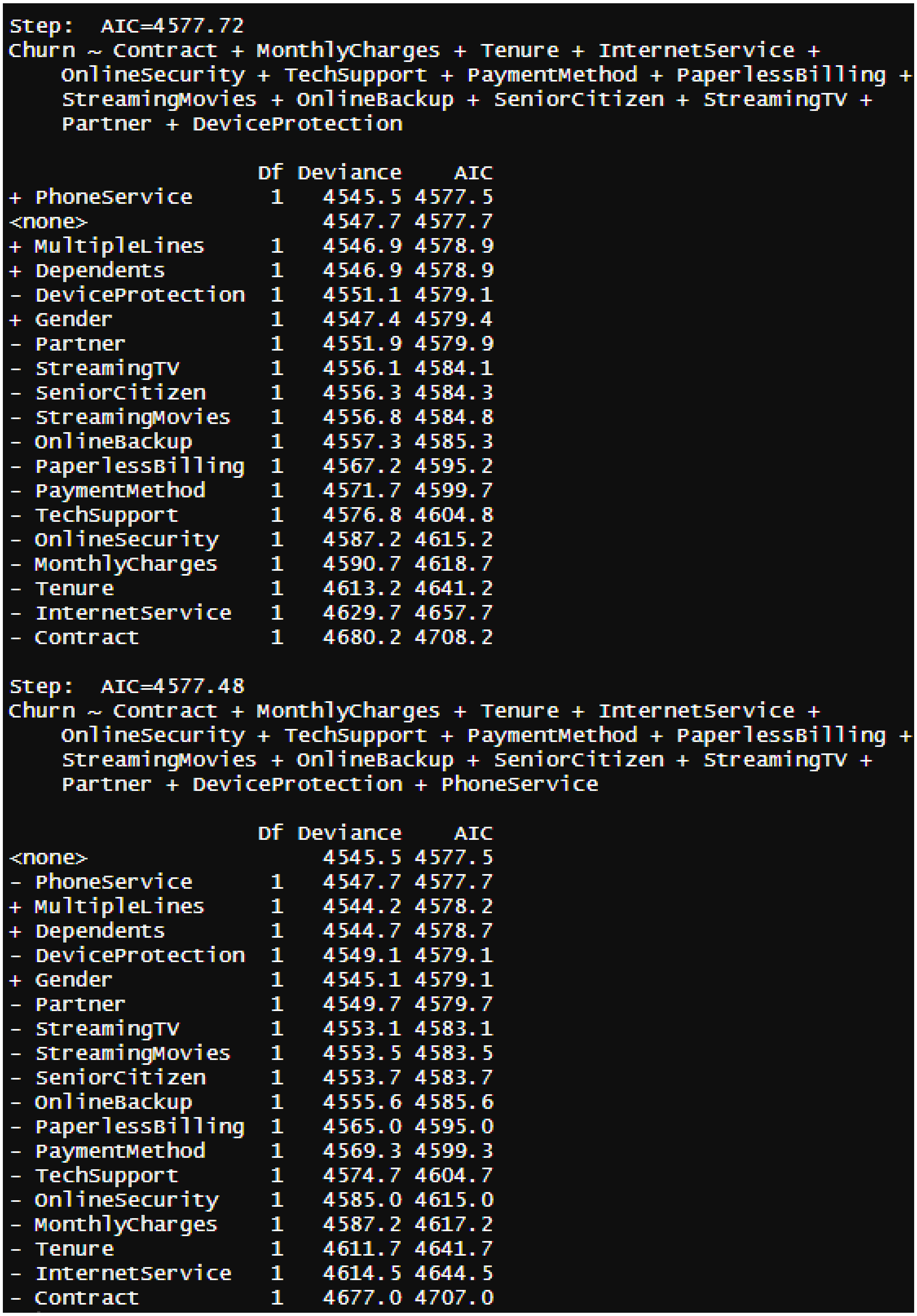
```
Step:  AIC=4665.65
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport

                   Df Deviance    AIC
+ PaymentMethod     1   4619.5 4635.5
+ PaperlessBilling  1   4626.2 4642.2
+ StreamingMovies   1   4634.0 4650.0
+ StreamingTV       1   4634.7 4650.7
+ OnlineBackup      1   4640.0 4656.0
+ SeniorCitizen     1   4640.3 4656.3
+ Dependents        1   4645.1 4661.1
+ Partner           1   4647.1 4663.1
+ PhoneService      1   4647.2 4663.2
+ DeviceProtection  1   4649.2 4665.2
<none>                  4651.6 4665.6
+ MultipleLines     1   4650.8 4666.8
+ Gender            1   4651.4 4667.4
- TechSupport       1   4689.0 4701.0
- OnlineSecurity    1   4707.8 4719.8
- Tenure            1   4741.3 4753.3
- MonthlyCharges    1   4750.0 4762.0
- InternetService   1   4771.2 4783.2
- Contract          1   4831.2 4843.2

Step:  AIC=4635.5
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport + PaymentMethod

                   Df Deviance    AIC
+ PaperlessBilling  1   4594.6 4612.6
+ StreamingMovies   1   4603.2 4621.2
+ StreamingTV       1   4604.7 4622.7
+ SeniorCitizen     1   4609.2 4627.2
+ OnlineBackup      1   4609.6 4627.6
+ Dependents        1   4613.8 4631.8
+ PhoneService      1   4615.3 4633.3
+ Partner           1   4616.1 4634.1
<none>                  4619.5 4635.5
+ DeviceProtection  1   4617.9 4635.9
+ MultipleLines     1   4618.2 4636.2
+ Gender            1   4619.1 4637.1
- PaymentMethod     1   4651.6 4665.6
- TechSupport       1   4652.2 4666.2
- OnlineSecurity    1   4671.6 4685.6
- Tenure            1   4690.3 4704.3
- MonthlyCharges    1   4716.8 4730.8
- InternetService   1   4739.0 4753.0
- Contract          1   4776.1 4790.1

Step:  AIC=4612.65
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling

                   Df Deviance    AIC
+ StreamingMovies   1   4580.6 4600.6
+ StreamingTV       1   4582.5 4602.5
+ OnlineBackup      1   4584.3 4604.3
+ SeniorCitizen     1   4586.2 4606.2
+ Dependents        1   4589.6 4609.6
+ PhoneService      1   4590.6 4610.6
+ Partner           1   4591.1 4611.1
<none>                  4594.6 4612.6
+ DeviceProtection  1   4593.3 4613.3
+ MultipleLines     1   4593.8 4613.8
+ Gender            1   4594.3 4614.3
- PaperlessBilling  1   4619.5 4635.5
- TechSupport       1   4626.1 4642.1
- PaymentMethod     1   4626.2 4642.2
- OnlineSecurity    1   4641.9 4657.9
- Tenure            1   4669.5 4685.5
- MonthlyCharges    1   4674.2 4690.2
- InternetService   1   4690.8 4706.8
- Contract          1   4739.9 4755.9
```



```
Step:  AIC=4600.57
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling + 
    StreamingMovies

                   Df Deviance    AIC
+ OnlineBackup      1   4570.3 4592.3
+ SeniorCitizen     1   4572.6 4594.6
+ StreamingTV       1   4574.0 4596.0
+ Dependents        1   4575.7 4597.7
+ Partner           1   4576.7 4598.7
+ DeviceProtection  1   4577.7 4599.7
+ PhoneService      1   4578.3 4600.3
<none>                  4580.6 4600.6
+ MultipleLines     1   4579.6 4601.6
+ Gender            1   4580.3 4602.3
- StreamingMovies   1   4594.6 4612.6
- PaperlessBilling  1   4603.2 4621.2
- PaymentMethod     1   4610.7 4628.7
- TechSupport       1   4614.7 4632.7
- OnlineSecurity    1   4626.4 4644.4
- MonthlyCharges    1   4632.8 4650.8
- Tenure            1   4663.0 4681.0
- InternetService   1   4664.1 4682.1
- Contract          1   4735.2 4753.2

Step:  AIC=4592.3
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling + 
    StreamingMovies + OnlineBackup

                   Df Deviance    AIC
+ SeniorCitizen     1   4562.4 4586.4
+ StreamingTV       1   4563.5 4587.5
+ Dependents        1   4565.8 4589.8
+ Partner           1   4566.8 4590.8
+ PhoneService      1   4567.4 4591.4
+ DeviceProtection  1   4567.8 4591.8
<none>                  4570.3 4592.3
+ MultipleLines     1   4569.3 4593.3
+ Gender            1   4569.9 4593.9
- OnlineBackup      1   4580.6 4600.6
- StreamingMovies   1   4584.3 4604.3
- PaperlessBilling  1   4593.4 4613.4
- PaymentMethod     1   4598.7 4618.7
- TechSupport       1   4603.2 4623.2
- OnlineSecurity    1   4613.2 4633.2
- MonthlyCharges    1   4626.8 4646.8
- Tenure            1   4639.0 4659.0
- InternetService   1   4660.9 4680.9
- Contract          1   4719.0 4739.0

Step:  AIC=4586.44
Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling + 
    StreamingMovies + OnlineBackup + SeniorCitizen

                   Df Deviance    AIC
+ StreamingTV       1   4555.5 4581.5
+ Partner           1   4558.4 4584.4
+ Dependents        1   4559.7 4585.7
+ PhoneService      1   4559.8 4585.8
+ DeviceProtection  1   4559.9 4585.9
<none>                  4562.4 4586.4
+ MultipleLines     1   4561.7 4587.7
+ Gender            1   4562.1 4588.1
- SeniorCitizen     1   4570.3 4592.3
- OnlineBackup      1   4572.6 4594.6
- StreamingMovies   1   4576.0 4598.0
- PaperlessBilling  1   4583.9 4605.9
- PaymentMethod     1   4590.0 4612.0
- TechSupport       1   4591.7 4613.7
- OnlineSecurity    1   4603.7 4625.7
- MonthlyCharges    1   4614.2 4636.2
- Tenure            1   4635.0 4657.0
- InternetService   1   4647.7 4669.7
- Contract          1   4704.3 4726.3
```

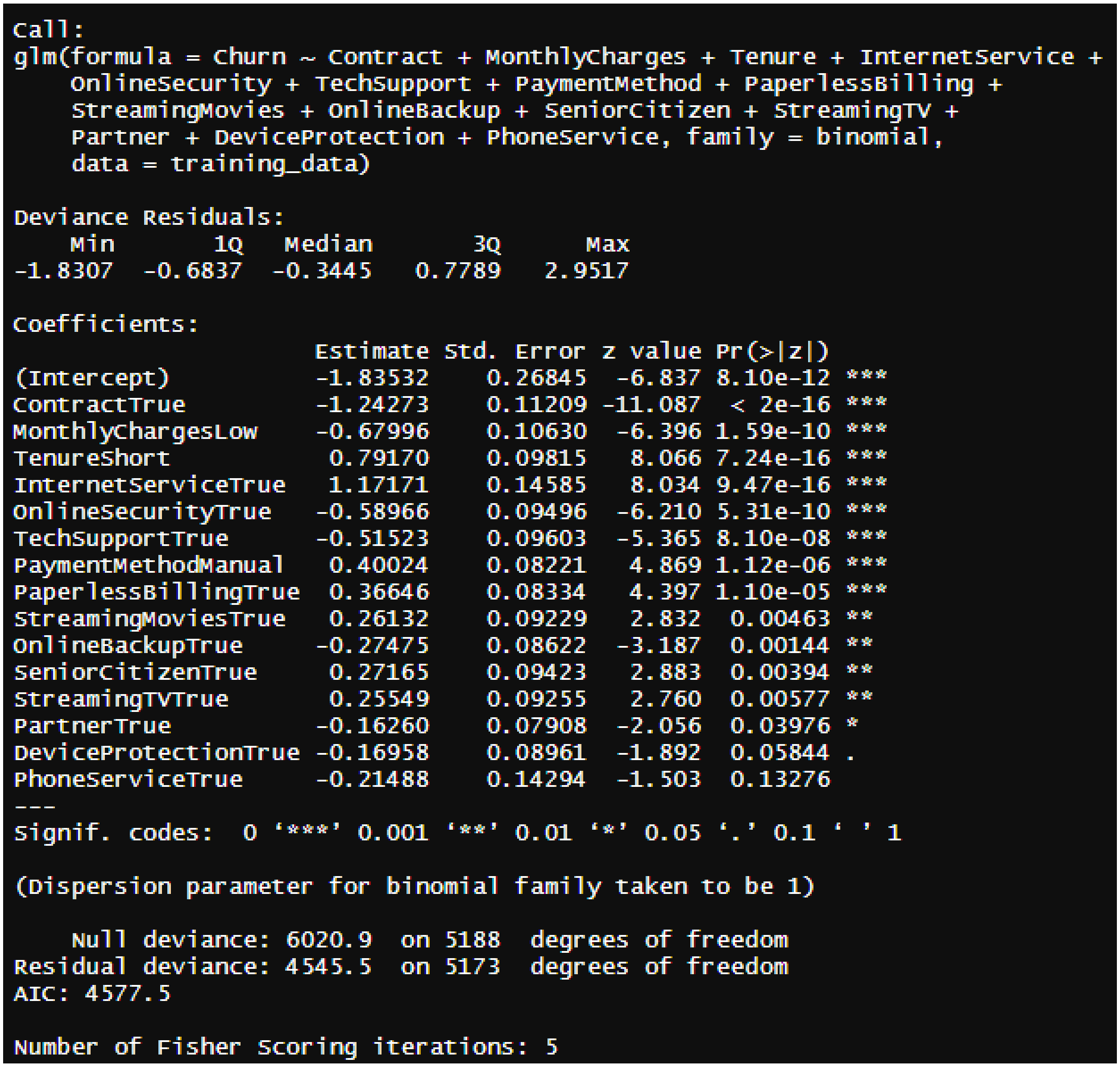


```
Step:  AIC=4581.5
Churn ~ Contract + MonthlyCharges + Tenure + InternetService +
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling +
    StreamingMovies + OnlineBackup + SeniorCitizen + StreamingTV

                   Df Deviance    AIC
+ Partner           1   4551.1 4579.1
+ DeviceProtection  1   4551.9 4579.9
+ Dependents        1   4552.6 4580.6
+ PhoneService      1   4553.5 4581.5
<none>                  4555.5 4581.5
+ MultipleLines     1   4554.7 4582.7
+ Gender            1   4555.1 4583.1
- StreamingTV       1   4562.4 4586.4
- StreamingMovies   1   4563.5 4587.5
- SeniorCitizen     1   4563.5 4587.5
- OnlineBackup      1   4565.8 4589.8
- PaperlessBilling  1   4575.4 4599.4
- PaymentMethod     1   4581.7 4605.7
- TechSupport       1   4586.8 4610.8
- OnlineSecurity    1   4595.9 4619.9
- MonthlyCharges    1   4596.0 4620.0
- Tenure            1   4631.9 4655.9
- InternetService   1   4635.3 4659.3
- Contract          1   4701.9 4725.9

Step:  AIC=4579.07
Churn ~ Contract + MonthlyCharges + Tenure + InternetService +
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling +
    StreamingMovies + OnlineBackup + SeniorCitizen + StreamingTV +
    Partner

                   Df Deviance    AIC
+ DeviceProtection  1   4547.7 4577.7
+ PhoneService      1   4549.1 4579.1
<none>                  4551.1 4579.1
+ MultipleLines     1   4550.1 4580.1
+ Dependents        1   4550.3 4580.3
+ Gender            1   4550.7 4580.7
- Partner           1   4555.5 4581.5
- StreamingTV       1   4558.4 4584.4
- StreamingMovies   1   4559.2 4585.2
- SeniorCitizen     1   4559.6 4585.6
- OnlineBackup      1   4561.0 4587.0
- PaperlessBilling  1   4571.1 4597.1
- PaymentMethod     1   4576.1 4602.1
- TechSupport       1   4582.4 4608.4
- OnlineSecurity    1   4590.6 4616.6
- MonthlyCharges    1   4592.4 4618.4
- Tenure            1   4620.0 4646.0
- InternetService   1   4630.3 4656.3
- Contract          1   4690.9 4716.9
```



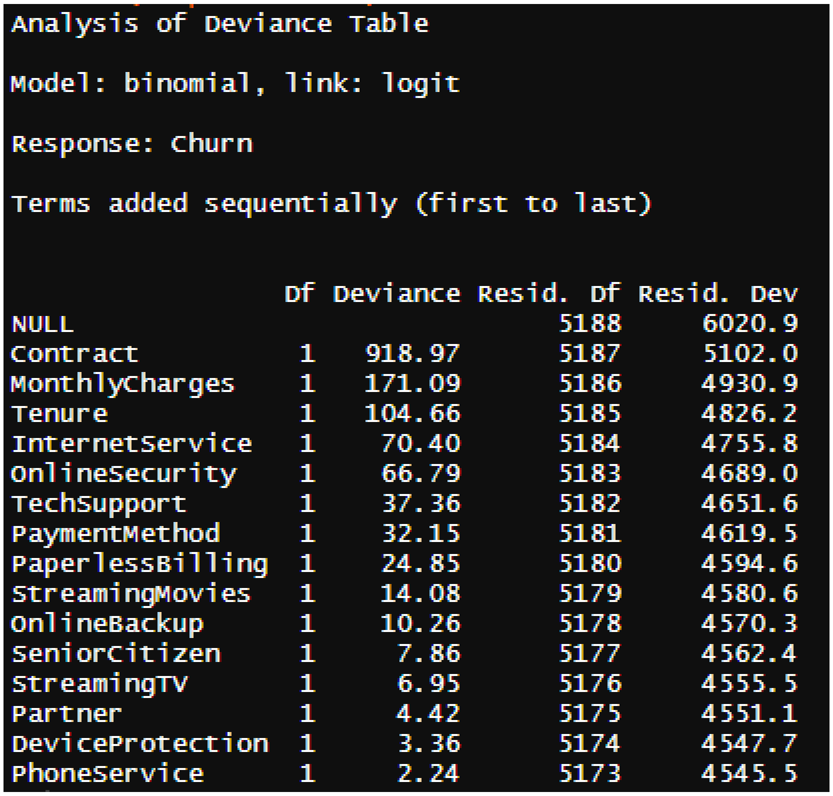
```
Step:  AIC=4577.72
Churn ~ Contract + MonthlyCharges + Tenure + InternetService +
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling +
    StreamingMovies + OnlineBackup + SeniorCitizen + StreamingTV +
    Partner + DeviceProtection

                   Df Deviance    AIC
+ PhoneService      1   4545.5 4577.5
<none>                  4547.7 4577.7
+ MultipleLines     1   4546.9 4578.9
+ Dependents        1   4546.9 4578.9
- DeviceProtection  1   4551.1 4579.1
+ Gender            1   4547.4 4579.4
- Partner           1   4551.9 4579.9
- StreamingTV       1   4556.1 4584.1
- SeniorCitizen     1   4556.3 4584.3
- StreamingMovies   1   4556.8 4584.8
- OnlineBackup      1   4557.3 4585.3
- PaperlessBilling  1   4567.2 4595.2
- PaymentMethod     1   4571.7 4599.7
- TechSupport       1   4576.8 4604.8
- OnlineSecurity    1   4587.2 4615.2
- MonthlyCharges    1   4590.7 4618.7
- Tenure            1   4613.2 4641.2
- InternetService   1   4629.7 4657.7
- Contract          1   4680.2 4708.2

Step:  AIC=4577.48
Churn ~ Contract + MonthlyCharges + Tenure + InternetService +
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling +
    StreamingMovies + OnlineBackup + SeniorCitizen + StreamingTV +
    Partner + DeviceProtection + PhoneService

                   Df Deviance    AIC
<none>                  4545.5 4577.5
- PhoneService      1   4547.7 4577.7
+ MultipleLines     1   4544.2 4578.2
+ Dependents        1   4544.7 4578.7
- DeviceProtection  1   4549.1 4579.1
+ Gender            1   4545.1 4579.1
- Partner           1   4549.7 4579.7
- StreamingTV       1   4553.1 4583.1
- StreamingMovies   1   4553.5 4583.5
- SeniorCitizen     1   4553.7 4583.7
- OnlineBackup      1   4555.6 4585.6
- PaperlessBilling  1   4565.0 4595.0
- PaymentMethod     1   4569.3 4599.3
- TechSupport       1   4574.7 4604.7
- OnlineSecurity    1   4585.0 4615.0
- MonthlyCharges    1   4587.2 4617.2
- Tenure            1   4611.7 4641.7
- InternetService   1   4614.5 4644.5
- Contract          1   4677.0 4707.0
```

Summary of the model.



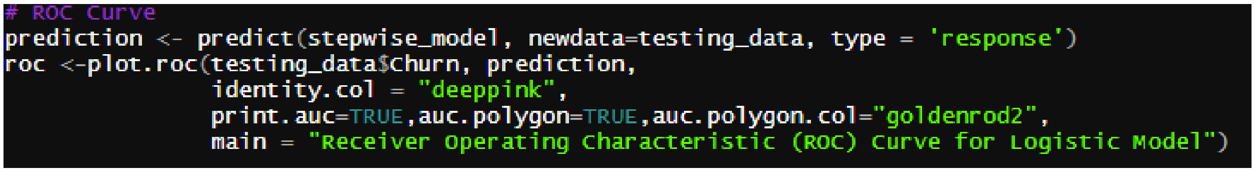
```
Call:
glm(formula = Churn ~ Contract + MonthlyCharges + Tenure + InternetService + 
    OnlineSecurity + TechSupport + PaymentMethod + PaperlessBilling + 
    StreamingMovies + OnlineBackup + SeniorCitizen + StreamingTV + 
    Partner + DeviceProtection + PhoneService, family = binomial, 
    data = training_data)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8307  -0.6837  -0.3445   0.7789   2.9517  

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)    
(Intercept)          -1.83532    0.26845  -6.837 8.10e-12 ***
ContractTrue         -1.24273    0.11209 -11.087  < 2e-16 ***
MonthlyChargesLow    -0.67996    0.10630  -6.396 1.59e-10 ***
TenureShort           0.79170    0.09815   8.066 7.24e-16 ***
InternetServiceTrue   1.17171    0.14585   8.034 9.47e-16 ***
OnlineSecurityTrue   -0.58966    0.09496  -6.210 5.31e-10 ***
TechSupportTrue      -0.51523    0.09603  -5.365 8.10e-08 ***
PaymentMethodManual   0.40024    0.08221   4.869 1.12e-06 ***
PaperlessBillingTrue  0.36646    0.08334   4.397 1.10e-05 ***
StreamingMoviesTrue   0.26132    0.09229   2.832  0.00463 ** 
OnlineBackupTrue     -0.27475    0.08622  -3.187  0.00144 ** 
SeniorCitizenTrue     0.27165    0.09423   2.883  0.00394 ** 
StreamingTVTrue       0.25549    0.09255   2.760  0.00577 ** 
PartnerTrue          -0.16260    0.07908  -2.056  0.03976 *  
DeviceProtectionTrue -0.16958    0.08961  -1.892  0.05844 .  
PhoneServiceTrue     -0.21488    0.14294  -1.503  0.13276    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 6020.9  on 5188  degrees of freedom
Residual deviance: 4545.5  on 5173  degrees of freedom
AIC: 4577.5

Number of Fisher Scoring iterations: 5
```

ANOVA of the model. This shows that as variables are added, deviance decreases. A decrease in deviance suggests a better fit for the model. Further, we can see that as more variables are used, the impact of those variables on the model decreases relative to the variables added earlier in the model. Finally, as a result of the change in deviance decreasing with additional variables, it is suggested that the variables left out of the model would not be useful.

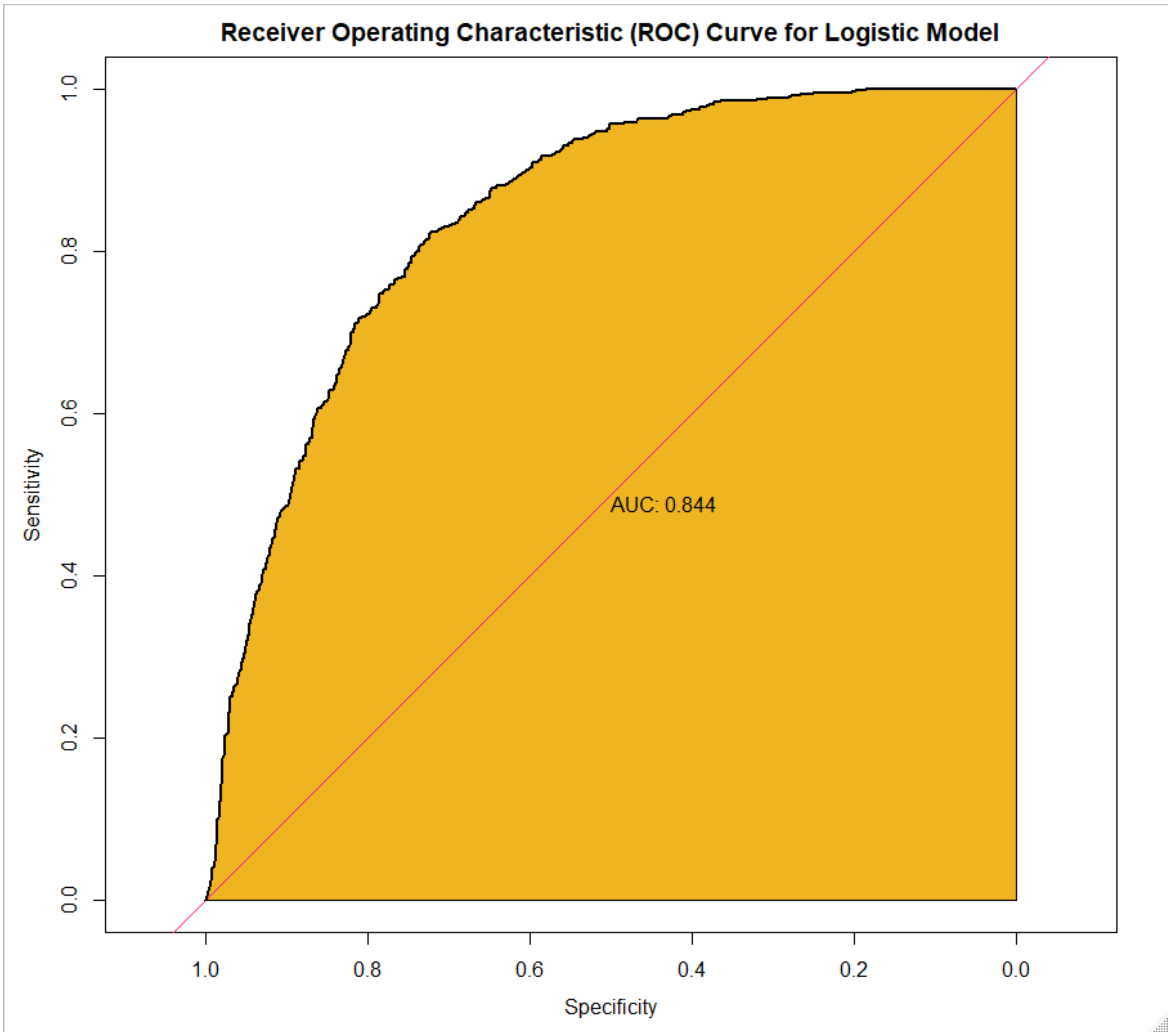


```
Analysis of Deviance Table

Model: binomial, link: logit

Response: Churn

Terms added sequentially (first to last)


                 Df Deviance Resid. Df Resid. Dev
NULL                              5188     6020.9
Contract          1   918.97      5187     5102.0
MonthlyCharges    1   171.09      5186     4930.9
Tenure            1   104.66      5185     4826.2
InternetService   1    70.40      5184     4755.8
OnlineSecurity    1    66.79      5183     4689.0
TechSupport       1    37.36      5182     4651.6
PaymentMethod     1    32.15      5181     4619.5
PaperlessBilling  1    24.85      5180     4594.6
StreamingMovies   1    14.08      5179     4580.6
OnlineBackup      1    10.26      5178     4570.3
SeniorCitizen     1     7.86      5177     4562.4
StreamingTV       1     6.95      5176     4555.5
Partner           1     4.42      5175     4551.1
DeviceProtection  1     3.36      5174     4547.7
PhoneService      1     2.24      5173     4545.5
```

Receiver Operating Characteristic (ROC) Curve for Logistic Model. As the AUC has a value of .844, we can see that the model has very high discrimination given that we are working with customer data.

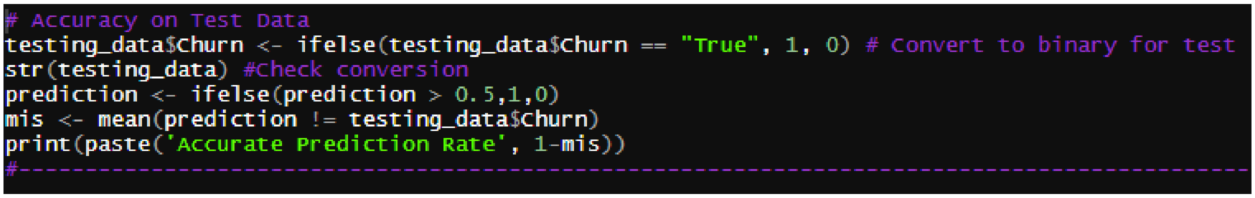


```
# ROC Curve
prediction <- predict(stepwise_model, newdata=testing_data, type = 'response')
roc <-plot.roc(testing_data$Churn, prediction,
               identity.col = "deeppink",
               print.auc=TRUE,auc.polygon=TRUE,auc.polygon.col="goldenrod2",
               main = "Receiver Operating Characteristic (ROC) Curve for Logistic Model")
```

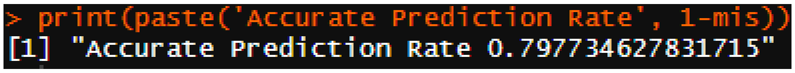


Receiver Operating Characteristic (ROC) Curve for Logistic Model

Accuracy of model on test data. Here we can see that the model accurately predicts customers who churn in the test data 79.77% of the time.



```
# Accuracy on Test Data
testing_data$Churn <- ifelse(testing_data$Churn == "True", 1, 0) # Convert to binary for test
str(testing_data) #Check conversion
prediction <- ifelse(prediction > 0.5,1,0)
mis <- mean(prediction != testing_data$Churn)
print(paste('Accurate Prediction Rate', 1-mis))
#------------------------------------------------------------------------------------------
```



```
> print(paste('Accurate Prediction Rate', 1-mis))
[1] "Accurate Prediction Rate 0.797734627831715"
```

Part L:

Using Multiple Correspondence Analysis is justified for a variety of reasons. First, the outcome variable is binary. Second, the predictor variables are also binary after having cleaned the data. While some data was not binary originally, most variables were either originally binary, or had three levels with two overlapping variables. Third, MCA is able to help distill a large number of variables into only a few dimensions. In this analysis, there were 18 predictor variables, so having only a few dimensions is greatly helpful.

Using Stepwise Logistic Regression by AIC is justified for multiple reasons. First, the outcome variable (churn) is binary, making logistic regression a good choice over other analyses such as linear regression. Second, using a stepwise strategy allows us to test the influence of each predictor variable on the model as we go, allowing the model to be corrected between each step. Third, minimizing AIC allows the model to improve by measuring the fit of the model as it goes in order to further improve at each step.

Part M:

In order to visually present my data, I first explained what the visualization was, then showed the code used to create the visualization. This allows viewers to be prepared for the visualized as well as to go back and get additional information if the want.

Within each visualization, following elements of storytelling were used to visually present the data:

1) Different color combinations to differentiate the data and to improve viewer understanding.
2) Different chart types were used to display the data visually in a variety of ways to improve viewer understanding.
3) Correct proportions and ranges were used throughout all data visualizations to improve viewer understanding.
4) The data used in all graphs and been previously cleaned, weighted, and scored before being used, resulting in consistent and correct values throughout the presentation.

Section 4:

Part N:

The phenomenon I wanted to detect was whether churn could be predicted using the data. The data was discriminating. The high AUC value from the ROC curve showing that the logistic stepwise by AIC model derived from the data was able to discriminate successfully. Further, when the logistic stepwise by AIC model was applied to the test data, it was able to accurate predict churn 79.77% of the time, further evidence that the data is discriminating. Further, the AIC of the logistic stepwise by AIC model was much lower than the null model, demonstrating a better fit to the data.

Part O:

The primary interaction found was between the Contract variable and the Internet Service variable. This interaction was found using the logistic regression method by reviewing the summary information of the model. These two variables had the greatest influence on the model, yet, in opposite directions. This suggests that there may be a significant interaction between being in a contract and having internet service. The MCA analysis also identified important variables by identifying dimensions that determine churn and then identifying which variables most great effect those dimensions. The MCA also identified the contract variable as being extremely influential.

Part P:

References:

Lê S, Josse J, Husson F (2008). “FactoMineR: A Package for Multivariate Analysis.” *Journal of Statistical Software*, 25(1), 1–18. doi: 10.18637/jss.v025.i01

Tuffery, S. (2018). Data Mining and Statistics for Decision Making. Wiley Series in Computational Statistics.