C744 Data Mining Assessment

Michiel Besseling

12/20/2019

Abstract

The purpose of this report is to apply data mining techniques to assist a telecommunications company in finding patterns and trends in customer retention from the given data set of customer records. The author will be discussing the selection of data mining software, exploring the distribution of individuals and variables in multidimensional vector space, selecting predictive models, and summarizing the findings. The ultimate goal will be to build a model to best predict whether a given client is likely to churn.

Section I: Tool Selection

This study will be done using R via R Studio. I have chosen to use R for a number of reasons. R is free and available on Windows or Mac platforms. R Studio is user friendly and has a number of additional features, such as R Markdown, which I am using to write this report. R also has a wide variety of additional packages that can be used along with its base functions. I will be employing many of these packages, such as dplyr, Tidyverse, FactomineR, ggplots2, and many more. One downside of R is that it cannot handle very large amounts of data (in the tens of millions +), but our data set is well within the capabilities of R.

To build a model to predict the churn rate of this particular company, I will first use multivariate correspondence analysis (MCA) and cluster analysis to describe the data set and find groups of commonly shared variables. I will then build a few models to predict churn rate using logistic regression. Finally I will analyze, compare, and possibly combine models to find the best performing model.

Section II: Data Exploration and Preperation

The goal of the data preparation is to load the data, load additional packages required, address any missing or abhorrent values, rename variable levels if too long or inconsistent, reduce the number of levels of each variable if possible, and export a .xls file.

We begin by loading the data and calling packages.

```r
data = read.csv("~/Downloads/School/WGU/C 744 Data Mining/WA_Fn-UseC_-Telco-Customer-Churn.csv")
library(ggplot2)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(plyr)
```

```
## -------------------------------------------------------------------------------
```

```
## You have loaded plyr after dplyr - this is likely to cause problems.
## If you need functions from both plyr and dplyr, please load plyr first, then dplyr:
## library(plyr); library(dplyr)
```

```
## -------------------------------------------------------------------------------
```

```
## 
## Attaching package: 'plyr'
```

```
## The following objects are masked from 'package:dplyr':
## 
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```r
library(FactoMineR)
library(factoextra)
```

```
## Welcome! Related Books: `Practical Guide To Cluster Analysis in R` at https://goo.gl/13EFCZ
```

```r
library(rpart)
library(rpart.plot)
library(MASS)
```

```
## 
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
## 
##     select
```

```r
library(caret)
```

```
## Loading required package: lattice
```

```r
library(ROCR)
```

```
## Loading required package: gplots
```

```
## 
## Attaching package: 'gplots'
```

```
## The following object is masked from 'package:stats':
## 
##     lowess
```

```r
library(DescTools)
```

```
## Registered S3 method overwritten by 'DescTools':
##   method         from 
##   reorder.factor gdata
```

```
## 
## Attaching package: 'DescTools'
```

```
## The following objects are masked from 'package:caret':
## 
##     MAE, RMSE
```

I first like to use the `str()` function to examine the data set.

```
str(data)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ customerID      : Factor w/ 7043 levels "0002-ORFBO","0003-MKNFE",..: 5376 3963 2565 5536 6512 655
##  $ gender          : Factor w/ 2 levels "Female","Male": 1 2 2 2 1 1 2 1 1 2 ...
##  $ SeniorCitizen   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Partner         : Factor w/ 2 levels "No","Yes": 2 1 1 1 1 1 1 1 2 1 ...
##  $ Dependents      : Factor w/ 2 levels "No","Yes": 1 1 1 1 1 1 2 1 1 2 ...
##  $ tenure          : int  1 34 2 45 2 8 22 10 28 62 ...
##  $ PhoneService    : Factor w/ 2 levels "No","Yes": 1 2 2 1 2 2 2 1 2 2 ...
##  $ MultipleLines   : Factor w/ 3 levels "No","No phone service",..: 2 1 1 2 1 3 3 2 3 1 ...
##  $ InternetService : Factor w/ 3 levels "DSL","Fiber optic",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ OnlineSecurity  : Factor w/ 3 levels "No","No internet service",..: 1 3 3 3 1 1 1 3 1 3 ...
##  $ OnlineBackup    : Factor w/ 3 levels "No","No internet service",..: 3 1 3 1 1 1 3 1 1 3 ...
##  $ DeviceProtection: Factor w/ 3 levels "No","No internet service",..: 1 3 1 3 1 3 1 1 3 1 ...
##  $ TechSupport     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 3 1 1 1 1 3 1 ...
##  $ StreamingTV     : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 3 1 3 1 ...
##  $ StreamingMovies : Factor w/ 3 levels "No","No internet service",..: 1 1 1 1 1 3 1 1 3 1 ...
##  $ Contract        : Factor w/ 3 levels "Month-to-month",..: 1 2 1 2 1 1 1 1 1 2 ...
##  $ PaperlessBilling: Factor w/ 2 levels "No","Yes": 2 1 2 1 2 2 2 1 2 1 ...
##  $ PaymentMethod   : Factor w/ 4 levels "Bank transfer (automatic)",..: 3 4 4 1 3 3 2 4 3 1 ...
##  $ MonthlyCharges  : num  29.9 57 53.9 42.3 70.7 ...
##  $ TotalCharges    : num  29.9 1889.5 108.2 1840.8 151.7 ...
##  $ Churn           : Factor w/ 2 levels "No","Yes": 1 1 2 1 2 2 1 1 2 1 ...
```

```
head(data)
```

```
##   customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 1 7590-VHVEG Female             0     Yes         No      1           No
## 2 5575-GNVDE   Male             0      No         No     34          Yes
## 3 3668-QPYBK   Male             0      No         No      2          Yes
## 4 7795-CFOCW   Male             0      No         No     45           No
## 5 9237-HQITU Female             0      No         No      2          Yes
## 6 9305-CDSKC Female             0      No         No      8          Yes
##      MultipleLines InternetService OnlineSecurity OnlineBackup DeviceProtection
## 1 No phone service             DSL             No          Yes               No
## 2               No             DSL            Yes           No              Yes
## 3               No             DSL            Yes          Yes               No
## 4 No phone service             DSL            Yes           No              Yes
## 5               No     Fiber optic             No           No               No
## 6              Yes     Fiber optic             No           No              Yes
##   TechSupport StreamingTV StreamingMovies       Contract PaperlessBilling
## 1          No          No              No Month-to-month              Yes
## 2          No          No              No       One year               No
## 3          No          No              No Month-to-month              Yes
## 4         Yes          No              No       One year               No
## 5          No          No              No Month-to-month              Yes
## 6          No         Yes             Yes Month-to-month              Yes
##               PaymentMethod MonthlyCharges TotalCharges Churn
## 1          Electronic check          29.85        29.85    No
## 2              Mailed check          56.95      1889.50    No
## 3              Mailed check          53.85       108.15   Yes
## 4 Bank transfer (automatic)          42.30      1840.75    No
## 5          Electronic check          70.70       151.65   Yes
## 6          Electronic check          99.65       820.50   Yes
```

There are 7,043 customer records with 21 variables. The response variable of interest is Churn, with “Yes” implying the customer has discontinued service with the company. This is a binary categorical variable.

The remaining 20 variables are the independent variables. To inspect the data, I first check whether each variable is correctly interpreted as a categorical or numerical variable. Most of these variables are categorical (factor), with most of those with three levels. The senior citizen category, consisting of only 1’s and 0’s, should be changed from an integer to a factor for consistency. Other than that, there are three numerical independent variables and 17 categorical independent variables. It would be reasonable to assume a decent amount of collinearity within these independent variables. For example, a customer’s TotalCharges value is likely to be strongly correlated with their tenure, and a customer is probably more likely to purchase OnlineBackup if they already have purchased OnlineSecurity.

Now we look at the data, checking for NA’s and possible inconsistencies.

```r
summary(data)
```

```
##       customerID      gender     SeniorCitizen    Partner    Dependents
##  0002-ORFBO:   1   Female:3488   Min.   :0.0000   No :3641   No :4933
##  0003-MKNFE:   1   Male  :3555   1st Qu.:0.0000   Yes:3402   Yes:2110
##  0004-TLHLJ:   1                 Median :0.0000
##  0011-IGKFF:   1                 Mean   :0.1621
##  0013-EXCHZ:   1                 3rd Qu.:0.0000
##  0013-MHZWF:   1                 Max.   :1.0000
##  (Other)   :7037
##      tenure      PhoneService          MultipleLines     InternetService
##  Min.   : 0.00   No : 682     No              :3390   DSL        :2421
##  1st Qu.: 9.00   Yes:6361     No phone service: 682   Fiber optic:3096
##  Median :29.00                Yes             :2971   No         :1526
##  Mean   :32.37
##  3rd Qu.:55.00
##  Max.   :72.00
##
##              OnlineSecurity              OnlineBackup
##  No                  :3498   No                  :3088
##  No internet service:1526   No internet service:1526
##  Yes                 :2019   Yes                 :2429
##
##
##
##
##             DeviceProtection              TechSupport
##  No                  :3095    No                  :3473
##  No internet service:1526    No internet service:1526
##  Yes                 :2422    Yes                 :2044
##
##
##
##
##               StreamingTV              StreamingMovies           Contract
##  No                  :2810   No                  :2785   Month-to-month:3875
##  No internet service:1526   No internet service:1526   One year      :1473
##  Yes                 :2707   Yes                 :2732   Two year      :1695
##
##
```

```
##
##
##  PaperlessBilling                   PaymentMethod  MonthlyCharges
##  No :2872         Bank transfer (automatic):1544   Min.   : 18.25
##  Yes:4171         Credit card (automatic)  :1522   1st Qu.: 35.50
##                   Electronic check         :2365   Median : 70.35
##                   Mailed check             :1612   Mean   : 64.76
##                                                    3rd Qu.: 89.85
##                                                    Max.   :118.75
##
##   TotalCharges    Churn
##  Min.   :  18.8   No :5174
##  1st Qu.: 401.4   Yes:1869
##  Median :1397.5
##  Mean   :2283.3
##  3rd Qu.:3794.7
##  Max.   :8684.8
##  NA's   :11
```

Fortunately, most of this data set looks relatively clean. All of the summaries above seem acceptable except the NA's in the TotalCharges variable. Let's examine these 11 particular observations.

```
subset(data, is.na(TotalCharges))
```

```
##      customerID gender SeniorCitizen Partner Dependents tenure PhoneService
## 489  4472-LVYGI Female             0     Yes        Yes      0           No
## 754  3115-CZMZD   Male             0      No        Yes      0          Yes
## 937  5709-LVOEQ Female             0     Yes        Yes      0          Yes
## 1083 4367-NUYAO   Male             0     Yes        Yes      0          Yes
## 1341 1371-DWPAZ Female             0     Yes        Yes      0           No
## 3332 7644-OMVMY   Male             0     Yes        Yes      0          Yes
## 3827 3213-VVOLG   Male             0     Yes        Yes      0          Yes
## 4381 2520-SGTTA Female             0     Yes        Yes      0          Yes
## 5219 2923-ARZLG   Male             0     Yes        Yes      0          Yes
## 6671 4075-WKNIU Female             0     Yes        Yes      0          Yes
## 6755 2775-SEFEE   Male             0      No        Yes      0          Yes
##         MultipleLines InternetService      OnlineSecurity        OnlineBackup
## 489  No phone service             DSL                 Yes                  No
## 754                No              No No internet service No internet service
## 937                No             DSL                 Yes                 Yes
## 1083              Yes              No No internet service No internet service
## 1341 No phone service             DSL                 Yes                 Yes
## 3332               No              No No internet service No internet service
## 3827              Yes              No No internet service No internet service
## 4381               No              No No internet service No internet service
## 5219               No              No No internet service No internet service
## 6671              Yes             DSL                  No                 Yes
## 6755              Yes             DSL                 Yes                 Yes
##         DeviceProtection         TechSupport         StreamingTV
## 489                  Yes                 Yes                 Yes
## 754  No internet service No internet service No internet service
## 937                  Yes                  No                 Yes
## 1083 No internet service No internet service No internet service
## 1341                 Yes                 Yes                 Yes
## 3332 No internet service No internet service No internet service
## 3827 No internet service No internet service No internet service
## 4381 No internet service No internet service No internet service
## 5219 No internet service No internet service No internet service
## 6671                 Yes                 Yes                 Yes
## 6755                  No                 Yes                  No
##          StreamingMovies Contract PaperlessBilling             PaymentMethod
## 489                   No Two year              Yes Bank transfer (automatic)
## 754  No internet service Two year               No              Mailed check
## 937                  Yes Two year               No              Mailed check
## 1083 No internet service Two year               No              Mailed check
## 1341                  No Two year               No   Credit card (automatic)
## 3332 No internet service Two year               No              Mailed check
## 3827 No internet service Two year               No              Mailed check
## 4381 No internet service Two year               No              Mailed check
## 5219 No internet service One year              Yes              Mailed check
## 6671                  No Two year               No              Mailed check
## 6755                  No Two year              Yes Bank transfer (automatic)
##      MonthlyCharges TotalCharges Churn
## 489           52.55           NA    No
## 754           20.25           NA    No
## 937           80.85           NA    No
## 1083          25.75           NA    No
## 1341          56.05           NA    No
## 3332          19.85           NA    No
## 3827          25.35           NA    No
## 4381          20.00           NA    No
## 5219          19.70           NA    No
## 6671          73.35           NA    No
## 6755          61.90           NA    No
```

All of the columns for these 11 observations, except the NA's, seem correctly filled out. The one column that they share in common is the tenure column, where all entries are zero. It seems reasonable to assume that these are all new customers who have not yet paid their first bill. Since the current minimum value for TotalCharges is 18.8 (far enough from zero), I will correct these observations by setting their TotalCharges entry to their MonthlyCharges, so it was as if these customers did pay their first bill. Then I run the `summary()` function (as we already did for the entire data set) to make sure that the NA's have been replaced and the minimum value has not changed.

```r
data$TotalCharges = ifelse(is.na(data$TotalCharges) == T, data$MonthlyCharges,
                           data$TotalCharges)
summary(data$TotalCharges)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    18.8   398.6  1394.5  2279.8  3786.6  8684.8
```

Next I will rename the levels of SeniorCitizen column as 1 = Yes and 0 = No, then check the results.

```r
data$SeniorCitizen = as.factor(ifelse(data$SeniorCitizen == 0 ,"No","Yes"))
table(data$SeniorCitizen)
```

```
##
##   No  Yes
## 5901 1142
```

Additionally, some of the plots I will use later on in this report use the names of each level of the categorical variables. In order to make the plots look neater, it is helpful to use as short of names as possible to describe each level. Thus I will change some of the names of the levels. For example, OnlineSecurity has a "No internet service” level, which I will rename as “DNA”" for “does not apply.”

```r
data$MultipleLines = as.factor(gsub("No phone service","DNA",data$MultipleLines))
data$OnlineSecurity = as.factor(gsub("No internet service","DNA",data$OnlineSecurity))
data$OnlineBackup = as.factor(gsub("No internet service","DNA",data$OnlineBackup))
data$DeviceProtection = as.factor(gsub("No internet service","DNA",data$DeviceProtection))
data$TechSupport = as.factor(gsub("No internet service","DNA",data$TechSupport))
data$StreamingTV   = as.factor(gsub("No internet service","DNA",data$StreamingTV))
data$StreamingMovies   = as.factor(gsub("No internet service","DNA",data$StreamingMovies))
```

Let’s also change the payment levels names as well as the contract to shorten the level names. The new names should be easy to decipher. Note that I changed my coding from the base function `gsub()` to the `mapvalues()` function found in the `dplyr` package. This is a good illustration of how packages in R are more user friendly than the base functions.

```r
data$PaymentMethod = mapvalues(data$PaymentMethod, c("Bank transfer (automatic)",
        "Credit card (automatic)", "Electronic check", "Mailed check"), c("Transfer","CC","ECheck", "Ma
data$Contract = mapvalues(data$Contract, c("Month-to-month", "One year",  "Two year" ),
                          c("MtM", "1year", "2year"))
data$InternetService = mapvalues(data$InternetService,"Fiber optic", "Fiber" )
levels(data$PaymentMethod)
```

```
## [1] "Transfer" "CC"       "ECheck"   "Mail"
```

```r
levels(data$Contract)
```

```
## [1] "MtM"   "1year" "2year"
```

```r
levels(data$InternetService)
```

```
## [1] "DSL"   "Fiber" "No"
```

This should be enough to have a clean and workable data set. I will rename it as churn (not to be confused with Churn, the target variable), while omitting the customer ID numbers. Let’s see how it looks.

```r
churn = data[ ,-1]
str(churn)
```

```
## 'data.frame':    7043 obs. of  20 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 ...
##  $ 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 "DNA","No","Yes": 1 2 2 1 2 3 3 1 3 2 ...
##  $ InternetService : Factor w/ 3 levels "DSL","Fiber",..: 1 1 1 1 2 2 2 1 2 1 ...
##  $ OnlineSecurity  : Factor w/ 3 levels "DNA","No","Yes": 2 3 3 3 2 2 2 3 2 3 ...
##  $ OnlineBackup    : Factor w/ 3 levels "DNA","No","Yes": 3 2 3 2 2 2 3 2 2 3 ...
##  $ DeviceProtection: Factor w/ 3 levels "DNA","No","Yes": 2 3 2 3 2 3 2 2 3 2 ...
##  $ TechSupport     : Factor w/ 3 levels "DNA","No","Yes": 2 2 2 3 2 2 2 2 3 2 ...
##  $ StreamingTV     : Factor w/ 3 levels "DNA","No","Yes": 2 2 2 2 2 3 3 2 3 2 ...
##  $ StreamingMovies : Factor w/ 3 levels "DNA","No","Yes": 2 2 2 2 2 3 2 2 3 2 ...
##  $ Contract        : Factor w/ 3 levels "MtM","1year",..: 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 "Transfer","CC",..: 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 ...
```

```
head(churn)
```

```
##   gender SeniorCitizen Partner Dependents tenure PhoneService MultipleLines
## 1 Female            No     Yes         No      1           No           DNA
## 2   Male            No      No         No     34          Yes            No
## 3   Male            No      No         No      2          Yes            No
## 4   Male            No      No         No     45           No           DNA
## 5 Female            No      No         No      2          Yes            No
## 6 Female            No      No         No      8          Yes           Yes
##   InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport
## 1             DSL             No          Yes               No          No
## 2             DSL            Yes           No              Yes          No
## 3             DSL            Yes          Yes               No          No
## 4             DSL            Yes           No              Yes         Yes
## 5           Fiber             No           No               No          No
## 6           Fiber             No           No              Yes          No
##   StreamingTV StreamingMovies Contract PaperlessBilling PaymentMethod
## 1          No              No      MtM              Yes        ECheck
## 2          No              No    1year               No          Mail
## 3          No              No      MtM              Yes          Mail
## 4          No              No    1year               No      Transfer
## 5          No              No      MtM              Yes        ECheck
## 6         Yes             Yes      MtM              Yes        ECheck
##   MonthlyCharges TotalCharges Churn
## 1          29.85        29.85    No
## 2          56.95      1889.50    No
## 3          53.85       108.15   Yes
## 4          42.30      1840.75    No
## 5          70.70       151.65   Yes
## 6          99.65       820.50   Yes
```

Lastly we export it as an .xls file. Typically, I would use the `xlsx` package for this purpose, but there seems to be a problem with R Studio finding the correct Javascript files for this package. I made the mistake of upgrading my operating system on my Mac to OS Catalina, and there have been many issues with software finding correct files paths. So I will export as a .csv file, then use Microsoft Excel to convert it to .xlsx format outside of R.

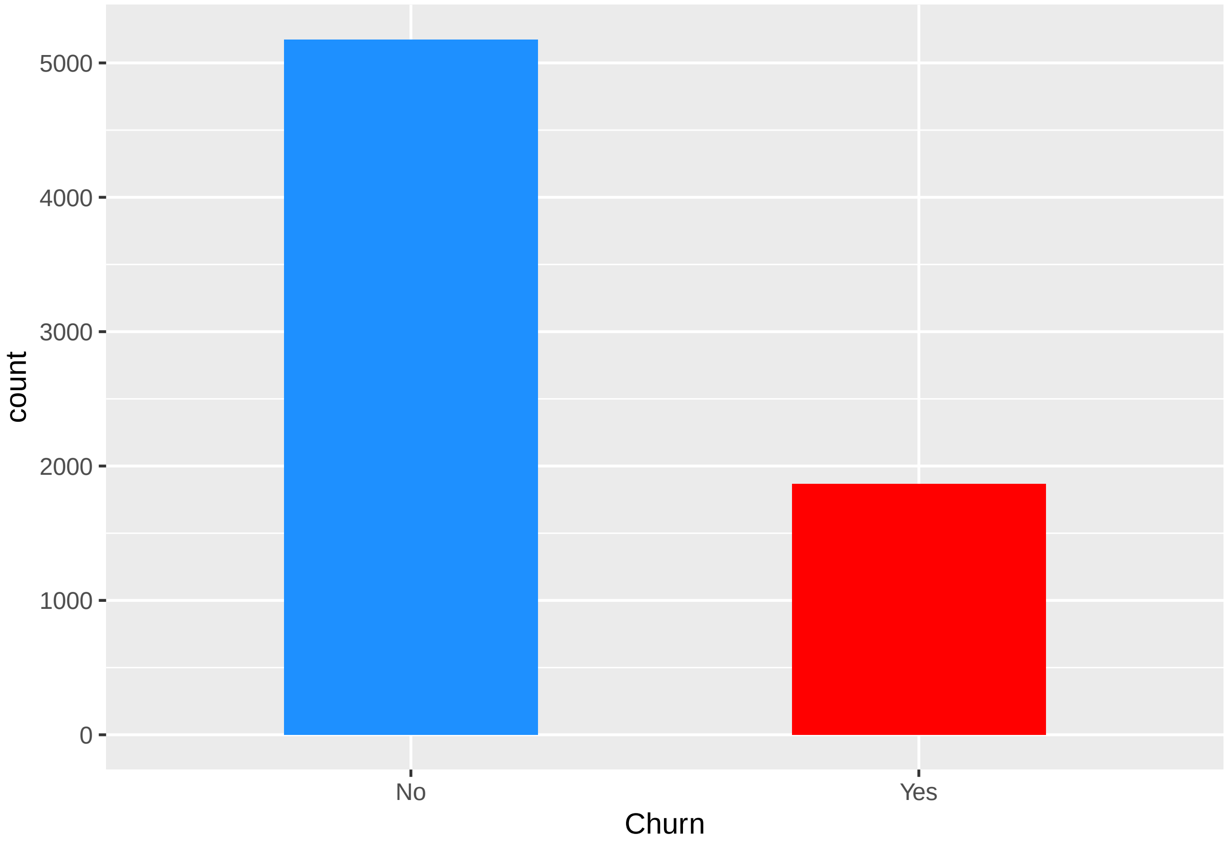
```
write.csv(churn, "churn_cleaned.csv")
```

Section III: Data Analysis

Univariate Distributions

Now we begin to look at our data. Since Churn is our target variable, let's look at its distribution using ggplot2.

```
p1 = ggplot(data=churn, aes(x=Churn)) +
  geom_bar(fill = c("dodgerblue","red"), width = .5)
p1
```



```
table(churn$Churn)
```

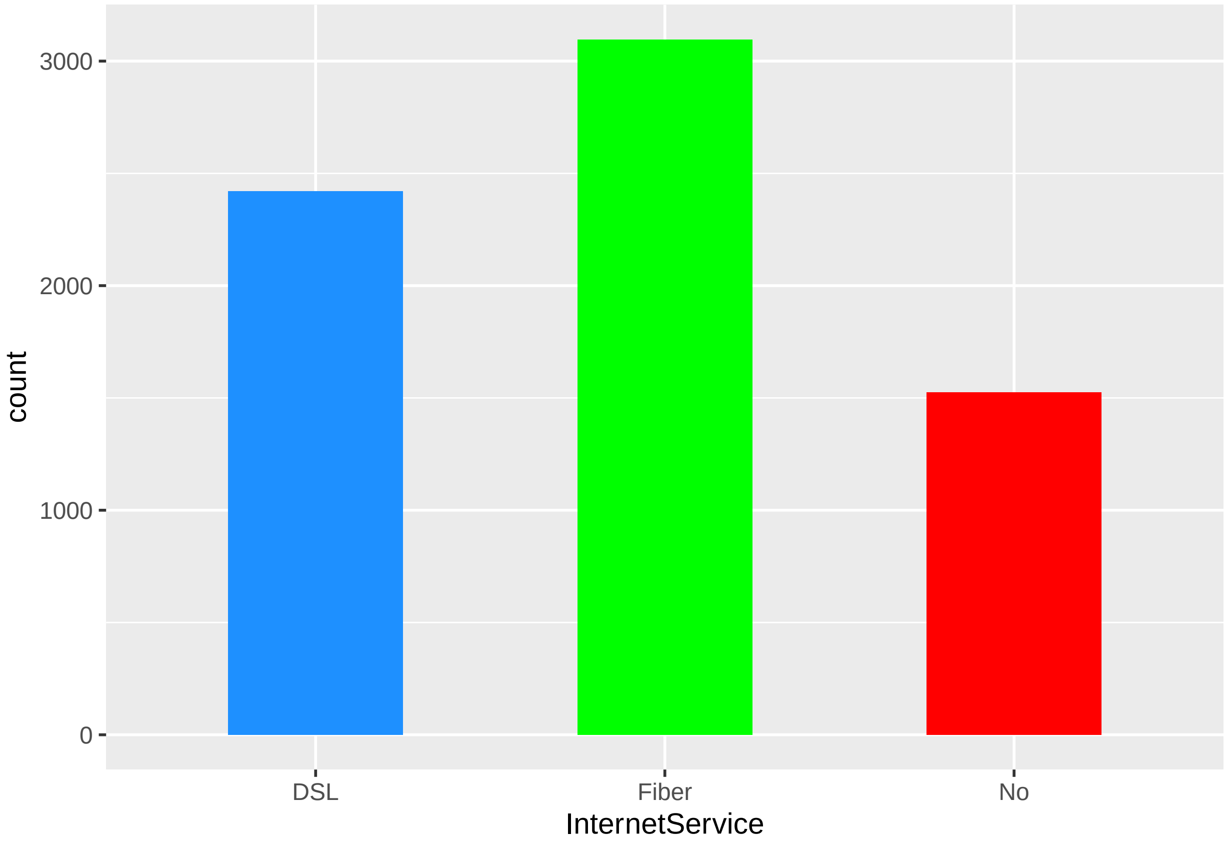
```
##
##   No  Yes
## 5174 1869
```

```
table(churn$Churn)/sum(table(churn$Churn))
```

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

So we see that about 26.54% of the 7,043 customers did leave the company. This is an important value since we can now build our first model. This model predicts that the churn rate of any randomly selected customer is 26.54%. It is now my job to find models which can do a better job of predicting the churn rate. We can then begin to understand the factors and the types of customers more likely to leave. The goal is that the company can address the potential churning customers' concerns and find ways to better improve customer retention.

```
phone = table(churn$PhoneService)
internet = table(churn$InternetService)
p2 = ggplot(data=churn, aes(x=InternetService)) +
  geom_bar(fill = c("dodgerblue","green","red"), width = .5)
p2
```



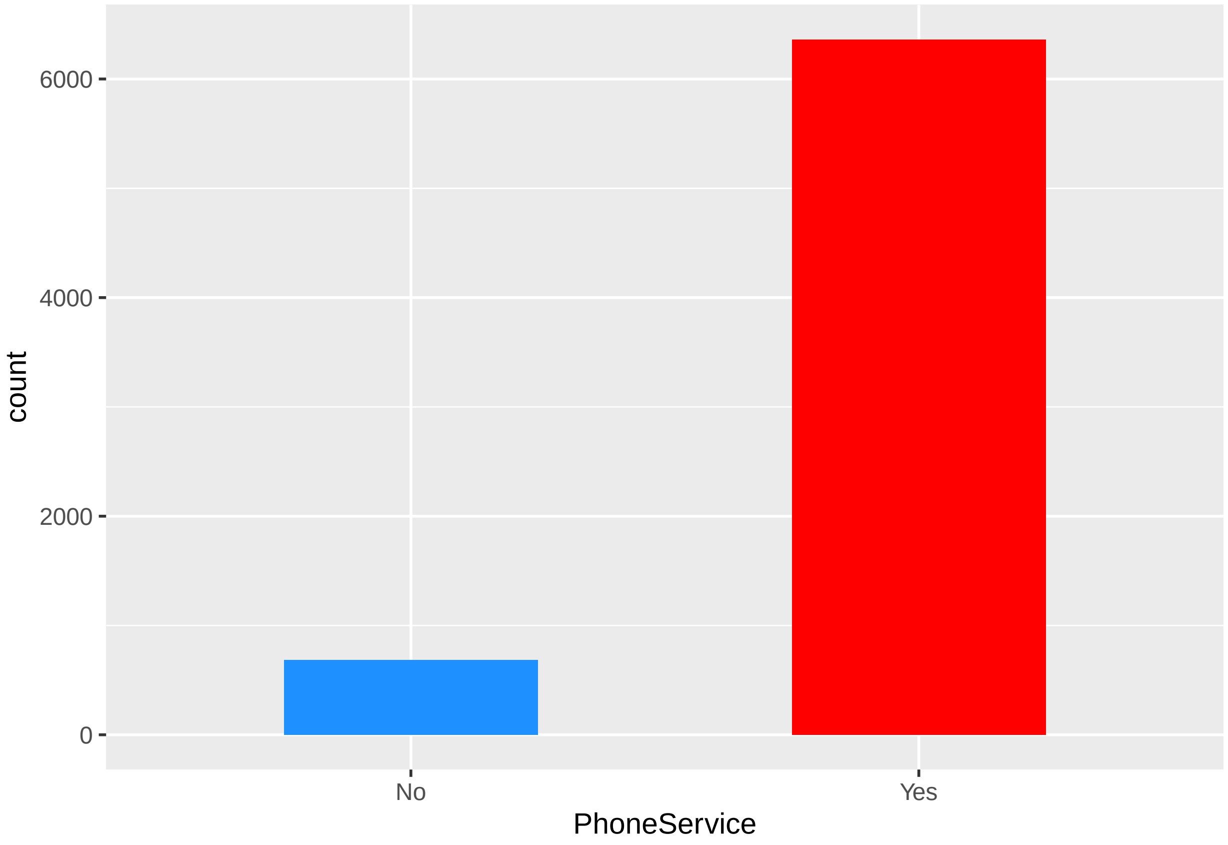
```
table(churn$InternetService)
```

```
##
##   DSL Fiber    No
##  2421  3096  1526
```

```
table(churn$InternetService)/sum(table(data$InternetService))
```

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

```
p3 = ggplot(data=churn, aes(x=PhoneService)) +
  geom_bar(fill = c("dodgerblue","red"), width = .5)
p3
```



```
print(list("Phone Service",phone))
```

```
## [[1]]
## [1] "Phone Service"
##
## [[2]]
##
##   No  Yes
##  682 6361
```

```
phone/sum(phone)
```

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

```
internet
```

```
##
##   DSL Fiber    No
##  2421  3096  1526
```

```
internet/sum(internet)
```

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

Over 90% of customers had phone service and 78.4% had some kind of internet service.

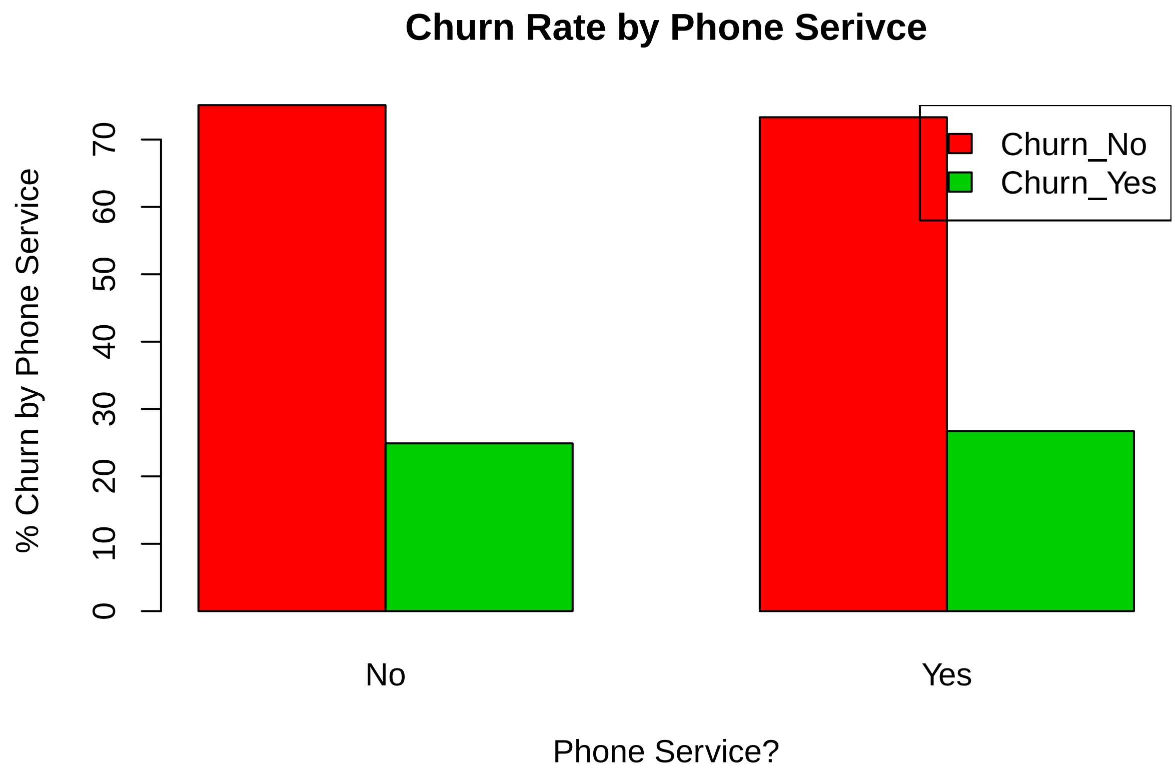
Bivariate Distributions

Let’s look at the conditional distribution of churn rate for each service.

```
c1 = table(churn$PhoneService, churn$Churn)
c2 = round(prop.table(c1,1)*100,1)
c2
```

```
##
##          No  Yes
##   No  75.1 24.9
##   Yes 73.3 26.7
```

```
barplot(t(c2), beside = T, main = "Churn Rate by Phone Serivce", col = 2:3, xlab = "Phone Service?", yla
legend('topright',fill = 2:3, c("Churn_No","Churn_Yes"))
```

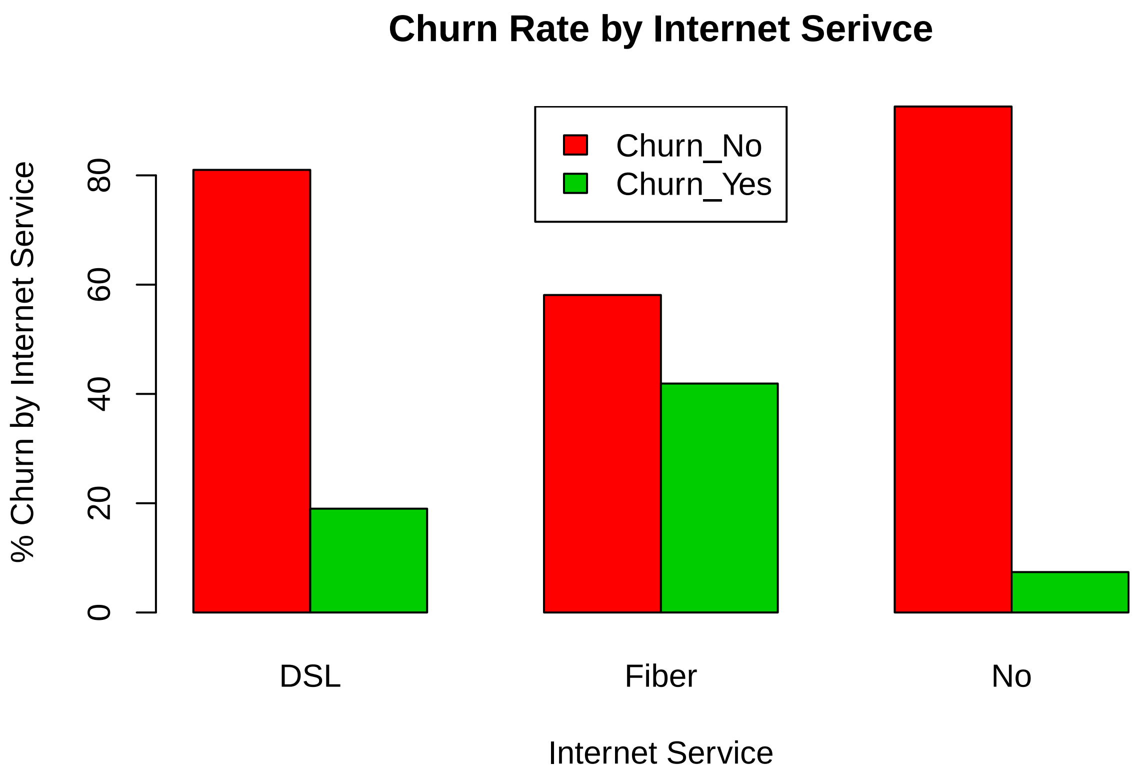


It appears that the churn rate is spread out evenly between the phone service.

```
c3 = table(churn$InternetService, churn$Churn)
c4 = round(prop.table(c3,1)*100,1)
c4
```

```
##
##            No  Yes
##   DSL   81.0 19.0
##   Fiber 58.1 41.9
##   No    92.6  7.4
```

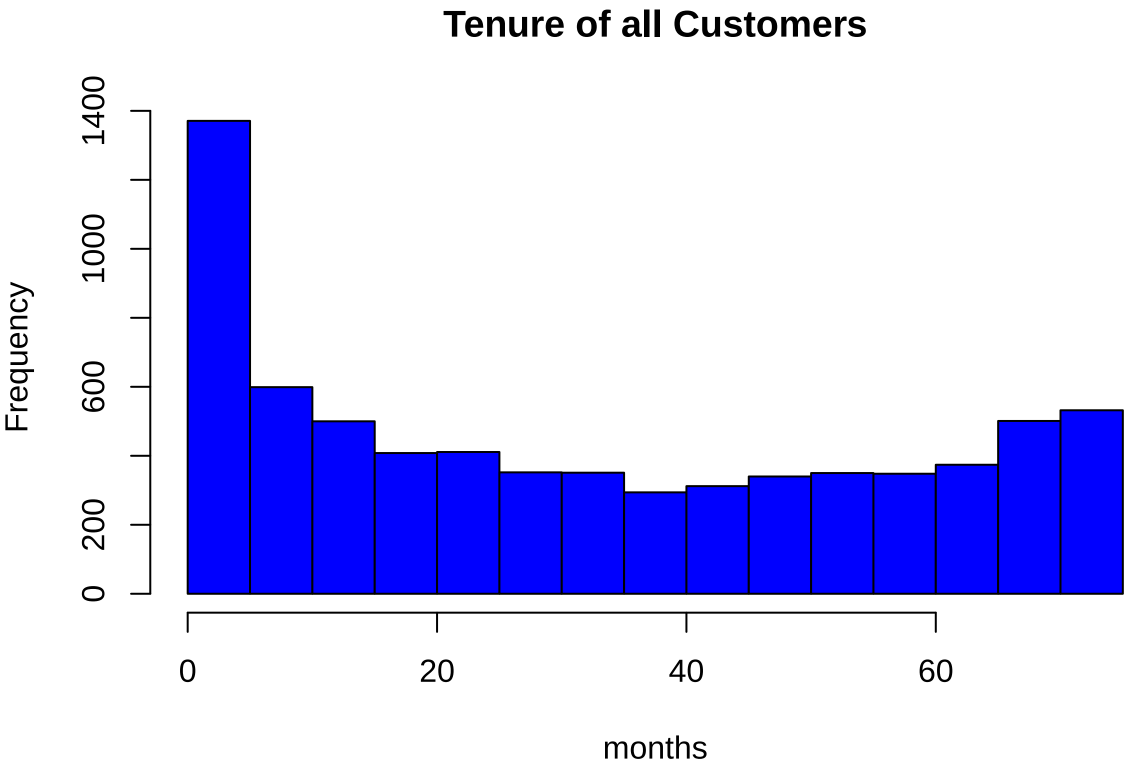
```
barplot(t(c4), beside = T, main = "Churn Rate by Internet Serivce", col = 2:3, xlab = "Internet Service
legend('top',fill = 2:3, c("Churn_No","Churn_Yes"))
```



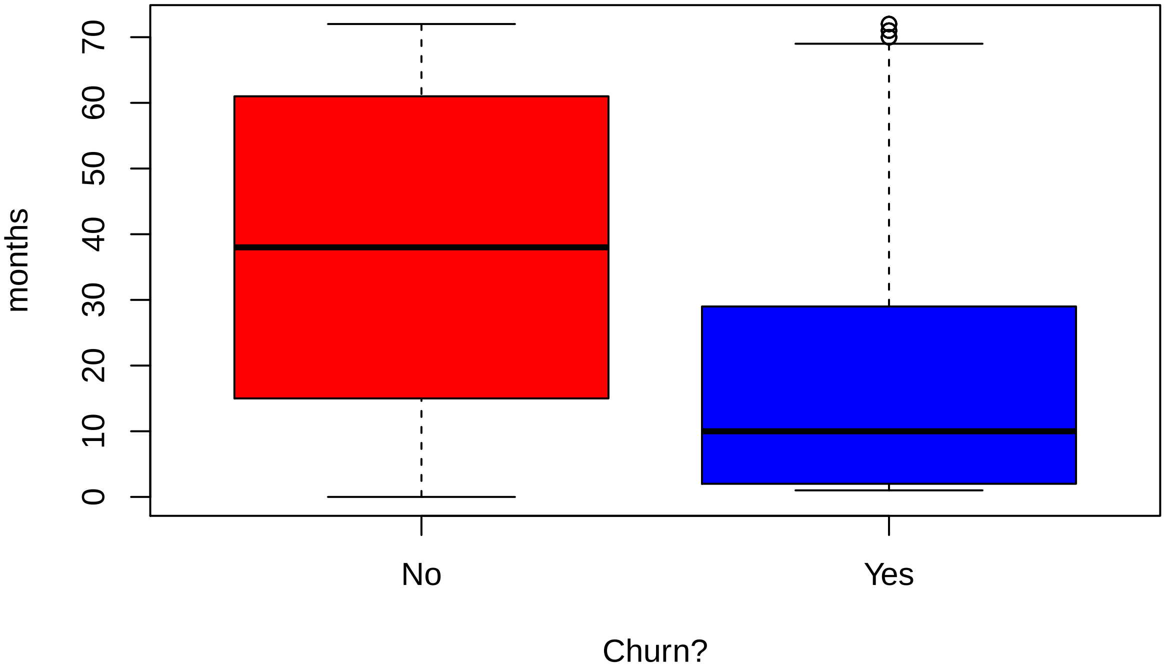
Now we see something suspicious. The fiber optic service appears to have a higher churn rate, 41.9% compared to the global 26.54% churn rate. Also, those with no internet service (phone service only), only have a 7.4% churn rate. DSL also has a lower churn rate than average at 19%.

Let's have a look at the tenure of the customers and its relation to churn rate.

```
hist(churn$tenure, main = "Tenure of all Customers", col = "blue", xlab = "months")
```



```
boxplot(churn$tenure~churn$Churn, col = c("red", "blue"), ylab = "months", xlab = "Churn?")
```



It does appear that those who left the company tended to have lower tenure values.

Testing and Training Data

In later steps I will be exploring some common features of customers and building models to predict the churn rate. First we will partition our data into 30% testing and 70% training data.

```r
set.seed(3.141592)
s = sample(1:nrow(churn),size = round(.7*nrow(churn)))
churn.train = churn[s,]
churn.test = churn[-s,]
```

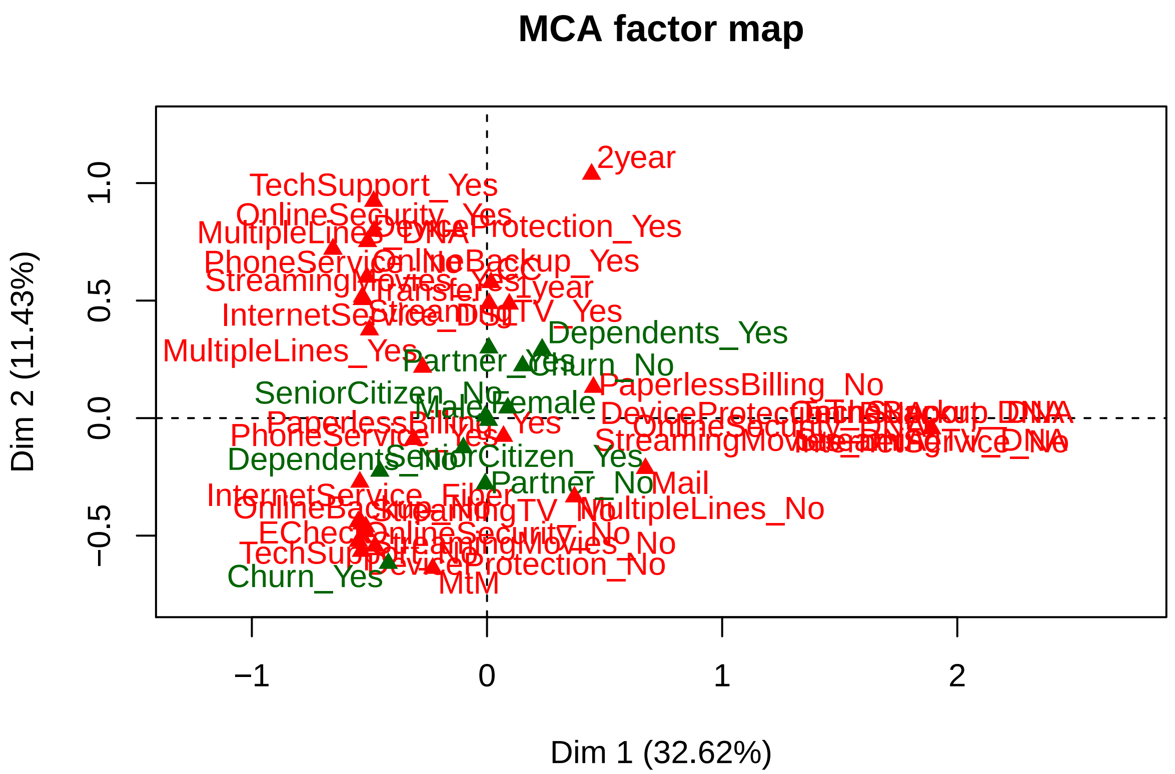
MCA

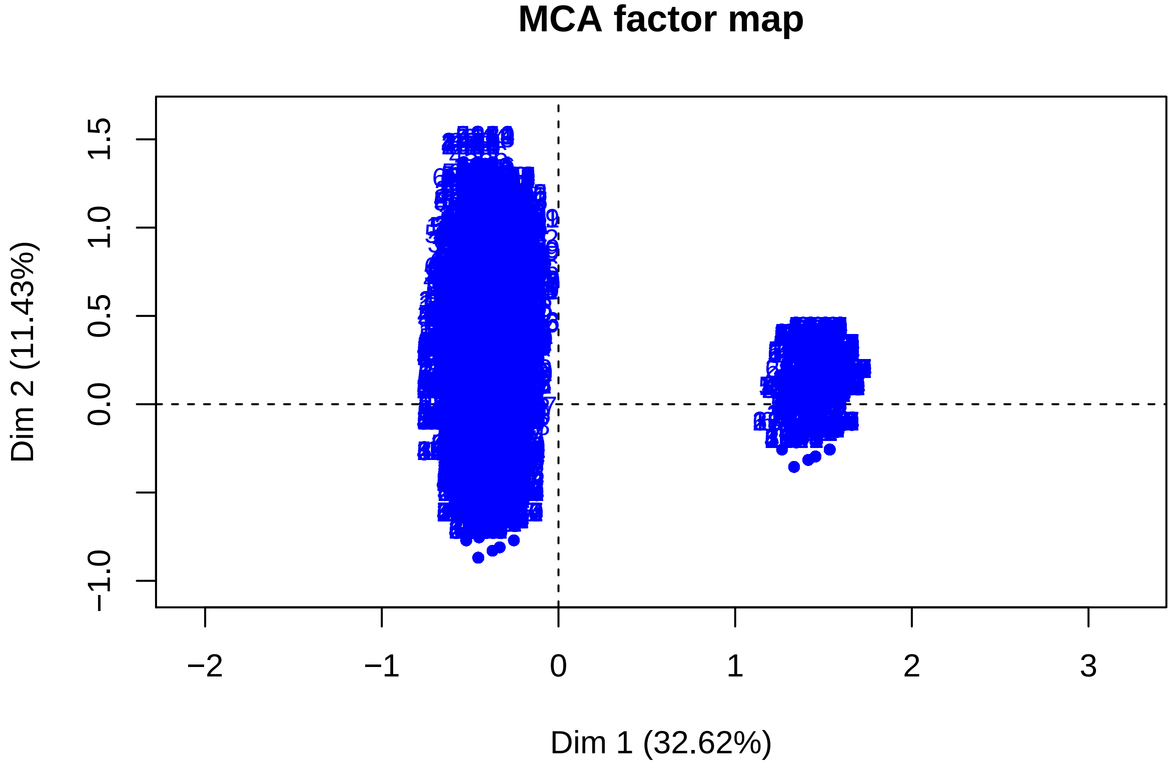
Here we use multiple correspondence analysis (MCA) to examine the relationship between the individuals and the variables. This technique uses linear algebra to reduce the number of dimensions of the data, which in this case is 20 dimensions. MCA can find clouds of individuals and clouds of variables using the notion of similarity. To individuals are similar if they share many characteristics, which in higher dimensional vector space would imply that the distance between the two points is relatively small. This makes the factor plots very useful and easy to interpret. On an MCA plot, if two individuals are close to one another, they have many features in common along the given two principle components. If two variables are close to one another, they share many of the same individuals. Additionally, the further a point is from the origin, the more leverage that individual or variable has on constructing that principle components. Conversely, points near the origin have little effect on the components, and we can interpret these points on having little effect on the construction of the principle components. So we look for clusters of both individuals and variables to interpret.

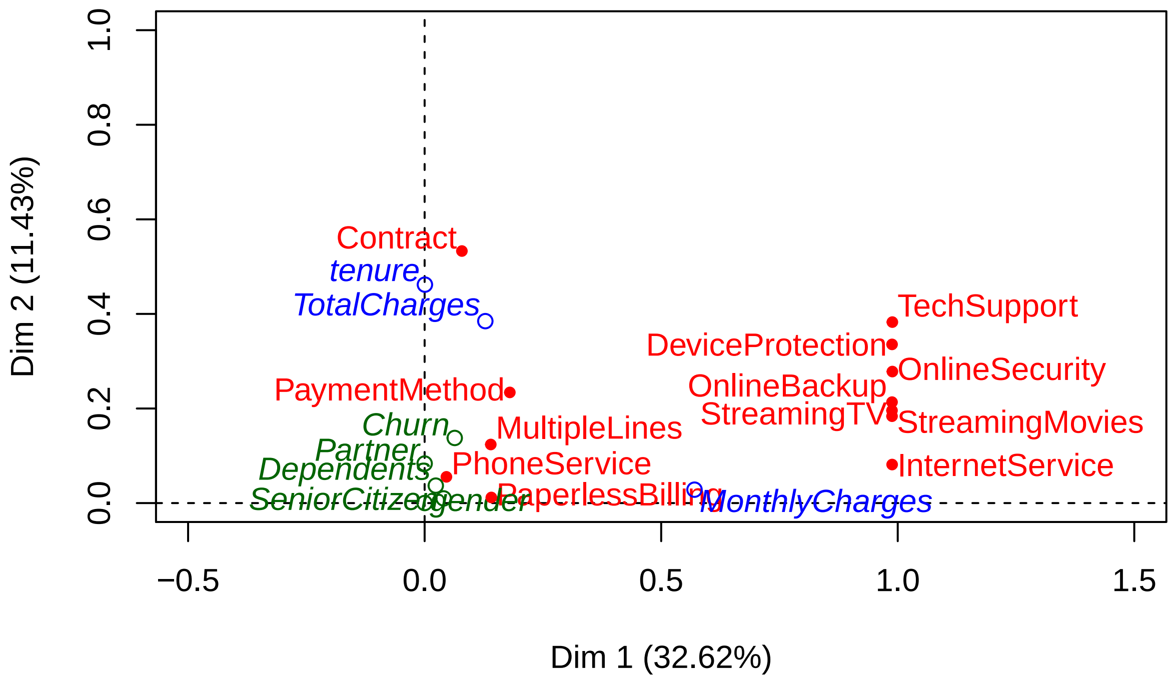
MCA may not be the most popular choice to examine the relationship between variables in this case, but I decided to experiment with it in this situation and I found the results consistent with my findings in decision trees (the results of which are not included in this report). MCA requires all variables to be categorical for the construction of the principle components; however quantitative variables can be represented as well. Variables can be set as active variables, which will be used to construct the principle components, or supplemental variables, which are not used in the construction of the axes. I have also set demographic data, such as Gender, Partner, and SeniorCitizen as qualitative supplemental variables. Other than using active categorical variables, there are no other assumptions necessary to perform MCA.

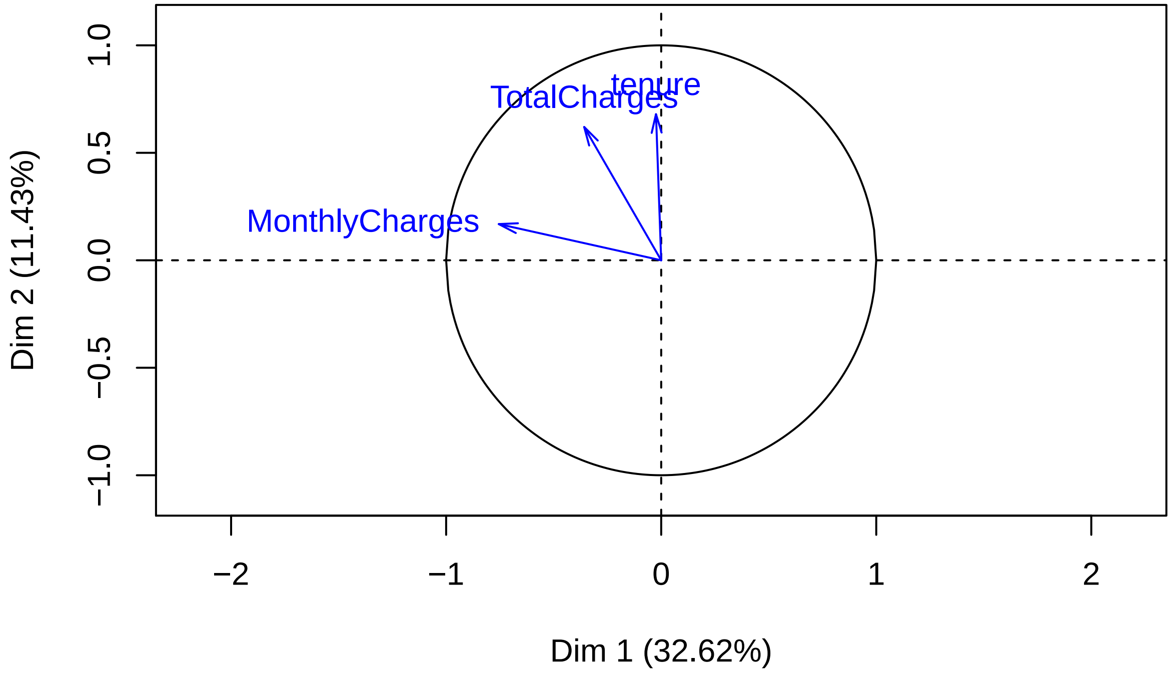
I will use the `FactomineR` and `factoextra` package and its `MCA` function.

```r
churn.mca = MCA(churn, quali.sup = c(1,2,3,4,20), quanti.sup = c(5,18,19), ncp=20)
```


MCA factor map
2year
TechSupport_Yes
OnlineSecurity_Yes
DeviceProtection_Yes
MultipleLines_DNA
PhoneService_No
OnlineBackup_Yes
StreamingMovies_Yes
CC
Transfer
1year
InternetService_DSL
StreamingTV_Yes
Dependents_Yes
MultipleLines_Yes
Partner_Yes
Churn_No
PaperlessBilling_No
SeniorCitizen_No
Male
Female
DeviceProtection_DNA
TechSupport_DNA
OnlineBackup_DNA
OnlineSecurity_DNA
StreamingMovies_DNA
StreamingTV_DNA
InternetService_No
PaperlessBilling_Yes
PhoneService_Yes
Dependents_No
SeniorCitizen_Yes
Mail
Partner_No
InternetService_Fiber
OnlineBackup_No
StreamingTV_No
MultipleLines_No
ECheck
OnlineSecurity_No
StreamingMovies_No
TechSupport_No
DeviceProtection_No
Churn_Yes
MtM
Dim 2 (11.43%)
1.0
0.5
0.0
−0.5
−1
0
1
2
Dim 1 (32.62%)


MCA factor map
Dim 2 (11.43%)
1.5
1.0
0.5
0.0
−1.0
−2
−1
0
1
2
3
Dim 1 (32.62%)




Supplementary variables on the MCA factor map

We should first analyze the eigenvalues of each of the principle components first before interpreting the results of these plots to ensure the dimensions are significant. However the FactomineR package automatically displays these four plots, so let's discuss them first.

The first three plots show the axes of the first two principle components, which together represents $32.71 + 11.28 = 43.99\%$ of the total inertia of all the principle components (the first component takes as high 19.1% of the inertia in this case if all dimensions are independent at a 95% confidence level).

The first plot above shows the position of each factor and level along the axes of the first two principle components (named Dim 1 and Dim 2 on the plots). The active variables are shown in red while the green variables represent the supplemental variables. Since this plot is so cluttered, we will zoom in and discuss it in more detail in the next chunk of code.

The second graph shows the position of individuals, which separated into at least two two distinct and well separated clouds. As we will see later, the cloud on the right side of the axis consists of the individuals with no internet service.

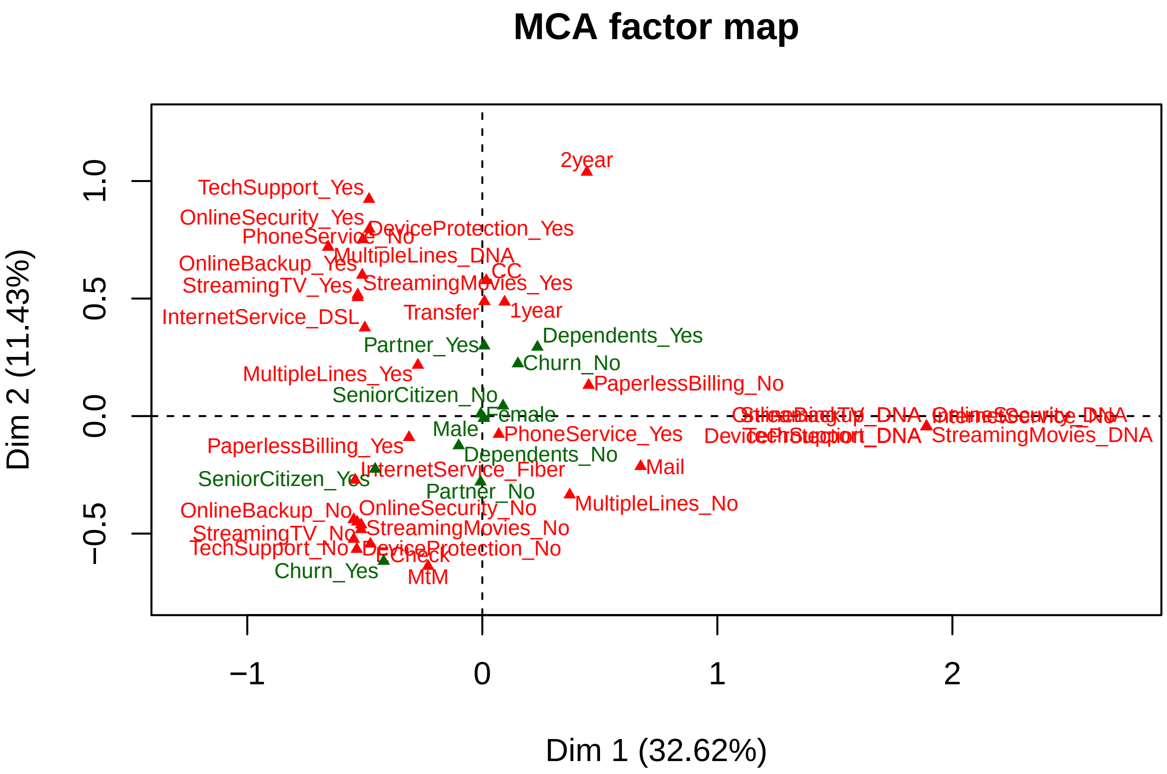
The third plot shows each of the variables (which in this map include all of the levels of each variable) and their position on the axes of the first two principle components. We see a cluster on the right consisting InternetService, StreamingMovies, etc. I will call the variables such as OnlineBackup, StreamingMovies, etc as additional features. The closeness of these variables indicates that they are strongly related. This should make sense since one's level of these additional features depends on whether they have Internet or not. More clearly, if a customer does not purchase InternetService they will also not have StreamingMovies, OnlineBackup, or any of these features. Because of the distance from the origin, it appears that the choice of InternetService may have the most influence on the model.

The fourth plot shows the representation of the supplemental, in this case quantitative, variables. The larger the radius, the better the representation of the variable is. All three variables are reasonably represented on the axes.

We can interpret the first principle component as predominantly characterized by internet usage and additional services for internet, such as TechSupport. The further along the horizontal axis, the more leverage the variable has along the axis of that principle component. Thus we see that the second principle component is characterized by tenure and its correlated values. The closer two variables are together in this plot, the stronger the collinearity. Additionally, the closer variables are to the origin, the weaker their leverage. So PaperlessBilling, SeniorCitizen, and PhoneService have little influence, at least in the first two components.

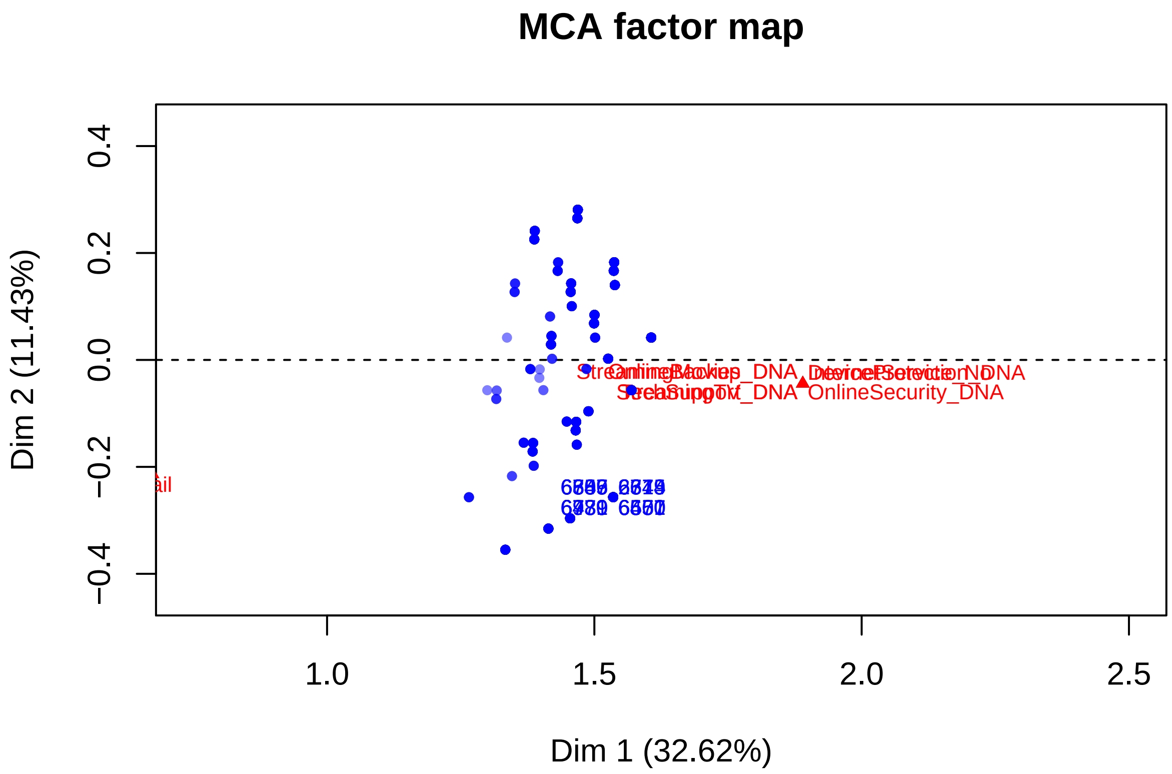
Let us know clean up the factor map. We see where our target variables lie on the plot as well.

```r
plot(churn.mca, cex = .65, invisible = "ind", autoLab = "y")
```



The Churn_Yes factor lies in the third quadrant at a healthy distance from the origin. The Churn_No factor lies in the first quadrant relatively close to the origin. Let us now zoom in on the cloud of variables and individuals. We first examine the large cloud of individuals on the right hand side.

```
plot(churn.mca, select = "cos2 16", cex = .7, autoLab = "yes", xlim=c(0.75,2.5), ylim = c(-.2,.2))
```



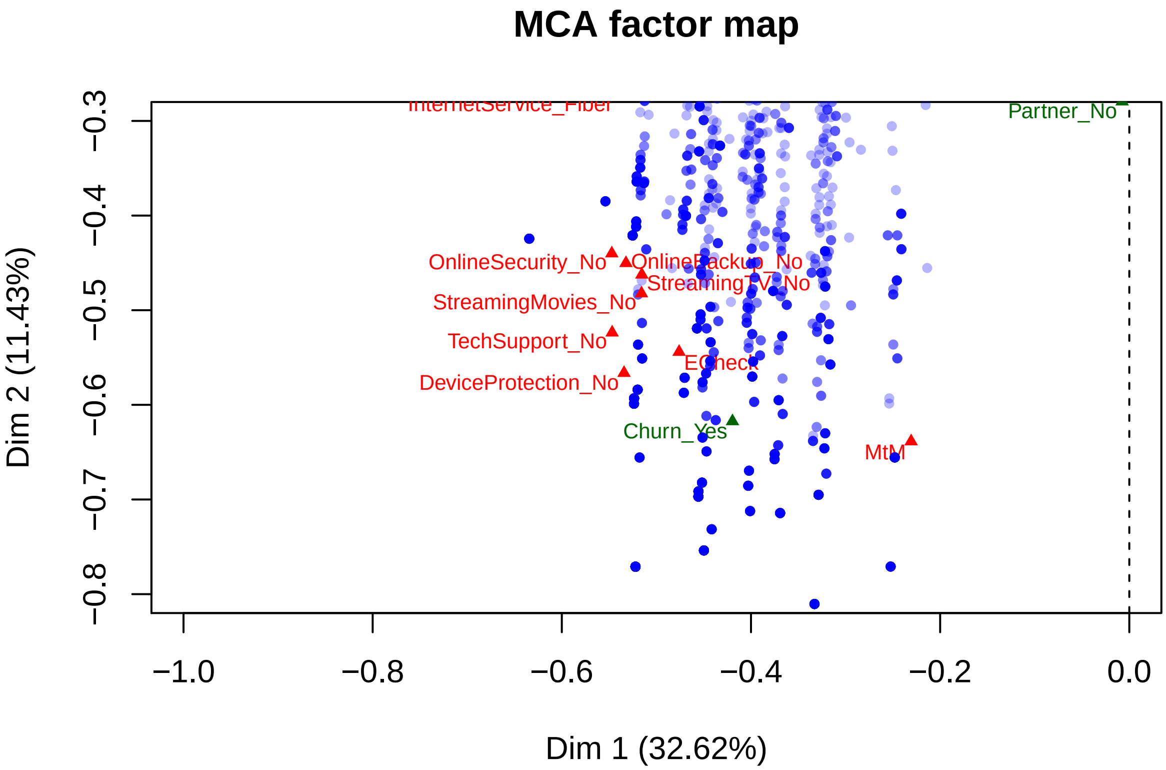
We have excluded some of the variables and individuals whose contributions are small. We first notice the clutter of red variables. The variables include InternetService_No, StreamingMovies_DNA (does not apply), and other DNA's for additional features. The extreme closeness suggests these variables are highly correlated, which makes sense because these additional features do not apply if a customer does not subscribe to internet service, as we can see in the table below.

```
table(churn$InternetService, churn$StreamingMovies)
```

```
##
##           DNA   No  Yes
##   DSL       0 1440  981
##   Fiber     0 1345 1751
##   No     1526    0    0
```

The plot below shows the third quadrant. Although it is difficult to discern, the Churn_Yes coordinate lies around (−0.3, −0.55).

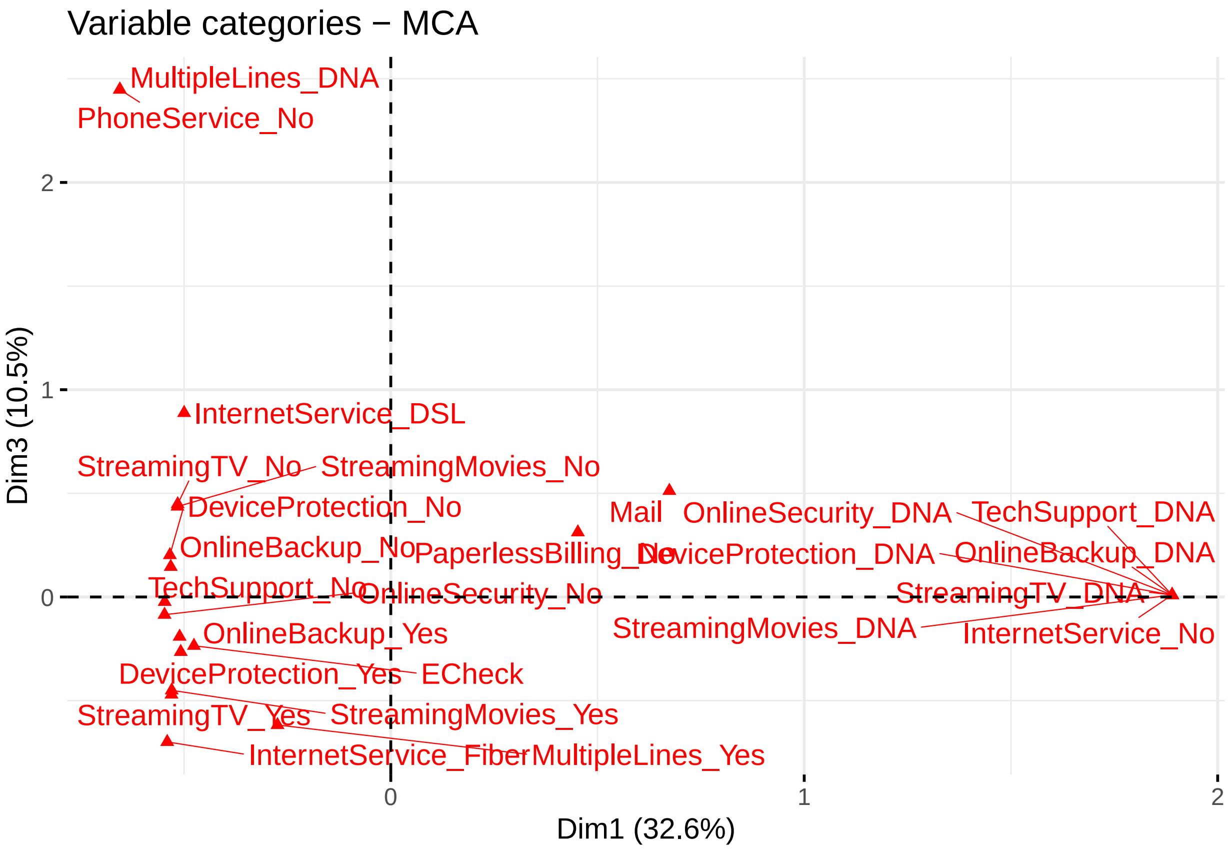
```
plot(churn.mca, select = "cos2 10", cex = .7, autoLab = "yes", xlim=c(-.6,-.4), ylim = c(-.8,-.3))
```



The above plot shows that churn is closely related to customers who are on month-to-month contracts, have no phone service with the company, pay using electronic checks, and opt for no additional features to their account such as device protection or streaming movies. The closeness of Fiber optic (top) suggests, as we have already seen, that the fiber optic is a factor in churn.

So far we have only looked at the first two components, which have the two highest proportions of variation. Below is a plot of the first and third principle components.

```
fviz_mca_var(churn.mca, axes = c(1,3), repel=TRUE, select.var = list(contrib = 25))
```



The third principle component (vertical axis) breaks up the types of services. Positive values indicate no phone service and using DSL.

But how many principle components should we use? We can create a scree plot which shows each principle component's eigenvalue or contribution to the total inertia.

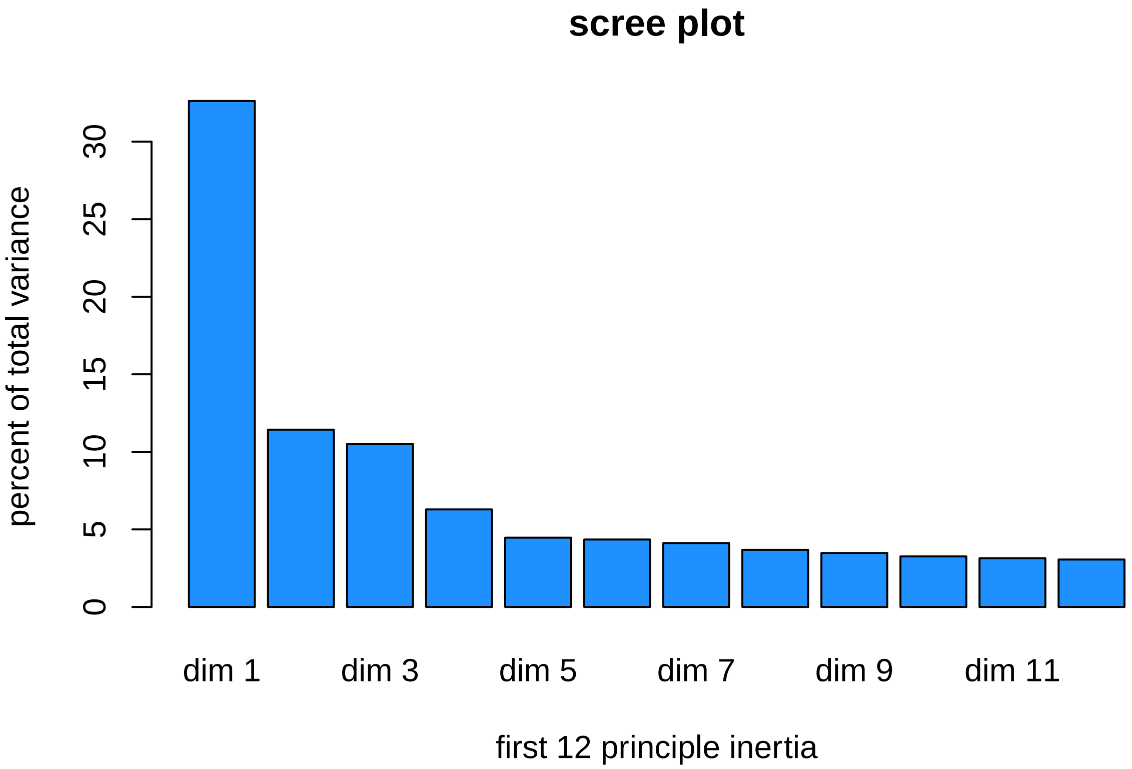
```
churn.mca$eig
```

```
##            eigenvalue percentage of variance cumulative percentage of variance
## dim 1   6.252601e-01           3.262227e+01                          32.62227
## dim 2   2.190901e-01           1.143079e+01                          44.05305
## dim 3   2.015365e-01           1.051495e+01                          54.56800
## dim 4   1.205085e-01           6.287401e+00                          60.85540
## dim 5   8.562765e-02           4.467530e+00                          65.32293
## dim 6   8.335440e-02           4.348925e+00                          69.67186
## dim 7   7.897697e-02           4.120538e+00                          73.79240
## dim 8   7.064390e-02           3.685769e+00                          77.47817
## dim 9   6.665695e-02           3.477754e+00                          80.95592
## dim 10  6.243879e-02           3.257676e+00                          84.21359
## dim 11  6.017378e-02           3.139502e+00                          87.35310
## dim 12  5.862512e-02           3.058702e+00                          90.41180
## dim 13  5.541925e-02           2.891439e+00                          93.30324
## dim 14  4.701236e-02           2.452819e+00                          95.75606
## dim 15  4.292744e-02           2.239693e+00                          97.99575
## dim 16  3.841480e-02           2.004251e+00                         100.00000
## dim 17  1.200936e-27           6.265751e-26                         100.00000
## dim 18  1.676835e-28           8.748707e-27                         100.00000
## dim 19  7.821521e-29           4.080794e-27                         100.00000
## dim 20  6.502758e-29           3.392743e-27                         100.00000
## dim 21 2.989199e-29          1.559582e-27                       100.00000
## dim 22 1.827147e-29          9.532940e-28                       100.00000
## dim 23 9.191394e-30          4.795510e-28                       100.00000
```

```
sum(churn.mca$eig[,2][1:20])
```

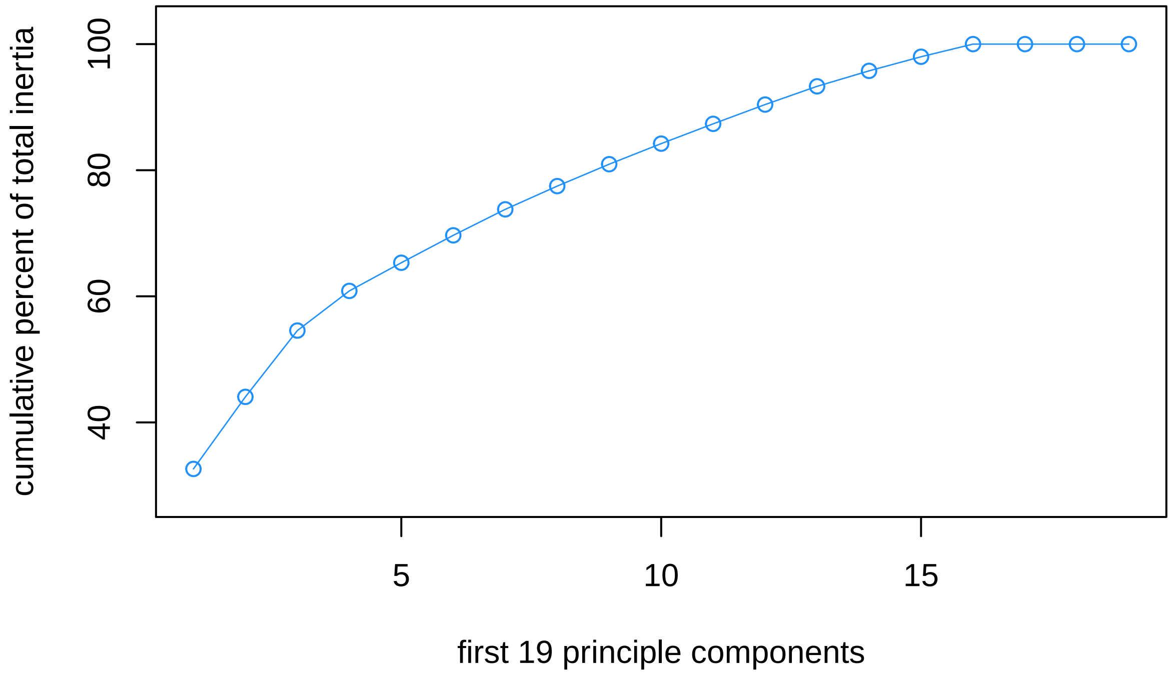
```
## [1] 100
```

```
barplot(churn.mca$eig[,2][1:12], main = "scree plot",
              ylab = "percent of total variance",
              xlab = "first 12 principle inertia", col = "dodgerblue")
```



We are looking for a big drop before the chart tapers off. This happens between the third and fourth components. We can also see this below using the cumulative percent of variation. Here we are looking for a bend in the plot where it begins to straighten out, which again occurs between the third and fourth component. We also note that the first three components account for nearly 47% of the variation.

```
plot(churn.mca$eig[,3][1:19], ylab = "cumulative percent of total inertia",
            xlab = "first 19 principle components", col = "dodgerblue", ylim =c(28,103))
lines(1:19,churn.mca$eig[,3][1:19], lwd = .7, col = "dodger blue")
```



One should concentrate on these first three principle components, where each component is a linear combination of all the variables. We will later be using hierarchical clustering on the first three components to create groups of individuals. We can see the components of each of the three components in descending importance.

```
dimdesc(churn.mca)
```

```
## $`Dim 1`
## $`Dim 1`$quanti
##                 correlation       p.value
## tenure          -0.02420463  4.222947e-02
## TotalCharges    -0.35813691 3.567750e-212
## MonthlyCharges  -0.75526530  0.000000e+00
##
## $`Dim 1`$quali
##                          R2       p.value
## InternetService  0.98820205  0.000000e+00
## OnlineSecurity   0.98869123  0.000000e+00
## OnlineBackup     0.98796931  0.000000e+00
## DeviceProtection 0.98801266  0.000000e+00
## TechSupport      0.98865227  0.000000e+00
## StreamingTV      0.98792375  0.000000e+00
## StreamingMovies  0.98791883  0.000000e+00
## PaymentMethod    0.18009531 8.789267e-303
## PaperlessBilling 0.14103072 9.360977e-235
## MultipleLines    0.13983001 5.453684e-231
## Contract         0.07876594 3.822578e-126
## Churn            0.06358943 1.326311e-102
## PhoneService     0.04602900  3.970995e-74
## SeniorCitizen    0.04019499  8.909699e-65
## Dependents       0.02353861  2.346357e-38
##
## $`Dim 1`$category
##                                          Estimate       p.value
## StreamingMovies=StreamingMovies_DNA    1.271839495  0.000000e+00
## StreamingTV=StreamingTV_DNA            1.271877343  0.000000e+00
## TechSupport=TechSupport_DNA            1.267365603  0.000000e+00
## DeviceProtection=DeviceProtection_DNA  1.270961271  0.000000e+00
## OnlineBackup=OnlineBackup_DNA          1.271127298  0.000000e+00
## OnlineSecurity=OnlineSecurity_DNA      1.267083739  0.000000e+00
## InternetService=InternetService_No     1.270486560  0.000000e+00
## PaperlessBilling=PaperlessBilling_No   0.302136089 9.360977e-235
## PaymentMethod=Mail                     0.488591710 1.246457e-223
## MultipleLines=MultipleLines_No         0.440994581 2.175287e-212
## Churn=Churn_No                         0.225804531 1.326311e-102
## Contract=2year                         0.270201479 3.907895e-101
## PhoneService=PhoneService_Yes          0.286826273  3.970995e-74
## SeniorCitizen=SeniorCitizen_No         0.215054300  8.909699e-65
## Dependents=Dependents_Yes              0.132419408  2.346357e-38
## Contract=1year                        -0.006331436  4.078737e-05
## Dependents=Dependents_No              -0.132419408  2.346357e-38
## SeniorCitizen=SeniorCitizen_Yes       -0.215054300  8.909699e-65
## MultipleLines=MultipleLines_DNA       -0.371220222  3.970995e-74
## PhoneService=PhoneService_No          -0.286826273  3.970995e-74
## MultipleLines=MultipleLines_Yes       -0.069774359  3.015563e-88
## Churn=Churn_Yes                       -0.225804531 1.326311e-102
## Contract=MtM                          -0.263870043 4.801035e-105
## OnlineSecurity=OnlineSecurity_Yes     -0.607130928 4.954057e-151
## TechSupport=TechSupport_Yes           -0.607980059 1.025326e-154
## PaymentMethod=ECheck                  -0.420643849 2.559509e-188
## InternetService=InternetService_DSL   -0.619092554 1.011371e-216
## DeviceProtection=DeviceProtection_Yes -0.625119167 1.559293e-224
## OnlineBackup=OnlineBackup_Yes         -0.627066968 3.939194e-228
## PaperlessBilling=PaperlessBilling_Yes -0.302136089 9.360977e-235
## StreamingMovies=StreamingMovies_No    -0.630363640 1.249890e-294
## StreamingTV=StreamingTV_Yes           -0.641838558 1.389218e-297
## StreamingTV=StreamingTV_No            -0.630038785 5.463214e-299
## StreamingMovies=StreamingMovies_Yes   -0.641475855 7.109674e-302
## TechSupport=TechSupport_No            -0.659385544  0.000000e+00
## DeviceProtection=DeviceProtection_No  -0.645842104  0.000000e+00
## OnlineBackup=OnlineBackup_No          -0.644060330  0.000000e+00
## OnlineSecurity=OnlineSecurity_No      -0.659952811  0.000000e+00
## InternetService=InternetService_Fiber -0.651394005  0.000000e+00
##
##
## $`Dim 2`
## $`Dim 2`$quanti
##                correlation      p.value
## tenure           0.6796492 0.000000e+00
## TotalCharges     0.6202696 0.000000e+00
## MonthlyCharges   0.1684138 5.718674e-46
##
## $`Dim 2`$quali
##                           R2       p.value
## OnlineSecurity   0.278079339  0.000000e+00
## OnlineBackup     0.213503739  0.000000e+00
## DeviceProtection 0.335573186  0.000000e+00
## TechSupport      0.382690740  0.000000e+00
## StreamingMovies  0.195794303  0.000000e+00
## Contract         0.533001821  0.000000e+00
## PaymentMethod    0.234016616  0.000000e+00
## StreamingTV      0.183721882 4.680632e-311
## Churn            0.137681022 8.458285e-229
## MultipleLines    0.123851097 7.486593e-203
## Partner          0.083590511 1.129048e-135
## InternetService  0.081488904 1.140546e-130
## PhoneService     0.055463408  2.305697e-89
## Dependents       0.036914105  1.533303e-59
## PaperlessBilling 0.011895939  4.347897e-20
## SeniorCitizen    0.009842565  7.118364e-17
##
## $`Dim 2`$category
##                                         Estimate       p.value
## Contract=2year                       0.347780368  0.000000e+00
## TechSupport=TechSupport_Yes          0.376493340  0.000000e+00
## DeviceProtection=DeviceProtection_Yes  0.329744380  0.000000e+00
## OnlineBackup=OnlineBackup_Yes        0.264391414 9.881313e-324
## OnlineSecurity=OnlineSecurity_Yes    0.323827932  0.000000e+00
## StreamingMovies=StreamingMovies_Yes  0.243235305 2.937236e-285
## StreamingTV=StreamingTV_Yes          0.236584842 2.235065e-267
## Churn=Churn_No                       0.196679107 8.458285e-229
## PaymentMethod=CC                     0.234228300 1.688450e-149
## Partner=Partner_Yes                  0.135406667 1.129048e-135
## InternetService=InternetService_DSL  0.166577464 1.215535e-119
## PaymentMethod=Transfer               0.192396624 7.120957e-108
## Contract=1year                       0.089213196 1.382802e-100
## MultipleLines=MultipleLines_DNA      0.242831508  2.305697e-89
## PhoneService=PhoneService_No         0.186374907  2.305697e-89
## Dependents=Dependents_Yes            0.098160879  1.533303e-59
## MultipleLines=MultipleLines_Yes      0.007664762  2.210796e-55
## PaperlessBilling=PaperlessBilling_No  0.051942888  4.347897e-20
## SeniorCitizen=SeniorCitizen_No       0.062993731  7.118364e-17
## SeniorCitizen=SeniorCitizen_Yes     -0.062993731  7.118364e-17
## PaperlessBilling=PaperlessBilling_Yes -0.051942888  4.347897e-20
## PaymentMethod=Mail                  -0.136041547  1.001188e-22
## Dependents=Dependents_No            -0.098160879  1.533303e-59
## PhoneService=PhoneService_Yes       -0.186374907  2.305697e-89
## InternetService=InternetService_Fiber -0.136640028  1.376554e-93
## Partner=Partner_No                  -0.135406667 1.129048e-135
## MultipleLines=MultipleLines_No      -0.250496270 4.151392e-170
## Churn=Churn_Yes                     -0.196679107 8.458285e-229
## StreamingTV=StreamingTV_No          -0.216195796 2.074882e-236
## PaymentMethod=ECheck                -0.290583376 3.409355e-250
## StreamingMovies=StreamingMovies_No  -0.224167271 1.140770e-254
## OnlineBackup=OnlineBackup_No        -0.227244953 8.586000e-266
## Contract=MtM                        -0.436993564  0.000000e+00
## TechSupport=TechSupport_No          -0.300473281  0.000000e+00
## DeviceProtection=DeviceProtection_No  -0.287091185  0.000000e+00
## OnlineSecurity=OnlineSecurity_No    -0.254570821  0.000000e+00
##
##
## $`Dim 3`
## $`Dim 3`$quanti
##                  correlation       p.value
## tenure          -0.2247324  2.448297e-81
## TotalCharges    -0.4347851 9.881313e-323
## MonthlyCharges  -0.5937049  0.000000e+00
##
## $`Dim 3`$quali
##                           R2       p.value
## PhoneService     0.642582743  0.000000e+00
## MultipleLines    0.742341322  0.000000e+00
## InternetService  0.486095412  0.000000e+00
## StreamingTV      0.165207525 9.045041e-277
## StreamingMovies  0.153983965 2.358596e-256
## PaymentMethod    0.082342877 8.670600e-131
## PaperlessBilling 0.067154063 1.903520e-108
## DeviceProtection 0.042120913  1.635439e-66
## OnlineBackup     0.021947291  1.188932e-34
## SeniorCitizen    0.018019019  1.111283e-29
## Partner          0.011095948  7.769640e-19
## OnlineSecurity   0.009329914  4.680090e-15
## Churn            0.007290749  7.092372e-13
## Contract         0.004743125  5.393315e-08
## Dependents       0.001528369  1.032304e-03
##
## $`Dim 3`$category
##                                           Estimate       p.value
## InternetService=InternetService_DSL    0.36897457  0.000000e+00
## MultipleLines=MultipleLines_DNA        0.81781291  0.000000e+00
## PhoneService=PhoneService_No           0.60843508  0.000000e+00
## StreamingTV=StreamingTV_No             0.20303621 6.181790e-221
## StreamingMovies=StreamingMovies_No     0.19652840 1.106821e-205
## PaymentMethod=Mail                     0.21852185 3.872356e-126
## PaperlessBilling=PaperlessBilling_No   0.11836640 1.903520e-108
## DeviceProtection=DeviceProtection_No   0.09872039  6.822779e-52
## SeniorCitizen=SeniorCitizen_No         0.08174747  1.111283e-29
## OnlineBackup=OnlineBackup_No           0.07076155  2.109973e-27
## Partner=Partner_No                     0.04731614  7.769640e-19
## OnlineSecurity=OnlineSecurity_Yes      0.05345435  6.314637e-14
## Churn=Churn_No                         0.04340830  7.092372e-13
## Contract=MtM                           0.03787373  6.800399e-08
## Dependents=Dependents_Yes              0.01915675  1.032304e-03
## PaymentMethod=CC                      -0.03797964  9.121504e-03
## Dependents=Dependents_No              -0.01915675  1.032304e-03
## MultipleLines=MultipleLines_No        -0.25938194  8.328787e-05
## Contract=2year                        -0.03558409  6.043283e-07
## PaymentMethod=Transfer                -0.06349390  2.434811e-07
## OnlineSecurity=OnlineSecurity_No      -0.04811254  9.385436e-13
## Churn=Churn_Yes                       -0.04340830  7.092372e-13
## Partner=Partner_Yes                   -0.04731614  7.769640e-19
## SeniorCitizen=SeniorCitizen_Yes       -0.08174747  1.111283e-29
## OnlineBackup=OnlineBackup_Yes         -0.08051922  1.210763e-31
## PaymentMethod=ECheck                  -0.11704831  2.220072e-45
## DeviceProtection=DeviceProtection_Yes -0.11098156  2.921955e-59
## PaperlessBilling=PaperlessBilling_Yes -0.11836640 1.903520e-108
## StreamingMovies=StreamingMovies_Yes   -0.20153732 1.315358e-212
## StreamingTV=StreamingTV_Yes           -0.20933671 8.577767e-230
## InternetService=InternetService_Fiber -0.34364597  0.000000e+00
## MultipleLines=MultipleLines_Yes       -0.55843097  0.000000e+00
## PhoneService=PhoneService_Yes         -0.60843508  0.000000e+00
```

If we look at the $Dim _$category for each of the three principle components, we see the factors responsible for each of the dimensions in descending order in terms of the p-value. Customers with high scores on the first principle component keep monthly charges low by not using additional services, and/or having phone only services. Additionally, they are more likely to not to use paperless billing and pay via mailed checks. Customers with positive values on the second principle components tend to pay more and subscribe to more additional services. Customers with positive values on the third principle components tend to use DSL and have moderate monthly bills. This supports the aforementioned graphical interpretations.

We can conclude from the MCA that there is different customer profiles, mostly separated by cost.

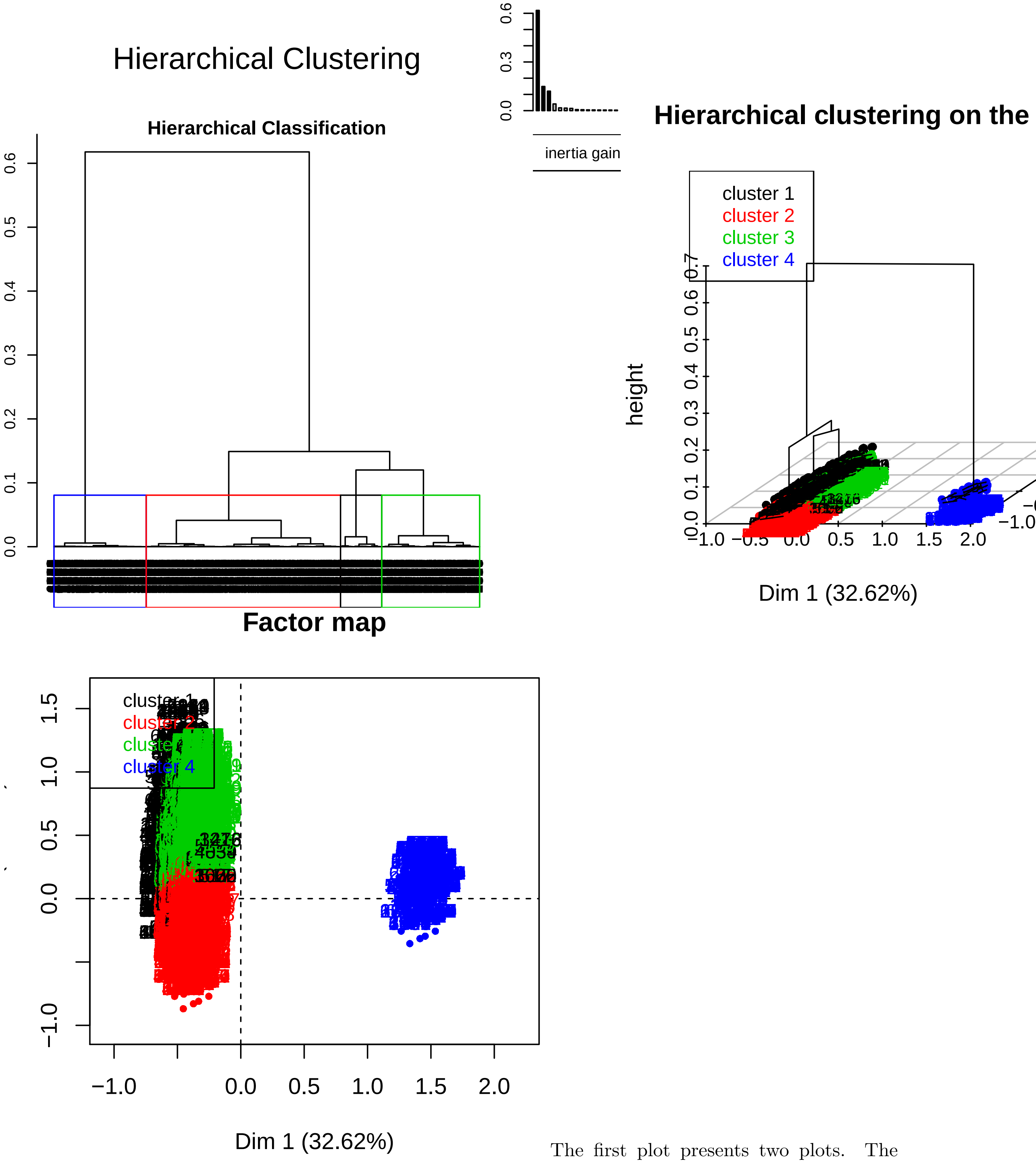
Hierarchical Clustering

Next we want to use the results from the MCA to create group our customers with similar characteristics. Then we would like to identify the clusters with the highest churn rates. For this I will employ the HCPC (hierarchical clustering of principle components). While most clustering require quantitative variables, HCPC uses the quantitative coordinates of the principle components from the MCA. The goal will be to create clusters where customers within each cluster are as similar as possible and each cluster is as different as possible from the other clusters.

I prefer this visual approach over, say, a decision tree. In a tree, the reader only see rules, often with important variables not even included in the nodes. Thus the reader cannot examine between the variables. In MCA plots, we will see the relationship of each variable, between each level of each variable, and between the individuals all on the same plot.

I have reduced the number of principle components to 3, as previously justified, and after some analysis I settled on choosing 4 cluster (that is, 4 different customer profiles). All variables, whether they active or supplemental, or qualitative or quantitative, as used in the construction of the clusters.

```r
churn.mca2 = MCA(churn, quali.sup = c(1,2,3,4,20),
            quanti.sup = c(5,18,19), ncp = 3, graph = FALSE)
churn.hcpc = HCPC(churn.mca2, nb.clust = 4)
```



The first plot presents two plots. The dendogram has each cluster of variables with their levels located in each colored box. Unfortunately the names are illegible. The bar graph on the top right shows the loss of inertia between the number of clusters. The first three bars are highlighted, and we see a big drop after the fourth bar followed by a slow decrease.

This suggests that 4 clusters is a reasonable choice, as partitioning into more clusters will decrease the inertia between clusters, hence the clusters themselves will be more similar to one another. The second plot show the results of the clusters of individuals on the MCA plot on the x-y-plane, and the tree dendogram stopping at the location of the variables on the MCA plane.

Here are the numeric results of the clustering.

```
churn.hcpc$desc.var
```

```
##
## Link between the cluster variable and the categorical variables (chi-square test)
## =================================================================================
##                        p.value df
## PhoneService      0.000000e+00  3
## MultipleLines     0.000000e+00  6
## InternetService   0.000000e+00  6
## OnlineSecurity    0.000000e+00  6
## OnlineBackup      0.000000e+00  6
## DeviceProtection  0.000000e+00  6
## TechSupport       0.000000e+00  6
## StreamingTV       0.000000e+00  6
## StreamingMovies   0.000000e+00  6
## Contract          0.000000e+00  6
## PaymentMethod     0.000000e+00  9
## Churn            1.955915e-206  3
## PaperlessBilling 2.187035e-169  3
## Partner           6.545736e-85  3
## SeniorCitizen     5.438407e-59  3
## Dependents        1.169433e-56  3
##
## Description of each cluster by the categories
## =============================================
## $`1`
##                                            Cla/Mod    Mod/Cla    Global
## InternetService=InternetService_DSL     28.7484511 100.000000 34.374556
## MultipleLines=MultipleLines_DNA        100.0000000  97.988506  9.683374
## PhoneService=PhoneService_No           100.0000000  97.988506  9.683374
## StreamingTV=StreamingTV_No              14.6263345  59.051724 39.897771
## StreamingMovies=StreamingMovies_No      14.2908438  57.183908 39.542808
## TechSupport=TechSupport_Yes             14.8238748  43.534483 29.021724
## OnlineSecurity=OnlineSecurity_Yes       14.7102526  42.672414 28.666761
## OnlineBackup=OnlineBackup_No            13.1152850  58.189655 43.844952
## DeviceProtection=DeviceProtection_No    12.6009693  56.034483 43.944342
## DeviceProtection=DeviceProtection_Yes   12.6341866  43.965517 34.388755
## OnlineSecurity=OnlineSecurity_No        11.4065180  57.327586 49.666335
## OnlineBackup=OnlineBackup_Yes           11.9802388  41.810345 34.488144
## TechSupport=TechSupport_No              11.3158652  56.465517 49.311373
## StreamingMovies=StreamingMovies_Yes     10.9077599  42.816092 38.790288
## PaperlessBilling=PaperlessBilling_No    10.7590529  44.396552 40.778078
## PaperlessBilling=PaperlessBilling_Yes    9.2783505  55.603448 59.221922
## StreamingMovies=StreamingMovies_DNA      0.0000000   0.000000 21.666903
## StreamingTV=StreamingTV_DNA              0.0000000   0.000000 21.666903
## TechSupport=TechSupport_DNA              0.0000000   0.000000 21.666903
## DeviceProtection=DeviceProtection_DNA    0.0000000   0.000000 21.666903
## OnlineBackup=OnlineBackup_DNA            0.0000000   0.000000 21.666903
## OnlineSecurity=OnlineSecurity_DNA       0.0000000   0.000000 21.666903
## InternetService=InternetService_No      0.0000000   0.000000 21.666903
## MultipleLines=MultipleLines_Yes         0.0000000   0.000000 42.183729
## MultipleLines=MultipleLines_No          0.4129794   2.011494 48.132898
## InternetService=InternetService_Fiber   0.0000000   0.000000 43.958540
## PhoneService=PhoneService_Yes           0.2200912   2.011494 90.316626
##                                             p.value     v.test
## InternetService=InternetService_DSL    0.000000e+00        Inf
## MultipleLines=MultipleLines_DNA        0.000000e+00        Inf
## PhoneService=PhoneService_No           0.000000e+00        Inf
## StreamingTV=StreamingTV_No             7.245782e-27  10.731440
## StreamingMovies=StreamingMovies_No     4.470764e-23   9.892858
## TechSupport=TechSupport_Yes            8.734744e-18   8.589502
## OnlineSecurity=OnlineSecurity_Yes      8.779493e-17   8.320227
## OnlineBackup=OnlineBackup_No           1.309950e-15   7.993656
## DeviceProtection=DeviceProtection_No   1.630419e-11   6.735790
## DeviceProtection=DeviceProtection_Yes  3.542628e-08   5.512256
## OnlineSecurity=OnlineSecurity_No       2.036596e-05   4.260840
## OnlineBackup=OnlineBackup_Yes          2.376665e-05   4.226204
## TechSupport=TechSupport_No             6.974656e-05   3.977149
## StreamingMovies=StreamingMovies_Yes    2.232606e-02   2.284774
## PaperlessBilling=PaperlessBilling_No   4.145402e-02   2.038960
## PaperlessBilling=PaperlessBilling_Yes  4.145402e-02  -2.038960
## StreamingMovies=StreamingMovies_DNA    5.282008e-79 -18.818951
## StreamingTV=StreamingTV_DNA            5.282008e-79 -18.818951
## TechSupport=TechSupport_DNA            5.282008e-79 -18.818951
## DeviceProtection=DeviceProtection_DNA  5.282008e-79 -18.818951
## OnlineBackup=OnlineBackup_DNA          5.282008e-79 -18.818951
## OnlineSecurity=OnlineSecurity_DNA      5.282008e-79 -18.818951
## InternetService=InternetService_No     5.282008e-79 -18.818951
## MultipleLines=MultipleLines_Yes       2.591627e-178 -28.472159
## MultipleLines=MultipleLines_No        6.893940e-185 -28.998392
## InternetService=InternetService_Fiber 1.157627e-188 -29.296253
## PhoneService=PhoneService_Yes          0.000000e+00       -Inf
##
## $`2`
##                                         Cla/Mod    Mod/Cla    Global
## Contract=MtM                          64.206452  91.943829 55.019168
## StreamingMovies=StreamingMovies_No    67.109515  69.068736 39.542808
## StreamingTV=StreamingTV_No            66.334520  68.883962 39.897771
## TechSupport=TechSupport_No            66.916211  85.883222 49.311373
## DeviceProtection=DeviceProtection_No  69.208401  79.157428 43.944342
## OnlineBackup=OnlineBackup_No          65.252591  74.464154 43.844952
## OnlineSecurity=OnlineSecurity_No      63.236135  81.744272 49.666335
## InternetService=InternetService_Fiber 60.077519  68.736142 43.958540
## PaymentMethod=ECheck                  64.270613  56.171471 33.579441
## Churn=Churn_Yes                       66.987694  46.267554 26.536987
## PhoneService=PhoneService_Yes         42.540481 100.000000 90.316626
## PaperlessBilling=PaperlessBilling_Yes 46.223927  71.249076 59.221922
## Partner=Partner_No                    47.541884  63.968958 51.696720
## Dependents=Dependents_No              44.070545  80.339985 70.041176
## MultipleLines=MultipleLines_No        46.106195  57.760532 48.132898
## SeniorCitizen=SeniorCitizen_Yes       54.028021  22.801183 16.214681
## InternetService=InternetService_DSL   34.944238  31.263858 34.374556
## PaymentMethod=Mail                    29.962779  17.849224 22.887974
## StreamingTV=StreamingTV_Yes           31.104544  31.116038 38.435326
## StreamingMovies=StreamingMovies_Yes   30.636896  30.931264 38.790288
## SeniorCitizen=SeniorCitizen_No        35.400780  77.198817 83.785319
## OnlineBackup=OnlineBackup_Yes         28.447921  25.535846 34.488144
## PaymentMethod=Transfer                23.575130  13.451589 21.922476
## PaymentMethod=CC                      22.273325  12.527716 21.610109
## Dependents=Dependents_Yes             25.213270  19.660015 29.958824
## OnlineSecurity=OnlineSecurity_Yes     24.467558  18.255728 28.666761
## Partner=Partner_Yes                   28.659612  36.031042 48.303280
## PaperlessBilling=PaperlessBilling_No  27.089136  28.750924 40.778078
## DeviceProtection=DeviceProtection_Yes 23.286540  20.842572 34.388755
## TechSupport=TechSupport_Yes           18.688845  14.116778 29.021724
## Contract=1year                        13.170401   7.169254 20.914383
## MultipleLines=MultipleLines_DNA        0.000000   0.000000  9.683374
## PhoneService=PhoneService_No           0.000000   0.000000  9.683374
## Churn=Churn_No                        28.102049  53.732446 73.463013
## Contract=2year                         1.415929   0.886918 24.066449
## StreamingMovies=StreamingMovies_DNA    0.000000   0.000000 21.666903
## StreamingTV=StreamingTV_DNA            0.000000   0.000000 21.666903
## TechSupport=TechSupport_DNA            0.000000   0.000000 21.666903
## DeviceProtection=DeviceProtection_DNA  0.000000   0.000000 21.666903
## OnlineBackup=OnlineBackup_DNA          0.000000   0.000000 21.666903
## OnlineSecurity=OnlineSecurity_DNA      0.000000   0.000000 21.666903
## InternetService=InternetService_No     0.000000   0.000000 21.666903
##                                             p.value     v.test
## Contract=MtM                           0.000000e+00        Inf
## StreamingMovies=StreamingMovies_No     0.000000e+00        Inf
## StreamingTV=StreamingTV_No             0.000000e+00        Inf
## TechSupport=TechSupport_No             0.000000e+00        Inf
## DeviceProtection=DeviceProtection_No   0.000000e+00        Inf
## OnlineBackup=OnlineBackup_No           0.000000e+00        Inf
## OnlineSecurity=OnlineSecurity_No       0.000000e+00        Inf
## InternetService=InternetService_Fiber 2.460685e-244  33.382499
## PaymentMethod=ECheck                  8.415755e-220  31.646712
## Churn=Churn_Yes                       3.596791e-190  29.414374
## PhoneService=PhoneService_Yes         4.166976e-154  26.444986
## PaperlessBilling=PaperlessBilling_Yes  2.308872e-60  16.388550
## Partner=Partner_No                     3.896963e-60  16.356697
## Dependents=Dependents_No               4.968052e-52  15.177713
## MultipleLines=MultipleLines_No         1.908333e-37  12.788204
## SeniorCitizen=SeniorCitizen_Yes        1.443289e-31  11.689444
## InternetService=InternetService_DSL    1.343464e-05  -4.352903
## PaymentMethod=Mail                     8.718327e-16  -8.043675
## StreamingTV=StreamingTV_Yes            1.082101e-23 -10.033853
## StreamingMovies=StreamingMovies_Yes    5.314620e-27 -10.760042
## SeniorCitizen=SeniorCitizen_No         1.443289e-31 -11.689444
## OnlineBackup=OnlineBackup_Yes          1.639796e-36 -12.619924
## PaymentMethod=Transfer                 4.650615e-44 -13.922096
## PaymentMethod=CC                       3.126644e-51 -15.056554
## Dependents=Dependents_Yes              4.968052e-52 -15.177713
## OnlineSecurity=OnlineSecurity_Yes      1.072640e-54 -15.575231
## Partner=Partner_Yes                    3.896963e-60 -16.356697
## PaperlessBilling=PaperlessBilling_No   2.308872e-60 -16.388550
## DeviceProtection=DeviceProtection_Yes  7.462569e-83 -19.282996
## TechSupport=TechSupport_Yes           2.208286e-112 -22.527974
## Contract=1year                        4.369921e-125 -23.788721
## MultipleLines=MultipleLines_DNA       4.166976e-154 -26.444986
## PhoneService=PhoneService_No          4.166976e-154 -26.444986
## Churn=Churn_No                        3.596791e-190 -29.414374
## Contract=2year                         0.000000e+00       -Inf
## StreamingMovies=StreamingMovies_DNA    0.000000e+00       -Inf
## StreamingTV=StreamingTV_DNA            0.000000e+00       -Inf
## TechSupport=TechSupport_DNA            0.000000e+00       -Inf
## DeviceProtection=DeviceProtection_DNA  0.000000e+00       -Inf
## OnlineBackup=OnlineBackup_DNA          0.000000e+00       -Inf
## OnlineSecurity=OnlineSecurity_DNA      0.000000e+00       -Inf
## InternetService=InternetService_No     0.000000e+00       -Inf
##
## $`3`
##                                        Cla/Mod   Mod/Cla    Global
## StreamingMovies=StreamingMovies_Yes   58.45534  75.50827 38.790288
## StreamingTV=StreamingTV_Yes           58.36720  74.70449 38.435326
## TechSupport=TechSupport_Yes           66.48728  64.25532 29.021724
## DeviceProtection=DeviceProtection_Yes 64.07927  73.38061 34.388755
## OnlineBackup=OnlineBackup_Yes         59.57184  68.41608 34.488144
## OnlineSecurity=OnlineSecurity_Yes     60.82219  58.06147 28.666761
## MultipleLines=MultipleLines_Yes       50.01683  70.26005 42.183729
## PhoneService=PhoneService_Yes         33.24949 100.00000 90.316626
## Contract=2year                        51.32743  41.13475 24.066449
## Contract=1year                        51.45961  35.83924 20.914383
## Partner=Partner_Yes                   40.24103  64.72813 48.303280
## InternetService=InternetService_Fiber 39.92248  58.43972 43.958540
## PaymentMethod=CC                      45.79501  32.95508 21.610109
## PaymentMethod=Transfer                45.46632  33.19149 21.922476
## Churn=Churn_No                        34.42211  84.20804 73.463013
## PaperlessBilling=PaperlessBilling_Yes 33.80484  66.66667 59.221922
## InternetService=InternetService_DSL   36.30731  41.56028 34.374556
## Dependents=Dependents_Yes             34.26540  34.18440 29.958824
## Dependents=Dependents_No              28.21812  65.81560 70.041176
## PaperlessBilling=PaperlessBilling_No  24.54735  33.33333 40.778078
## OnlineSecurity=OnlineSecurity_No      25.35735  41.93853 49.666335
## PaymentMethod=ECheck                  21.01480  23.49882 33.579441
## OnlineBackup=OnlineBackup_No          21.63212  31.58392 43.844952
## Churn=Churn_Yes                       17.87052  15.79196 26.536987
## TechSupport=TechSupport_No            21.76792  35.74468 49.311373
## StreamingTV=StreamingTV_No            19.03915  25.29551 39.897771
## StreamingMovies=StreamingMovies_No    18.59964  24.49173 39.542808
## PaymentMethod=Mail                    13.58561  10.35461 22.887974
## Partner=Partner_No                    20.48888  35.27187 51.696720
## DeviceProtection=DeviceProtection_No  18.19063  26.61939 43.944342
## MultipleLines=MultipleLines_No        18.55457  29.73995 48.132898
## MultipleLines=MultipleLines_DNA        0.00000   0.00000  9.683374
## PhoneService=PhoneService_No           0.00000   0.00000  9.683374
## StreamingMovies=StreamingMovies_DNA    0.00000   0.00000 21.666903
## StreamingTV=StreamingTV_DNA            0.00000   0.00000 21.666903
## TechSupport=TechSupport_DNA            0.00000   0.00000 21.666903
## DeviceProtection=DeviceProtection_DNA  0.00000   0.00000 21.666903
## OnlineBackup=OnlineBackup_DNA          0.00000   0.00000 21.666903
## OnlineSecurity=OnlineSecurity_DNA      0.00000   0.00000 21.666903
## InternetService=InternetService_No     0.00000   0.00000 21.666903
## Contract=MtM                          12.56774  23.02600 55.019168
##                                             p.value     v.test
## StreamingMovies=StreamingMovies_Yes    0.000000e+00        Inf
## StreamingTV=StreamingTV_Yes            0.000000e+00        Inf
## TechSupport=TechSupport_Yes            0.000000e+00        Inf
## DeviceProtection=DeviceProtection_Yes  0.000000e+00        Inf
## OnlineBackup=OnlineBackup_Yes          0.000000e+00        Inf
## OnlineSecurity=OnlineSecurity_Yes     2.528992e-267  34.930858
## MultipleLines=MultipleLines_Yes       5.954287e-216  31.365642
## PhoneService=PhoneService_Yes         3.597817e-113  22.608217
## Contract=2year                        5.091068e-101  21.337533
## Contract=1year                         4.651134e-85  19.543877
## Partner=Partner_Yes                    1.248913e-73  18.151522
## InternetService=InternetService_Fiber  1.015049e-57  16.014323
## PaymentMethod=CC                       2.464419e-49  14.764967
## PaymentMethod=Transfer                 2.850066e-48  14.598997
## Churn=Churn_No                         1.714275e-43  13.828556
## PaperlessBilling=PaperlessBilling_Yes  5.086245e-17   8.384683
## InternetService=InternetService_DSL    1.548838e-16   8.252678
## Dependents=Dependents_Yes              4.743039e-07   5.036426
## Dependents=Dependents_No               4.743039e-07  -5.036426
## PaperlessBilling=PaperlessBilling_No   5.086245e-17  -8.384683
## OnlineSecurity=OnlineSecurity_No       1.753585e-17  -8.509054
## PaymentMethod=ECheck                   6.403988e-33 -11.951141
## OnlineBackup=OnlineBackup_No           7.823695e-43 -13.718906
## Churn=Churn_Yes                        1.714275e-43 -13.828556
## TechSupport=TechSupport_No             7.819519e-51 -14.995816
## StreamingTV=StreamingTV_No             1.253645e-62 -16.702653
## StreamingMovies=StreamingMovies_No     8.284241e-67 -17.267363
## PaymentMethod=Mail                     3.188339e-67 -17.322390
## Partner=Partner_No                     1.248913e-73 -18.151522
## DeviceProtection=DeviceProtection_No   8.005944e-85 -19.516142
## MultipleLines=MultipleLines_No         3.866746e-93 -20.471483
## MultipleLines=MultipleLines_DNA       3.597817e-113 -22.608217
## PhoneService=PhoneService_No          3.597817e-113 -22.608217
## StreamingMovies=StreamingMovies_DNA   5.841003e-275 -35.430265
## StreamingTV=StreamingTV_DNA           5.841003e-275 -35.430265
## TechSupport=TechSupport_DNA           5.841003e-275 -35.430265
## DeviceProtection=DeviceProtection_DNA 5.841003e-275 -35.430265
## OnlineBackup=OnlineBackup_DNA         5.841003e-275 -35.430265
## OnlineSecurity=OnlineSecurity_DNA     5.841003e-275 -35.430265
## InternetService=InternetService_No    5.841003e-275 -35.430265
## Contract=MtM                          5.084050e-282 -35.885818
##
## $`4`
##                                          Cla/Mod    Mod/Cla    Global
## StreamingMovies=StreamingMovies_DNA   100.000000 100.000000 21.666903
## StreamingTV=StreamingTV_DNA           100.000000 100.000000 21.666903
## TechSupport=TechSupport_DNA           100.000000 100.000000 21.666903
## DeviceProtection=DeviceProtection_DNA 100.000000 100.000000 21.666903
## OnlineBackup=OnlineBackup_DNA         100.000000 100.000000 21.666903
## OnlineSecurity=OnlineSecurity_DNA     100.000000 100.000000 21.666903
## InternetService=InternetService_No    100.000000 100.000000 21.666903
## PaperlessBilling=PaperlessBilling_No   37.604457  70.773263 40.778078
## MultipleLines=MultipleLines_No         34.926254  77.588467 48.132898
## PaymentMethod=Mail                     45.967742  48.558322 22.887974
## Churn=Churn_No                         27.309625  92.595020 73.463013
## PhoneService=PhoneService_Yes          23.989939 100.000000 90.316626
## Contract=2year                         37.640118  41.808650 24.066449
## SeniorCitizen=SeniorCitizen_No         24.978817  96.592398 83.785319
## Dependents=Dependents_Yes              30.473934  42.136304 29.958824
## Contract=1year                         24.711473  23.853211 20.914383
## Dependents=Dependents_No               17.899858  57.863696 70.041176
## SeniorCitizen=SeniorCitizen_Yes         4.553415   3.407602 16.214681
## MultipleLines=MultipleLines_Yes        11.511276  22.411533 42.183729
## Contract=MtM                           13.522581  34.338139 55.019168
## MultipleLines=MultipleLines_DNA         0.000000   0.000000  9.683374
## PhoneService=PhoneService_No            0.000000   0.000000  9.683374
## Churn=Churn_Yes                         6.046014   7.404980 26.536987
## PaymentMethod=ECheck                    5.158562   7.994758 33.579441
## PaperlessBilling=PaperlessBilling_Yes  10.692879  29.226737 59.221922
## OnlineSecurity=OnlineSecurity_Yes       0.000000   0.000000 28.666761
## TechSupport=TechSupport_Yes             0.000000   0.000000 29.021724
## StreamingMovies=StreamingMovies_Yes     0.000000   0.000000 38.790288
## StreamingMovies=StreamingMovies_No      0.000000   0.000000 39.542808
## StreamingTV=StreamingTV_Yes             0.000000   0.000000 38.435326
## StreamingTV=StreamingTV_No              0.000000   0.000000 39.897771
## TechSupport=TechSupport_No              0.000000   0.000000 49.311373
## DeviceProtection=DeviceProtection_Yes   0.000000   0.000000 34.388755
## DeviceProtection=DeviceProtection_No    0.000000   0.000000 43.944342
## OnlineBackup=OnlineBackup_Yes           0.000000   0.000000 34.488144
## OnlineBackup=OnlineBackup_No            0.000000   0.000000 43.844952
## OnlineSecurity=OnlineSecurity_No        0.000000   0.000000 49.666335
## InternetService=InternetService_Fiber   0.000000   0.000000 43.958540
## InternetService=InternetService_DSL     0.000000   0.000000 34.374556
##                                             p.value     v.test
## StreamingMovies=StreamingMovies_DNA    0.000000e+00        Inf
## StreamingTV=StreamingTV_DNA            0.000000e+00        Inf
## TechSupport=TechSupport_DNA            0.000000e+00        Inf
## DeviceProtection=DeviceProtection_DNA  0.000000e+00        Inf
## OnlineBackup=OnlineBackup_DNA          0.000000e+00        Inf
## OnlineSecurity=OnlineSecurity_DNA      0.000000e+00        Inf
## InternetService=InternetService_No     0.000000e+00        Inf
## PaperlessBilling=PaperlessBilling_No  5.263538e-159  26.867532
## MultipleLines=MultipleLines_No        4.110091e-155  26.532308
## PaymentMethod=Mail                    2.052246e-143  25.498423
## Churn=Churn_No                         6.584621e-98  20.999812
## PhoneService=PhoneService_Yes          2.469267e-77  18.614110
## Contract=2year                         4.861679e-69  17.561451
## SeniorCitizen=SeniorCitizen_No         3.833827e-68  17.443846
## Dependents=Dependents_Yes              1.666761e-30  11.479795
## Contract=1year                         1.608594e-03   3.154344
## Dependents=Dependents_No               1.666761e-30 -11.479795
## SeniorCitizen=SeniorCitizen_Yes        3.833827e-68 -17.443846
## MultipleLines=MultipleLines_Yes        7.164590e-74 -18.182020
## Contract=MtM                          2.322052e-75 -18.369106
## MultipleLines=MultipleLines_DNA       2.469267e-77 -18.614110
## PhoneService=PhoneService_No          2.469267e-77 -18.614110
## Churn=Churn_Yes                       6.584621e-98 -20.999812
## PaymentMethod=ECheck                 9.539848e-151 -26.151242
## PaperlessBilling=PaperlessBilling_Yes 5.263538e-159 -26.867532
## OnlineSecurity=OnlineSecurity_Yes    1.084465e-259 -34.424499
## TechSupport=TechSupport_Yes          1.239114e-263 -34.686957
## StreamingMovies=StreamingMovies_Yes   0.000000e+00       -Inf
## StreamingMovies=StreamingMovies_No    0.000000e+00       -Inf
## StreamingTV=StreamingTV_Yes           0.000000e+00       -Inf
## StreamingTV=StreamingTV_No            0.000000e+00       -Inf
## TechSupport=TechSupport_No            0.000000e+00       -Inf
## DeviceProtection=DeviceProtection_Yes  0.000000e+00       -Inf
## DeviceProtection=DeviceProtection_No   0.000000e+00       -Inf
## OnlineBackup=OnlineBackup_Yes          0.000000e+00       -Inf
## OnlineBackup=OnlineBackup_No           0.000000e+00       -Inf
## OnlineSecurity=OnlineSecurity_No       0.000000e+00       -Inf
## InternetService=InternetService_Fiber  0.000000e+00       -Inf
## InternetService=InternetService_DSL    0.000000e+00       -Inf
##
##
## Link between the cluster variable and the quantitative variables
## ================================================================
##                     Eta2 P-value
## tenure         0.3017368       0
## MonthlyCharges 0.7849691       0
## TotalCharges   0.5319944       0
##
## Description of each cluster by quantitative variables
## =====================================================
## $`1`
##                    v.test Mean in category Overall mean sd in category
## TotalCharges    -9.530634       1502.43865   2279.79899     1340.12850
## MonthlyCharges -20.731257         42.31516     64.76169       11.47535
##                Overall sd      p.value
## TotalCharges   2266.56924 1.563263e-21
## MonthlyCharges   30.08791 1.809984e-95
##
## $`2`
##                   v.test Mean in category Overall mean sd in category
## MonthlyCharges  21.29221         74.42655     64.76169       16.27628
## TotalCharges   -24.13935       1454.37456   2279.79899     1487.32278
## tenure         -36.98662         18.66814     32.37115       17.81497
##                Overall sd       p.value
## MonthlyCharges   30.08791 1.340588e-100
## TotalCharges   2266.56924 9.660896e-129
## tenure           24.55774 1.879320e-299
##
## $`3`
##                  v.test Mean in category Overall mean sd in category Overall sd
## TotalCharges   60.11937       4758.45090   2279.79899     2088.68038 2266.56924
## MonthlyCharges 48.49021         91.30031     64.76169       16.20431   30.08791
## tenure         42.66285         51.42884     32.37115       19.39772   24.55774
##                 p.value
## TotalCharges          0
## MonthlyCharges        0
## tenure                0
##
## $`4`
##                    v.test Mean in category Overall mean sd in category
## tenure          -3.27795         30.54718     32.37115      24.348525
## TotalCharges   -31.48792        662.69056   2279.79899     555.344986
## MonthlyCharges -64.07509         21.07919     64.76169       2.163512
##                Overall sd       p.value
## tenure           24.55774  1.045640e-03
## TotalCharges   2266.56924 1.271376e-217
## MonthlyCharges   30.08791  0.000000e+00
```

This output begins by showing us the variables most responsible for forming the clusters in descending importance. PhoneService, InternetService, and the additional features make up the first few variables. As previously mentioned, these variables are highly correlated, especially the since DNA (does not apply) categories are perfectly correlated with having no internet or having no phone service. Lengths of contracts and PaperlessBilling also have a large effect on forming clusters.

Now let's see the 4 customer profiles.

Cluster 1 consists of customers who have DSL internet and no phone service. These customers tend to subscribe to a variety mixes of additional features. 100% of customers with no phone service belong to cluster 1 while 97.9885% of customers in cluster 1 have no phone service.

Cluster 2 consists of customers who have FiberOptic internet, use streaming TV and movies, do not have TechSupport, and on month-to-month or one year contracts. This cluster is most associated with churn. 60.077% of those with fiber optic internet service belong to cluster 2 and 68.7361% of those in cluster 2 use fiber optic. More importantly for our study, 66.987% of those who churn are found in cluster 2 while 46.2675% of customers in cluster 2 have churned.

Cluster 3 consists who loyal, high-valued customers. They tend to subscribe to more additional features, have longer contracts, and longer tenure. They tend to have partners and pay by bank transfers or credit cards. 100% of customers in cluster 3 have phone service.

Cluster 4, the large separated cloud on the right side of the MCA plot, are the phone service only customers. 100% of the customers have phone service, although only 23.9899% of all those with phone service are in cluster 4. They also tend to be more loyal customers with longer contracts.

In the next chunk of code I make attach the clusters of each to a new data frame.

```
clust = churn.hcpc$data.clust$clust
churn.with.clusters = cbind(churn, clust )
head(churn.with.clusters)
```

```
##   gender SeniorCitizen Partner Dependents tenure PhoneService MultipleLines
## 1 Female            No     Yes         No      1           No           DNA
## 2   Male            No      No         No     34          Yes            No
## 3   Male            No      No         No      2          Yes            No
## 4   Male            No      No         No     45           No           DNA
## 5 Female            No      No         No      2          Yes            No
## 6 Female            No      No         No      8          Yes           Yes
##   InternetService OnlineSecurity OnlineBackup DeviceProtection TechSupport
## 1             DSL             No          Yes               No          No
## 2             DSL            Yes           No              Yes          No
## 3             DSL            Yes          Yes               No          No
## 4             DSL            Yes           No              Yes         Yes
## 5           Fiber             No           No               No          No
## 6           Fiber             No           No              Yes          No
##   StreamingTV StreamingMovies Contract PaperlessBilling PaymentMethod
## 1          No              No      MtM              Yes        ECheck
## 2          No              No    1year               No          Mail
## 3          No              No      MtM              Yes          Mail
## 4          No              No    1year               No      Transfer
## 5          No              No      MtM              Yes        ECheck
## 6         Yes             Yes      MtM              Yes        ECheck
##   MonthlyCharges TotalCharges Churn clust
## 1          29.85        29.85    No     1
## 2          56.95      1889.50    No     2
## 3          53.85       108.15   Yes     2
## 4          42.30      1840.75    No     1
## 5          70.70       151.65   Yes     2
## 6          99.65       820.50   Yes     2
```

In looking at the data frame, it also helps to relate the clusters to the individuals. For example, the first customer has been classified as belonging to the first cluster. We confirm that she does indeed have DSL internet and no phone service, like the description of the first cluster previously mentioned.

Lastly we can directly look at the churn rates for each cluster.

```r
t=table(churn.with.clusters$clust, churn.with.clusters$Churn)
t
```

```
## 
##        No  Yes
##   1  526  170
##   2 1454 1252
##   3 1781  334
##   4 1413  113
```

```r
prop.table(t,1)
```

```
## 
##             No       Yes
##   1 0.7557471 0.2442529
##   2 0.5373245 0.4626755
##   3 0.8420804 0.1579196
##   4 0.9259502 0.0740498
```

```r
churn.mca.df = data.frame(cbind(churn.mca$ind$coord , Churn = as.integer(churn[ ,20])-1))
str(churn.mca.df)
```

```
## 'data.frame':    7043 obs. of  21 variables:
##  $ Dim.1 : num  -0.632 -0.204 -0.319 -0.452 -0.454 ...
##  $ Dim.2 : num  -0.237429 -0.000785 -0.288069 0.711054 -0.869182 ...
##  $ Dim.3 : num  1.147 0.4363 0.3791 1.2785 -0.0337 ...
##  $ Dim.4 : num  0.6173 -0.7451 -0.6428 -0.0146 -0.0453 ...
##  $ Dim.5 : num  -0.2215 0.6471 0.0123 0.2328 -0.0601 ...
##  $ Dim.6 : num  -0.0097 0.09382 -0.00486 0.52492 -0.04353 ...
##  $ Dim.7 : num  0.40597 -0.1535 -0.02818 0.27187 0.00272 ...
##  $ Dim.8 : num  -0.4914 0.155 -0.6447 0.49 0.0507 ...
##  $ Dim.9 : num  -0.0642 0.3762 -0.1952 0.1743 -0.1708 ...
##  $ Dim.10: num  -0.1457 0.1366 0.1334 0.1569 -0.0408 ...
##  $ Dim.11: num  -0.1221 0.2072 -0.0272 0.0714 -0.1266 ...
##  $ Dim.12: num  -0.0304 -0.5059 -0.2912 -0.1795 -0.1115 ...
##  $ Dim.13: num  0.183 -0.105 -0.384 0.373 0.186 ...
##  $ Dim.14: num  -0.0171 0.05367 -0.01497 0.03609 -0.00277 ...
##  $ Dim.15: num  -0.0848 -0.143 -0.1373 0.393 0.1004 ...
##  $ Dim.16: num  0.0356 0.1757 -0.091 -0.0311 0.2335 ...
##  $ Dim.17: num  2.90e-12 7.20e-15 -1.07e-14 -1.36e-14 6.76e-15 ...
##  $ Dim.18: num  5.23e-15 -3.92e-14 3.85e-14 -1.03e-12 -8.26e-15 ...
##  $ Dim.19: num  -1.81e-14 -2.08e-14 1.08e-14 -2.14e-13 5.05e-14 ...
##  $ Dim.20: num  4.21e-16 -1.37e-15 -2.46e-14 -1.62e-14 -3.30e-14 ...
##  $ Churn : num  0 0 1 0 1 1 0 0 1 0 ...
```

Cluster 2 has the highest churn rate, far greater than the global 26.5%. Cluster 5 has the lowest churn rate and seem the most loyal customers. Cluster 2 has a much higher churn rate than the global rate of 26.5%. The other three clusters have lower rates, with cluster 4 having the lowest churn rate.

Section IV: Predictive Model Selection

Here we choose between three models- one variable model, full model, and a model with a few interactions. We will run binary logistic regression for each of the three models. I ran many different models before settling on these three to compare. The models I ran included using both the MCA components and the four clusters as previously mentioned. These models did perform reasonably well, but I did find that the interaction model generated the best results.

Logistic regression is used when the response variable of interest is categorical, like in our case. Logistic regression can use either categorical (binary or multinomial) or numerical independent variables, again like in our case. The assumptions on the distribution of independent variables in logistic regression is lighter than many other techniques. Here we assume each variable belongs to an exponential family (Gaussian, exponential, Poisson, to name a few), which we will be assuming for our subsequent models. However, interpreting these more complex logistic model is more difficult, but in our case each model will ultimately predict the probability of each customer of churning. We will use those probabilities and choose a cutoff value to determine whether a given customer is classified as a "churn" or not. The untransformed y-variable logistic regression is called "logit," or log-odds, and the general equation is written in the form:

$log\frac{p}{(1-p)} = \beta_0 + \beta_{1i}X_1 + \cdots + \beta_{ki}X_k$, where $log()$ is the natural logarithm.

First we partition the data into 70% training data ans 30% testing data. We use `set.seed()` for reproducibility.

```r
set.seed(3.141592)
s = sample(1:nrow(churn),size = round(.7*nrow(churn)))
churn.train = churn[s,]
churn.test = churn[-s,]
churn.with.clusters.train = churn.with.clusters[s, ]
churn.with.clusters.test = churn.with.clusters[-s, ]
churn.mca.df.train = churn.mca.df[s, ]
churn.mca.df.test = churn.mca.df[-s, ]
```

My first model (model 1) will consist of the InternetService and its three levels.

```r
churn1.glm = glm(Churn~InternetService, family = binomial(link = logit), data = churn.train)
contrasts(churn$InternetService)
```

```
##       Fiber No
## DSL       0  0
## Fiber     1  0
## No        0  1
```

```r
summary(churn1.glm)
```

```
##
## Call:
## glm(formula = Churn ~ InternetService, family = binomial(link = logit),
##     data = churn.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0272  -1.0272  -0.6389   1.3355   2.3134
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -1.48551    0.06250 -23.770   <2e-16 ***
## InternetServiceFiber  1.12127    0.07632  14.691   <2e-16 ***
## InternetServiceNo    -1.11918    0.13571  -8.247   <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: 5627.6  on 4929  degrees of freedom
## Residual deviance: 5081.8  on 4927  degrees of freedom
## AIC: 5087.8
##
## Number of Fisher Scoring iterations: 5
```

I will be using the AIC (Alaike Information Criteria) to determine the best model. Although the number is difficult to interpret, the lower the AIC, the better the model assuming one model is nested in the other. Model 1 has an AIC of 5165.3.

Although the output of the coefficients is difficult to read directly, it does tell us that there is internet service type does have a significance effect on the churn rate. However, this model is not very useful in prediction, since it would merely use the conditional probabilities to predict a customer's probability of leaving.

```r
c5 = table(churn.train$InternetService, churn.train$Churn)
c6 = round(prop.table(c3,1)*100,2)
c5
```

```
##
##           No  Yes
##   DSL   1387  314
##   Fiber 1271  883
##   No    1001   74
```

```r
c6
```

```
##
##            No   Yes
##   DSL   81.04 18.96
##   Fiber 58.11 41.89
##   No    92.60  7.40
```

We can also see some different pseudo r^2 values. It should be noted that in logistic regression there is no true r^2 like the one we find in linear regression using ordinary least squares. However, these values are still useful when comparing models.

```r
PseudoR2(churn1.glm, which = c( "McFadden", "McFaddenAdj", "Nagelkerke", "CoxSnell"))
```

```
##    McFadden McFaddenAdj  Nagelkerke    CoxSnell
##  0.09699145  0.09592528  0.15397848  0.10480735
```

Now let's try the next model. Model 2 will be the full model, without interaction effects.

```r
churn2.glm = glm(Churn~., family = binomial(link = logit), data = churn.train)
summary(churn2.glm)
```

```
##
## Call:
## glm(formula = Churn ~ ., family = binomial(link = logit), data = churn.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8445  -0.6720  -0.2981   0.7248   3.4982
##
## Coefficients: (7 not defined because of singularities)
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)          -9.355e-01  2.478e+00  -0.377 0.705811
## genderMale           -2.991e-02  7.762e-02  -0.385 0.700018
## SeniorCitizenYes      2.295e-01  1.006e-01   2.282 0.022506 *
## PartnerYes            7.617e-02  9.266e-02   0.822 0.411071
## DependentsYes        -2.176e-01  1.067e-01  -2.039 0.041419 *
## tenure               -6.449e-02  7.555e-03  -8.536  < 2e-16 ***
## PhoneServiceYes      -6.963e-01  9.611e-01  -0.724 0.468768
## MultipleLinesNo      -1.695e-01  2.107e-01  -0.805 0.421047
## MultipleLinesYes             NA         NA      NA       NA
## InternetServiceFiber  5.598e-01  9.543e-01   0.587 0.557499
## InternetServiceNo     3.488e-01  2.472e+00   0.141 0.887782
## OnlineSecurityNo      4.542e-01  2.149e-01   2.114 0.034542 *
## OnlineSecurityYes            NA         NA      NA       NA
## OnlineBackupNo        3.048e-01  2.101e-01   1.451 0.146867
## OnlineBackupYes              NA         NA      NA       NA
## DeviceProtectionNo    8.500e-02  2.105e-01   0.404 0.686320
## DeviceProtectionYes          NA         NA      NA       NA
## TechSupportNo         3.048e-01  2.156e-01   1.414 0.157368
## TechSupportYes               NA         NA      NA       NA
## StreamingTVNo        -2.136e-02  3.910e-01  -0.055 0.956448
## StreamingTVYes               NA         NA      NA       NA
## StreamingMoviesNo    -1.388e-01  3.881e-01  -0.358 0.720674
## StreamingMoviesYes           NA         NA      NA       NA
## Contract1year        -6.530e-01  1.288e-01  -5.068 4.03e-07 ***
## Contract2year        -1.328e+00  2.041e-01  -6.505 7.76e-11 ***
## PaperlessBillingYes   3.089e-01  8.985e-02   3.438 0.000586 ***
## PaymentMethodCC      -5.919e-02  1.350e-01  -0.438 0.661052
## PaymentMethodECheck   2.249e-01  1.136e-01   1.980 0.047659 *
## PaymentMethodMail    -7.443e-02  1.378e-01  -0.540 0.589107
## MonthlyCharges        7.960e-03  3.798e-02   0.210 0.833977
## TotalCharges          3.812e-04  8.547e-05   4.460 8.20e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
```

```
##
##     Null deviance: 5627.6  on 4929  degrees of freedom
## Residual deviance: 4076.3  on 4906  degrees of freedom
## AIC: 4124.3
##
## Number of Fisher Scoring iterations: 6
```

The first thing of note is that the AIC of 4117.7 has reduced drastically from the first model, suggesting this model is better. The next we notice is the list of variables, including their p-values. Typically, we can reduce the model by eliminating variables that are not significant. The variables gender, Partner, PhoneService, OnlineBackup, and TechSupport are not significant at a 5% level.

Looking at the r^2 values, we see a vast improvement over the last model.

```
PseudoR2(churn2.glm, which = c( "McFadden", "McFaddenAdj", "Nagelkerke", "CoxSnell"))
```

```
##    McFadden McFaddenAdj  Nagelkerke    CoxSnell
##   0.2756656   0.2671362   0.3966330   0.2699731
```

Next we see how well the model can predict our test data. The code below is my way of being able to change the model and data set to test different models. I will use the model to predict the probability of each individual in the training data set. As of now, I will use a 50% cutoff level to classify whether each customer is a "churn" or not. We then create a confusion matrix and calculate the misclassification rate.

```
churn.glm = churn2.glm ####set it and forget it
train.data = churn.train####set it and forget it
test.data = churn.test####set it and forget it
pred = predict(churn.glm, test.data, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
pred.train = predict(churn.glm, train.data, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
head(pred.train) ##some of the predicitions
```

```
##        2821        3770         652        5095        6692        6842
## 0.009048798 0.054635365 0.611663320 0.074810466 0.039062892 0.241578878
```

```
Classified_Churn.test = ifelse(pred > .5, "1","0")
Classified_Churn.train = ifelse(pred.train > .5, "1","0")
pred_churn.test = cbind(churn.test, pred, Classified_Churn.test)
pred_churn.train = cbind(churn.train, pred.train, Classified_Churn.train)
pred.train = predict(churn.glm, train.data, type = "response")
```

```
## Warning in predict.lm(object, newdata, se.fit, scale = 1, type = if (type == :
## prediction from a rank-deficient fit may be misleading
```

```
Classified_Churn.test = ifelse(pred > .5, "1","0")
Classified_Churn.train = ifelse(pred.train > .5, "1","0")
pred_churn.test = cbind(churn.test, pred, Classified_Churn.test)
pred_churn.train = cbind(churn.train, pred.train, Classified_Churn.train)
#head(pred_churn.test)
conf.matrix.train = xtabs(~Churn + Classified_Churn.train, data = pred_churn.train )
conf.matrix.train
```

```
##      Classified_Churn.train
## Churn    0    1
##   No  3298  361
##   Yes  593  678
```

```
print(noquote(c("misclassification error is", 1-sum(diag(conf.matrix.train)/sum(conf.matrix.train)))))
```

```
## [1] misclassification error is 0.193509127789047
```

```
pred3 = prediction(pred, test.data$Churn)
roc1 = performance(pred3, "tpr","fpr")#####
roc = roc1####
auc = performance(pred3, "auc")
auc = unlist(slot(auc, "y.values"[[1]])[1])
auc = round(auc, 4)
auc1 = auc####
```

Beside the overall error rate, it is advantage for the company to consider the specificity, or rate of correctly classified churned customers of all those who did churn. For example, we correctly predicted 750 of the 750 + 572 = 1322 customers who did churn. The model missed 572 customers at a rate of 43.3%.

```
false_pos.train = conf.matrix.train[2,1]/rowSums(conf.matrix.train)[2]
false_pos.train
```

```
##       Yes
## 0.4665618
```

We will repeat this process for the testing data to see how well the model performs on data which it has not seen before.

```
conf.matrix.test = xtabs(~Churn + Classified_Churn.test, data = pred_churn.test )
conf.matrix.test
```

```
##      Classified_Churn.test
## Churn    0    1
##   No  1375  140
##   Yes  270  328
```

```
false_pos.test = conf.matrix.test[2,1]/rowSums(conf.matrix.test)[2]
print(noquote(c("misclassification error is", 1-sum(diag(conf.matrix.test)/sum(conf.matrix.test)))))
```

```
## [1] misclassification error is 0.194036914339801
```

```
false_pos.test = conf.matrix.test[2,1]/rowSums(conf.matrix.test)[2]
false_pos.test
```

```
##      Yes
## 0.451505
```

```
pred3 = prediction(pred, test.data$Churn)
roc2 = performance(pred3, "tpr","fpr")#####
roc = roc2####
auc = performance(pred3, "auc")
auc = unlist(slot(auc, "y.values"[[1]])[1])
auc = round(auc, 4)
auc2 = auc####
```

The test data closely matches the training data, suggesting this is a good model.

I am saving this model to create an ROC curve in the next section.

```
pred3 = prediction(pred, test.data$Churn)
roc2 = performance(pred3, "tpr","fpr")#####
roc = roc2####
auc = performance(pred3, "auc")
auc = unlist(slot(auc, "y.values"[[1]])[1])
auc = round(auc, 4)
auc = auc2####
```

After this model, I tried many different models trying to improve the misclassification rate of 19.56% testing and 19.31% training. One typical strategy would be dropping insignificant variable, such as gender, and rerunning the model. However, every model I tried behaved more poorly when I removed gender. Although the MCA before also found little influence of gender, there could be an interaction effect with another variable that influences the logistic regression in some way.

For the last model, I will run logistic regression on the MCA components. I first ran all the variables and reduced to the significant dimensions.

```
churn3.glm = glm(Churn~ Dim.1 + Dim.2 + Dim.4 + Dim.15 + Dim.16, family = binomial(link = logit), data =
churn.glm = churn3.glm ####set it and forget it
train.data = churn.mca.df.train####set it and forget it
test.data = churn.mca.df.test####set it and forget it
summary(churn.glm)
```

```
##
## Call:
## glm(formula = Churn ~ Dim.1 + Dim.2 + Dim.4 + Dim.15 + Dim.16,
##     family = binomial(link = logit), data = churn.mca.df.train)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5183  -0.6948  -0.3175   0.9036   2.9129
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.54798    0.04953 -31.254  < 2e-16 ***
## Dim.1       -1.01278    0.06809 -14.874  < 2e-16 ***
## Dim.2       -2.32702    0.10223 -22.762  < 2e-16 ***
## Dim.4        1.44093    0.10902  13.217  < 2e-16 ***
## Dim.15       1.34796    0.18517   7.280 3.35e-13 ***
## Dim.16      -1.60802    0.21528  -7.469 8.06e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5627.6  on 4929  degrees of freedom
## Residual deviance: 4320.9  on 4924  degrees of freedom
## AIC: 4332.9
##
## Number of Fisher Scoring iterations: 5
```

The AIC of 4314.2 is not quite as good as the 4117.7 from the previous model.

```
pred = predict(churn.glm, test.data, type = "response")
pred.train = predict(churn.glm, train.data, type = "response")
#head(pred)
```

```
Classified_Churn.test = ifelse(pred > .5, "1","0")
Classified_Churn.train = ifelse(pred.train > .5, "1","0")
pred_churn.test = cbind(churn.test, pred, Classified_Churn.test)
pred_churn.train = cbind(churn.train, pred.train, Classified_Churn.train)
pred_all = predict(churn.glm, churn.mca.df, type = "response")#####
#head(pred_churn.test)
conf.matrix.train = xtabs(~Churn + Classified_Churn.train, data = pred_churn.train )
conf.matrix.train
```

```
##      Classified_Churn.train
## Churn    0    1
##   No  3270  389
##   Yes  686  585
```

```
print(noquote(c("misclassification error is", 1-sum(diag(conf.matrix.train)/sum(conf.matrix.train)))))
```

```
## [1] misclassification error is 0.218052738336714
```

```
false_pos.train = conf.matrix.train[2,1]/rowSums(conf.matrix.train)[2]
false_pos.train
```

```
##       Yes
## 0.5397325
```

```
conf.matrix.test = xtabs(~Churn + Classified_Churn.test, data = pred_churn.test )
conf.matrix.test
```

```
##      Classified_Churn.test
## Churn    0    1
##   No  1363  152
##   Yes  304  294
```

```
1-sum(diag(conf.matrix.test))/sum(conf.matrix.test)
```

```
## [1] 0.2158069
```

```
false_pos.test = conf.matrix.test[2,1]/rowSums(conf.matrix.test)[2]
print(noquote(c("misclassification error is", 1-sum(diag(conf.matrix.test)/sum(conf.matrix.test)))))
```

```
## [1] misclassification error is 0.215806909607194
```

```
false_pos.test = conf.matrix.test[2,1]/rowSums(conf.matrix.test)[2]
false_pos.test
```

```
##       Yes
## 0.5083612
```

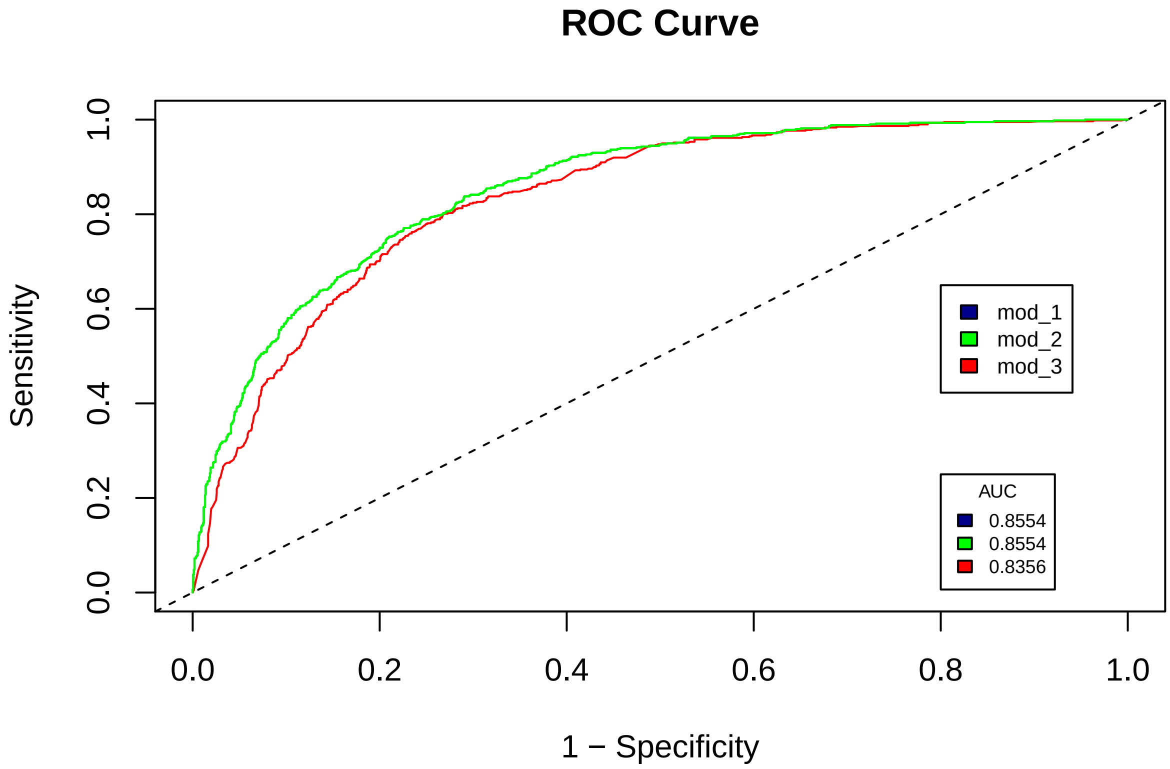
```
pred3 = prediction(pred, test.data$Churn)
roc3 = performance(pred3, "tpr","fpr")#####
roc = roc3####
auc = performance(pred3, "auc")
auc = unlist(slot(auc, "y.values"[[1]])[1])
auc = round(auc, 4)
auc3 = auc####
```

The misclassification rates, 21.5% and 22.1%, are much higher than the previous model.

Model Performance

We can plot all three models on a ROC curve, using the ROCR package.

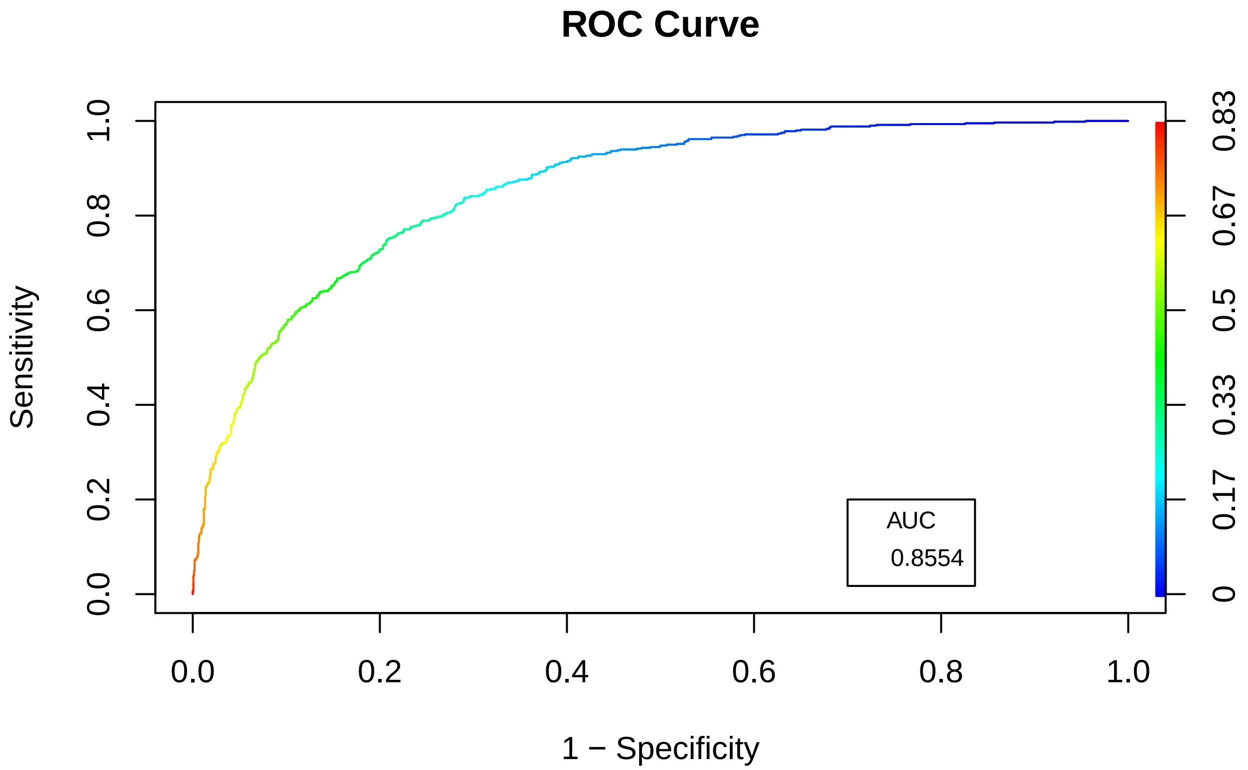
```
plot(roc1, main ="ROC Curve", xlab = "1 - Specificity", ylab = "Sensitivity", col ="darkblue")
abline(a=0,b=1, lty = 2)
legend(.8,.25, c(auc1,auc2,auc3), title= "AUC", cex = .6, fill=c("darkblue","green","red"))

###Adding curves#######
plot(roc2, add = T, col = "green")
plot(roc3, add = T, col = "red")
plot(roc2, add = T, col = "green")
#plot(roc5, add = T, col = "darksalmon")
legend(.8,.65, fill = c("darkblue","green", "red"), c("mod_1","mod_2" ,"mod_3"), cex = .7)
```



Using this chart, we can judge which model provides the better overall performance. The horizontal axis represent the proportion of misclassified churned customers while the vertical axis represents the sensitivity, or proportion of not churned customers who are incorrectly labeled as churned. These values depend upon the cutoff level used in predicting each individual based on their probabilities of churn (currently we are using 0.5 as the cutoff, but that will change). We are looking for the curve with the most area under the curve (AUC) and the one that rises the quickest. Clearly model 2 performs better than the other two models, with an area under the curve of 0.841, greater than the other two models.

To interpret the plot, let's examine a point on the from the second model, say (0.2,0..75). This means that the second model would miss about 20% of those who churned but would correctly classify about 75% of those who did not churn at the given cutoff value. Now that we are selecting the second model to predict churn, let's add the the differing cutoff levels.

```
plot(roc2, colorize = TRUE, main ="ROC Curve", xlab = "1 - Specificity", ylab = "Sensitivity")
legend(.7,.2, auc2, title= "AUC", cex = .75)
```



The color of the curve includes the different cutoff values. For instance, the (0.2,0..75) corresponds to the color green (towards the light blue end) on the spectrum, which corresponds to a cutoff value of roughly 33%. We can try to find the cutoff value that reduces the error as much as possible (which I have done and can reduce the misclassification rates as low as 19.2% 19.3% for the testing and training data respectively), or try to find the cutoff value that misses as few churned customers while not over classifying too many customers.

Because the company should want to favor identifying the potential churned customers over correctly identifying those who do not churn, we want to choose a cutoff value that corresponds to a bump or sharp turn on the ROC chart. The best point for this is approximately (0.22,0.75), which corresponds to a cut off value of 31% (which I found manually).

So our model will use the predict a customer's probability of churning based on the values of the 20 variables. If the customer has a 31% chance or greater of churning, they will be predicted as churning.

```
cutoff = .31
Classified_Churn = ifelse(pred > cutoff, "1","0")
Classified_Churn.train = ifelse(pred.train > cutoff, "1","0")
Classified_Churn_all = ifelse(pred_all > cutoff, "1","0")
pred_churn.test = cbind(test.data, pred, Classified_Churn)
pred_churn.train = cbind(train.data, pred.train, Classified_Churn.train)
#head(pred_churn.test)
conf.matrix.train = xtabs(~Churn + Classified_Churn.train, data = pred_churn.train )
conf.matrix.train
```

```
##      Classified_Churn.train
## Churn    0    1
##     0 2728  931
##     1  316  955
```

```
print(noquote(c("misclassification error is", 1-sum(diag(conf.matrix.train)/sum(conf.matrix.train)))))
```

```
## [1] misclassification error is 0.252941176470588
```

```
false_pos.train = conf.matrix.train[2,1]/rowSums(conf.matrix.train)[2]
false_pos.train
```

```
##         1
## 0.2486231
```

```
conf.matrix = xtabs(~Churn + Classified_Churn, data = pred_churn.test )
conf.matrix
```

```
##      Classified_Churn
## Churn    0    1
##     0 1154  361
##     1  141  457
```

```
1-sum(diag(conf.matrix))/sum(conf.matrix)
```

```
## [1] 0.2375769
```

```
false_pos.test = conf.matrix[2,1]/rowSums(conf.matrix)[2]
print(noquote(c("false positive rate", false_pos.test)))
```

```
##                                       1
## false positive rate   0.235785953177258
```

So using a cutoff of 31% we have correctly classified 418 customers as churned in our testing data, much better than the 305 previously classified customers using a 50% level. We only misclassify 23.6% of the churned customers. We do, however, over classify 377/(377+418) = 47.4% of customers as churned who were not churned. Because of this, our classifications error is increased 23.3% but it is a worthy trade off.

Summary

We have cleaned the data set, used descriptive techniques to cluster customers, and found a model to predict the churn of customers. We conclude by writing one more data set which includes the the original data, the cluster of each customer that we classified them to, the predicted churn using the logistic regression model, and the coordinates of the dimension using MCA.

```
final_set = cbind(churn, clust, pred_all, Classified_Churn_all,churn.mca.df[, -20])
write.csv(final_set ,"C744_FInal_Data_Set.csv")
```