SIM. Assignment 2: Telco Customer Churn

Adrià Casanova, Víctor Garcia, Zhengyong Ji

2024-01-05

Contents

- Profiling: AN EXTRA POINT TO BE DONE. WE CAN DO IT LATTER. USE THE SCRIPT PROVIDED BY MVA ???
- Canviar labels eixos numerical variables annex
- SI GENEREM UN WORD, L'HEM DE JUSTIFICAR I HEM DE NUMERAR-NE LES PÀGINES ABANS DE CONVERTIR-LO A PDF.
- #Pàgines actual (sense annex): 41/50

0. Introduction

In this project, we will study the data set “Telco Customer Churn”, which can be found at https://www.kaggle.com/datasets/blastchar/telco-customer-churn. Our goal is to analyze the correlation between the amount of customers who left within the last month (Churn) and different features that describe the customer and the services he/she/they has signed up for. Then, we will build a logistic model that will allow us to predict the variable Churn.

All members have contributed equally to all parts of the project.

We start by taking a first general look at the dataset.

```r
head(df)
```

```
##   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
```

```r
dim(df)
```

```
## [1] 7043   21
```

```r
summary(df)
```

```
##       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
```

The data set contains 7043 observations of 21 variables.

1. Data preparation

The first part of the project consisted on doing some basic data preparation to ensure that data is ready for the next sections.

Firstly, we checked that all datatypes were consistent with the metadata and declared “SeniorCitizen” as a factor, as it represented a qualitative concept.

```
df$SeniorCitizen <- factor(df$SeniorCitizen, labels = c("Yes", "No"))
```

Secondly, we discretized all numeric variables by splitting data into 4 categories. Their boundaries were obtained simply by dividing the total range in 4 equal intervals and the distribution was checked using histograms to ensure that they were similar to the original variables.

```
df$c.tenure <- df$tenure # Create a new variable called Categorical.tenure
m.tenure <- max(df$tenure, na.rm = TRUE)
df$c.tenure <- replace(df$c.tenure, df$tenure <= m.tenure/4, m.tenure/4)
for (i in 1:3) {
  idx <- (m.tenure*i/4 < df$tenure) & (df$tenure <= m.tenure*(i+1)/4)
  df$c.tenure <- replace(df$c.tenure, idx, m.tenure*(i+1)/4)
}
min(df$tenure, na.rm = TRUE)
```

```
## [1] 0
```

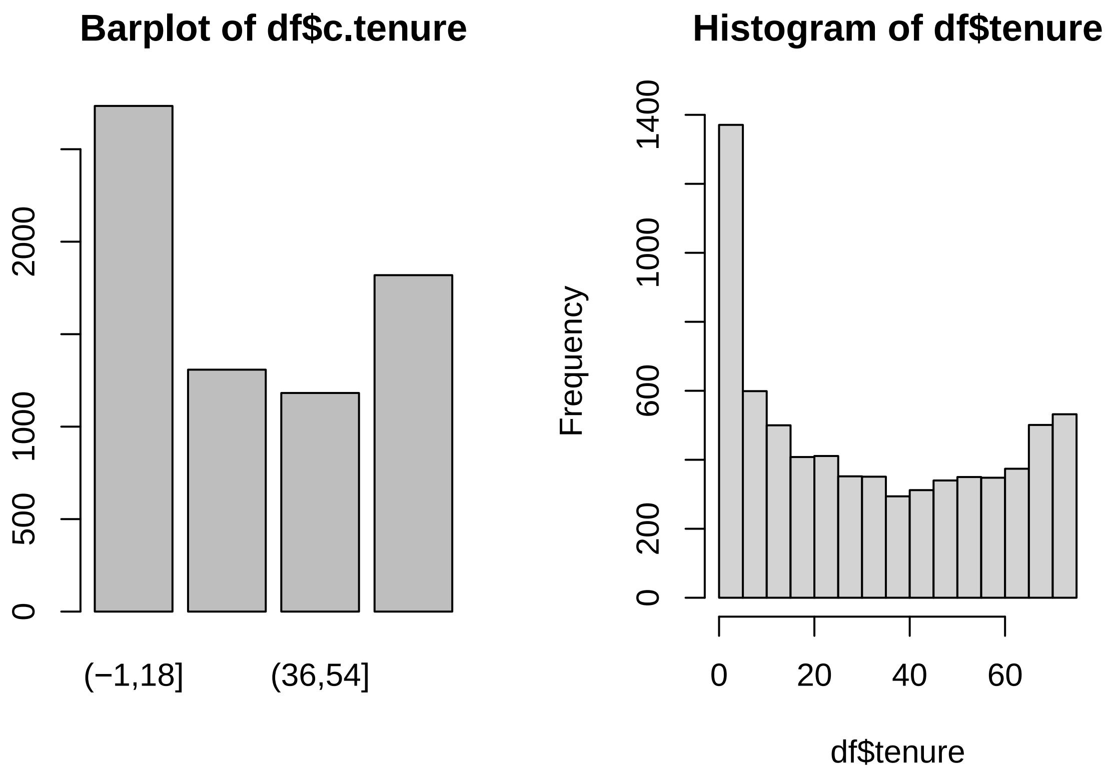
```
breakpts <- seq(m.tenure/4, m.tenure, m.tenure/4); breakpts
```

```
## [1] 18 36 54 72
```

```
df$c.tenure <- factor(df$c.tenure, labels = c("(-1,18]", "(18,36]",
                                              "(36,54]", "(54,72]"))
summary(df$c.tenure)
```

```
## (-1,18] (18,36] (36,54] (54,72]
##    2734    1308    1182    1819
```

```
par(mfrow=c(1,2))
plot(df$c.tenure, main = "Barplot of df$c.tenure")
hist(df$tenure)
```



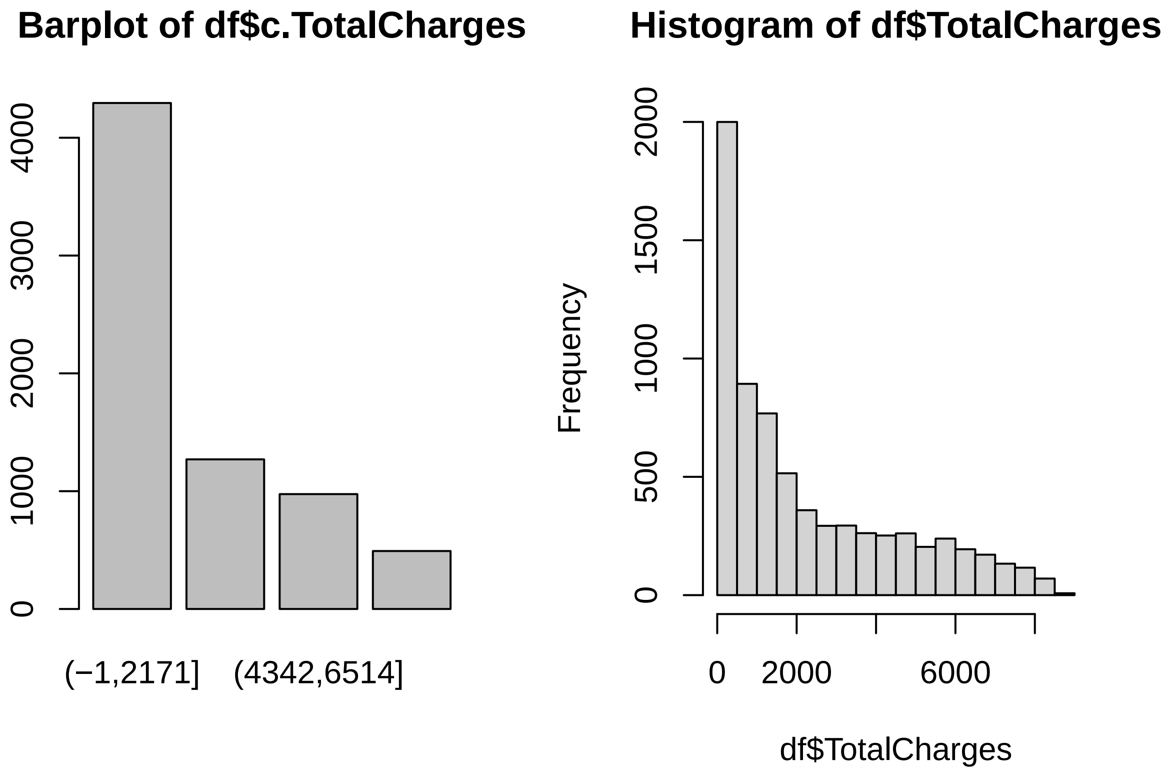
```
df$c.TotalCharges <- df$TotalCharges
m.TotalCharges <- max(df$TotalCharges, na.rm = TRUE)
df$c.TotalCharges <- replace(df$c.TotalCharges, df$TotalCharges <= m.TotalCharges/4,
↪  m.TotalCharges/4)
for (i in 1:3) {
  idx <- (m.TotalCharges*i/4 < df$TotalCharges) & (df$TotalCharges <=
                                                     m.TotalCharges*(i+1)/4)
  df$c.TotalCharges <- replace(df$c.TotalCharges, idx, m.TotalCharges*(i+1)/4)
}
breakpts <- seq(m.TotalCharges/4, m.TotalCharges, m.TotalCharges/4); breakpts
```

```
## [1] 2171.2 4342.4 6513.6 8684.8
```

```
df$c.TotalCharges <- factor(df$c.TotalCharges, labels = c("(-1,2171]",
                                                          "(2171,4342]",
                                                          "(4342,6514]",
                                                          "(6514,8685]"))
summary(df$c.TotalCharges)
```

```
##   (-1,2171] (2171,4342] (4342,6514] (6514,8685]        NA's
##        4295        1270         975         492          11
```

```
par(mfrow=c(1,2))
plot(df$c.TotalCharges, main = "Barplot of df$c.TotalCharges")
hist(df$TotalCharges)
```



```r
df$c.MonthlyCharges <- df$MonthlyCharges
m.MonthlyCharges <- max(df$MonthlyCharges, na.rm = TRUE)
df$c.MonthlyCharges <- replace(df$c.MonthlyCharges, df$MonthlyCharges <=
↪  m.MonthlyCharges/4, m.MonthlyCharges/4)
for (i in 1:3) {
  idx <- (m.MonthlyCharges*i/4 < df$MonthlyCharges) & (df$MonthlyCharges <=
                                                      m.MonthlyCharges*(i+1)/4)
  df$c.MonthlyCharges <- replace(df$c.MonthlyCharges, idx,
                                 m.MonthlyCharges*(i+1)/4)
}
min(df$MonthlyCharges, na.rm = TRUE)
```

```
## [1] 18.25
```

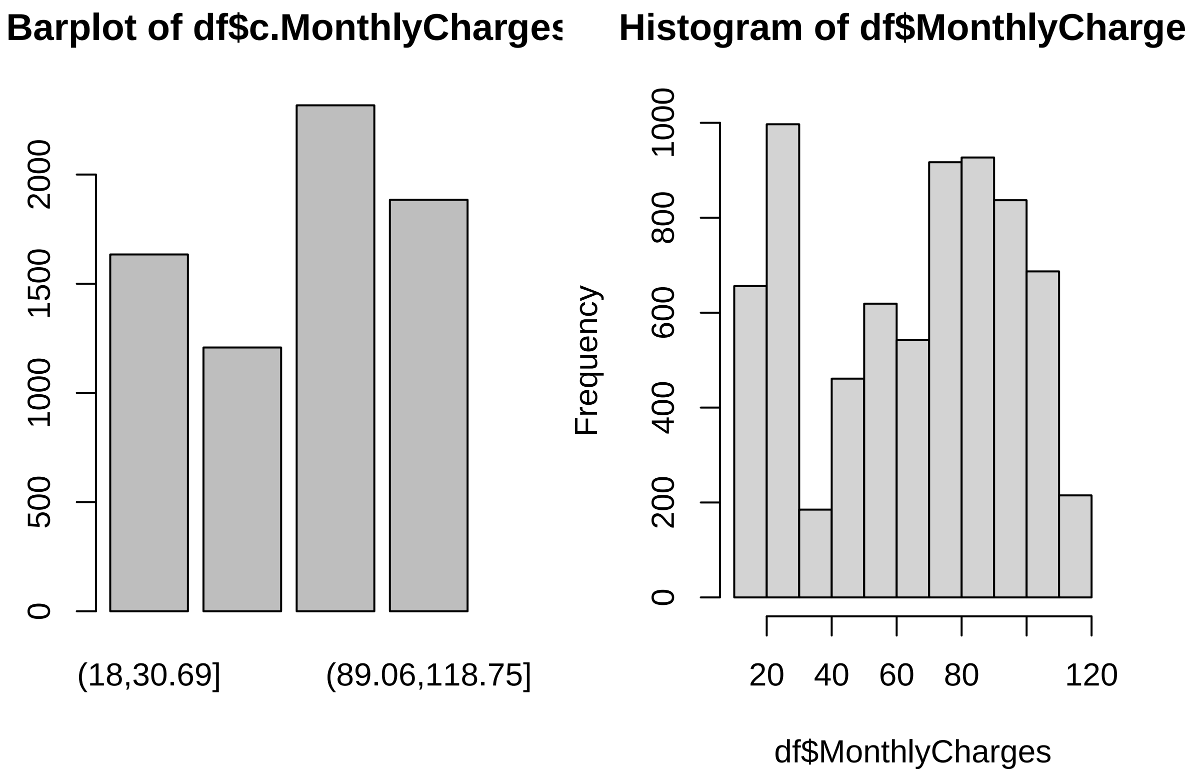
```r
breakpts <- seq(m.MonthlyCharges/4, m.MonthlyCharges, m.MonthlyCharges/4)
breakpts
```

```
## [1]  29.6875  59.3750  89.0625 118.7500
```

```r
df$c.MonthlyCharges <- factor(df$c.MonthlyCharges, labels = c("(18,30.69]",
                                                          "(30.69,59.38]",
                                                          "(59.38,89.06]",
                                                          "(89.06,118.75]"))
summary(df$c.MonthlyCharges)
```

```
##     (18,30.69]  (30.69,59.38]  (59.38,89.06] (89.06,118.75]
##           1634           1208           2317           1884
```

```
par(mfrow=c(1,2))
plot(df$c.MonthlyCharges, main = "Barplot of df$c.MonthlyCharges")
hist(df$MonthlyCharges)
```



```
par(mfrow=c(1,1))
```

Lastly, we identified categorical and numerical variables for later use.

```
numeric_val_idx = which(sapply(df, is.numeric))
numeric_val = names(df)[numeric_val_idx]
# The only numerical features that we have are tenure, MonthlyCharges and TotalChages.

# So the remaining will be categorical features.
categoric_val_idx = which(sapply(df, is.factor))
categoric_val = names(df)[categoric_val_idx]
```

2. Exploratory Data Analysis (EDA)

EDA was done mainly automatically using the “DataExplorer” library. It plots, for each variable, the distribution of numeric variables, the proportion of individuals in each category and the amount of missing values, among other metadata.

The main conclusions of this section are: 1- Using the QQ plots and distribution plots we see that no numerical variable is normally distributed. This was also checked visually and with Kolmogorov-Smirnov tests, a more suitable approach than Shappiro-Wilk for large samples.

2- Our database is not balanced in some categories, like PhoneService (10% of “No”) or SeniorCitizen(16% of “No”). This is specially relevant for the target, “Churn”, that has 73% of cases of “No”, so individuals that churned will be more difficult to predict.

3- Qualitative variables have a maximum of 4 levels, so all of them may be suitable for modeling without any aggregation.

5- Some categories, like “OnlineSecurity” or “OnlineBackup”, are not applicable if the client does not have an internet connection. Consequently, there is a special level for those cases that contains around 22% of the clients.

```r
# Basic EDA
summary(df)
```

```
##       customerID     gender     SeniorCitizen Partner    Dependents
##  0002-ORFBO:   1   Female:3488   Yes:5901      No :3641   No :4933  
##  0003-MKNFE:   1   Male  :3555   No :1142      Yes:3402   Yes:2110  
##  0004-TLHLJ:   1                                                    
##  0011-IGKFF:   1                                                    
##  0013-EXCHZ:   1                                                    
##  0013-MHZWF:   1                                                    
##  (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         c.tenure        c.TotalCharges
##  Min.   :  18.8   No :5174   (-1,18]:2734   (-1,2171]  :4295
##  1st Qu.: 401.4   Yes:1869   (18,36]:1308   (2171,4342]:1270
##  Median :1397.5              (36,54]:1182   (4342,6514]: 975
##  Mean   :2283.3              (54,72]:1819   (6514,8685]: 492
##  3rd Qu.:3794.7                             NA's       :  11
##  Max.   :8684.8
##  NA's   :11
##        c.MonthlyCharges
##  (18,30.69]    :1634
##  (30.69,59.38] :1208
##  (59.38,89.06] :2317
##  (89.06,118.75]:1884
##
##
##
```

```
# Completed EDA
#create_report(df, output_file = "Telco.html")
```

```
# tests
ks.test(df$TotalCharges, "pnorm")
```

```
## Warning in ks.test.default(df$TotalCharges, "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  df$TotalCharges
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(df$MonthlyCharges, "pnorm")
```

```
## Warning in ks.test.default(df$MonthlyCharges, "pnorm"): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
```

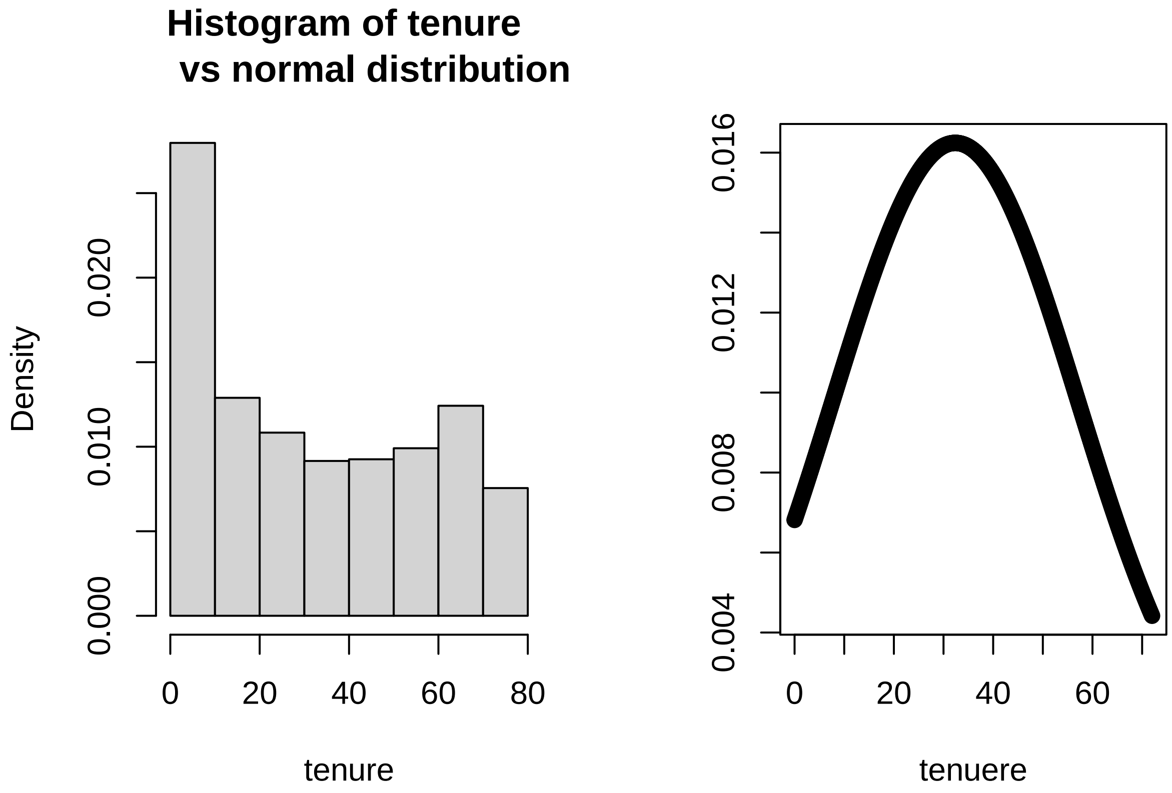
```
##
## data:  df$MonthlyCharges
## D = 1, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
ks.test(df$tenure, "pnorm")
```

```
## Warning in ks.test.default(df$tenure, "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

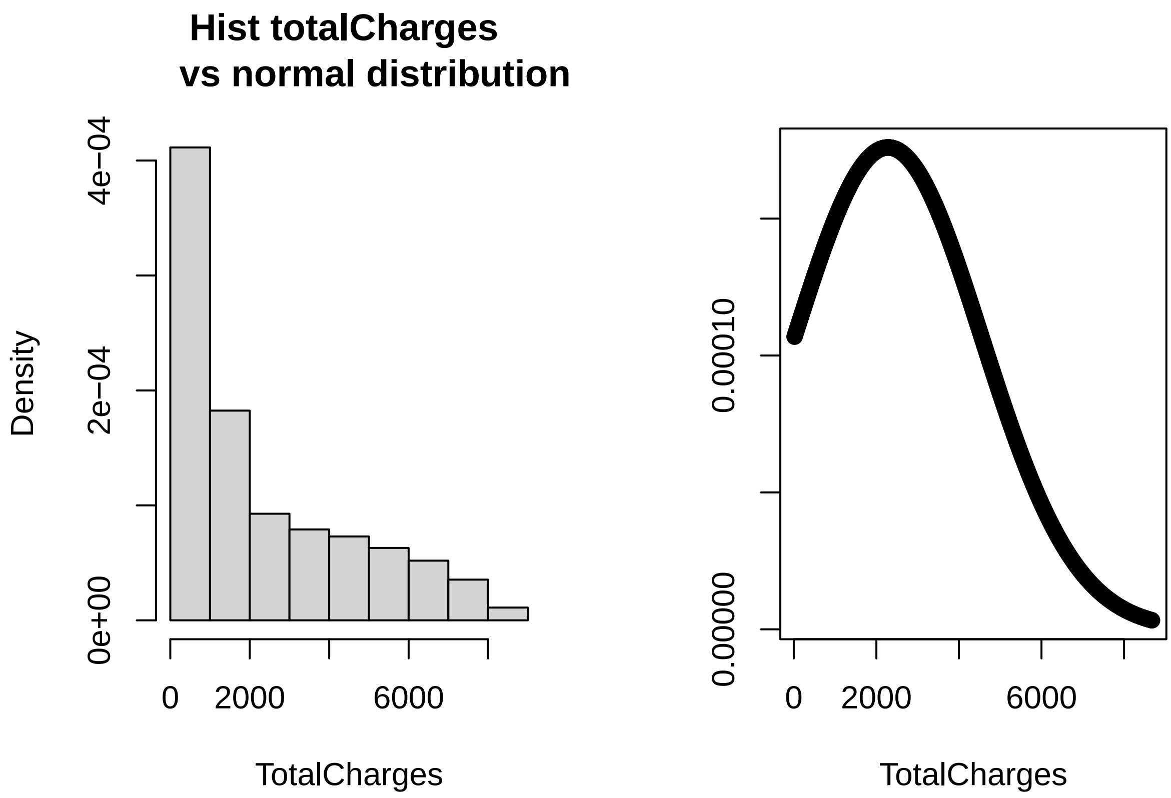
```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  df$tenure
## D = 0.88865, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```
# plots
par(mfrow=c(1,2))
hist(df$tenure, prob = TRUE, breaks = 10, main = 'Histogram of tenure
     vs normal distribution', xlab = 'tenure')
x <- seq(min(df$tenure), max(df$tenure), by = .1)
y <- dnorm(x, mean = mean(df$tenure), sd = sd(df$tenure))
plot(x,y, xlab = 'tenuere', ylab = '')
```

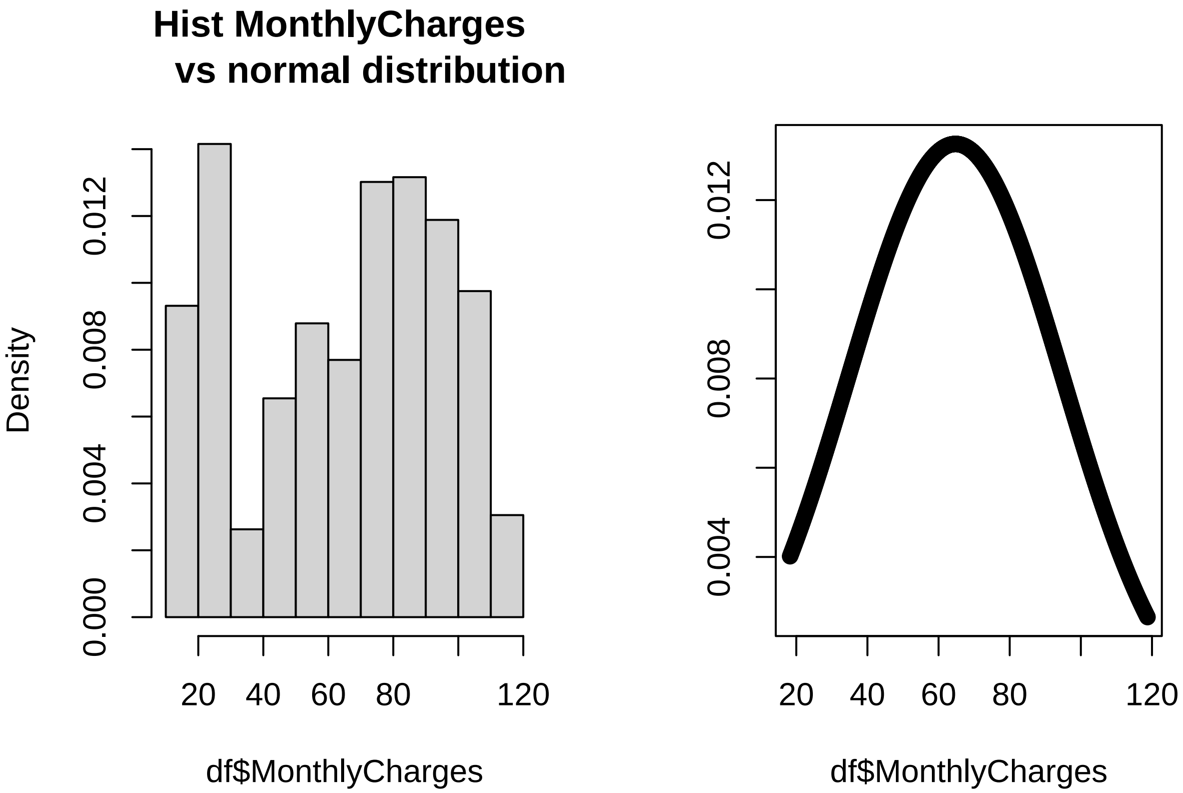


```
hist(df$TotalCharges, prob = TRUE, breaks = 10, main = 'Hist totalCharges
     vs normal distribution', xlab = 'TotalCharges')
```

```
x <- seq(min(df$TotalCharges, na.rm = TRUE), max(df$TotalCharges, na.rm = TRUE),
         by = 10)
y <- dnorm(x, mean = mean(df$TotalCharges, na.rm = TRUE), sd = sd(df$TotalCharges, na.rm
↪  = TRUE))
plot(x,y, xlab = 'TotalCharges', ylab = '')
```



```
hist(df$MonthlyCharges, prob = TRUE, breaks = 10, main = 'Hist MonthlyCharges
     vs normal distribution', xlab = 'df$MonthlyCharges')
x <- seq(min(df$MonthlyCharges, na.rm = TRUE), max(df$MonthlyCharges, na.rm = TRUE),
         by = .1)
y <- dnorm(x, mean = mean(df$MonthlyCharges, na.rm = TRUE), sd = sd(df$MonthlyCharges,
↪  na.rm = TRUE))
plot(x,y, xlab = 'df$MonthlyCharges', ylab = '')
```



```r
par(mfrow=c(1,1))
```

3. Data Quality Report

In this section we analysed the missing values, outliers and errors of numeric variables to increase the quality of data before modeling.

To start with, we detected that only “TotalCharges”, and hence “c.TotalCharges”, has a total of 22 missing observations. However, all of them correspond to new clients who have not receive their first invoice yet, so “TotalCharges” can not have a value. In other words, they are “not applicable cases”. We naturally impute this observations with 0.

```r
# Distribution of missings in df per variable
apply(sapply(df, is.na), 2, sum)
```

```
##       customerID           gender    SeniorCitizen          Partner
##                0                0                0                0
##       Dependents           tenure     PhoneService    MultipleLines
##                0                0                0                0
##  InternetService   OnlineSecurity     OnlineBackup DeviceProtection
##                0                0                0                0
##      TechSupport      StreamingTV  StreamingMovies         Contract
##                0                0                0                0
## PaperlessBilling    PaymentMethod   MonthlyCharges     TotalCharges
##                0                0                0               11
##            Churn         c.tenure   c.TotalCharges c.MonthlyCharges
##                0                0               11                0
```

```r
# Distribution of missings in df per individual
table(apply(sapply(df, is.na), 1, sum))
```

```
##
##    0    2
## 7032   11
```

```r
# Check that all missings in "TotalCharges" correspond to individuals tenure = 0
TotalCharges.na <- which(is.na(df$TotalCharges))
sum(TotalCharges.na == which(df$tenure == 0)) == length(TotalCharges.na)
```

```
## [1] TRUE
```

```r
# So we transform them after creating a new numeric variable with all the missings of the
↪  database
df$n.na <- apply(sapply(df, is.na), 1, sum)

df$TotalCharges[TotalCharges.na] = 0
df$c.TotalCharges[TotalCharges.na] = "(-1,2171]"
```

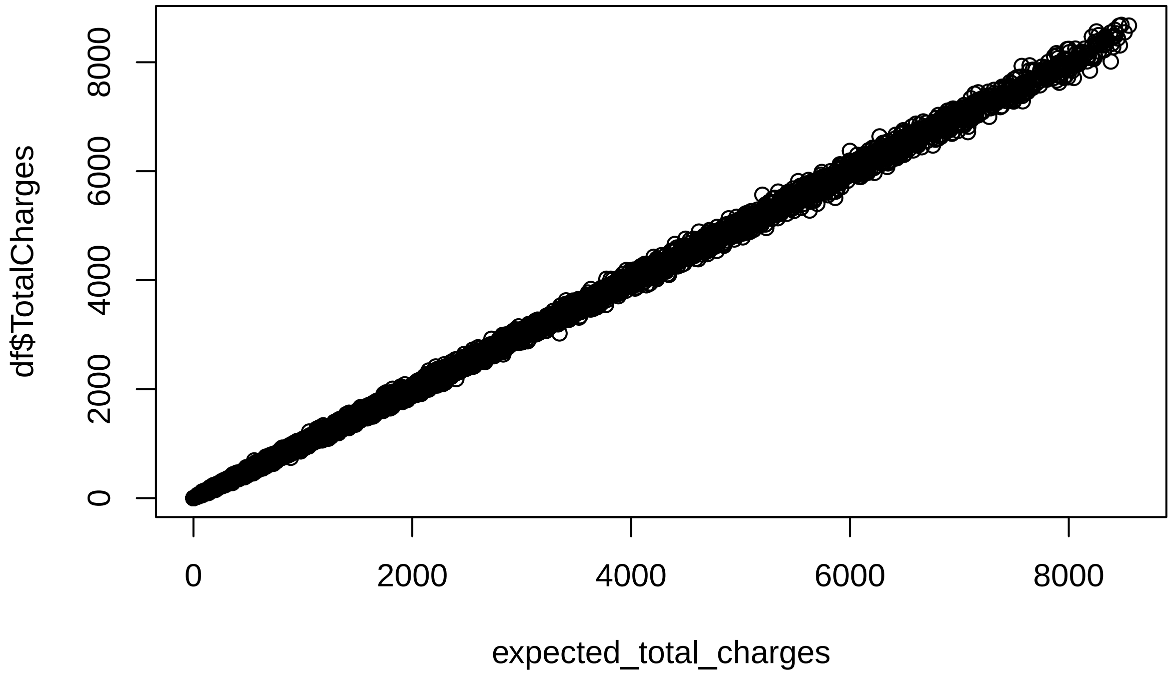
Secondly, we detected data inconsistencies. For categorical values, we checked the EDA automatic reports and the summaries to ensure that all qualitative variables categories were meaningful and that there was not any misspelling errors. We also checked that all values of numeric variables were positive and reasonable.

Additionally, for “TotalCharges” we ensured that all the values were correct by manually calculating the value and comparing it to the actual total charge.

```r
# Expected total charges as the product of monthly charges and tenure
expected_total_charges = df$MonthlyCharges * df$tenure

# Plot them against the actual total charges
```

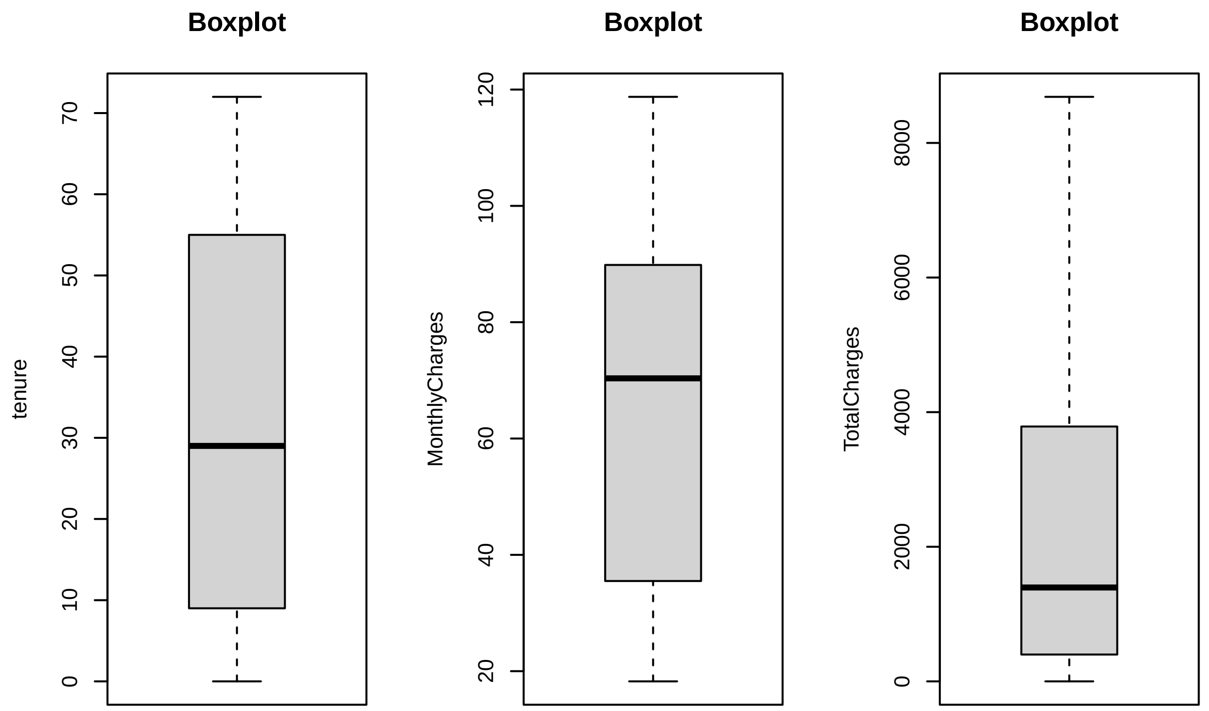
```
plot(expected_total_charges, df$TotalCharges)
```



```
# There are no outliers, so TotalCharges is consistent.
```

Thirdly, we analysed univariate outliers in numeric variables using Boxplots and the typical thresholds: 1.5 * IQR(interquartile range) for mild outliers and 3 * IQR for severe outliers. As there were not any we considered that all points were suitable for our models.

```
par(mfrow=c(1, length(numeric_val_idx)))
for (var in as.numeric(numeric_val_idx)) {
  Boxplot(df[,var], ylab = names(df)[var], main = "Boxplot")
}
```



```r
par(mfrow=c(1,1))
```

3.1 In depth analysis of missing values

Next, we will compute for every group of individuals the mean of missing values. Then we will rank the groups according to the computed mean.

```r
# c.TotalCharges has missings, so it doesn't make sense to compute the mean
# of missings in its categories

interesting_cat_idx <- categoric_val_idx[-c(1,20)]
k = 0
for (i in interesting_cat_idx){
  k <- k + length(levels(df[,i]))
}
groups.na <- matrix(0, k, 2)
l = 1
for (idx in interesting_cat_idx) {
  categories.na <- tapply(df$n.na, df[,idx], mean)
  for (j in seq(length(categories.na))) {
    groups.na[l + j - 1,] <- c(categories.na[j],
                               paste(names(df)[idx], levels(df[,idx])[j],
                                     sep = "."))
  }
  l <- l + j
}
```

```r
groups.na.df <- data.frame(na.perc = groups.na[,1], group = groups.na[,2])
groups.na.df[order(groups.na.df$na.perc, decreasing = TRUE),]
```

```
##                 na.perc                                   group
## 37   0.0117994100294985                       Contract.Two year
## 8    0.0104265402843602                          Dependents.Yes
## 43  0.00992555831265509              PaymentMethod.Mailed check
## 46  0.00804681784930505                        c.tenure.(-1,18]
## 16  0.00786369593709043                      InternetService.No
## 18  0.00786369593709043      OnlineSecurity.No internet service
## 21  0.00786369593709043        OnlineBackup.No internet service
## 24  0.00786369593709043    DeviceProtection.No internet service
## 27  0.00786369593709043         TechSupport.No internet service
## 30  0.00786369593709043         StreamingTV.No internet service
## 33  0.00786369593709043     StreamingMovies.No internet service
## 50  0.00734394124847001             c.MonthlyCharges.(18,30.69]
## 9   0.00586510263929619                         PhoneService.No
## 12  0.00586510263929619          MultipleLines.No phone service
## 38  0.00557103064066852                     PaperlessBilling.No
## 6   0.00529100529100529                             Partner.Yes
## 44  0.00425202937765752                                Churn.No
## 14  0.00413052457662123                     InternetService.DSL
## 19  0.00396235760277365                      OnlineSecurity.Yes
## 28  0.00391389432485323                         TechSupport.Yes
## 3   0.00372818166412472                       SeniorCitizen.Yes
## 2   0.00337552742616034                             gender.Male
## 51   0.0033112582781457          c.MonthlyCharges.(30.69,59.38]
## 25  0.00330305532617671                    DeviceProtection.Yes
## 22  0.00329353643474681                        OnlineBackup.Yes
## 31  0.00295530107129664                         StreamingTV.Yes
## 11  0.00294985250737463                        MultipleLines.No
## 32  0.00287253141831239                      StreamingMovies.No
## 1   0.00286697247706422                           gender.Female
## 10  0.00282974375098255                        PhoneService.Yes
## 13  0.00269269606193201                       MultipleLines.Yes
## 40  0.00259067357512953 PaymentMethod.Bank transfer (automatic)
## 52  0.00258955545964609          c.MonthlyCharges.(59.38,89.06]
## 39  0.00143850395588588                    PaperlessBilling.Yes
## 36  0.00135777325186694                       Contract.One year
## 41  0.00131406044678055   PaymentMethod.Credit card (automatic)
## 5   0.00109859928591046                              Partner.No
## 34 0.000732064421669107                     StreamingMovies.Yes
## 29 0.000711743772241993                          StreamingTV.No
## 20 0.000647668393782383                         OnlineBackup.No
## 23 0.000646203554119548                     DeviceProtection.No
## 26 0.000575871004894904                          TechSupport.No
## 17 0.000571755288736421                       OnlineSecurity.No
## 4                     0                        SeniorCitizen.No
## 7                     0                           Dependents.No
## 15                    0             InternetService.Fiber optic
## 35                    0                  Contract.Month-to-month
## 42                    0           PaymentMethod.Electronic check
## 45                    0                               Churn.Yes
## 47                        0                          c.tenure.(18,36]
## 48                        0                          c.tenure.(36,54]
## 49                        0                          c.tenure.(54,72]
## 53                        0              c.MonthlyCharges.(89.06,118.75]
```

The groups with the highest proportion of missing data are made of those individuals who:

- Have a two-year contract
- Have dependents
- Pay with a mailed check

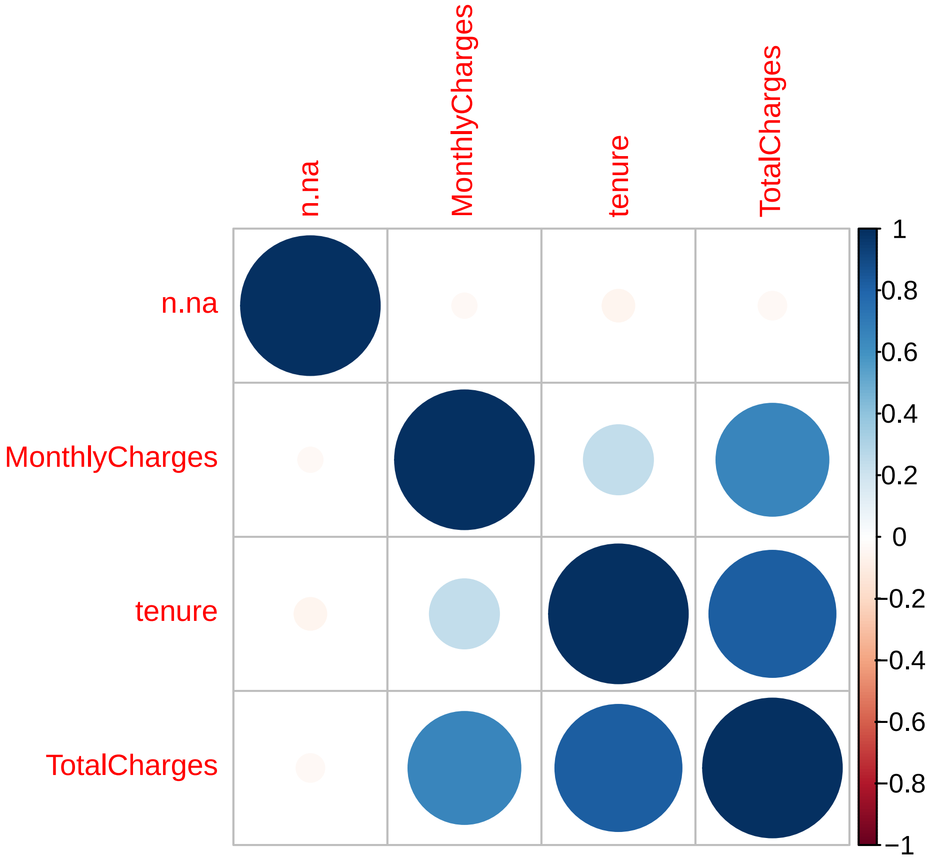
Since the set of individuals with missing data is exactly that of the new clients, we conclude that recently incorporated clients tend to: sign a two-year contract, have dependents and pay with a mailed check.

We can compute as well the pearson correlation coefficient between “n.na” and the numerical variables.

```
# Creation of the correlation matrix
corr_mat <- cor(df[,c(numeric_val_idx, 25)],)
corr_mat
```

```
##                     tenure MonthlyCharges TotalCharges        n.na
## tenure          1.00000000     0.24789986   0.82617840 -0.05213467
## MonthlyCharges  0.24789986     1.00000000   0.65117383 -0.03068535
## TotalCharges    0.82617840     0.65117383   1.00000000 -0.03977955
## n.na           -0.05213467    -0.03068535  -0.03977955  1.00000000
```

```
corrplot(corr_mat, order = 'hclust', tl.cex = 0.9)
```



n.na is independent to the rest of numerical variables, probably because it evaluates to 0 in most observations.

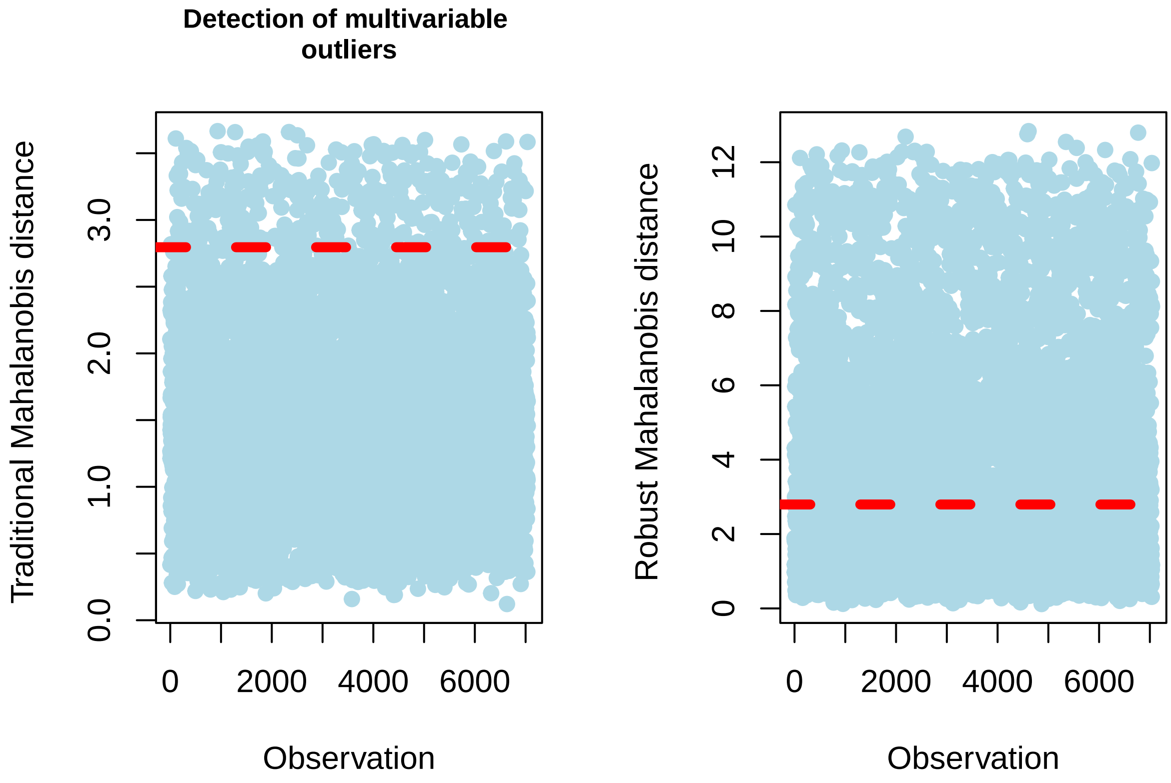
3.2 Multivariate outliers

In this section we focused on detecting the multivariate outliers using “Moutlier”. We discovered 344 multivariate outliers, about 5% of the individuals, as it was expected. We decided to maintain them and only remove them in the modeling step if they turned out to be influential points.

```
set.seed(123)
res.mout <- Moutlier(df[,numeric_val_idx], quantile = 0.95, plot= FALSE)

# Visual representation
par(mfrow=c(1,2), cex.main=0.8)
plot(res.mout$md, col="lightblue", pch = 19, main = 'Detection of multivariable
outliers', xlab= 'Observation',
     ylab ='Traditional Mahalanobis distance ')
abline(h = res.mout$cutoff, col = "red", lwd = 5, lty = 2)

plot(res.mout$rd, col="lightblue", pch = 19, xlab= 'Observation',
     ylab ='Robust Mahalanobis distance ')
abline(h = res.mout$cutoff, col = "red", lwd = 5, lty = 2)
```



```
par(mfrow=c(1,1), cex.main=1)

# Identification of the outliers
outliers = which(res.mout$md>res.mout$cutoff & res.mout$rd > res.mout$cutoff)
length(outliers)
```

```
## [1] 344
```

```r
length(outliers)/dim(df)[1]*100
```

```
## [1] 4.884282
```

4. Profiling of the target and feature selection

Numeric variables' correlations

We analysed the pearson correlation coefficient to detect variables that were highly related and not include them all in the model. In the correlation plot of section 3.1 we see that "TotalCharges" is highly correlated with "MonthlyCharges" and "tenure" as the first one is calculated as the product of the others.

Profiling of the target

Later on, we profiled the target Churn using a custom function "profiling()" created in the Multivariate Analysis subject of the Master's degree. This method expands "catdes()" and performs many plots and tests according to the type of each variable. We will focus on plots and the given tests' results: Chiˆ2, ANOVA and Kruskal-Wallis, which can be found in the annex.

```
# Analysis of all variables except the ID
profiling(df[-c(grep("customerID", names(df)), grep("Churn", names(df)))], df$Churn,
↪  "Churn")
```

The most relevant conclusions are: - Some variables are not significant, like Gender (Chiˆ2 p-value=0.4866) or Phone service (Chiˆ2 p-value=0.3388). Consequently, we state that churn is independent of the client's gender and whether he/she/they has a phone service contracted.

```
profiling(df[c(grep("gender", names(df)),grep("PhoneService", names(df)))], df$Churn,
↪  "Churn")
```

- There are variables like "MultipleLines" that even being significant (Chiˆ2 p-value=0.003464) the difference among levels is small, as we can see in the plots

```
profiling(df[grep("MultipleLines", names(df))], df$Churn, "Churn")
```

- The rest of variables, including the discretized, have a small p-value (< 2.2e-16) in the Chiˆ2, ANOVA or Kruskal-Wallis test, according to their type, and have at least one level where the target's distribution is different than in the rest. For example, 40% of people that did not have an online backup churned, while only 22% of customers having the backup did.

```
profiling(df[grep("OnlineBackup", names(df))], df$Churn, "Churn")
```

Feature Selection

Finally, we decided which variables were suitable to be included in the model.

The id was removed, since it will not give us any knowledge nor be useful to predict the target.

```
df$customerID <- NULL
```

We then computed the relationship between all the variables and the target with the "catdes()" method and chose the most relevant of them for the target's explanation.

All p-values of the Chi-squared test for categorical variables are very low, less than 0.001. The 6 variables with the lowest p-value are Contract, OnlineSecurity, TechSupport, c.tenure, InternetService, PaymentMethod. Note that the list includes a discretized numerical variable.

```
# Correlation between all variables and our qualitative target Churn.
res.cat = catdes(df, grep("Churn", names(df)))

# Most important categorical variables, sorted by p value
res.cat$test.chi2
```

```
##                          p.value df
## Contract         5.863038e-258  2
## OnlineSecurity   2.661150e-185  2
## TechSupport      1.443084e-180  2
## c.tenure         4.192004e-178  3
## InternetService  9.571788e-160  2
## PaymentMethod    3.682355e-140  3
## OnlineBackup     2.079759e-131  2
## DeviceProtection 5.505219e-122  2
## StreamingMovies   2.667757e-82  2
## StreamingTV       5.528994e-82  2
## c.MonthlyCharges  8.977393e-72  3
## PaperlessBilling  2.614597e-58  1
## Dependents        3.276083e-43  1
## c.TotalCharges    3.057813e-39  3
## SeniorCitizen     9.477904e-37  1
## Partner           1.519037e-36  1
## MultipleLines     3.464383e-03  2
```

As for numeric variables, “tenure” has the smallest p-value in the F-test, much lower than those of discrete variables. As we have already seen, there is a high correlation between “MonthlyCharges”, “tenure” and “TotalCharges” so we will only include in the models “TotalCharges” or “MonthlyCharges” together with “tenure”.

```
res.cat$quanti.var
```

```
##                      Eta2       P-value
## tenure         0.12406504 7.999058e-205
## TotalCharges   0.03933251  2.127212e-63
## MonthlyCharges 0.03738671  2.706646e-60
```

Profiling of the target with the selected categorical features

Lastly, we decided to make an extensive profiling of the six categorical variables that we could use in the model in order to understand them better. The main conclusions for each variable were:

- Contract: The probability of churning is decreased when the contract term increases. For example, if a costumer has a month contract and changes it to an annual the probability of not churning increases from 0.58 to 0.89.
- InternetService: People that do not have an internet service do not usually churn (7%). However, if they had a Fiber optic connection, the probability to churn increases (42%). This could be explained by the fact that users with a fast internet connection try to get the best offer for the service, but it would be necessary to make a market analysis to validate this hypothesis.
- OnlineSecurity: The probability of churning is small when the customer has online security. However, having an internet connection or not seems a more interesting feature than the variable itself, as the “No internet service” level has the smallest p-value.
- TechSupport: Having tech support increases the probability of not churning from 60% to 84% (when compared with not having it, although having internet service). Having internet service or not is, again, a more relevant feature.
- c.tenure: Loyalty is important, since people tend to churn less when they have spent longer with the service. For example, people who have spent less than 1.5 years has churned 44% of times, but only 8% of those who have stayed for more than 4.5 years have churned.

- PaymentMethod: The proportion of people that churned is very similar in all types of payment except for “Electronic check”. In this level, the proportion of churns is 45%, 18% higher than the global average.

```
# Global proportions of Churn categories
proportions(table(df$Churn))
```

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

```
# Calculate the indexes of the variables to investigate
names = c("Contract", "OnlineSecurity", "TechSupport", "c.tenure", "InternetService",
↪  "PaymentMethod")
index = NULL

for (i in 1:length(names)) {
  ind = grep(names[i], colnames(df))
  index = append(index, ind)
}
index = append(index, grep("Churn", names(df)))

# Profiling of only those variables
res.cat2 = catdes(df[,index], length(index))

res.cat2$category
```

```
## $No
##                                          Cla/Mod  Mod/Cla   Global
## Contract=Two year                       97.16814 31.83224 24.06645
## c.tenure=(54,72]                        92.02859 32.35408 25.82706
## InternetService=No                      92.59502 27.30963 21.66690
## TechSupport=No internet service         92.59502 27.30963 21.66690
## OnlineSecurity=No internet service      92.59502 27.30963 21.66690
## Contract=One year                       88.73048 25.26092 20.91438
## OnlineSecurity=Yes                      85.38881 33.32045 28.66676
## TechSupport=Yes                         84.83366 33.51372 29.02172
## PaymentMethod=Credit card (automatic)   84.75690 24.93235 21.61011
## InternetService=DSL                     81.04089 37.92037 34.37456
## PaymentMethod=Bank transfer (automatic) 83.29016 24.85504 21.92248
## PaymentMethod=Mailed check              80.89330 25.20294 22.88797
## c.tenure=(36,54]                        81.97970 18.72826 16.78262
## c.tenure=(18,36]                        77.29358 19.54001 18.57163
## PaymentMethod=Electronic check          54.71459 25.00966 33.57944
## InternetService=Fiber optic             58.10724 34.77000 43.95854
## c.tenure=(-1,18]                        55.59620 29.37766 38.81869
## TechSupport=No                          58.36453 39.17665 49.31137
## OnlineSecurity=No                       58.23328 39.36993 49.66634
## Contract=Month-to-month                 57.29032 42.90684 55.01917
##                                               p.value    v.test
## Contract=Two year                       3.588830e-187  29.178937
## c.tenure=(54,72]                        2.745248e-113  22.620153
## InternetService=No                       6.584621e-98  20.999812
## TechSupport=No internet service          6.584621e-98  20.999812
## OnlineSecurity=No internet service       6.584621e-98  20.999812
## Contract=One year                        3.593041e-57  15.935502
## OnlineSecurity=Yes                       1.606459e-50  14.947938
## TechSupport=Yes                          1.323174e-46  14.334963
## PaymentMethod=Credit card (automatic)    6.408166e-32  11.758206
## InternetService=DSL                      2.545367e-26  10.614727
## PaymentMethod=Bank transfer (automatic)  1.180908e-24  10.250207
## PaymentMethod=Mailed check               3.226893e-15   7.881803
## c.tenure=(36,54]                         6.217772e-14   7.503412
## c.tenure=(18,36]                         4.375264e-04   3.516348
## PaymentMethod=Electronic check          1.790860e-136 -24.864755
## InternetService=Fiber optic             2.289126e-148 -25.941138
## c.tenure=(-1,18]                        7.876341e-159 -26.852547
## TechSupport=No                          1.899538e-183 -28.883947
## OnlineSecurity=No                       6.171504e-190 -29.396034
## Contract=Month-to-month                 3.620915e-283 -35.959308
##
## $Yes
##                                           Cla/Mod   Mod/Cla   Global
## Contract=Month-to-month                 42.709677 88.550027 55.01917
## OnlineSecurity=No                       41.766724 78.170144 49.66634
## TechSupport=No                          41.635474 77.367576 49.31137
## c.tenure=(-1,18]                        44.403804 64.954521 38.81869
## InternetService=Fiber optic             41.892765 69.395399 43.95854
## PaymentMethod=Electronic check          45.285412 57.303371 33.57944
## c.tenure=(18,36]                        22.706422 15.890851 18.57163
## c.tenure=(36,54]                        18.020305 11.396469 16.78262
## PaymentMethod=Mailed check              19.106700 16.479401 22.88797
## PaymentMethod=Bank transfer (automatic) 16.709845 13.804173 21.92248
## InternetService=DSL                     18.959108 24.558587 34.37456
## PaymentMethod=Credit card (automatic)   15.243101 12.413055 21.61011
## TechSupport=Yes                         15.166341 16.586410 29.02172
## OnlineSecurity=Yes                      14.611194 15.783842 28.66676
## Contract=One year                       11.269518  8.881755 20.91438
## InternetService=No                       7.404980  6.046014 21.66690
## TechSupport=No internet service          7.404980  6.046014 21.66690
## OnlineSecurity=No internet service       7.404980  6.046014 21.66690
## c.tenure=(54,72]                         7.971413  7.758159 25.82706
## Contract=Two year                        2.831858  2.568218 24.06645
##                                               p.value     v.test
## Contract=Month-to-month                 3.620915e-283  35.959308
## OnlineSecurity=No                       6.171504e-190  29.396034
## TechSupport=No                          1.899538e-183  28.883947
## c.tenure=(-1,18]                        7.876341e-159  26.852547
## InternetService=Fiber optic             2.289126e-148  25.941138
## PaymentMethod=Electronic check          1.790860e-136  24.864755
## c.tenure=(18,36]                         4.375264e-04  -3.516348
## c.tenure=(36,54]                         6.217772e-14  -7.503412
## PaymentMethod=Mailed check               3.226893e-15  -7.881803
## PaymentMethod=Bank transfer (automatic)  1.180908e-24 -10.250207
## InternetService=DSL                      2.545367e-26 -10.614727
## PaymentMethod=Credit card (automatic)    6.408166e-32 -11.758206
## TechSupport=Yes                          1.323174e-46 -14.334963
## OnlineSecurity=Yes                       1.606459e-50 -14.947938
## Contract=One year                        3.593041e-57 -15.935502
## InternetService=No                       6.584621e-98 -20.999812
## TechSupport=No internet service          6.584621e-98 -20.999812
## OnlineSecurity=No internet service       6.584621e-98 -20.999812
## c.tenure=(54,72]                        2.745248e-113 -22.620153
## Contract=Two year                       3.588830e-187 -29.178937
```

```
# Another visualization of the profiling
#profiling(df[,index], df$Churn, "Churn")
```

5. Modeling

Data splitting

First, let's split the dataset into training and testing set. We have decided that 70% of the data will be used for training.

```
set.seed(123)

sampling = sample.split(df$Churn, SplitRatio = 0.7)
train = subset(df, sampling == TRUE)
test = subset(df, sampling == FALSE)
```

Modeling only with numerical variables

As we mentioned, there is a strong correlation between {tenure, MonthlyCharges} and TotalCharges, as the second one is simply the product of the variables in the first set. Hence, we will build two models, one for each set of variables, and keep the best one.

```
m0.set1 = glm (Churn ~ tenure + MonthlyCharges, data = train, family = binomial)
# Checking the Anova test, both variables are significant to our model.
# Hence, we won't remove any of them.
Anova(m0.set1, test = "LR")
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: Churn
##                LR Chisq Df Pr(>Chisq)
## tenure          1071.50  1  < 2.2e-16 ***
## MonthlyCharges   583.55  1  < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
m0.set2 = glm (Churn ~ TotalCharges, data = train, family = binomial)

BIC(m0.set1, m0.set2)
```

```
##         df      BIC
## m0.set1  3 4444.286
## m0.set2  2 5504.792
```

Checking the Bayesian criterion, the set {tenure, MonthlyCharges} has a much lower value and its variables are significant. Hence, we'll choose this set of variables for further analysis.

We also check possible transformation for our model m0.set1.

```
m0.log = glm (Churn ~ tenure + log(MonthlyCharges), data = train, family = binomial)
m0.sqrt = glm (Churn ~ sqrt(tenure) + MonthlyCharges, data = train, family = binomial)

BIC (m0.set1, m0.log, m0.sqrt)
```

```
##         df      BIC
## m0.set1  3 4444.286
## m0.log   3 4465.685
## m0.sqrt  3 4397.700
```

We have tried several transformations for both variables (sqrt, log, exp, etc), but BIC shows that the best model is the one with sqrt on tenure.

Discretized variables might create a better model, so we study this possibility.

```
m1 = glm (Churn ~ c.tenure + MonthlyCharges, data = train, family = binomial)

BIC(m1, m0.sqrt)
```

```
##         df      BIC
## m1       5 4585.287
## m0.sqrt  3 4397.700
```

Checking the AIC and BIC parameters, we decided to keep the numerical version of tenure. We have checked as well the model with MonthlyCharges discretized, but the AIC is worse once more.

Residual analysis only with numerical variables

It is important to look for influential points in the model that could worsen it. "influencePlot()" computes the Cook's distance of each point, so that we can compare them with the threshold studied in the course.

```
# Check influential points
influent = influencePlot(m0.sqrt)[3]; influent
```

```
##             CookD
## 269  0.0059784386
## 431  0.0061523136
## 3827 0.0005332286
## 4381 0.0004744407
```

```
# Calculate D's threshold
D_thresh <- 2/sqrt(dim(train)[1]); D_thresh
```

```
## [1] 0.02848436
```

The Cook's distances obtained from "influencePlot()" are smaller than our threshold, so we will not remove any point.

Adding factor main effects to the model

After being satisfied with our final model based on numerical variables, we add categorical variables to it in decreasing relevance order.

```
# Adding {Contract}
m2 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract,
          data = train, family = binomial)

# Adding {contract} indeed reduces the BIC of our model.
BIC(m0.sqrt, m2)
```

```
##         df      BIC
## m0.sqrt  3 4397.700
## m2       5 4217.893
```

```
# Adding {InternetService}
m3 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService,
          data = train, family = binomial)

# Adding {InternetService} indeed reduces the BIC of our model.
BIC(m2,m3)
```

```
##    df      BIC
## m2  5 4217.893
## m3  7 4190.321
```

We have figured out in the profiling section that {InternetService} and {OnlineSecurity, TechSupport} have some levels that are strongly correlated. Specifically, when "InternetService" = "No", "OnlineSecurity and"TechSupport" can't be given a value, so they are declared as "No intervet service".

To avoid multicollinearity and NA's, we need to decide which variable to keep.

```
# Adding {TechSupport, OnlineSecurity}
m4 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + OnlineSecurity
          + TechSupport, data = train, family = binomial)

BIC(m3, m4)
```

```
##    df      BIC
## m3  7 4190.321
## m4  8 4153.715
```

The BIC criterion for m4 is smaller, but taking into account that "InternetService" is more correlated with the target variable and the difference in the BIC is not that significant, we decided to keep m3, with "InternetService".

```
# Adding {PaymentMethod}
m5 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService
          + PaymentMethod, data = train, family = binomial)

# Adding {PaymentMethod} indeed reduces the BIC of our model.
BIC(m3, m5)
```

```
##    df      BIC
## m3  7 4190.321
## m5 10 4174.603
```

Residual analysis with categorical variables

We repeat the residual analysis performed earlier with our current model.

```
# Check influential points
influent = influencePlot(m5)[3]; influent
```

```
##             CookD
## 269  0.0055072625
## 937  0.0002690920
## 4273 0.0054160023
## 6755 0.0002337059
```

```
# Calculate D's threshold
D_thresh <- 2/sqrt(dim(train)[1]); D_thresh
```

```
## [1] 0.02848436
```

As before, the Cook's distances obtained from "influencePlot()" are smaller than our threshold, so we will not remove any point.

Adding interactions to the model

Sometimes interactions between dependent variables improve a model, so let us see how they work in our case. To start with, we check all possible interactions and execute "step()" to end up with the most relevant ones.

```
# Check all possible interactions of model m5
m6 = glm (Churn ~ (sqrt(tenure) + MonthlyCharges + Contract + InternetService
          + PaymentMethod)^2, data = train, family = binomial)

# Use step function to find the combination that minimizes the AIC.
step(m6)
```

```
## Start:  AIC=4081.75
## Churn ~ (sqrt(tenure) + MonthlyCharges + Contract + InternetService +
##     PaymentMethod)^2
##
##                                  Df Deviance    AIC
## - Contract:PaymentMethod          6   4006.0 4076.0
## - sqrt(tenure):InternetService    2   4000.4 4078.4
## - sqrt(tenure):MonthlyCharges     1   3999.7 4079.7
## - InternetService:PaymentMethod   6   4010.3 4080.3
## - MonthlyCharges:PaymentMethod    3   4004.7 4080.7
## <none>                                3999.7 4081.7
## - Contract:InternetService        4   4013.3 4087.3
## - MonthlyCharges:Contract         2   4009.6 4087.6
## - sqrt(tenure):Contract           2   4009.6 4087.6
## - sqrt(tenure):PaymentMethod      3   4012.2 4088.2
## - MonthlyCharges:InternetService  2   4031.4 4109.4
##
## Step:  AIC=4076.02
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +
##     PaymentMethod + sqrt(tenure):MonthlyCharges + sqrt(tenure):Contract +
##     sqrt(tenure):InternetService + sqrt(tenure):PaymentMethod +
##     MonthlyCharges:Contract + MonthlyCharges:InternetService +
##     MonthlyCharges:PaymentMethod + Contract:InternetService +
##     InternetService:PaymentMethod
##
##                                  Df Deviance    AIC
## - sqrt(tenure):InternetService    2   4006.9 4072.9
## - InternetService:PaymentMethod   6   4015.8 4073.8
## - sqrt(tenure):MonthlyCharges     1   4006.0 4074.0
## - MonthlyCharges:PaymentMethod    3   4010.9 4074.9
## <none>                                4006.0 4076.0
## - Contract:InternetService        4   4020.0 4082.0
## - sqrt(tenure):Contract           2   4016.5 4082.5
## - sqrt(tenure):PaymentMethod      3   4018.6 4082.6
## - MonthlyCharges:Contract         2   4016.6 4082.6
## - MonthlyCharges:InternetService  2   4037.4 4103.4
##
## Step:  AIC=4072.85
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +
##     PaymentMethod + sqrt(tenure):MonthlyCharges + sqrt(tenure):Contract +
##     sqrt(tenure):PaymentMethod + MonthlyCharges:Contract + MonthlyCharges:InternetService +
##     MonthlyCharges:PaymentMethod + Contract:InternetService +
##     InternetService:PaymentMethod
```

```
##
##                                   Df Deviance    AIC
## - sqrt(tenure):MonthlyCharges     1   4006.9 4070.9
## - InternetService:PaymentMethod   6   4017.3 4071.3
## - MonthlyCharges:PaymentMethod    3   4011.8 4071.8
## <none>                                4006.9 4072.9
## - sqrt(tenure):PaymentMethod      3   4020.1 4080.1
## - MonthlyCharges:Contract         2   4018.4 4080.4
## - sqrt(tenure):Contract           2   4018.5 4080.5
## - Contract:InternetService        4   4025.3 4083.3
## - MonthlyCharges:InternetService  2   4038.1 4100.1
##
## Step:  AIC=4070.87
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +
##     PaymentMethod + sqrt(tenure):Contract + sqrt(tenure):PaymentMethod +
##     MonthlyCharges:Contract + MonthlyCharges:InternetService +
##     MonthlyCharges:PaymentMethod + Contract:InternetService +
##     InternetService:PaymentMethod
##
##                                   Df Deviance    AIC
## - InternetService:PaymentMethod   6   4017.3 4069.3
## - MonthlyCharges:PaymentMethod    3   4011.8 4069.8
## <none>                                4006.9 4070.9
## - sqrt(tenure):PaymentMethod      3   4020.6 4078.6
## - MonthlyCharges:Contract         2   4018.6 4078.6
## - sqrt(tenure):Contract           2   4018.8 4078.8
## - Contract:InternetService        4   4025.3 4081.3
## - MonthlyCharges:InternetService  2   4041.3 4101.3
##
## Step:  AIC=4069.3
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +
##     PaymentMethod + sqrt(tenure):Contract + sqrt(tenure):PaymentMethod +
##     MonthlyCharges:Contract + MonthlyCharges:InternetService +
##     MonthlyCharges:PaymentMethod + Contract:InternetService
##
##                                   Df Deviance    AIC
## - MonthlyCharges:PaymentMethod    3   4022.7 4068.7
## <none>                                4017.3 4069.3
## - sqrt(tenure):PaymentMethod      3   4029.1 4075.1
## - sqrt(tenure):Contract           2   4028.5 4076.5
## - MonthlyCharges:Contract         2   4028.8 4076.8
## - Contract:InternetService        4   4037.1 4081.1
## - MonthlyCharges:InternetService  2   4049.3 4097.3
##
## Step:  AIC=4068.74
## Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService +
##     PaymentMethod + sqrt(tenure):Contract + sqrt(tenure):PaymentMethod +
##     MonthlyCharges:Contract + MonthlyCharges:InternetService +
##     Contract:InternetService
##
##                                   Df Deviance    AIC
## <none>                                4022.7 4068.7
## - sqrt(tenure):PaymentMethod      3   4031.4 4071.4
## - sqrt(tenure):Contract           2   4033.0 4075.0
## - MonthlyCharges:Contract         2   4035.2 4077.2
## - Contract:InternetService        4   4042.4 4080.4
## - MonthlyCharges:InternetService  2   4055.1 4097.1
##
## Call:  glm(formula = Churn ~ sqrt(tenure) + MonthlyCharges + Contract +
##     InternetService + PaymentMethod + sqrt(tenure):Contract +
##     sqrt(tenure):PaymentMethod + MonthlyCharges:Contract + MonthlyCharges:InternetService +
##     Contract:InternetService, family = binomial, data = train)
##
## Coefficients:
##                                      (Intercept)
##                                         1.339900
##                                     sqrt(tenure)
##                                        -0.394175
##                                   MonthlyCharges
##                                        -0.018996
##                                 ContractOne year
##                                        -3.421120
##                                 ContractTwo year
##                                        -3.924823
##                       InternetServiceFiber optic
##                                        -1.692574
##                                InternetServiceNo
##                                        -2.818093
##             PaymentMethodCredit card (automatic)
##                                        -0.271243
##                    PaymentMethodElectronic check
##                                         0.193796
##                        PaymentMethodMailed check
##                                         0.090168
##                    sqrt(tenure):ContractOne year
##                                         0.117605
##                    sqrt(tenure):ContractTwo year
##                                         0.428931
## sqrt(tenure):PaymentMethodCredit card (automatic)
##                                         0.053065
##       sqrt(tenure):PaymentMethodElectronic check
##                                         0.055748
##           sqrt(tenure):PaymentMethodMailed check
##                                        -0.104976
##                  MonthlyCharges:ContractOne year
##                                         0.039591
##                  MonthlyCharges:ContractTwo year
##                                        -0.004834
##        MonthlyCharges:InternetServiceFiber optic
##                                         0.041717
##                 MonthlyCharges:InternetServiceNo
##                                         0.064961
##      ContractOne year:InternetServiceFiber optic
##                                        -2.259452
##      ContractTwo year:InternetServiceFiber optic
##                                        -0.687971
##               ContractOne year:InternetServiceNo
##                                           0.989583
##                 ContractTwo year:InternetServiceNo
##                                          -0.411802
##
## Degrees of Freedom: 4929 Total (i.e. Null);  4907 Residual
## Null Deviance:       5704
## Residual Deviance: 4023  AIC: 4069
```

Even though “step()” recommends not to add any interaction, we see how the ones with the smallest AIC perform. That is, we add the interactions between “sqrt(tenure)” and “PaymentMethod” or “Contract”.

```
m7 = glm(Churn ~ sqrt(tenure) * PaymentMethod + sqrt(tenure) * Contract +
          MonthlyCharges + InternetService + PaymentMethod, data = train,
          family = binomial)

BIC (m5, m7)
```

```
##    df      BIC
## m5 10 4174.603
## m7 15 4194.054
```

According to the BIC criterion, no improvement is obtained.

Now we will add the interaction with the highest AIC instead, “MonthlyCharges:InternetService”.

```
m8 = glm(Churn ~ sqrt(tenure) +  Contract + MonthlyCharges * InternetService
         + PaymentMethod, data = train, family = binomial)

BIC(m5,m8)
```

```
##    df      BIC
## m5 10 4174.603
## m8 12 4164.660
```

```
summary(m8)
```

```
##
## Call:
## glm(formula = Churn ~ sqrt(tenure) + Contract + MonthlyCharges *
##     InternetService + PaymentMethod, family = binomial, data = train)
##
## Coefficients:
##                                            Estimate Std. Error z value Pr(>|z|)
## (Intercept)                                0.750286   0.297440   2.522  0.01165
## sqrt(tenure)                              -0.348759   0.023305 -14.965  < 2e-16
## ContractOne year                          -0.841360   0.128380  -6.554 5.61e-11
## ContractTwo year                          -1.728150   0.203049  -8.511  < 2e-16
## MonthlyCharges                            -0.009692   0.005254  -1.845  0.06507
## InternetServiceFiber optic                -1.504022   0.481191  -3.126  0.00177
## InternetServiceNo                         -2.268409   1.568535  -1.446  0.14812
## PaymentMethodCredit card (automatic)       0.011553   0.136266   0.085  0.93243
## PaymentMethodElectronic check              0.453998   0.113412   4.003 6.25e-05
## PaymentMethodMailed check                 -0.154366   0.138451  -1.115  0.26487
## MonthlyCharges:InternetServiceFiber optic  0.034298   0.006722   5.103 3.35e-07
## MonthlyCharges:InternetServiceNo           0.049096   0.075285   0.652  0.51431
##
## (Intercept)                               *
## sqrt(tenure)                            ***
## ContractOne year                        ***
## ContractTwo year                        ***
## MonthlyCharges                          .
## InternetServiceFiber optic              **
## InternetServiceNo
## PaymentMethodCredit card (automatic)
## PaymentMethodElectronic check           ***
## PaymentMethodMailed check
## MonthlyCharges:InternetServiceFiber optic ***
## MonthlyCharges:InternetServiceNo
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 5704.4  on 4929  degrees of freedom
## Residual deviance: 4062.6  on 4918  degrees of freedom
## AIC: 4086.6
##
## Number of Fisher Scoring iterations: 6
```

The BIC improved from 4174 to 4164, but with the cost of 2 degrees of freedom. Adding the interaction between “MonthlyCharges” and “InternetService” is a trade-off between simplicity and accuracy. At this point, after having added many variables, we value more simplicity, so we will not add this interaction.

Trying link function probit

We are interested in the effect of changing the link function of the logistic regression to probit.

```
m9 = glm (Churn ~ sqrt(tenure) + MonthlyCharges + Contract + InternetService
          + PaymentMethod, data = train, family = binomial(link = "probit"))

BIC(m5, m9)
```

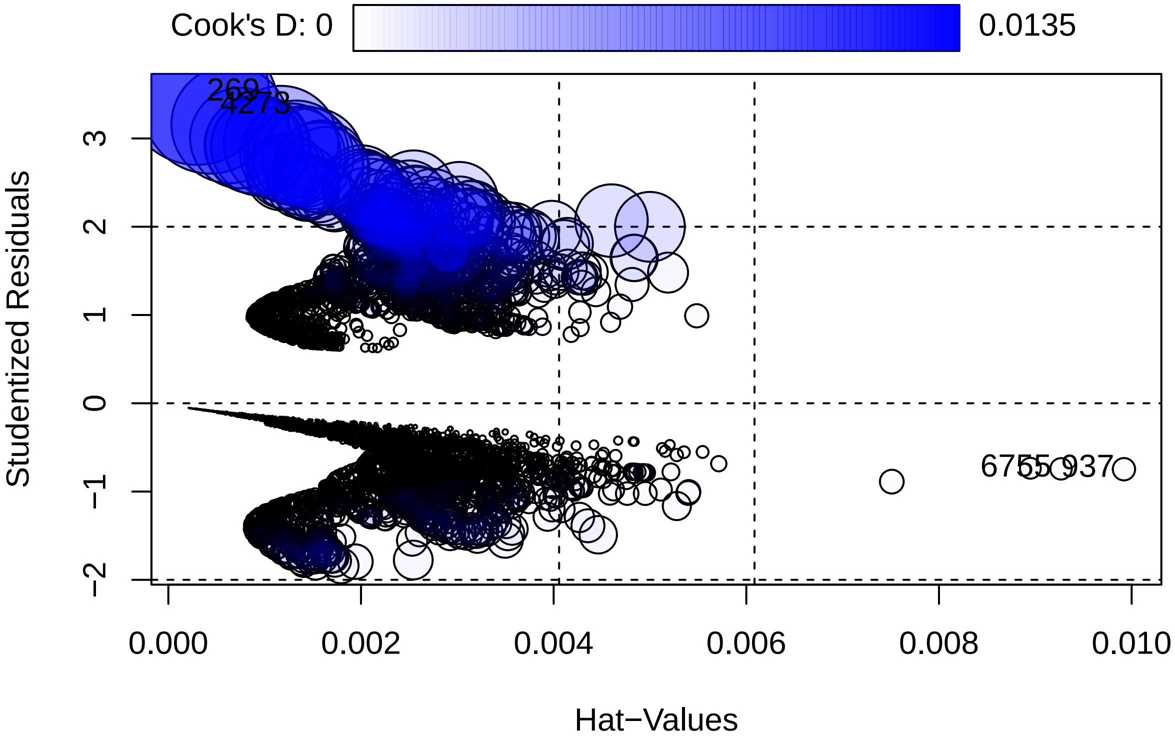
```
##    df      BIC
## m5 10 4174.603
## m9 10 4177.774
```

Sadly, based on the BIC criterion, no improvement is obtained.

Final residual analysis

We will perform now a final residual analysis.

```
# Check influential points
influent = influencePlot(m9)[3]; influent
```



```
##              CookD
## 269  0.0134537052
## 937  0.0003234289
## 4273 0.0119056685
## 6755 0.0002981121
```

```
# Calculate D's threshold
D_thresh <- 2/sqrt(dim(train)[1]); D_thresh
```

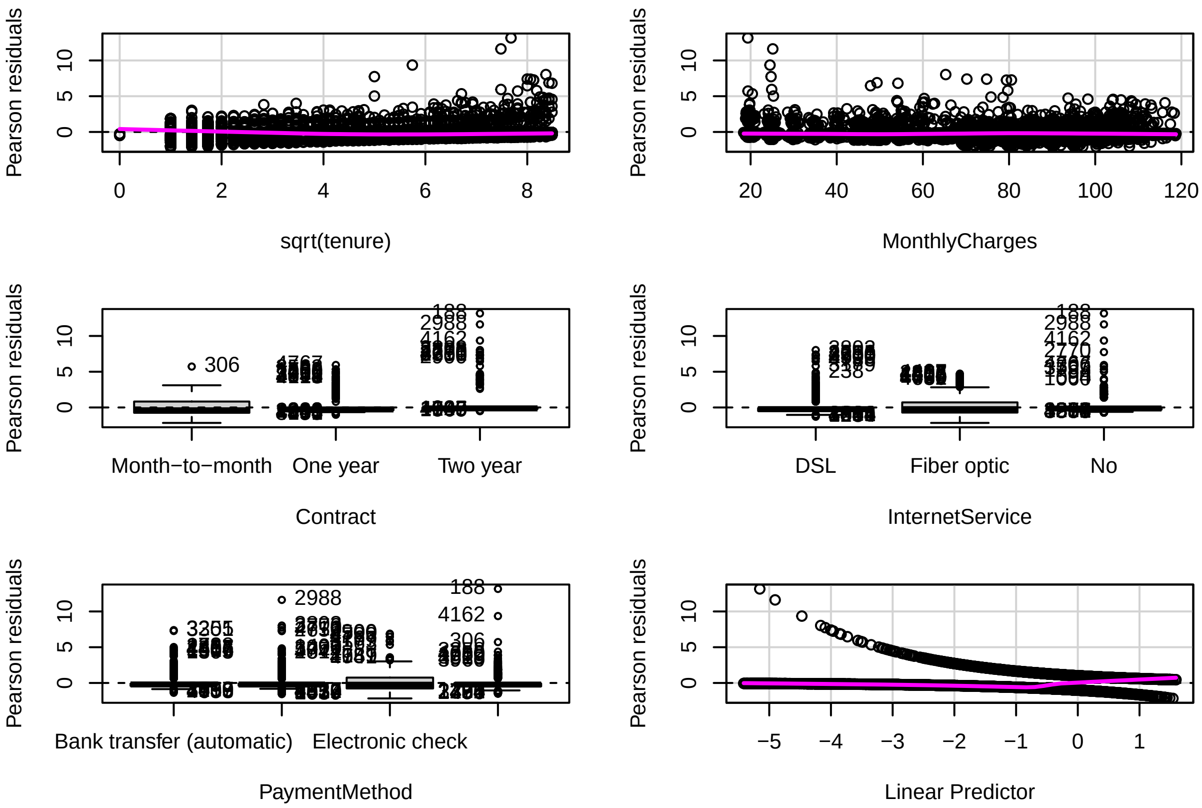
```
## [1] 0.02848436
```

Observations 269 and 4273 may be influential points, but both of them are smaller than the threshold.

6. Goodness of fit

NOT FINISHED FROM HERE UNTIL THE ANNEX

```r
residualPlots(m5)
```



```
##                  Test stat Pr(>|Test stat|)
## sqrt(tenure)        21.586        3.384e-06 ***
## MonthlyCharges      21.877        2.907e-06 ***
## Contract
## InternetService
## PaymentMethod
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model prediction

```r
# First, we compute the probability of Churn for each observation (from test)
# with predict function.
predictions = predict(m5, test, type = "response")

# Then, for those that have a probability higher than 0.5, we can consider
# Churn == "Yes"
probability = ifelse(predictions >= 0.5, "Yes", "No")

# Finally, compute the Confusion Matrix of predicted result.
```

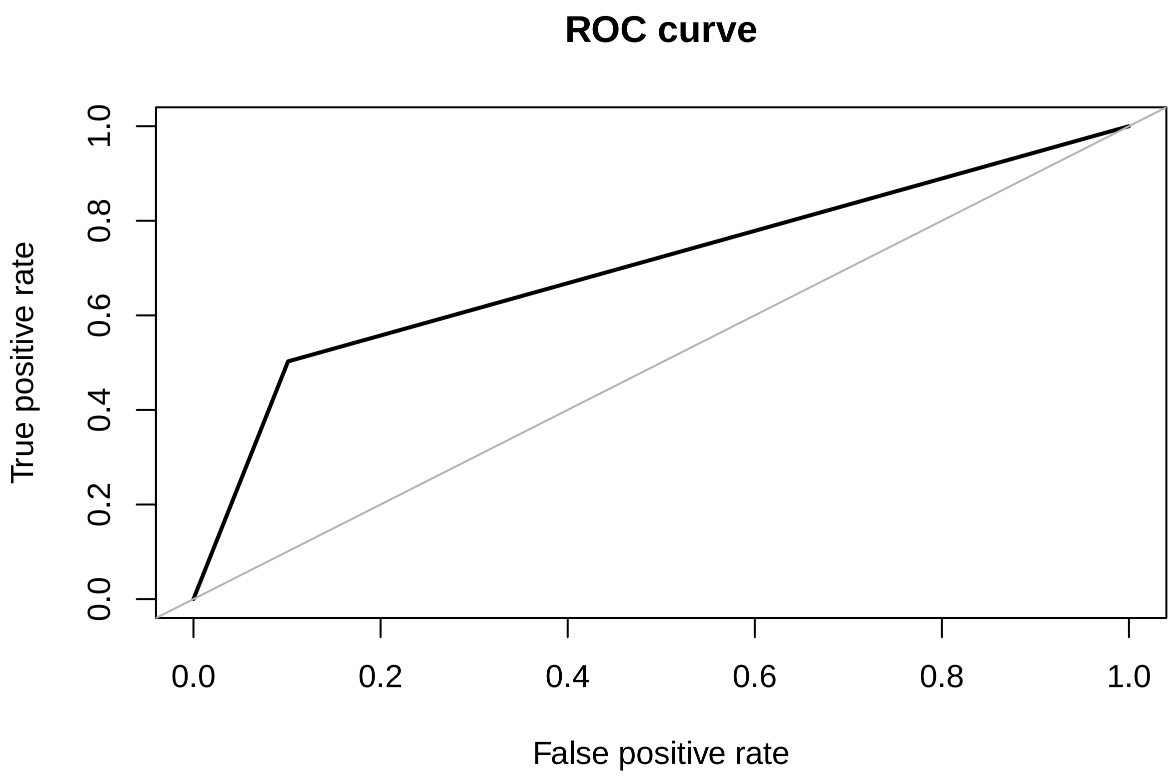
```
CM = table(test$Churn, probability, dnn = c("Actual Churn", "Predicted Churn")); CM
```

```
##             Predicted Churn
## Actual Churn   No  Yes
##          No  1395  157
##          Yes  279  282
```

```
accuracy = sum(diag(CM))/dim(test)[1]*100; accuracy
```

```
## [1] 79.36583
```

```
roc.curve(test$Churn, probability)
```



```
## Area under the curve (AUC): 0.701
```

```
library(DescTools)
PseudoR2(m5, which = "McFadden")
```

```
##  McFadden
## 0.2830884
```

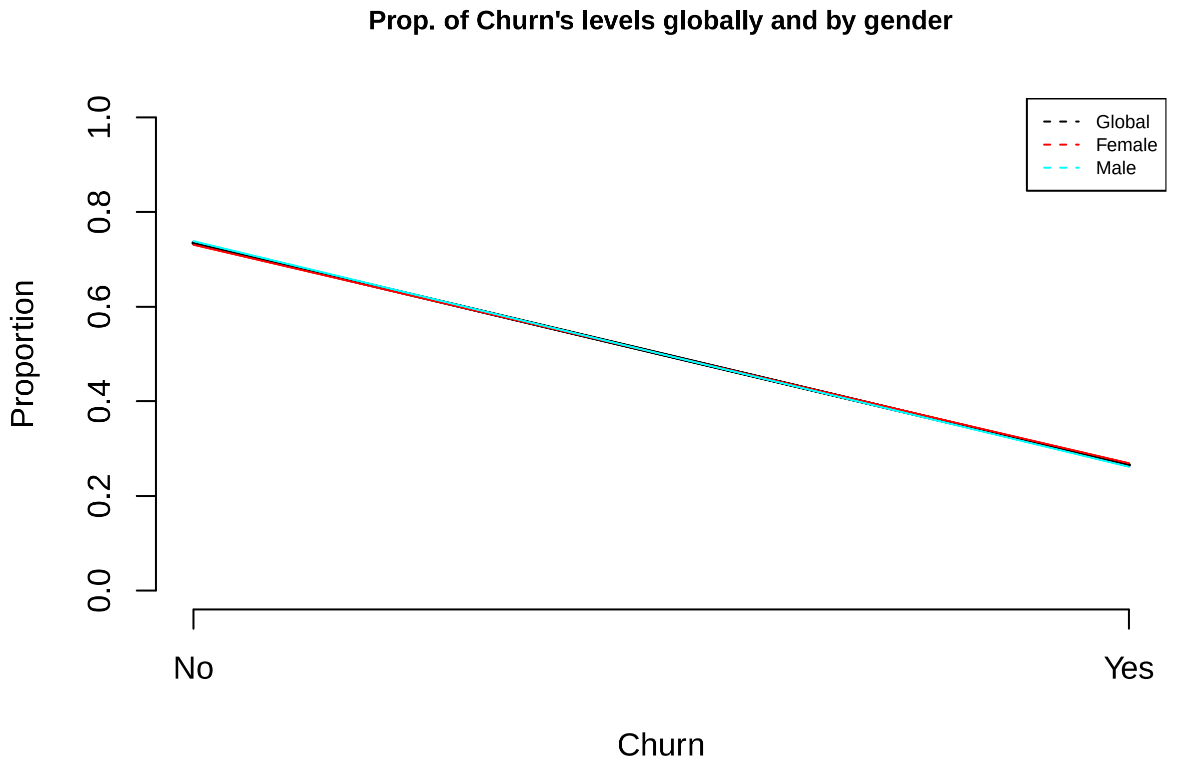
7. Model interpretation

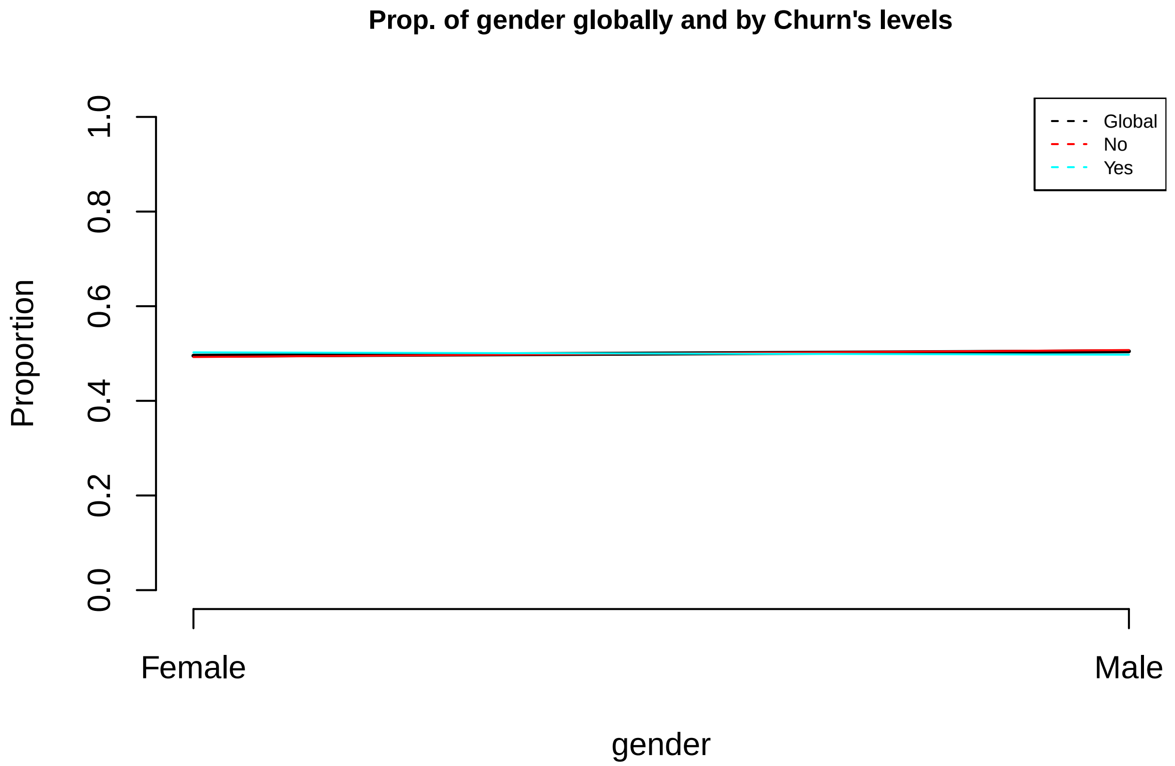
Annex

Expanded profiling of the target with the "profiling()" method

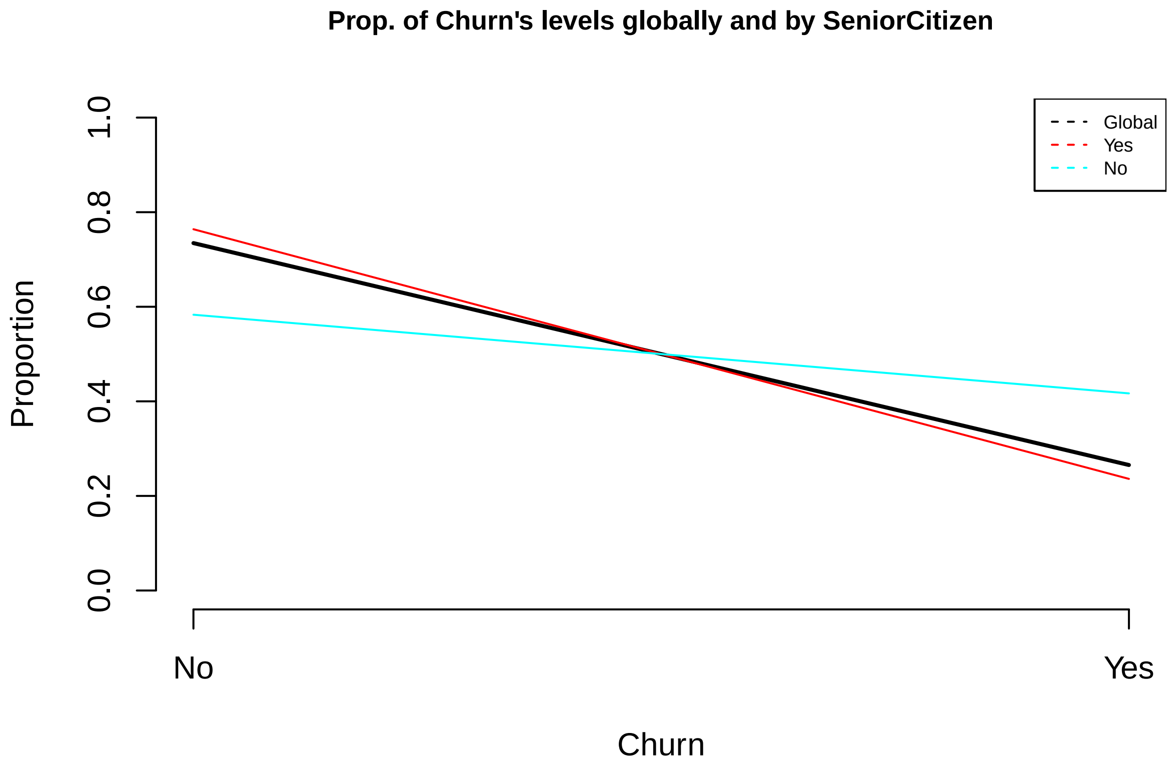
```
# Analysis of all variables except the ID
profiling(df[-c(grep("customerID", names(df)), grep("Churn", names(df)))], df$Churn,
↪  "Churn")
```

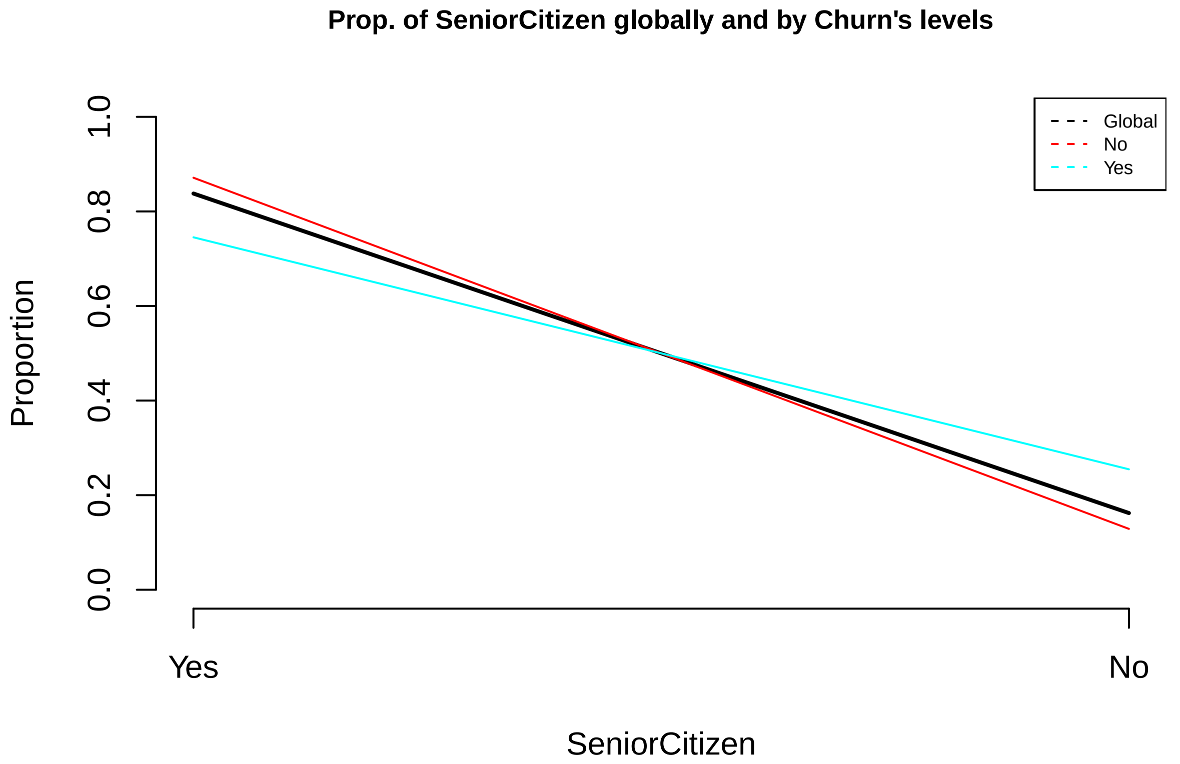
```
## [1] "Variable gender"
## [1] "Categories=" "Female"      "Male"
```



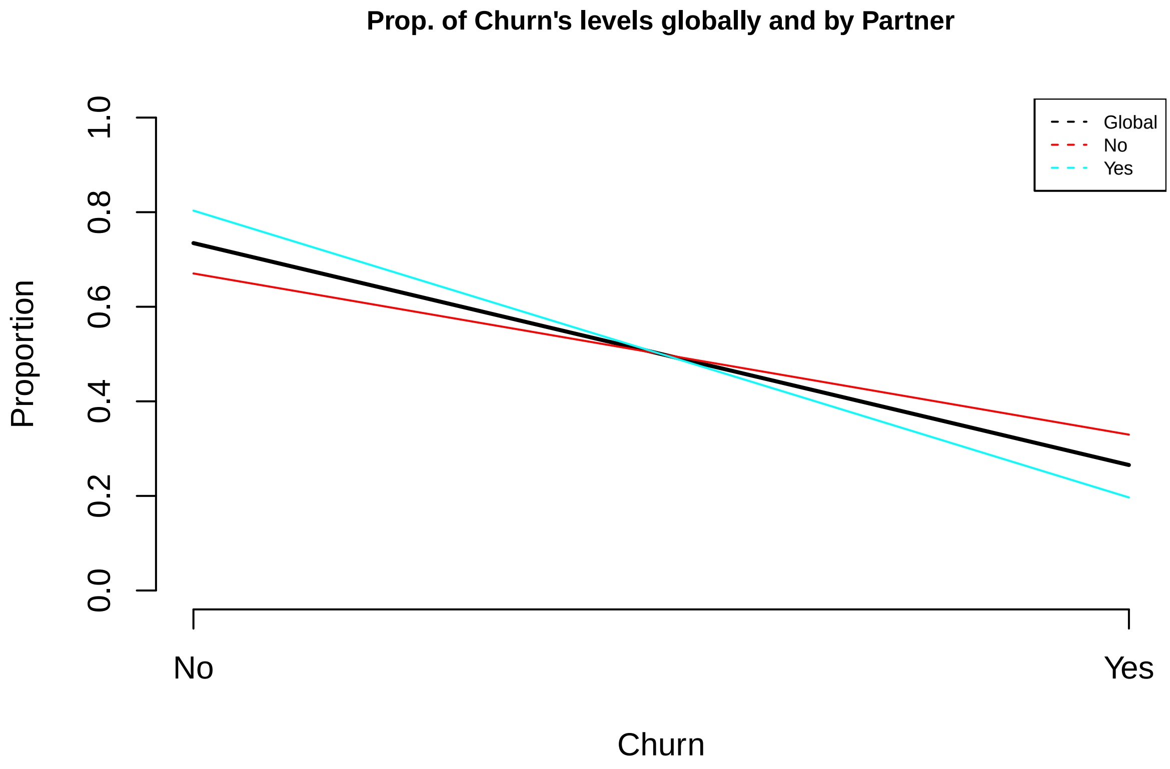


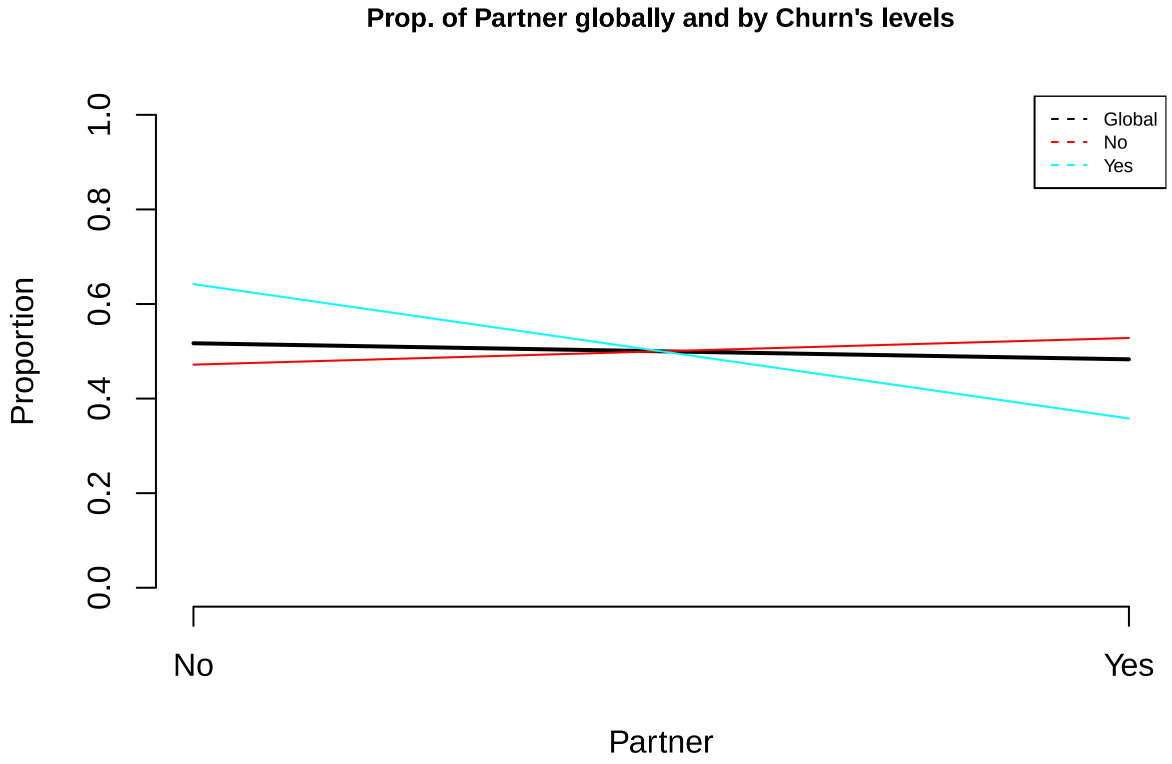
```
## [1] "Cross Table:"
##         P
##            No  Yes
##   Female 2549  939
##   Male   2625  930
## [1] "Distributions by columns:"
##
## P        Female      Male
##   No  0.7307913 0.7383966
##   Yes 0.2692087 0.2616034
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dades[, k] and as.factor(P)
## X-squared = 0.48408, df = 1, p-value = 0.4866
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P        Female      Male
##   No  0.4926556 0.5073444
##   Yes 0.5024077 0.4975923
##
## $vtest
##      Xquali
## P          Female       Male
##   No  -0.7227493  0.7227493
##   Yes  0.7227493 -0.7227493
##
## $pval
##      Xquali
## P        Female     Male
##   No  0.234917 0.234917
##   Yes 0.234917 0.234917
##
## [1] "Variable SeniorCitizen"
## [1] "Categories=" "Yes"         "No"
```



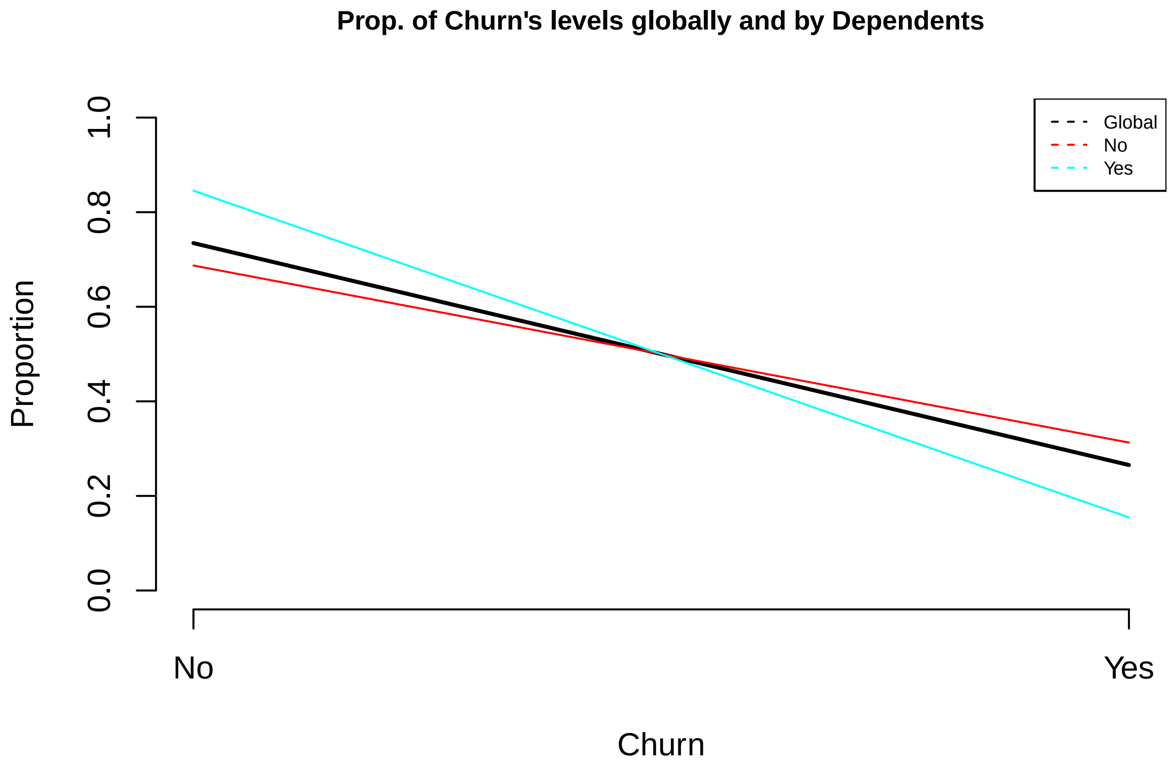


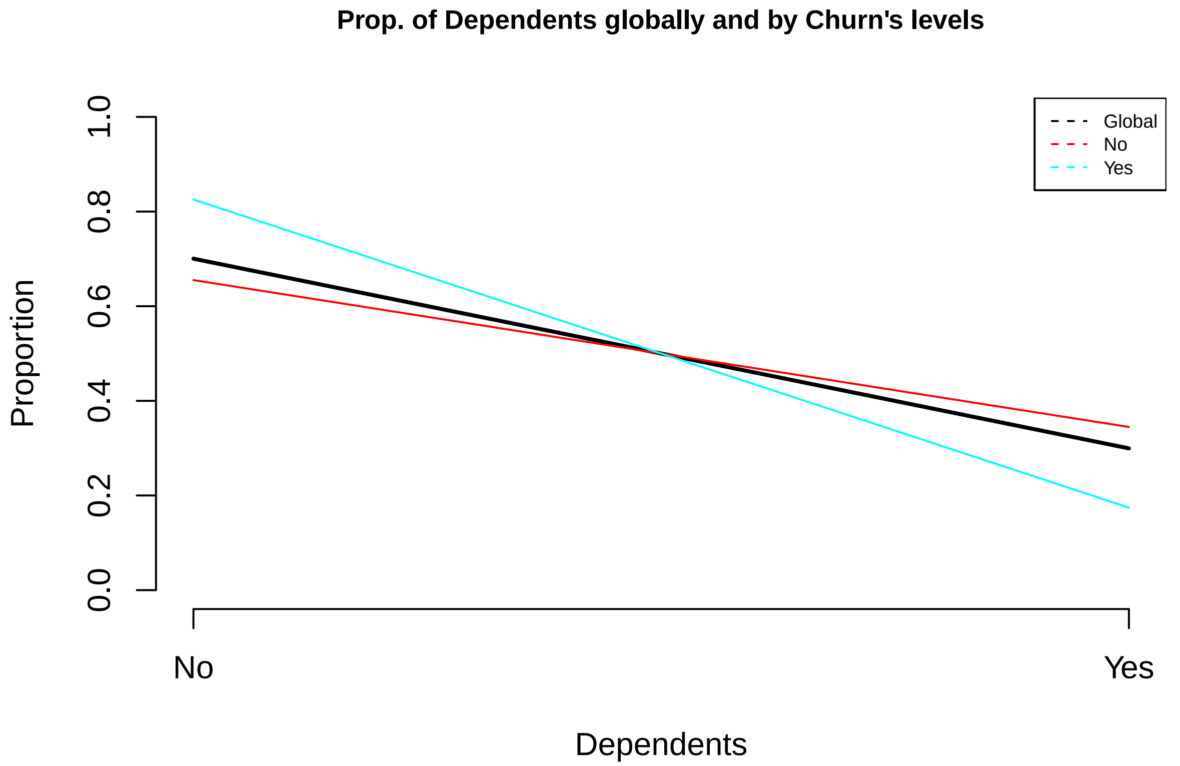
```
## [1] "Cross Table:"
##      P
##         No  Yes
##   Yes 4508 1393
##   No   666  476
## [1] "Distributions by columns:"
##
## P           Yes        No
##   No  0.7639383 0.5831874
##   Yes 0.2360617 0.4168126
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dades[, k] and as.factor(P)
## X-squared = 159.43, df = 1, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P           Yes        No
##   No  0.8712795 0.1287205
##   Yes 0.7453184 0.2546816
##
## $vtest
##      Xquali
## P           Yes        No
##   No   12.66302 -12.66302
##   Yes -12.66302  12.66302
##
## $pval
##      Xquali
## P              Yes           No
##   No  4.738952e-37 0.000000e+00
##   Yes 0.000000e+00 4.738952e-37
##
## [1] "Variable Partner"
## [1] "Categories=" "No"          "Yes"
```



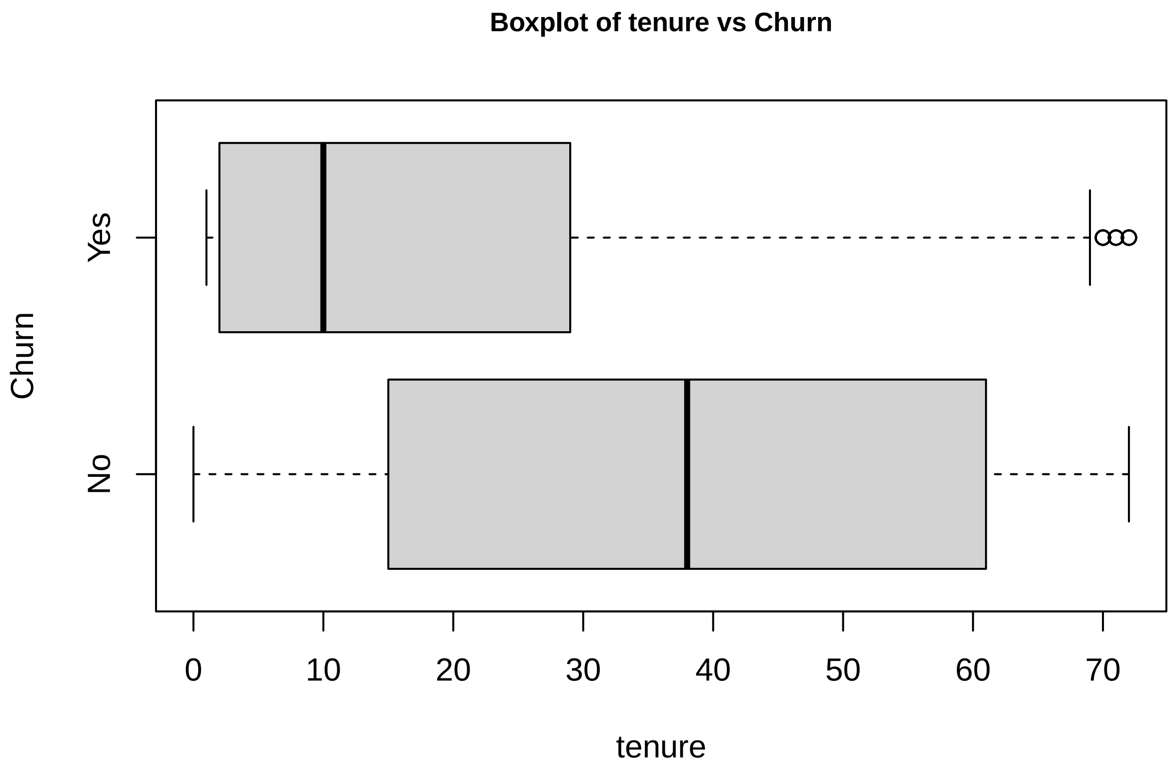


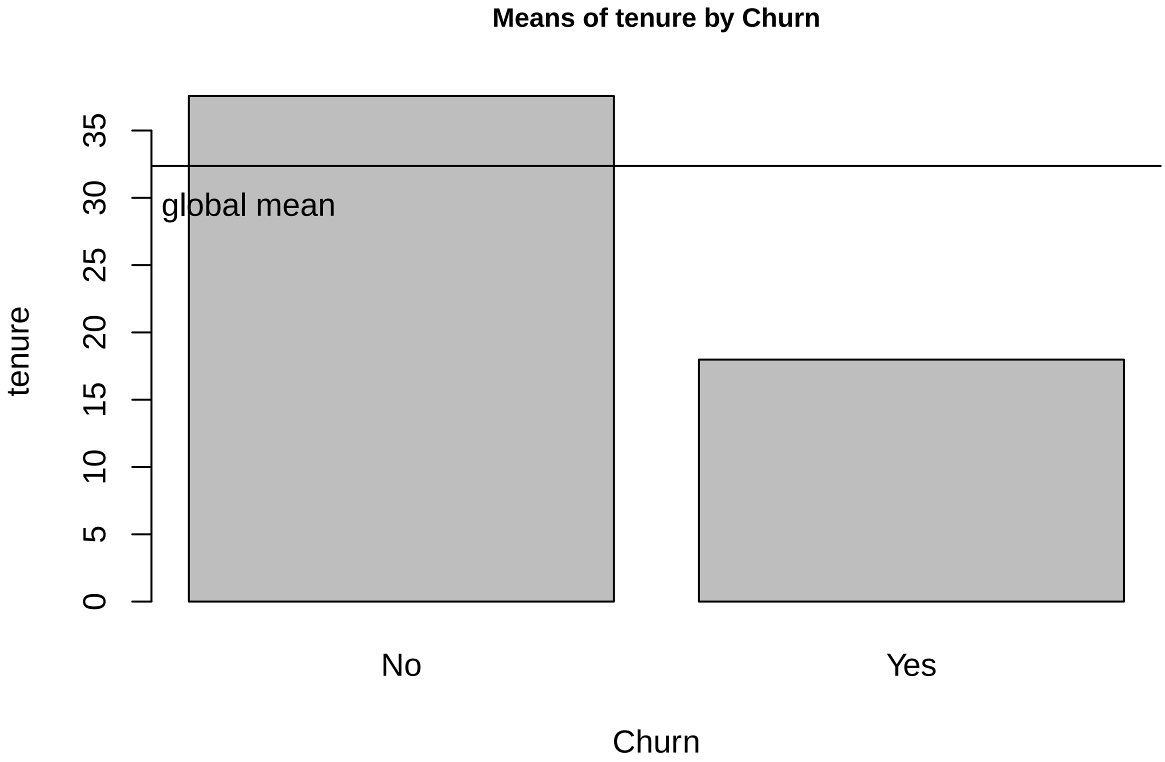
```
## [1] "Cross Table:"
##      P
##         No  Yes
##   No  2441 1200
##   Yes 2733  669
## [1] "Distributions by columns:"
##
## P            No       Yes
##   No  0.6704202 0.8033510
##   Yes 0.3295798 0.1966490
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dades[, k] and as.factor(P)
## X-squared = 158.73, df = 1, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P            No       Yes
##   No  0.4717820 0.5282180
##   Yes 0.6420546 0.3579454
##
## $vtest
##      Xquali
## P            No       Yes
##   No  -12.62595  12.62595
##   Yes  12.62595 -12.62595
##
## $pval
##      Xquali
## P               No          Yes
##   No  0.000000e+00 7.595183e-37
##   Yes 7.595183e-37 0.000000e+00
##
## [1] "Variable Dependents"
## [1] "Categories=" "No"          "Yes"
```



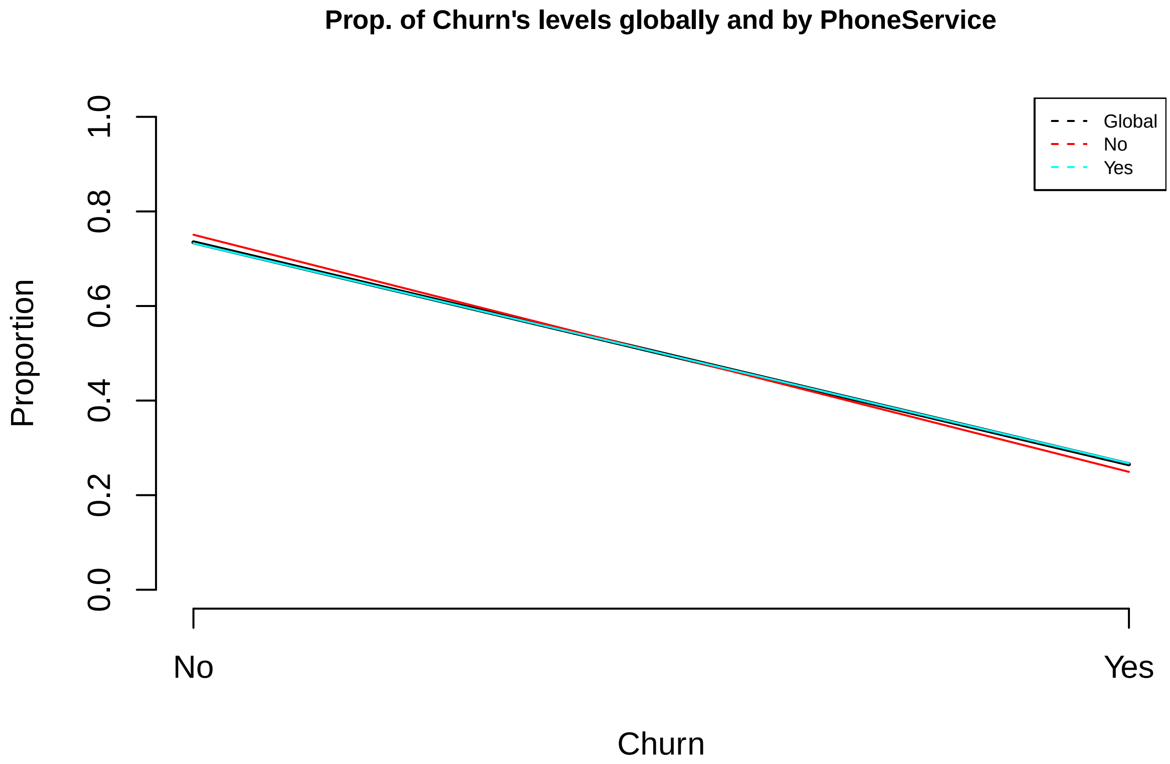


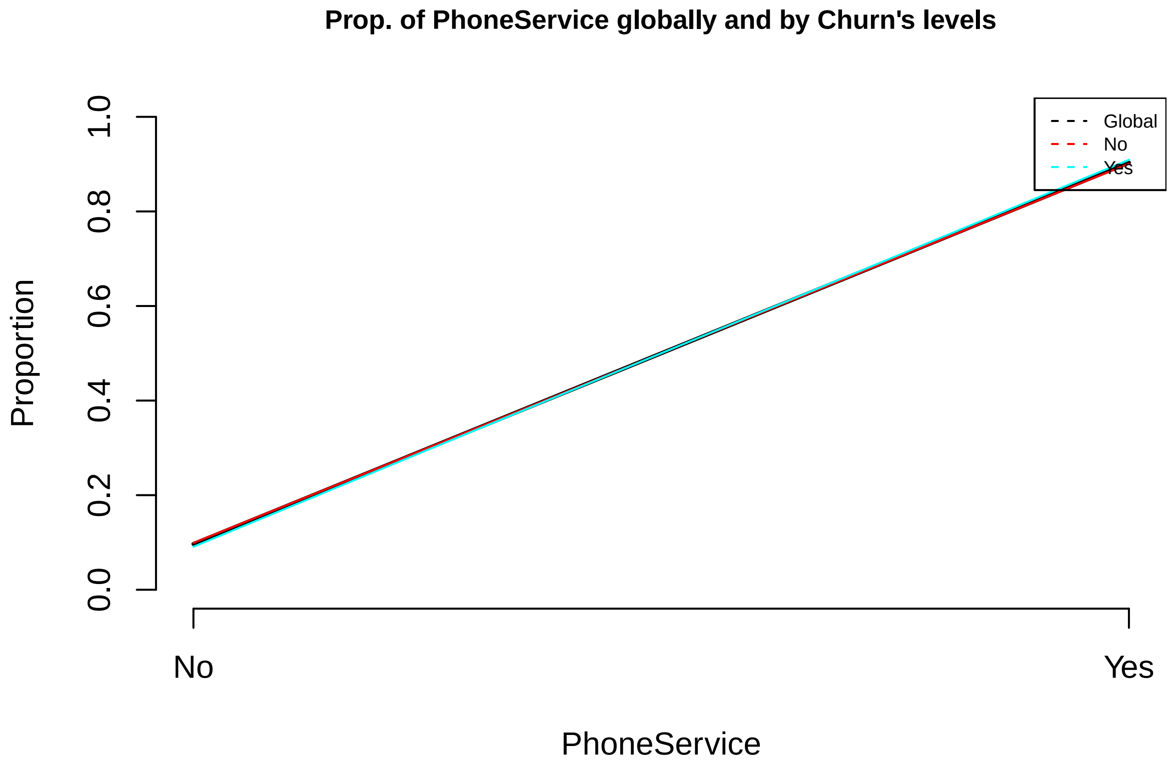
```
## [1] "Cross Table:"
##      P
##         No  Yes
##   No  3390 1543
##   Yes 1784  326
## [1] "Distributions by columns:"
##
## P             No       Yes
##   No  0.6872086 0.8454976
##   Yes 0.3127914 0.1545024
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dades[, k] and as.factor(P)
## X-squared = 189.13, df = 1, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P             No       Yes
##   No  0.6551991 0.3448009
##   Yes 0.8255752 0.1744248
##
## $vtest
##      Xquali
## P            No       Yes
##   No  -13.78188  13.78188
##   Yes  13.78188 -13.78188
##
## $pval
##      Xquali
## P               No          Yes
##   No  0.000000e+00 1.638041e-43
##   Yes 1.638041e-43 0.000000e+00
##
## [1] "Analysis by level of : tenure"
```



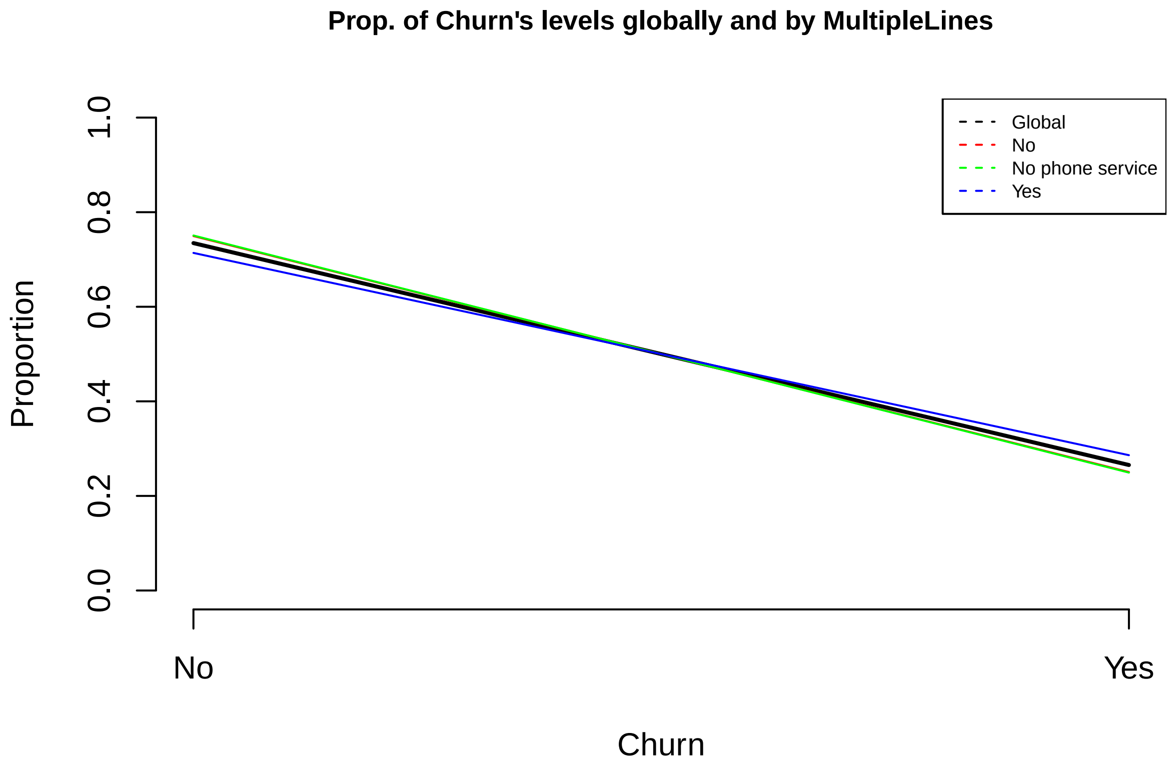


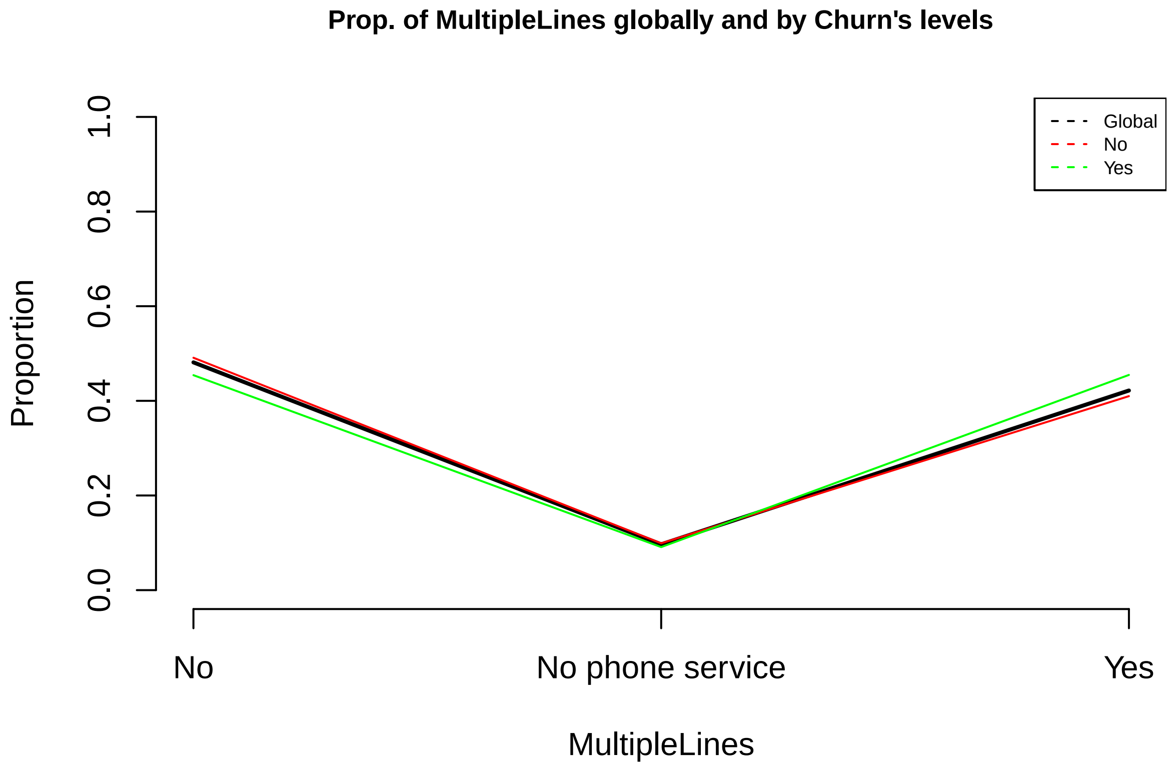
```
## [1] "Statistics by group:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    0.00   15.00   38.00   37.57   61.00   72.00 
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##    1.00    2.00   10.00   17.98   29.00   72.00 
## [1] "p-valueANOVA: 1.19549454726051e-232"
## [1] "p-value Kruskal-Wallis: 2.41914018186156e-208"
## [1] "p-values ValorsTest: "
##            No           Yes 
## 2.081921e-181  0.000000e+00 
## [1] "Variable PhoneService"
## [1] "Categories=" "No"          "Yes"
```


Prop. of Churn's levels globally and by PhoneService
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
Global
No
Yes
No
Yes
Churn

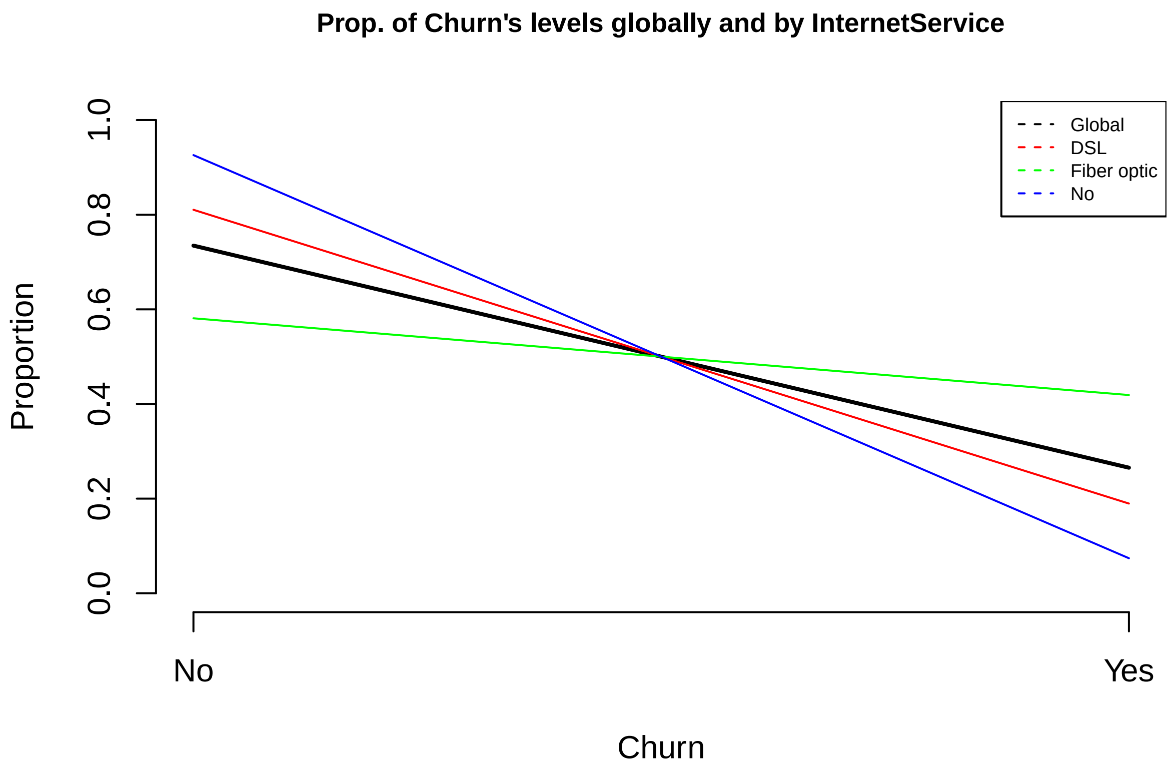


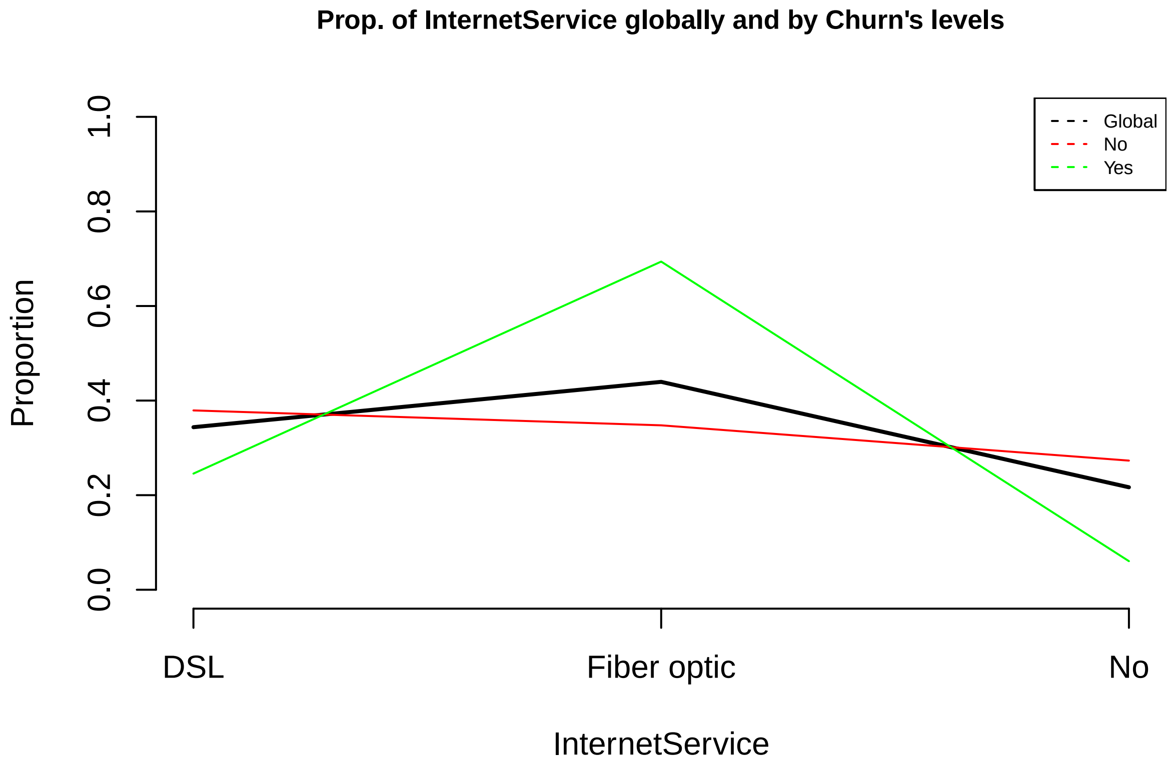
```
## [1] "Cross Table:"
##      P
##         No  Yes
##   No   512  170
##   Yes 4662 1699
## [1] "Distributions by columns:"
##
## P            No       Yes
##   No  0.7507331 0.7329036
##   Yes 0.2492669 0.2670964
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dades[, k] and as.factor(P)
## X-squared = 0.91503, df = 1, p-value = 0.3388
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P             No        Yes
##   No  0.09895632 0.90104368
##   Yes 0.09095773 0.90904227
##
## $vtest
##      Xquali
## P             No       Yes
##   No   1.002202 -1.002202
##   Yes -1.002202  1.002202
##
## $pval
##      Xquali
## P             No       Yes
##   No  0.1581231 0.1581231
##   Yes 0.1581231 0.1581231
##
## [1] "Variable MultipleLines"
## [1] "Categories="      "No"               "No phone service" "Yes"
```



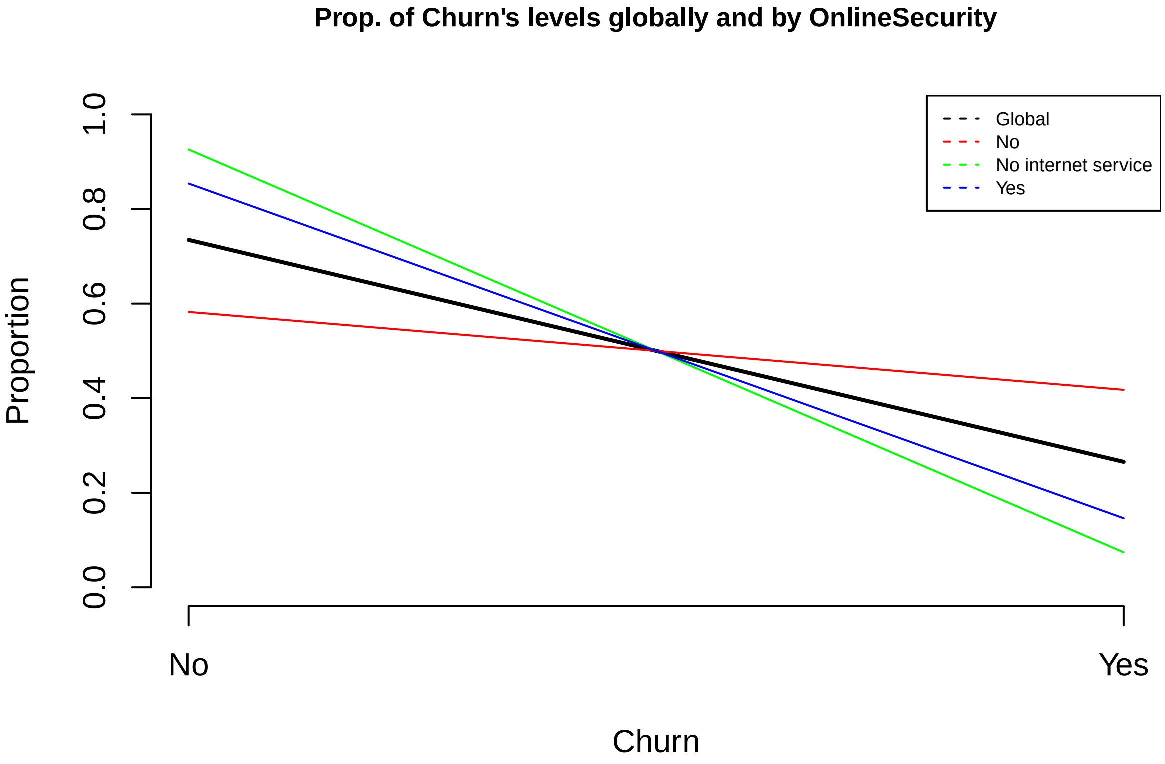


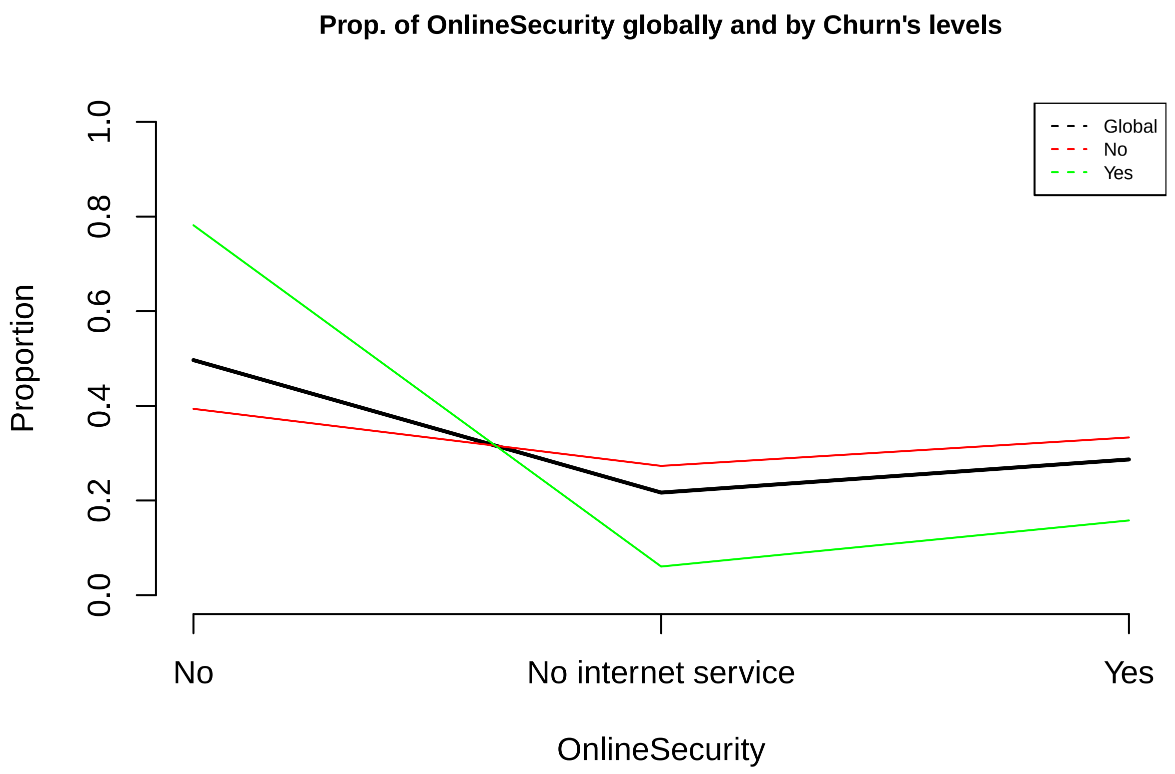
```
## [1] "Cross Table:"
##                    P
##                       No  Yes
##   No               2541  849
##   No phone service  512  170
##   Yes              2121  850
## [1] "Distributions by columns:"
##
## P            No No phone service       Yes
##   No  0.7495575        0.7507331 0.7139010
##   Yes 0.2504425        0.2492669 0.2860990
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 11.33, df = 2, p-value = 0.003464
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P             No No phone service        Yes
##   No  0.49110939       0.09895632 0.40993429
##   Yes 0.45425361       0.09095773 0.45478866
##
## $vtest
##      Xquali
## P            No No phone service       Yes
##   No   2.733239         1.002202 -3.365474
##   Yes -2.733239        -1.002202  3.365474
##
## $pval
##      Xquali
## P               No No phone service          Yes
##   No  0.0031357380     0.1581230658 0.0003820611
##   Yes 0.0031357380     0.1581230658 0.0003820611
##
## [1] "Variable InternetService"
## [1] "Categories=" "DSL"         "Fiber optic" "No"
```



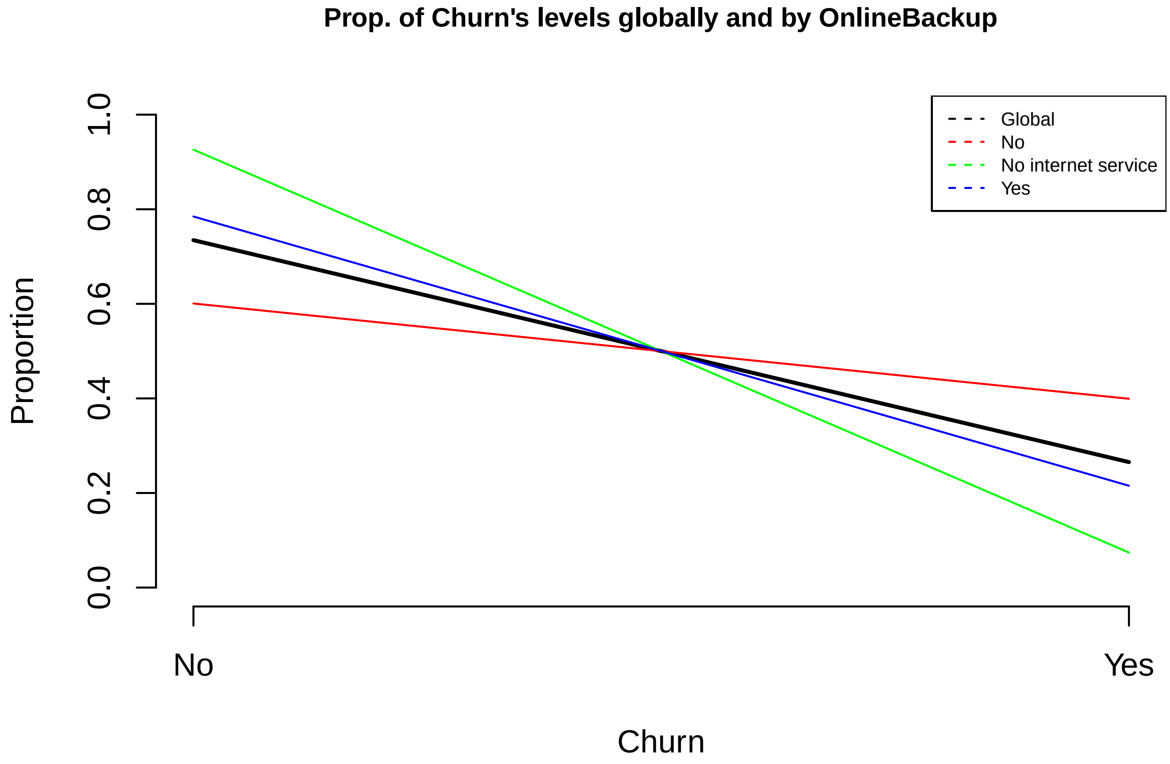


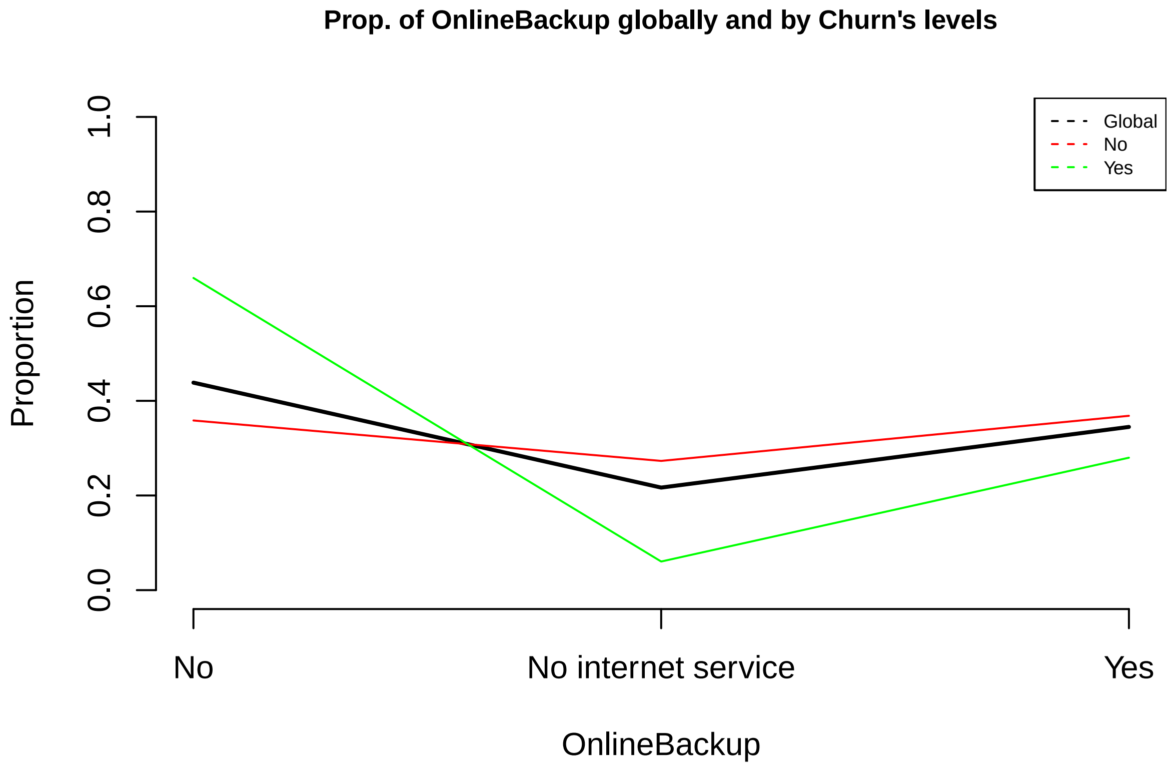
```
## [1] "Cross Table:"
##              P
##                 No  Yes
##   DSL         1962  459
##   Fiber optic 1799 1297
##   No          1413  113
## [1] "Distributions by columns:"
##
## P           DSL Fiber optic        No
##   No  0.8104089   0.5810724 0.9259502
##   Yes 0.1895911   0.4189276 0.0740498
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 732.31, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P            DSL Fiber optic         No
##   No  0.37920371  0.34770004 0.27309625
##   Yes 0.24558587  0.69395399 0.06046014
##
## $vtest
##      Xquali
## P           DSL Fiber optic        No
##   No   10.42434   -25.84981  19.12516
##   Yes -10.42434    25.84981 -19.12516
##
## $pval
##      Xquali
## P               DSL   Fiber optic            No
##   No   9.598875e-26  0.000000e+00  7.795425e-82
##   Yes  0.000000e+00 1.222462e-147  0.000000e+00
##
## [1] "Variable OnlineSecurity"
## [1] "Categories="         "No"                  "No internet service"
## [4] "Yes"
```



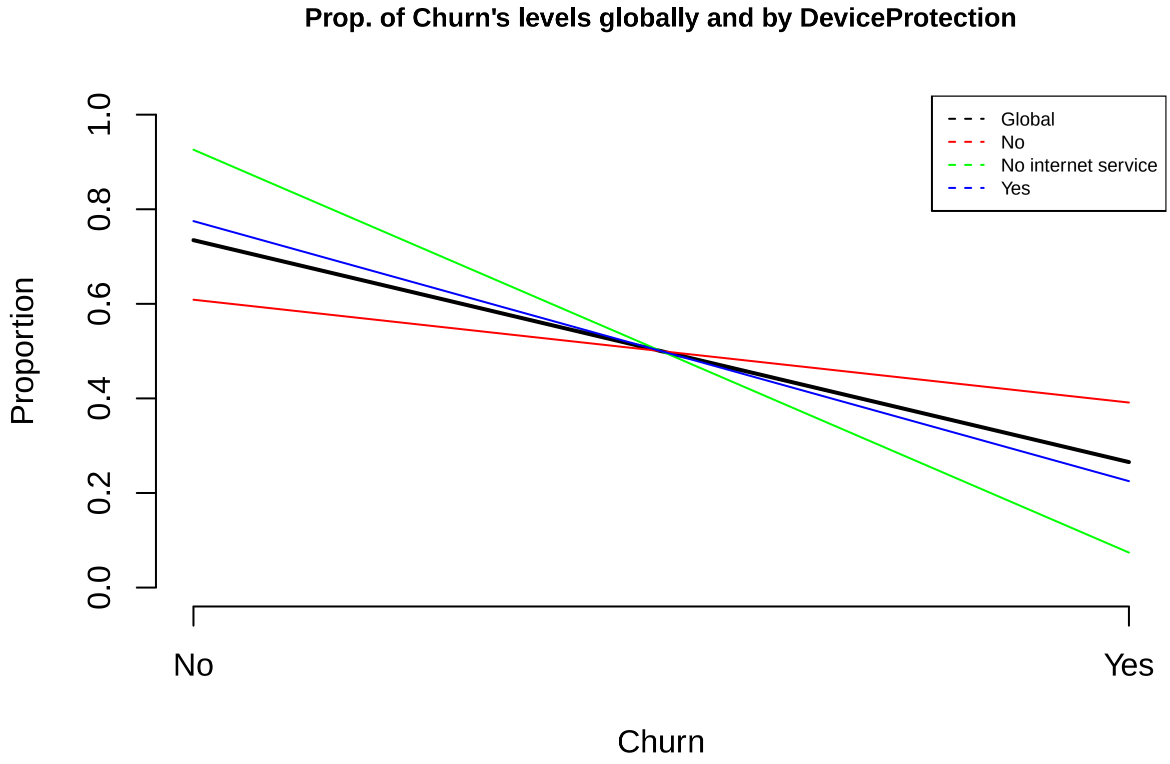


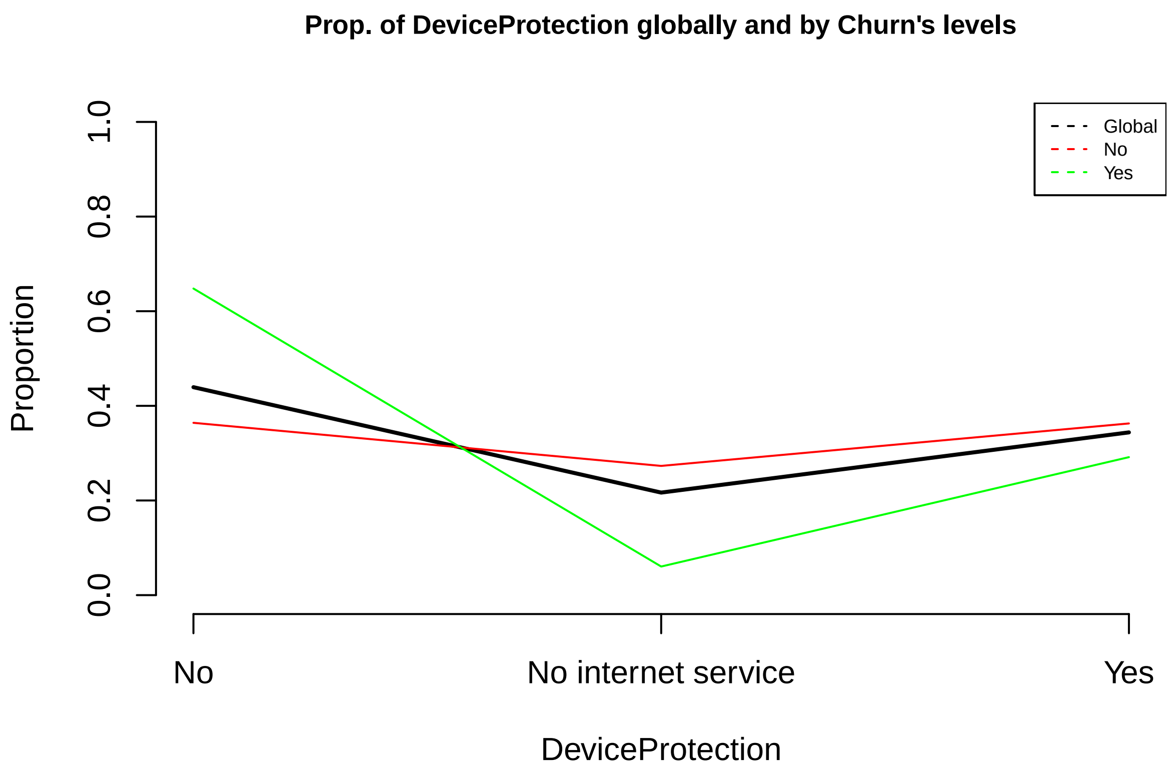
```
## [1] "Cross Table:"
##                        P
##                           No  Yes
##   No                    2037 1461
##   No internet service 1413  113
##   Yes                   1724  295
## [1] "Distributions by columns:"
##
## P             No No internet service       Yes
##   No  0.5823328           0.9259502 0.8538881
##   Yes 0.4176672           0.0740498 0.1461119
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 850, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P              No No internet service        Yes
##   No  0.39369927          0.27309625 0.33320448
##   Yes 0.78170144          0.06046014 0.15783842
##
## $vtest
##      Xquali
## P            No No internet service       Yes
##   No  -28.75497            19.12516  14.36975
##   Yes  28.75497           -19.12516 -14.36975
##
## $pval
##      Xquali
## P                No No internet service           Yes
##   No   0.000000e+00        7.795425e-82  4.005837e-47
##   Yes 3.925582e-182        0.000000e+00  0.000000e+00
##
## [1] "Variable OnlineBackup"
## [1] "Categories="         "No"                  "No internet service"
## [4] "Yes"
```



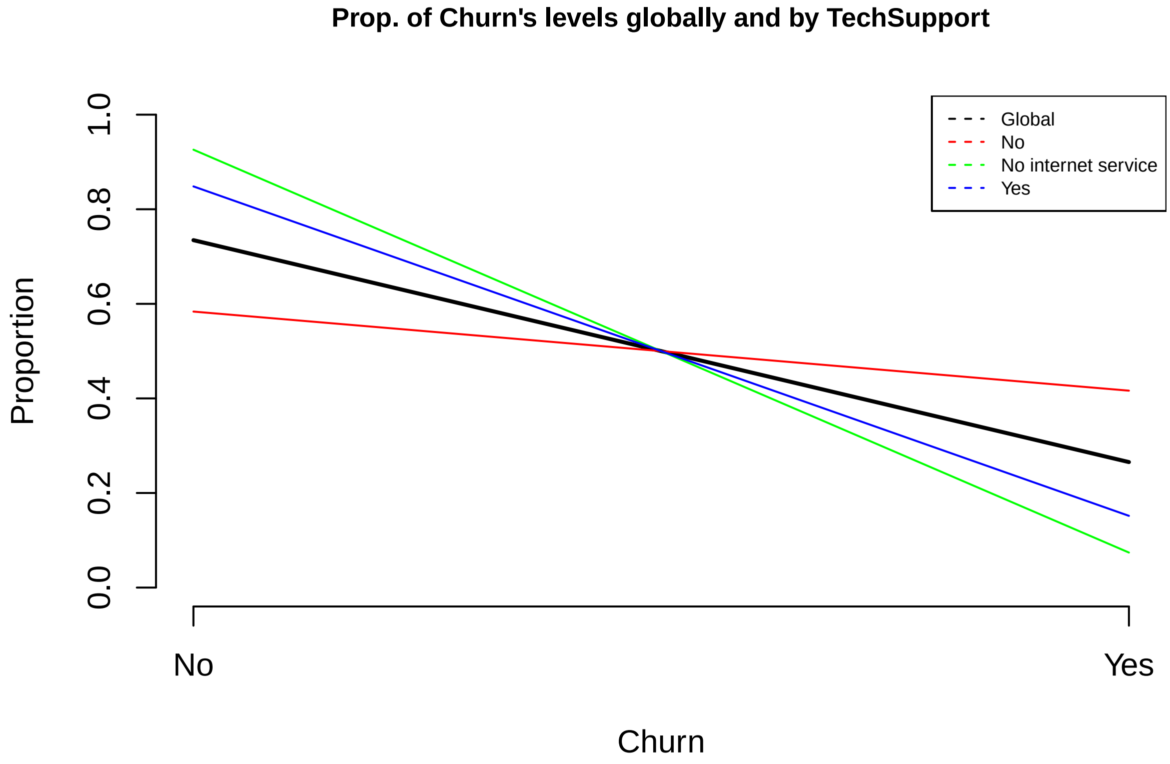


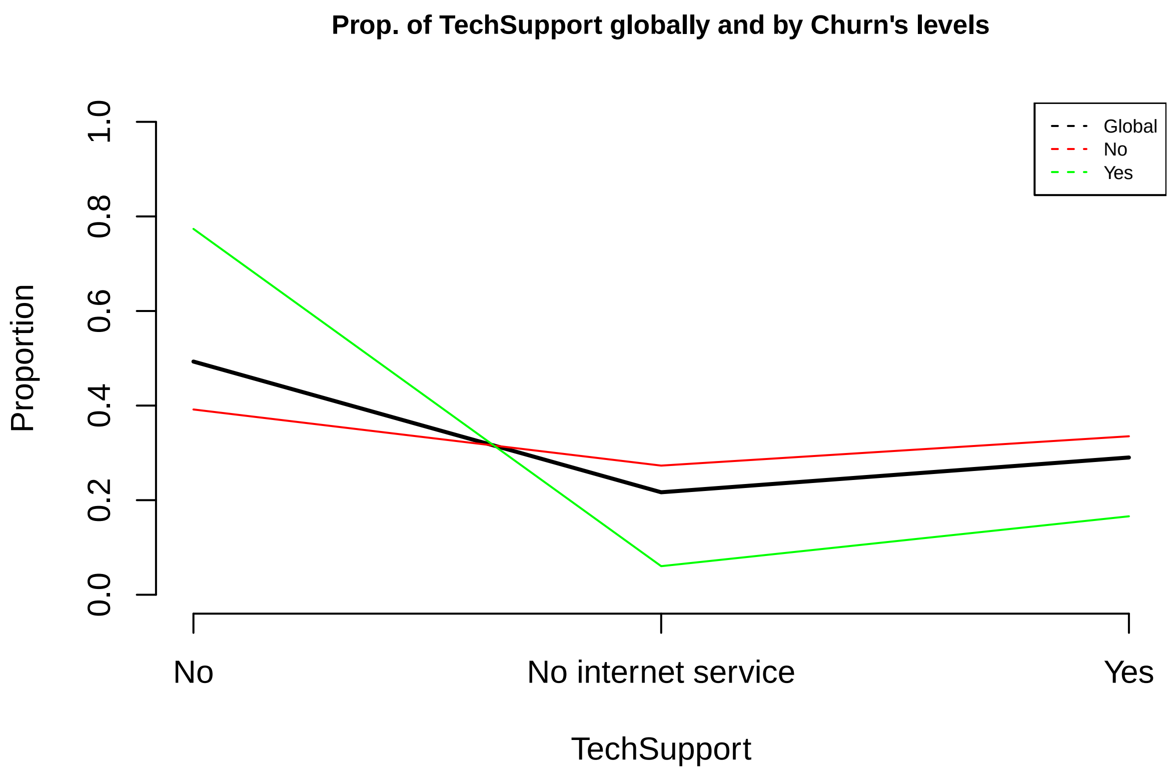
```
## [1] "Cross Table:"
##                       P
##                          No  Yes
##   No                   1855 1233
##   No internet service 1413  113
##   Yes                  1906  523
## [1] "Distributions by columns:"
##
## P             No No internet service       Yes
##   No  0.6007124           0.9259502 0.7846851
##   Yes 0.3992876           0.0740498 0.2153149
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 601.81, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P              No No internet service        Yes
##   No  0.35852339          0.27309625 0.36838036
##   Yes 0.65971108          0.06046014 0.27982879
##
## $vtest
##      Xquali
## P             No No internet service        Yes
##   No  -22.491687           19.125155   6.903041
##   Yes  22.491687          -19.125155  -6.903041
##
## $pval
##      Xquali
## P                No No internet service           Yes
##   No   0.000000e+00        7.795425e-82  2.545045e-12
##   Yes 2.502984e-112        0.000000e+00  2.545075e-12
##
## [1] "Variable DeviceProtection"
## [1] "Categories="         "No"                  "No internet service"
## [4] "Yes"
```



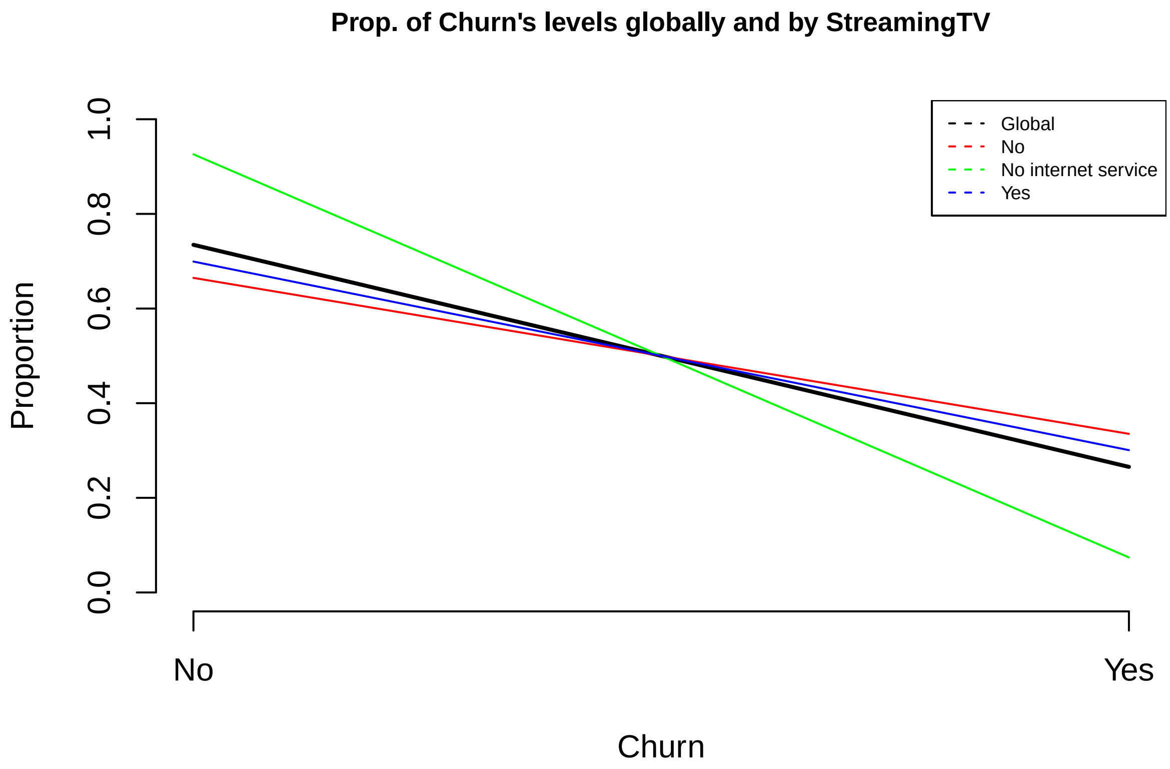


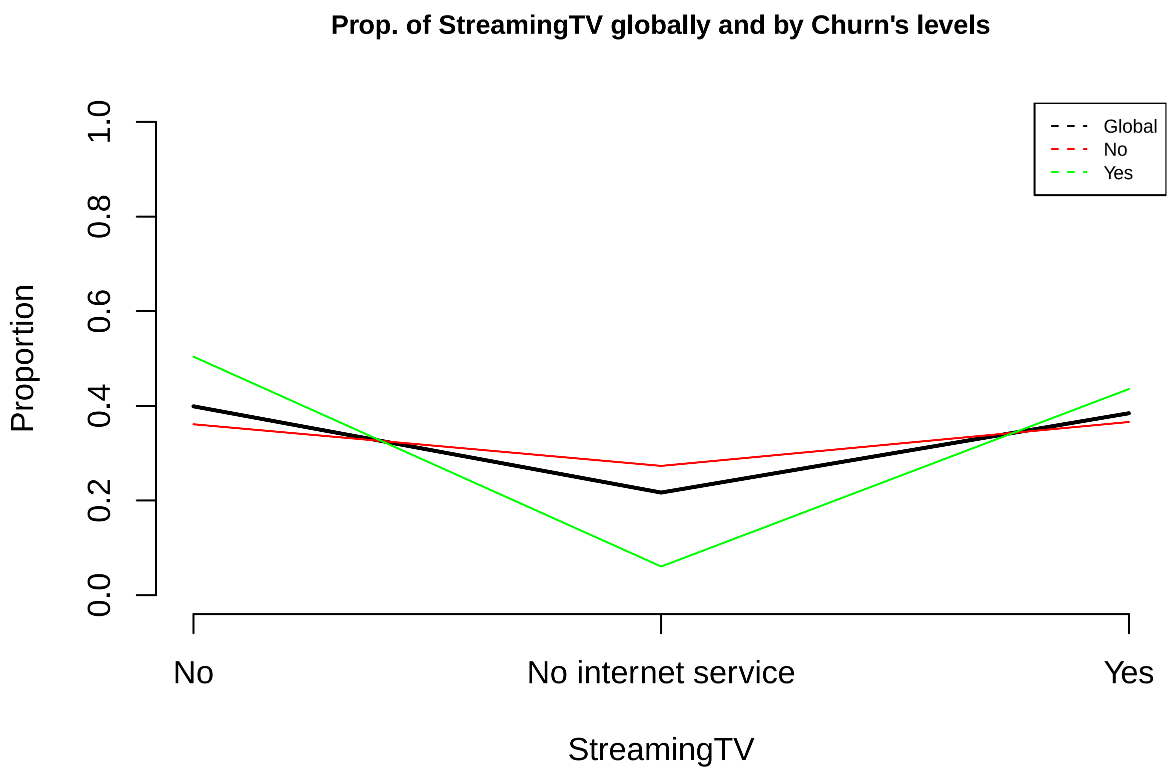
```
## [1] "Cross Table:"
##                      P
##                         No  Yes
##   No                  1884 1211
##   No internet service 1413  113
##   Yes                 1877  545
## [1] "Distributions by columns:"
##
## P            No No internet service       Yes
##   No  0.6087237           0.9259502 0.7749794
##   Yes 0.3912763           0.0740498 0.2250206
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 558.42, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P             No No internet service        Yes
##   No  0.36412833          0.27309625 0.36277542
##   Yes 0.64794007          0.06046014 0.29159979
##
## $vtest
##      Xquali
## P              No No internet service        Yes
##   No  -21.188888           19.125155   5.552301
##   Yes  21.188888          -19.125155  -5.552301
##
## $pval
##      Xquali
## P                 No No internet service           Yes
##   No   0.000000e+00        7.795425e-82  1.409671e-08
##   Yes 6.045963e-100        0.000000e+00  1.409671e-08
##
## [1] "Variable TechSupport"
## [1] "Categories="         "No"                  "No internet service"
## [4] "Yes"
```



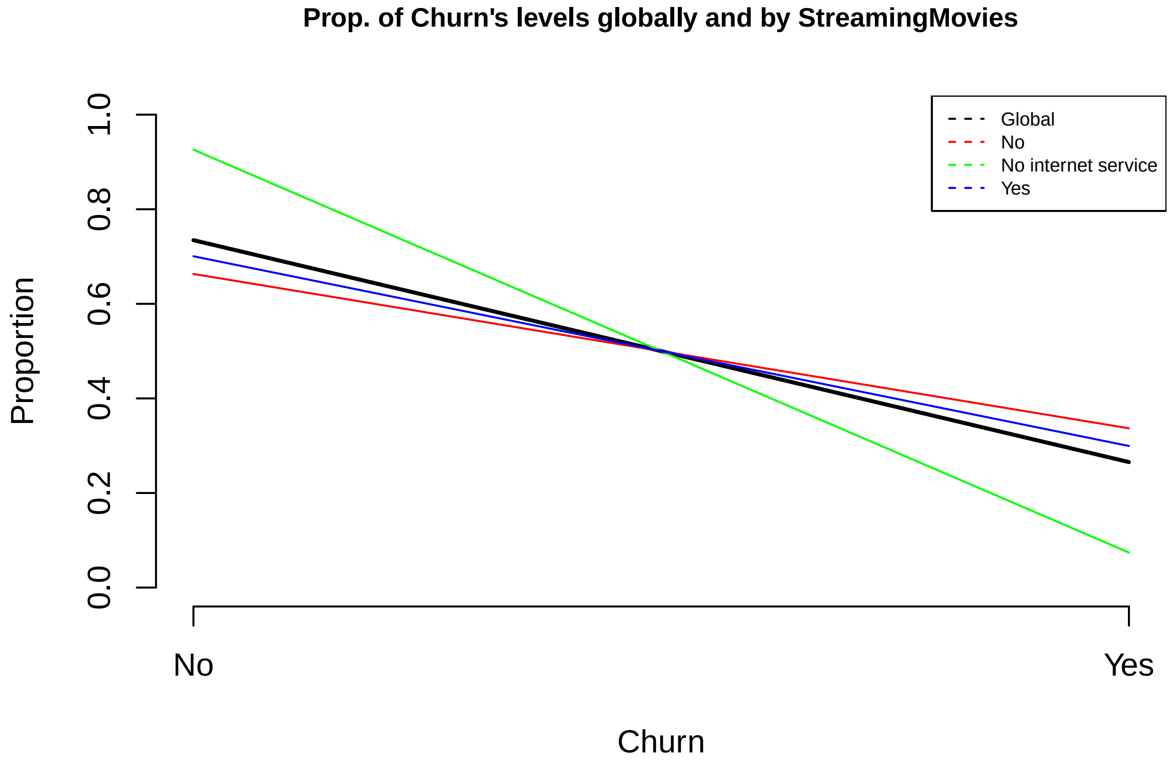


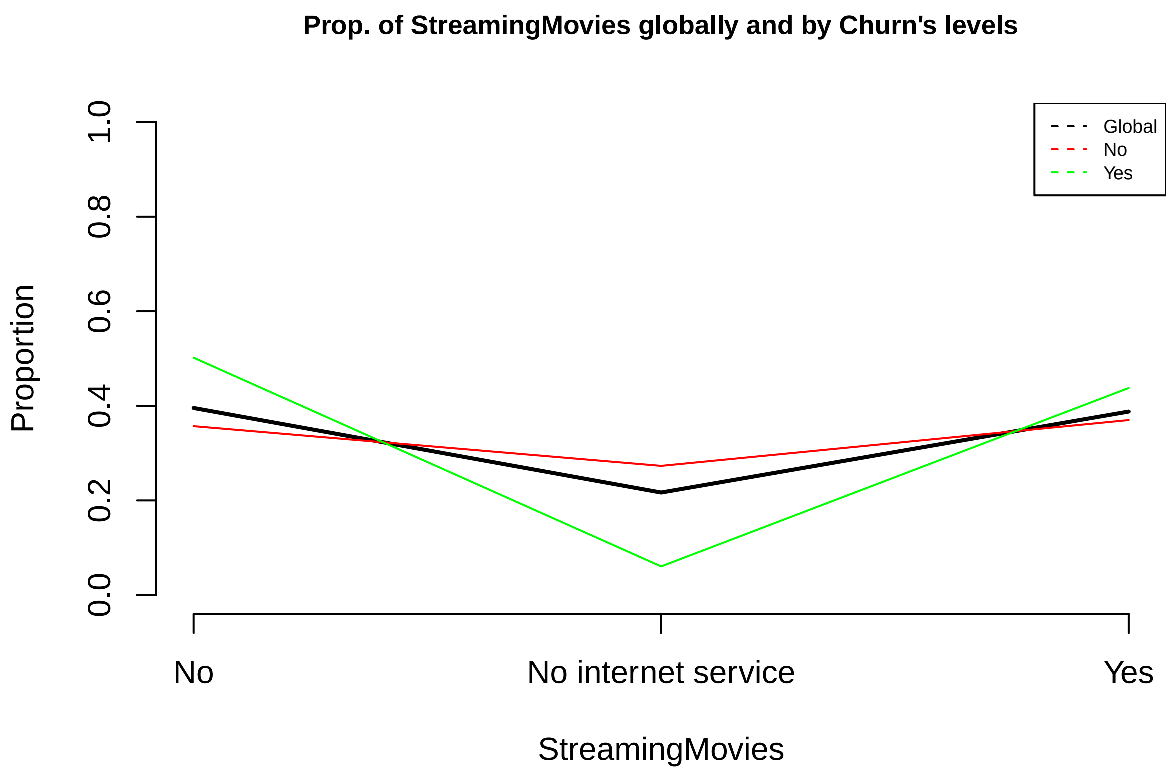
```
## [1] "Cross Table:"
##                       P
##                         No  Yes
##   No                  2027 1446
##   No internet service 1413  113
##   Yes                 1734  310
## [1] "Distributions by columns:"
##
## P            No No internet service       Yes
##   No  0.5836453           0.9259502 0.8483366
##   Yes 0.4163547           0.0740498 0.1516634
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 828.2, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P             No No internet service        Yes
##   No  0.39176652          0.27309625 0.33513722
##   Yes 0.77367576          0.06046014 0.16586410
##
## $vtest
##      Xquali
## P            No No internet service       Yes
##   No  -28.30547            19.12516  13.81983
##   Yes  28.30547           -19.12516 -13.81983
##
## $pval
##      Xquali
## P                No No internet service           Yes
##   No   0.000000e+00        7.795425e-82  9.676286e-44
##   Yes 1.479823e-176        0.000000e+00  0.000000e+00
##
## [1] "Variable StreamingTV"
## [1] "Categories="         "No"                  "No internet service"
## [4] "Yes"
```



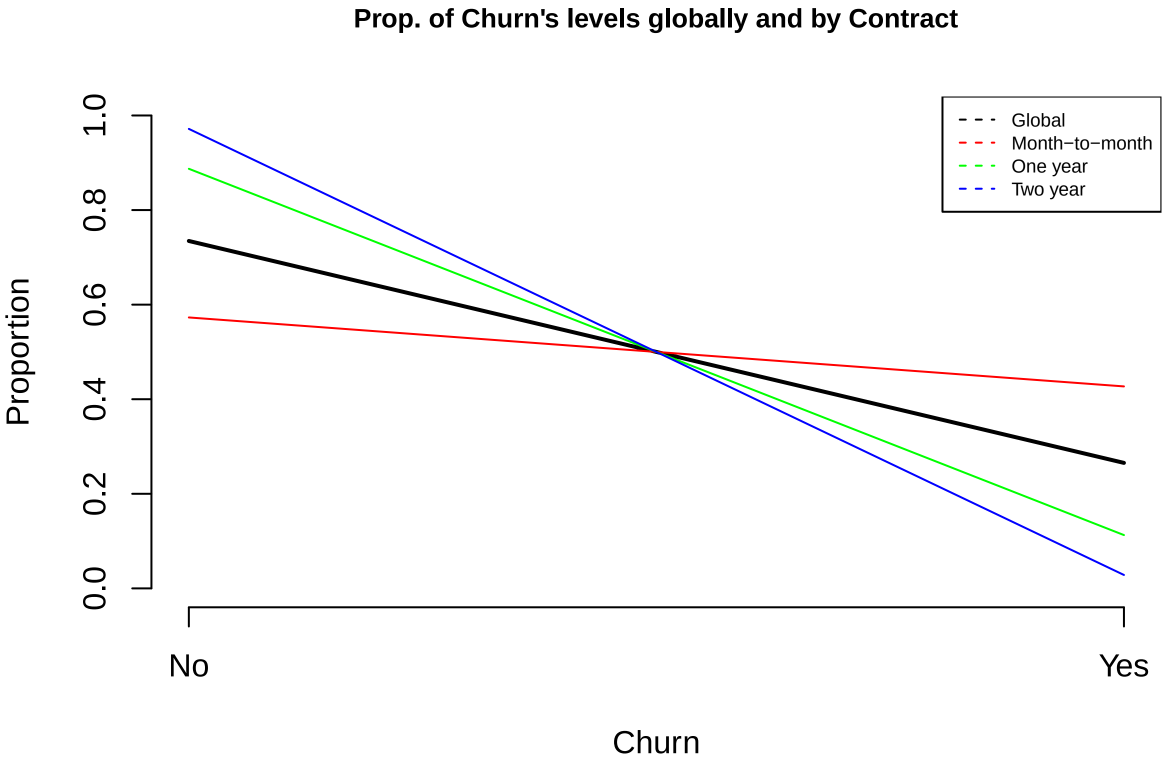


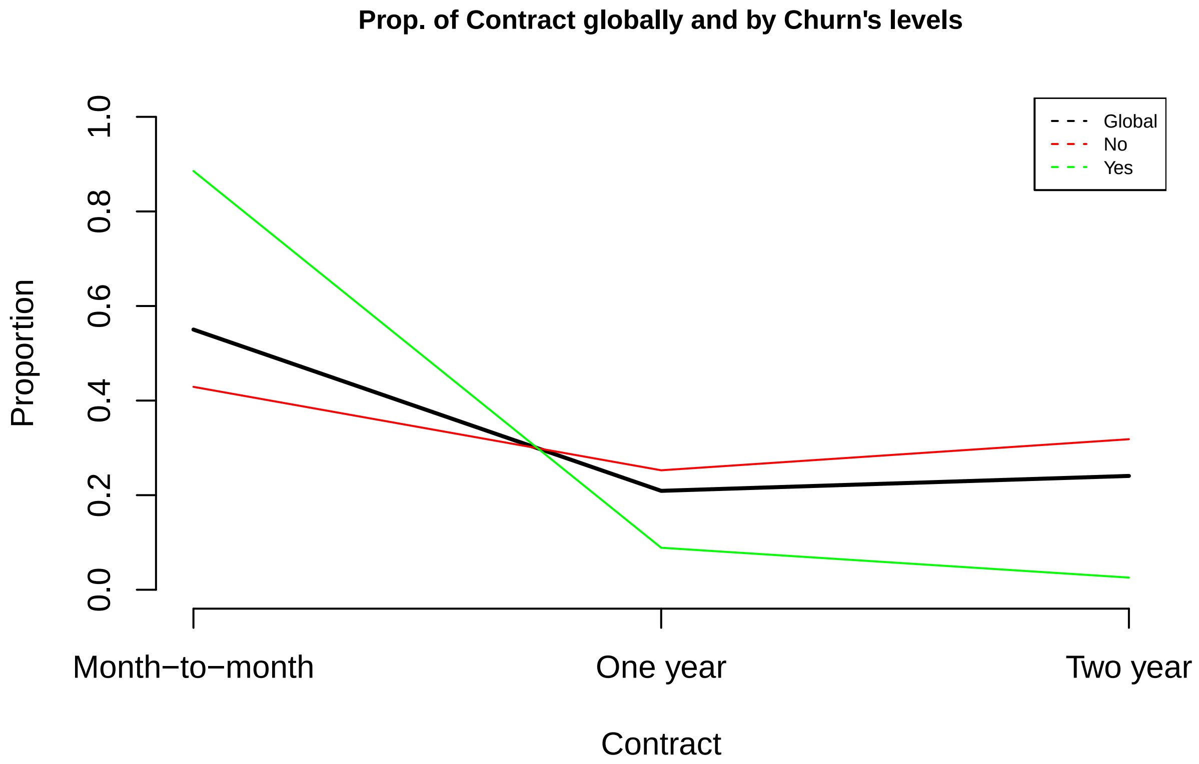
```
## [1] "Cross Table:"
##                        P
##                           No  Yes
##   No                    1868  942
##   No internet service 1413  113
##   Yes                   1893  814
## [1] "Distributions by columns:"
##
## P            No No internet service       Yes
##   No  0.6647687           0.9259502 0.6992981
##   Yes 0.3352313           0.0740498 0.3007019
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 374.2, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P              No No internet service        Yes
##   No  0.36103595          0.27309625 0.36586780
##   Yes 0.50401284          0.06046014 0.43552702
##
## $vtest
##      Xquali
## P             No No internet service        Yes
##   No  -10.818954           19.125155  -5.306236
##   Yes  10.818954          -19.125155   5.306236
##
## $pval
##      Xquali
## P               No No internet service          Yes
##   No  0.000000e+00        7.795425e-82 5.595609e-08
##   Yes 1.399774e-27        0.000000e+00 5.595609e-08
##
## [1] "Variable StreamingMovies"
## [1] "Categories="         "No"                 "No internet service"
## [4] "Yes"
```



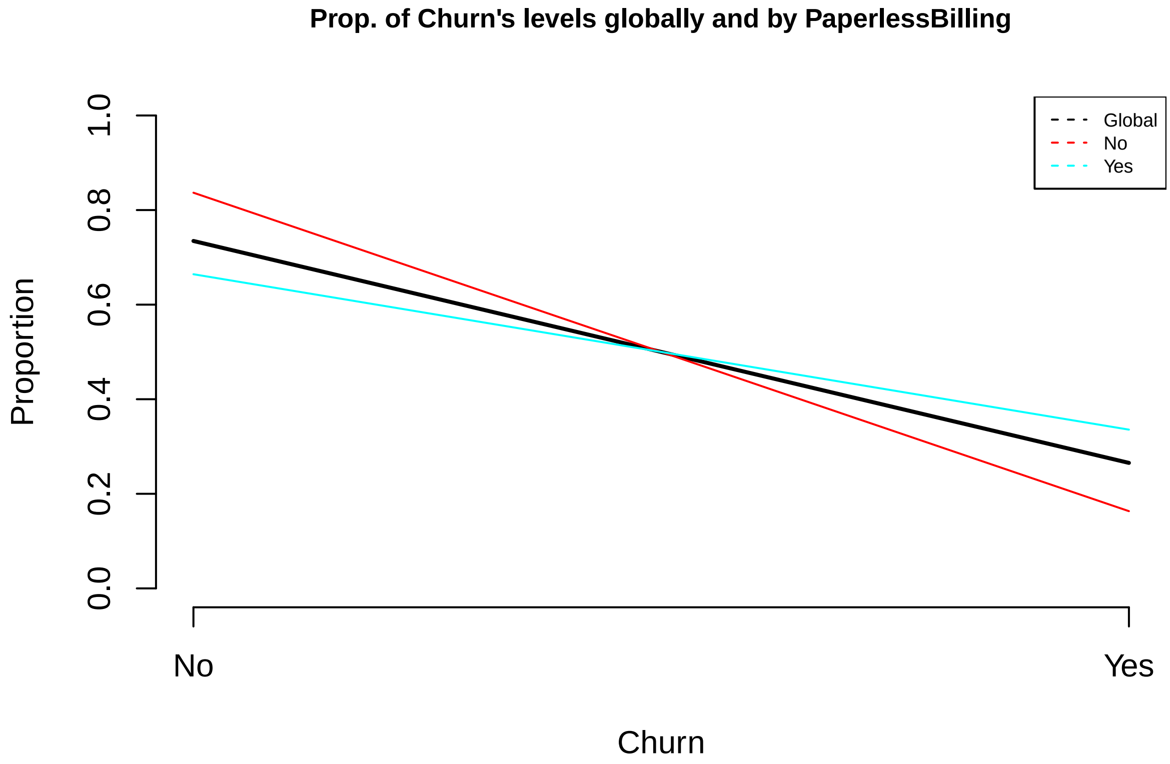


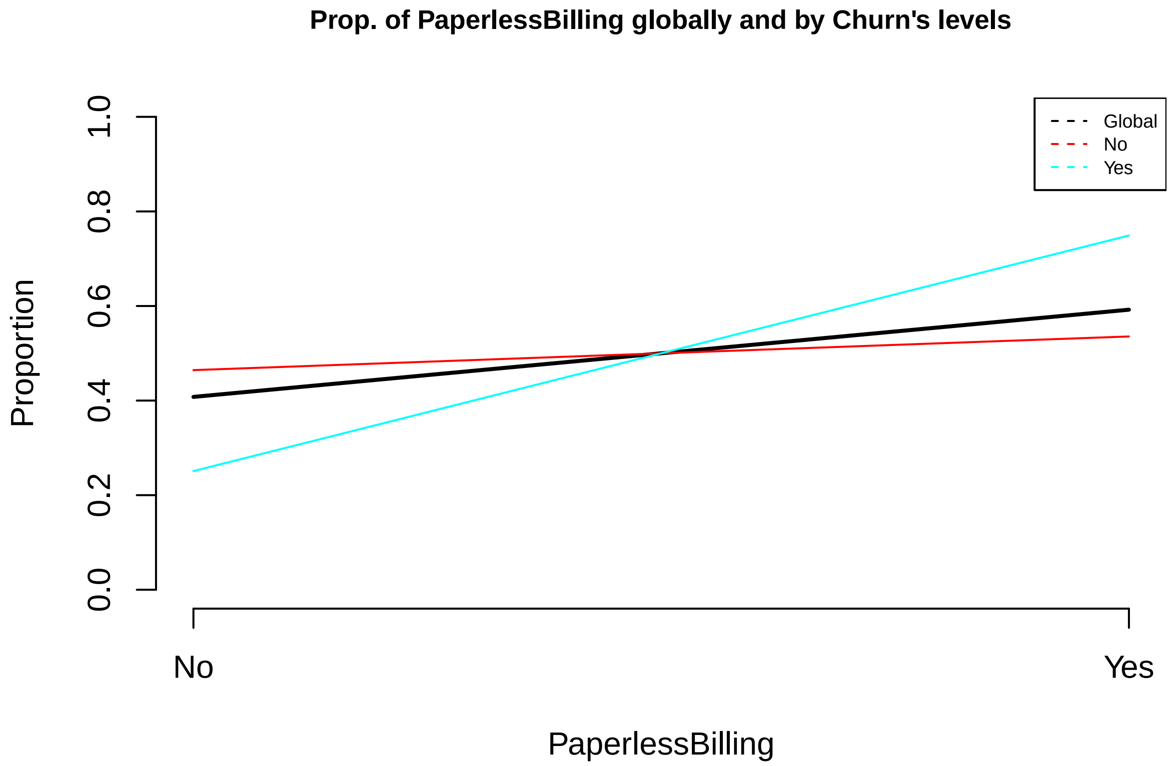
```
## [1] "Cross Table:"
##                        P
##                           No  Yes
##   No                    1847  938
##   No internet service 1413  113
##   Yes                   1914  818
## [1] "Distributions by columns:"
##
## P             No No internet service       Yes
##   No  0.6631957           0.9259502 0.7005857
##   Yes 0.3368043           0.0740498 0.2994143
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 375.66, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P              No No internet service        Yes
##   No  0.35697719          0.27309625 0.36992656
##   Yes 0.50187266          0.06046014 0.43766720
##
## $vtest
##      Xquali
## P              No No internet service         Yes
##   No  -10.980853           19.125155  -5.151298
##   Yes  10.980853          -19.125155   5.151298
##
## $pval
##      Xquali
## P                No No internet service          Yes
##   No  0.000000e+00        7.795425e-82 1.293448e-07
##   Yes 2.362211e-28        0.000000e+00 1.293448e-07
##
## [1] "Variable Contract"
## [1] "Categories="    "Month-to-month" "One year"       "Two year"
```



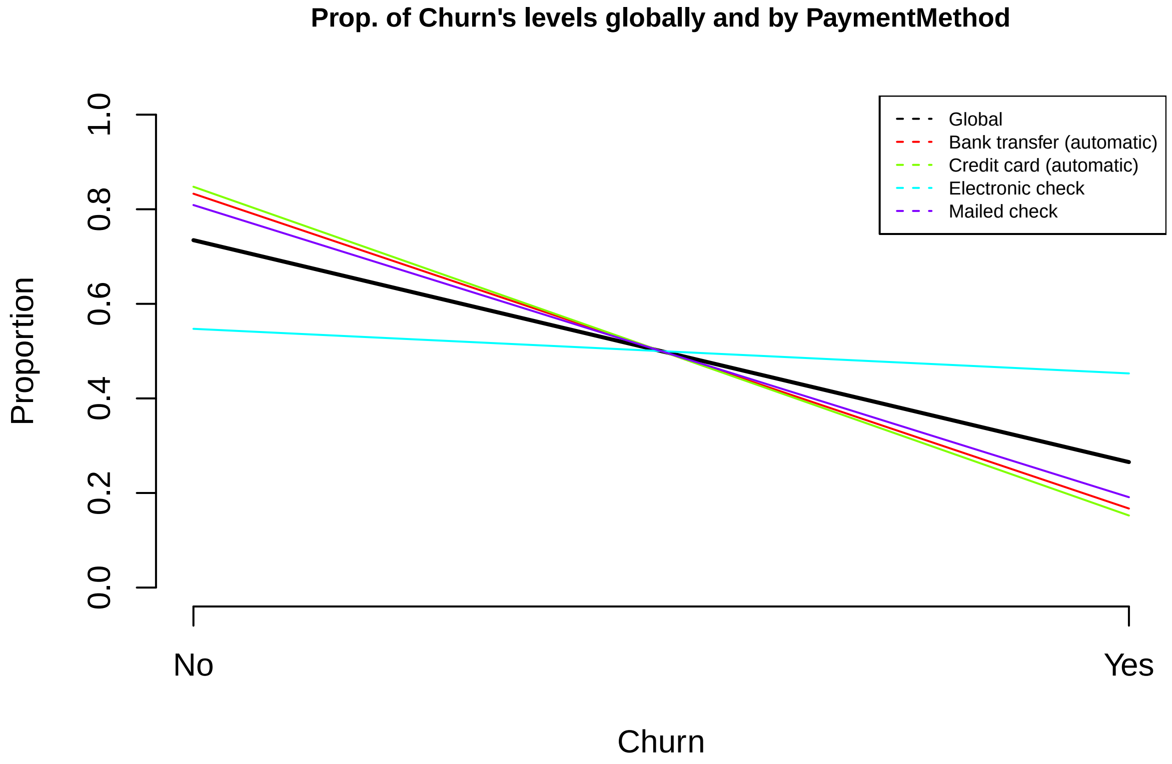


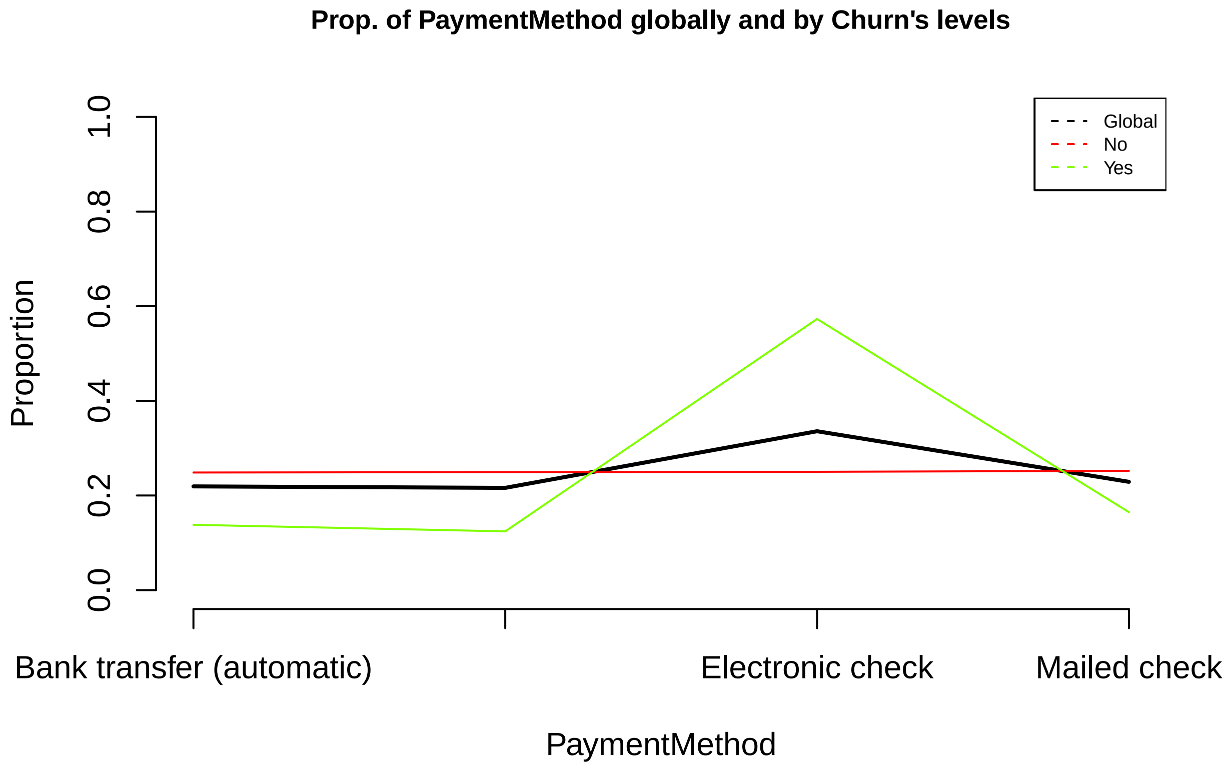
```
## [1] "Cross Table:"
##                 P
##                    No  Yes
##   Month-to-month 2220 1655
##   One year       1307  166
##   Two year       1647   48
## [1] "Distributions by columns:"
##
## P     Month-to-month   One year   Two year
##   No      0.57290323 0.88730482 0.97168142
##   Yes     0.42709677 0.11269518 0.02831858
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 1184.6, df = 2, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P     Month-to-month   One year   Two year
##   No      0.42906842 0.25260920 0.31832238
##   Yes     0.88550027 0.08881755 0.02568218
##
## $vtest
##      Xquali
## P      Month-to-month  One year  Two year
##   No       -33.99728  14.92312  25.36589
##   Yes       33.99728 -14.92312 -25.36589
##
## $pval
##      Xquali
## P      Month-to-month      One year      Two year
##   No    0.000000e+00  1.165649e-50 3.001022e-142
##   Yes  1.221803e-253  0.000000e+00  0.000000e+00
##
## [1] "Variable PaperlessBilling"
## [1] "Categories=" "No"          "Yes"
```



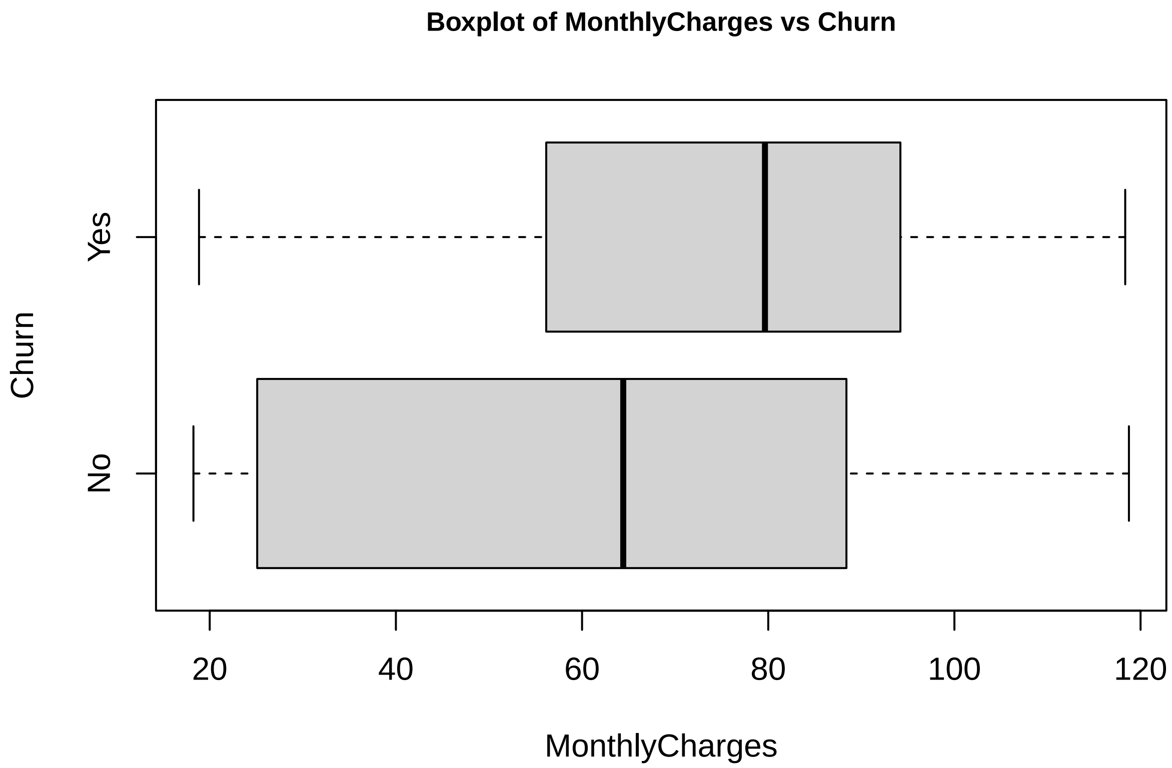


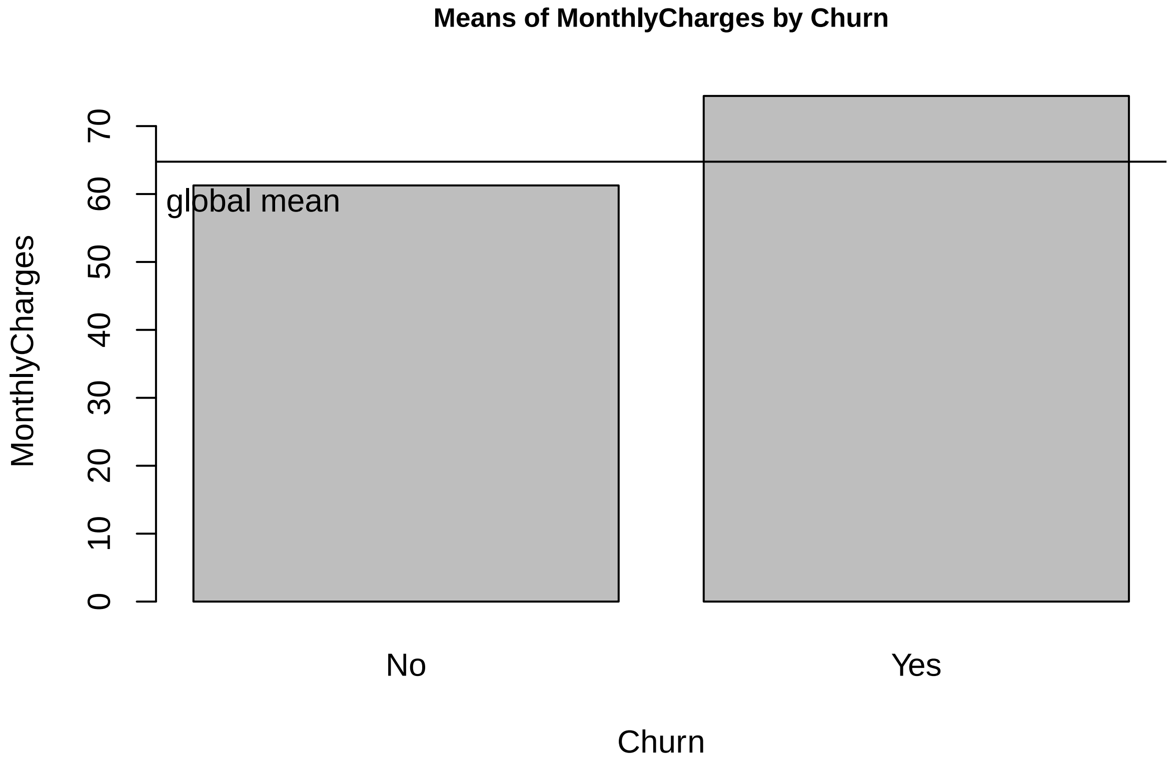
```
## [1] "Cross Table:"
##      P
##         No  Yes
##   No  2403  469
##   Yes 2771 1400
## [1] "Distributions by columns:"
##
## P            No       Yes
##   No  0.8366992 0.6643491
##   Yes 0.1633008 0.3356509
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test with Yates' continuity correction
##
## data:  dades[, k] and as.factor(P)
## X-squared = 258.28, df = 1, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P            No       Yes
##   No  0.4644376 0.5355624
##   Yes 0.2509363 0.7490637
##
## $vtest
##      Xquali
## P            No       Yes
##   No   16.09848 -16.09848
##   Yes -16.09848  16.09848
##
## $pval
##      Xquali
## P               No          Yes
##   No  1.307299e-58 0.000000e+00
##   Yes 0.000000e+00 1.307299e-58
##
## [1] "Variable PaymentMethod"
## [1] "Categories="               "Bank transfer (automatic)"
## [3] "Credit card (automatic)"   "Electronic check"
## [5] "Mailed check"
```



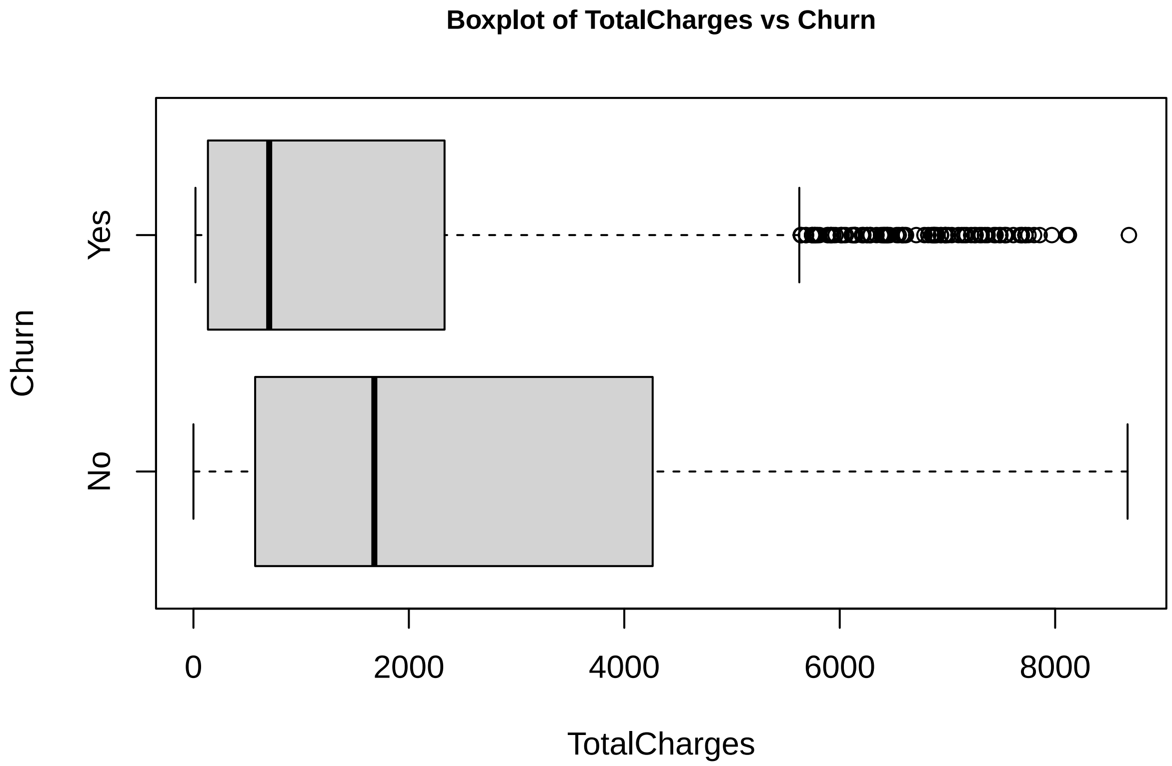


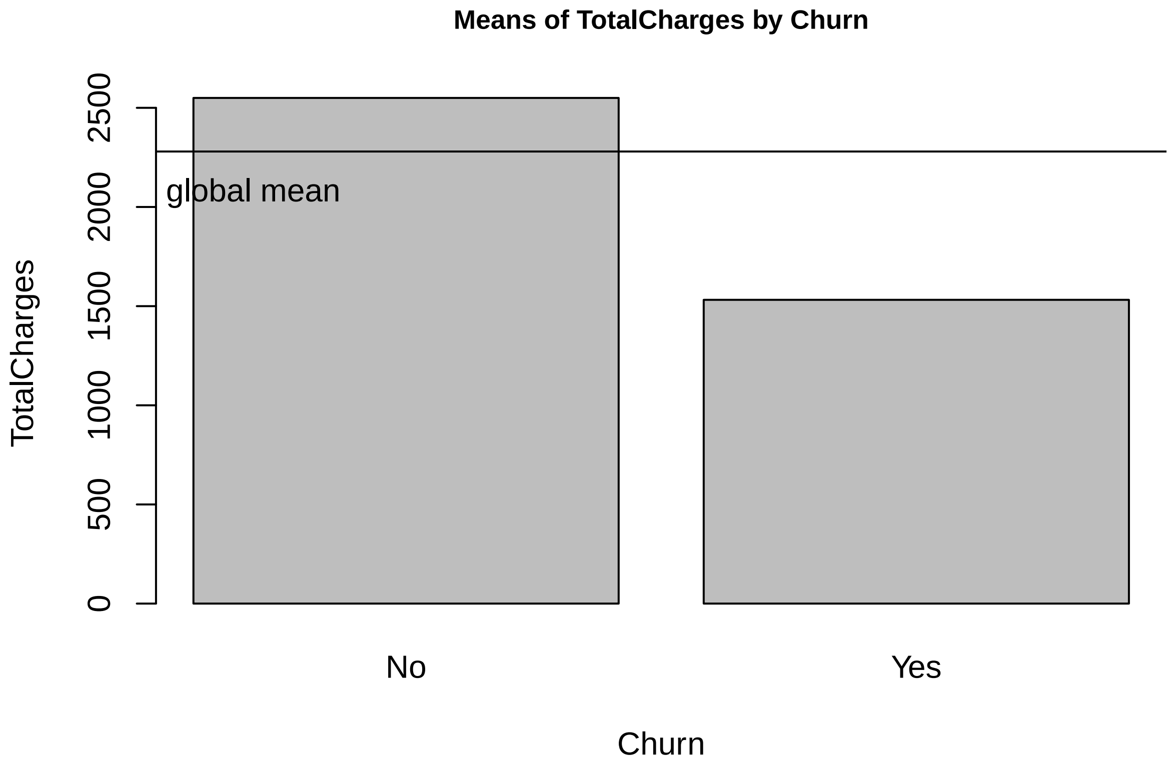
```
## [1] "Cross Table:"
##                            P
##                               No  Yes
##   Bank transfer (automatic) 1286  258
##   Credit card (automatic)   1290  232
##   Electronic check          1294 1071
##   Mailed check              1304  308
## [1] "Distributions by columns:"
##
## P     Bank transfer (automatic) Credit card (automatic) Electronic check
##   No                  0.8329016               0.8475690        0.5471459
##   Yes                 0.1670984               0.1524310        0.4528541
##
## P     Mailed check
##   No     0.8089330
##   Yes    0.1910670
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 648.14, df = 3, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P     Bank transfer (automatic) Credit card (automatic) Electronic check
##   No                  0.2485504               0.2493235        0.2500966
##   Yes                 0.1380417               0.1241306        0.5730337
##      Xquali
## P     Mailed check
##   No     0.2520294
##   Yes    0.1647940
##
## $vtest
##      Xquali
## P     Bank transfer (automatic) Credit card (automatic) Electronic check
##   No                   9.897550               11.270950       -25.337801
##   Yes                 -9.897550              -11.270950        25.337801
##      Xquali
## P     Mailed check
##   No      7.694261
##   Yes    -7.694261
##
## $pval
##      Xquali
## P     Bank transfer (automatic) Credit card (automatic) Electronic check
##   No               2.132984e-23            9.129469e-30     0.000000e+00
##   Yes              0.000000e+00            0.000000e+00    6.123943e-142
##      Xquali
## P      Mailed check
##   No   7.115733e-15
##   Yes  7.105427e-15
##
## [1] "Analysis by level of : MonthlyCharges"
```


Boxplot of MonthlyCharges vs Churn
Churn
Yes
No
20
40
60
80
100
120
MonthlyCharges

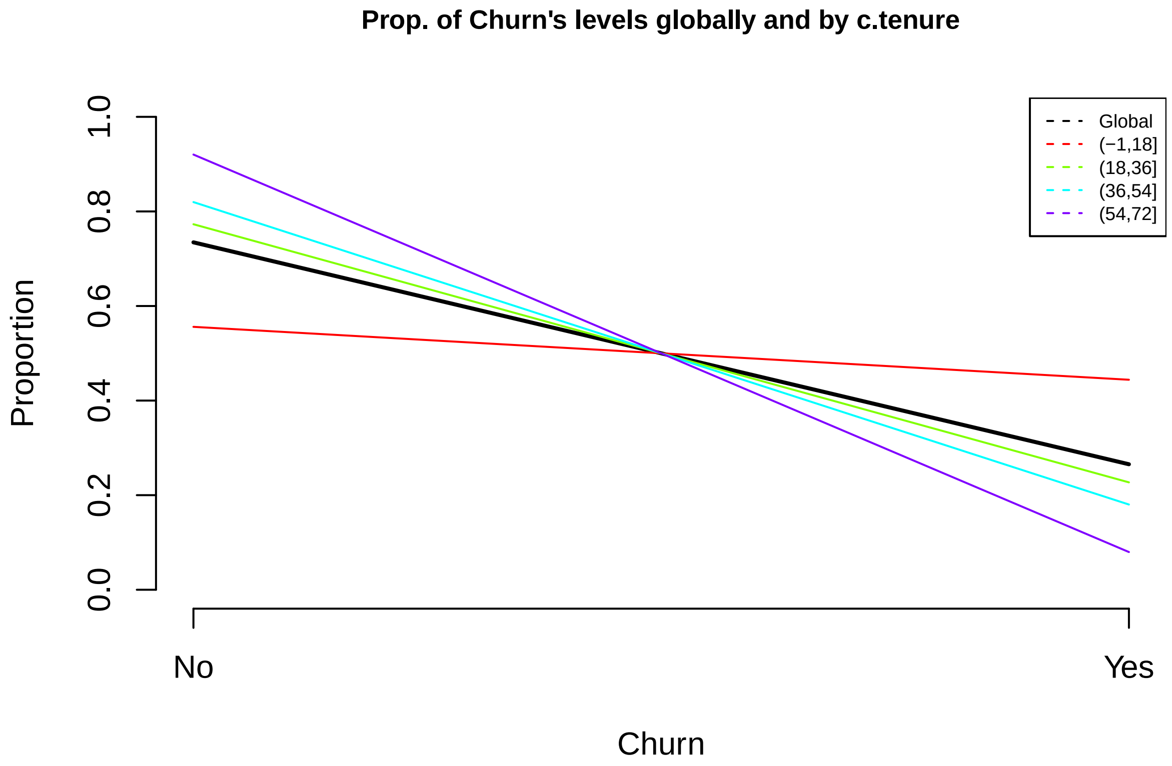


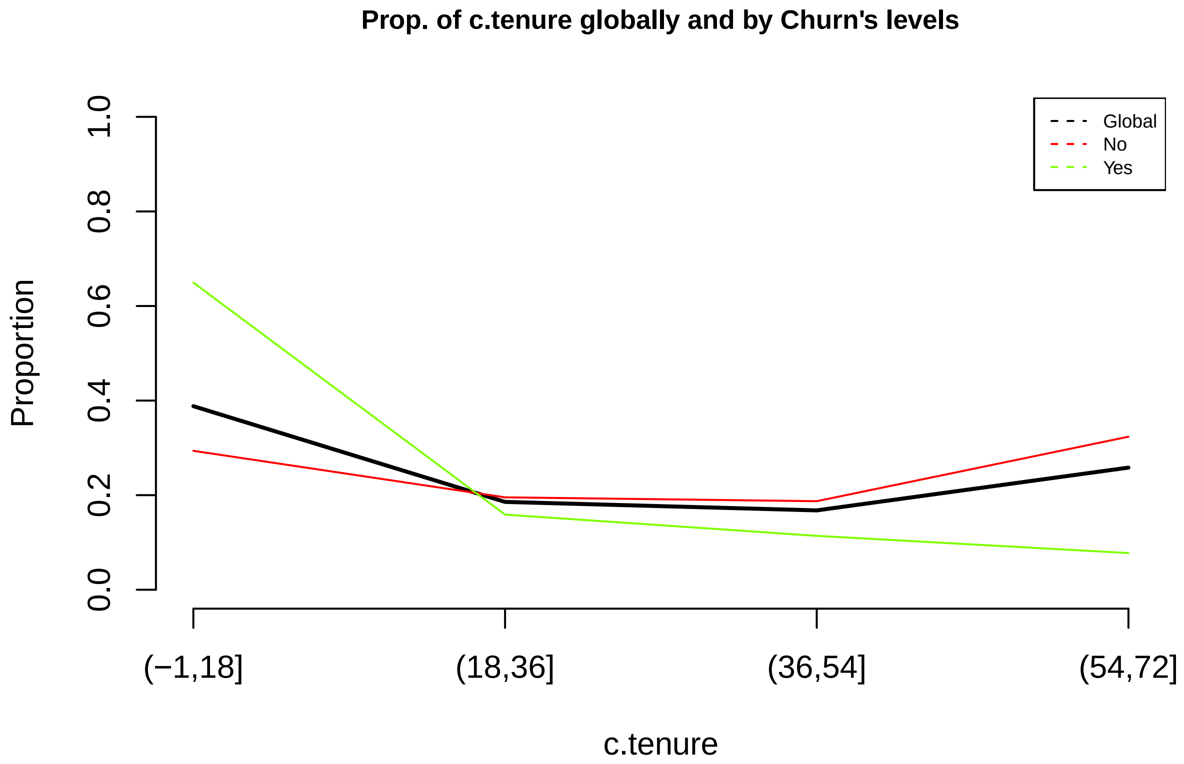
```
## [1] "Statistics by group:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.25   25.10   64.42   61.27   88.40  118.75
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.85   56.15   79.65   74.44   94.20  118.35
## [1] "p-valueANOVA: 8.59244933154708e-73"
## [1] "p-value Kruskal-Wallis: 3.31128554878381e-54"
## [1] "p-values ValorsTest: "
##           No          Yes
## 0.000000e+00 1.861643e-58
## [1] "Analysis by level of : TotalCharges"
```


Boxplot of TotalCharges vs Churn
Churn
Yes
No
0
2000
4000
6000
8000
TotalCharges



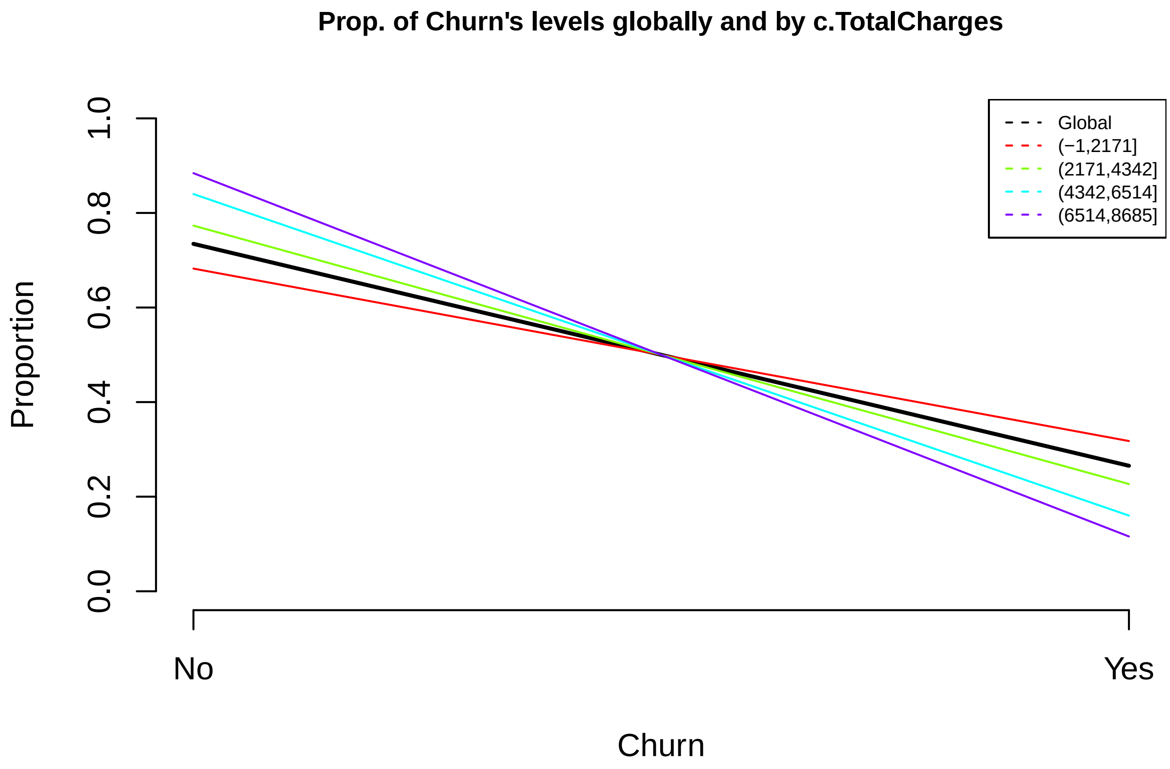
```
## [1] "Statistics by group:"
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##     0.0   572.9  1679.5  2549.9  4262.9  8672.5 
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
##   18.85  134.50  703.55 1531.80 2331.30 8684.80 
## [1] "p-valueANOVA: 5.90258060907269e-75"
## [1] "p-value Kruskal-Wallis: 5.68430392462642e-83"
## [1] "p-values ValorsTest: "
##           No          Yes 
## 2.476582e-61 0.000000e+00 
## [1] "Variable c.tenure"
## [1] "Categories=" "(-1,18]"     "(18,36]"     "(36,54]"     "(54,72]"
```

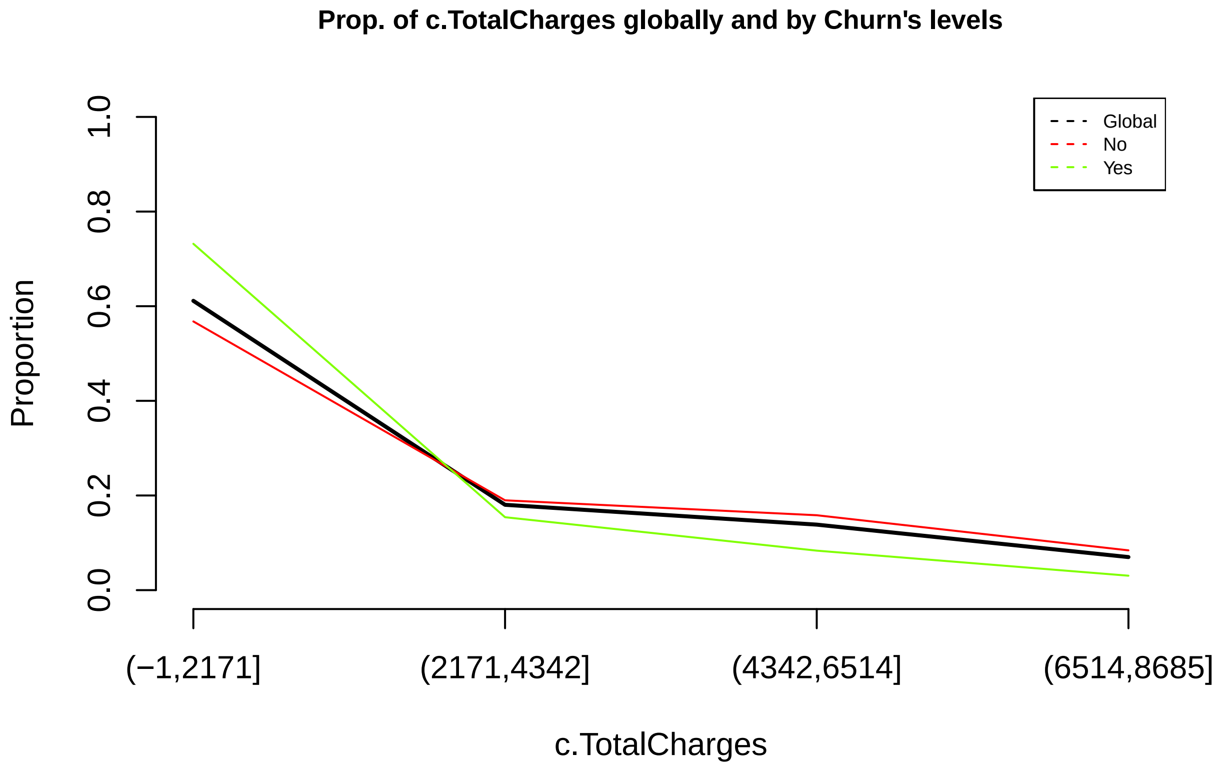

Prop. of Churn's levels globally and by c.tenure
Proportion
0.0
0.2
0.4
0.6
0.8
1.0
No
Yes
Churn
Global
(−1,18]
(18,36]
(36,54]
(54,72]



```
## [1] "Cross Table:"
##          P
##             No  Yes
##   (-1,18] 1520 1214
##   (18,36] 1011  297
##   (36,54]  969  213
##   (54,72] 1674  145
## [1] "Distributions by columns:"
##
## P        (-1,18]    (18,36]    (36,54]    (54,72]
##   No  0.55596196 0.77293578 0.81979695 0.92028587
##   Yes 0.44403804 0.22706422 0.18020305 0.07971413
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 823.12, df = 3, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P        (-1,18]    (18,36]    (36,54]    (54,72]
##   No  0.29377658 0.19540008 0.18728257 0.32354078
##   Yes 0.64954521 0.15890851 0.11396469 0.07758159
##
```

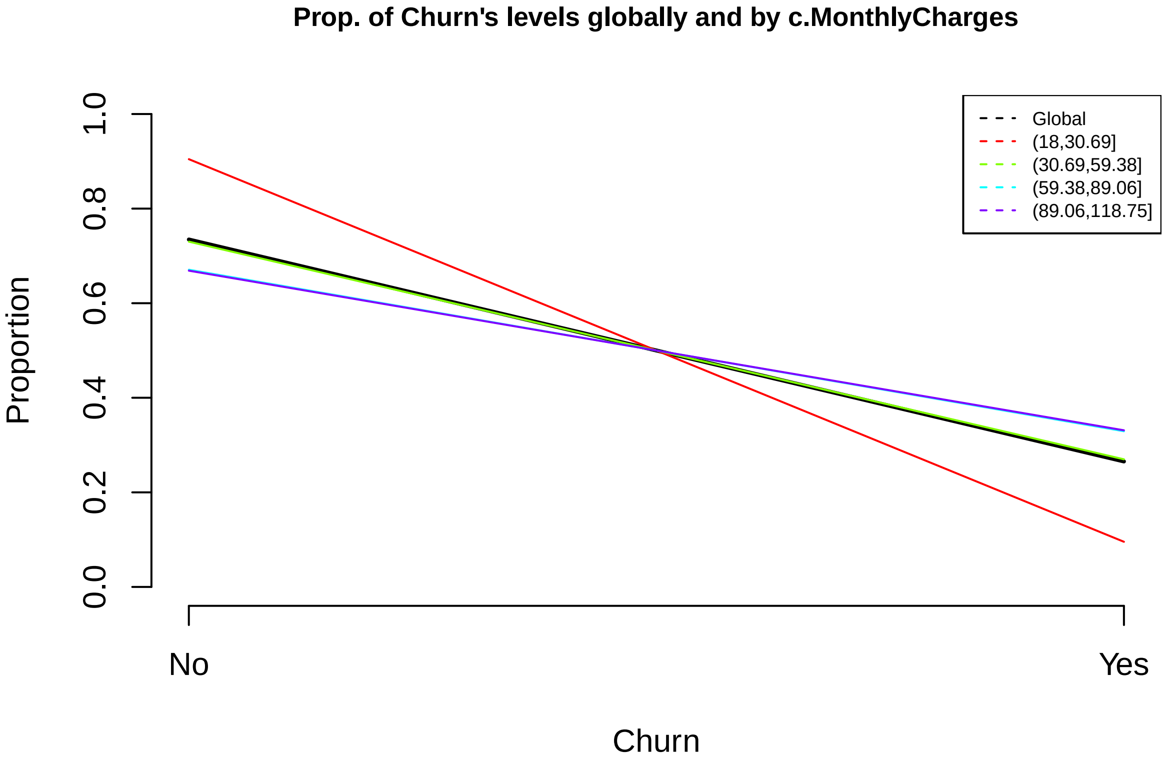
```
## $vtest
##      Xquali
## P         (-1,18]    (18,36]    (36,54]    (54,72]
##   No  -27.050598   3.477112   7.269625  20.822929
##   Yes  27.050598  -3.477112  -7.269625 -20.822929
##
## $pval
##      Xquali
## P            (-1,18]       (18,36]       (36,54]       (54,72]
##   No   0.000000e+00  2.534231e-04  1.802435e-13  1.341373e-96
##   Yes 1.879067e-161  2.534231e-04  1.801892e-13  0.000000e+00
##
## [1] "Variable c.TotalCharges"
## [1] "Categories=" "(-1,2171]"   "(2171,4342]" "(4342,6514]" "(6514,8685]"
```

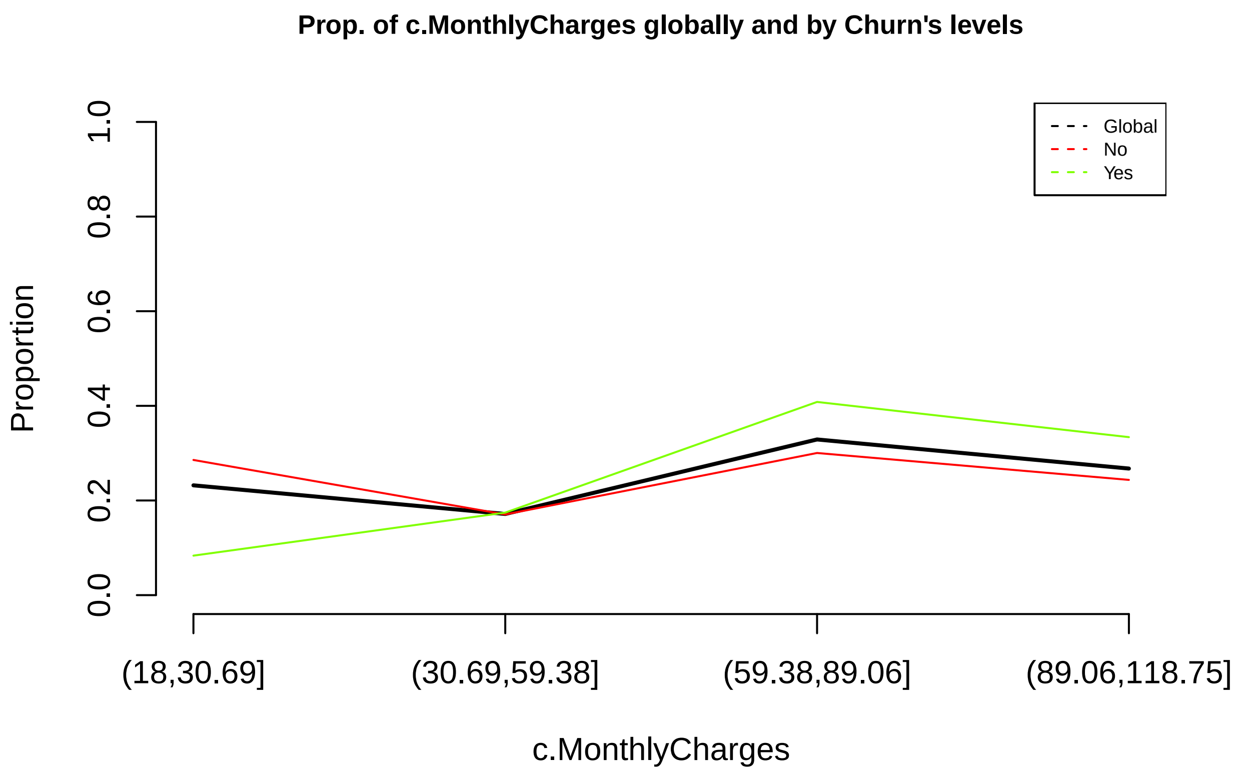




```
## [1] "Cross Table:"
##              P
##                 No  Yes
##   (-1,2171]   2938 1368
##   (2171,4342]  982  288
##   (4342,6514]  819  156
##   (6514,8685]  435   57
## [1] "Distributions by columns:"
##
## P     (-1,2171] (2171,4342] (4342,6514] (6514,8685]
##   No  0.6823038   0.7732283   0.8400000   0.8841463
##   Yes 0.3176962   0.2267717   0.1600000   0.1158537
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 182.13, df = 3, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P      (-1,2171] (2171,4342] (4342,6514] (6514,8685]
##   No  0.56783920  0.18979513  0.15829146  0.08407422
##   Yes 0.73194222  0.15409310  0.08346709  0.03049759
##
```

```
## $vtest
##       Xquali
## P      (-1,2171] (2171,4342] (4342,6514] (6514,8685]
##   No  -12.474952    3.441018    8.028134    7.788175
##   Yes  12.474952   -3.441018   -8.028134   -7.788175
##
## $pval
##       Xquali
## P        (-1,2171]  (2171,4342]  (4342,6514]  (6514,8685]
##   No  0.000000e+00 2.897645e-04 4.948298e-16 3.399196e-15
##   Yes 5.113421e-36 2.897645e-04 4.440892e-16 3.441691e-15
##
## [1] "Variable c.MonthlyCharges"
## [1] "Categories="    "(18,30.69]"     "(30.69,59.38]"  "(59.38,89.06]"
## [5] "(89.06,118.75]"
```





```
## [1] "Cross Table:"
##                  P
##                     No  Yes
##   (18,30.69]      1478  156
##   (30.69,59.38]    882  326
##   (59.38,89.06]   1554  763
##   (89.06,118.75]  1260  624
## [1] "Distributions by columns:"
##
## P     (18,30.69] (30.69,59.38] (59.38,89.06] (89.06,118.75]
##   No  0.90452876    0.73013245    0.67069486     0.66878981
##   Yes 0.09547124    0.26986755    0.32930514     0.33121019
## [1] "Chi^2 test: "
##
##  Pearson's Chi-squared test
##
## data:  dades[, k] and as.factor(P)
## X-squared = 332.54, df = 3, p-value < 2.2e-16
##
## [1] "ValorTestXquali:"
## $rowpf
##      Xquali
## P     (18,30.69] (30.69,59.38] (59.38,89.06] (89.06,118.75]
##   No  0.28565906    0.17046772    0.30034789     0.24352532
##   Yes 0.08346709    0.17442483    0.40823970     0.33386838
##
```

```
## $vtest
##      Xquali
## P      (18,30.69] (30.69,59.38] (59.38,89.06] (89.06,118.75]
##   No   17.7490901    -0.3889736    -8.5089368     -7.5625505
##   Yes -17.7490901     0.3889736     8.5089368      7.5625505
## 
## $pval
##      Xquali
## P       (18,30.69] (30.69,59.38] (59.38,89.06] (89.06,118.75]
##   No  8.758458e-71  3.486478e-01  0.000000e+00   1.976197e-14
##   Yes 0.000000e+00  3.486478e-01  8.776773e-18   1.976207e-14
## 
## [1] "P.values per class: No"
##           gender    SeniorCitizen          Partner       Dependents 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##     PhoneService    MultipleLines  InternetService   OnlineSecurity 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##     OnlineBackup DeviceProtection      TechSupport      StreamingTV 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##  StreamingMovies         Contract PaperlessBilling    PaymentMethod 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##   MonthlyCharges         c.tenure   c.TotalCharges c.MonthlyCharges 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##           tenure     TotalCharges 
##        2.08e-181         2.48e-61 
## [1] "P.values per class: Yes"
##           gender    SeniorCitizen          Partner       Dependents 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##           tenure     PhoneService    MultipleLines  InternetService 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##   OnlineSecurity     OnlineBackup DeviceProtection      TechSupport 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##      StreamingTV  StreamingMovies         Contract PaperlessBilling 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
##    PaymentMethod     TotalCharges         c.tenure   c.TotalCharges 
##         0.00e+00         0.00e+00         0.00e+00         0.00e+00 
## c.MonthlyCharges   MonthlyCharges 
##         0.00e+00         1.86e-58
```