Data mining report 1

Karol Pustelnik
index 249828

April 14, 2020

Contents

1 Short description of problem

In the report I will analyse clients from telephone network. I will focus mainly on those who had chosen international plan. In the report, my main goal is to answer the question: Why did some clients resign from services?

2 Description of data mining techniques

I will use basic techniques like:

- Aggregate indicators
- Basic data visualization
- Functions of my own or known from lecture
- dplyr package
- ggplot2 package

3 Results

Our data is information about some phone network clients. Lets take a glimpse on our data set:

```
dane <- read.csv(file = "churn.txt")
library(dplyr)
library(ggplot2)
```

```
dane1<-select(dane,-4) #Column 4 removal
attach(dane1)
daneInt<-dane1[which(dane$Int.l.Plan=="yes"),] #Creating subset of data
attach(daneInt) #setting column names

## The following objects are masked from dane1:
##
##     Account.Length, Area.Code, Churn., CustServ.Calls, Day.Calls,
##     Day.Charge, Day.Mins, Eve.Calls, Eve.Charge, Eve.Mins, Int.l.Plan,
##     Intl.Calls, Intl.Charge, Intl.Mins, Night.Calls, Night.Charge,
##     Night.Mins, State, VMail.Message, VMail.Plan

is.factor(Int.l.Plan) #checking types of variables

## [1] TRUE

is.numeric(CustServ.Calls)

## [1] TRUE
```

To begin data analysis, let's find aggregate indicators. I use function from data mining lecture.

```
my.summary <- function(X)
{
  wynik <- c(min(X),quantile(X,0.25), median(X), mean(X), quantile(X,0.75), max(X), var
  names(wynik) <- c("min", "Q1", "median", "mean", "Q3", "max", "var", "sd", "IQR")
  return(wynik)
}
my.summary(Intl.Calls)

##       min        Q1    median      mean        Q3       max       var        sd
##  1.000000  3.000000  4.000000  4.609907  6.000000 20.000000  6.915678  2.629768
##       IQR
##  3.000000
```

```
my.summary(Intl.Charge)

##       min        Q1    median      mean        Q3       max       var        sd
## 0.3500000 2.4300000 2.9200000 2.8699071 3.2900000 5.4000000 0.5302034 0.7281507
##       IQR
## 0.8600000

my.summary(Intl.Mins)

##       min        Q1    median      mean        Q3       max       var        sd
##  1.300000  9.000000 10.800000 10.628173 12.200000 20.000000  7.278055  2.697787
##       IQR
##  3.200000

table(dane1$Int.l.Plan)

##
##   no  yes
## 3010  323
```

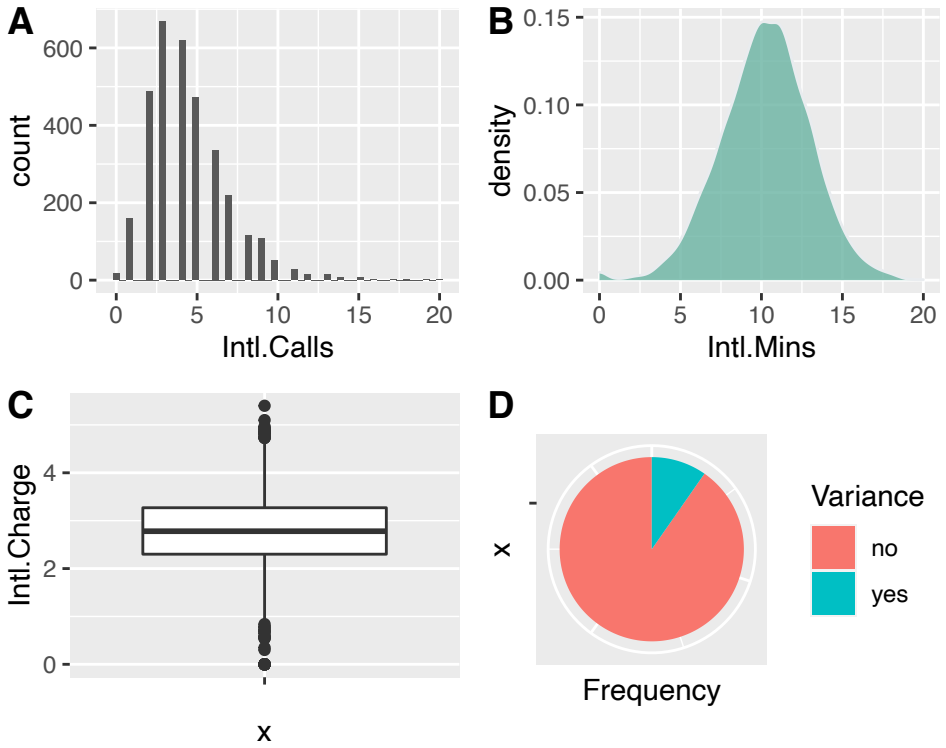
Now let's plot some graphs that will shed more light on our data.

```
p1<-ggplot(dane1, aes(x=Intl.Calls)) +
  geom_histogram(bins=nclass.FD(dane1$Intl.Calls))

p2<-ggplot(dane1,aes(x=Intl.Mins)) +
  geom_density(fill="#69b3a2", color="#e9ecef", alpha=0.8)

p3<-ggplot(dane1, aes(x="", y=Intl.Charge)) +
  geom_boxplot()

Int_numbers<-data.frame(table(dane$Int.l.Plan))
Frequency<-Int_numbers$Freq
Variance<-Int_numbers$Var1
p4<-ggplot(Int_numbers, aes(x="", y=Frequency, fill=Variance)) +
  geom_bar(stat="identity", width=1) +
  coord_polar("y", start=0)
p5<-ggplot(dane1, aes(x=Intl.Mins, y=Intl.Charge)) +
  geom_point( color="#69b3a2")
ggarrange(p1, p2, p3, p4 + rremove("x.text"),
          labels = c("A", "B", "C","D","E"), #naming plots
          ncol = 2, nrow = 2)
```



It's too early to draw conclusions. Let's see if some of the variables are correlated. To do this, I use:

```
numeric_data<-Filter(is.numeric, dane1) #subsetting data
numeric_data<-data.frame(numeric_data)
n<-length(numeric_data)
n

## [1] 16

output<-matrix(ncol=n,nrow=n) #creating a loop to fill up matrix
  for (i in 1:n)
    for(j in 1:n)
      if(i!=j)
          output[i,j]<-cor(numeric_data[i],numeric_data[j])
output[is.na(output)]<-0
correlation_matrix<-output>0.95 #keeping only significant correlation
correlation_matrix

##        [,1]  [,2]  [,3]  [,4]  [,5]  [,6]  [,7]  [,8]  [,9] [,10] [,11] [,12]
##  [1,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [2,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [3,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [4,] FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE
##  [5,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [6,] FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [7,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE
##  [8,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##  [9,] FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE FALSE FALSE FALSE
```

```
## [10,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE
## [11,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [12,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE  TRUE FALSE FALSE
## [13,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [14,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [15,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
## [16,] FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE
##       [,13] [,14] [,15] [,16]
##  [1,] FALSE FALSE FALSE FALSE
##  [2,] FALSE FALSE FALSE FALSE
##  [3,] FALSE FALSE FALSE FALSE
##  [4,] FALSE FALSE FALSE FALSE
##  [5,] FALSE FALSE FALSE FALSE
##  [6,] FALSE FALSE FALSE FALSE
##  [7,] FALSE FALSE FALSE FALSE
##  [8,] FALSE FALSE FALSE FALSE
##  [9,] FALSE FALSE FALSE FALSE
## [10,] FALSE FALSE FALSE FALSE
## [11,] FALSE FALSE FALSE FALSE
## [12,] FALSE FALSE FALSE FALSE
## [13,] FALSE FALSE  TRUE FALSE
## [14,] FALSE FALSE FALSE FALSE
## [15,]  TRUE FALSE FALSE FALSE
## [16,] FALSE FALSE FALSE FALSE
```

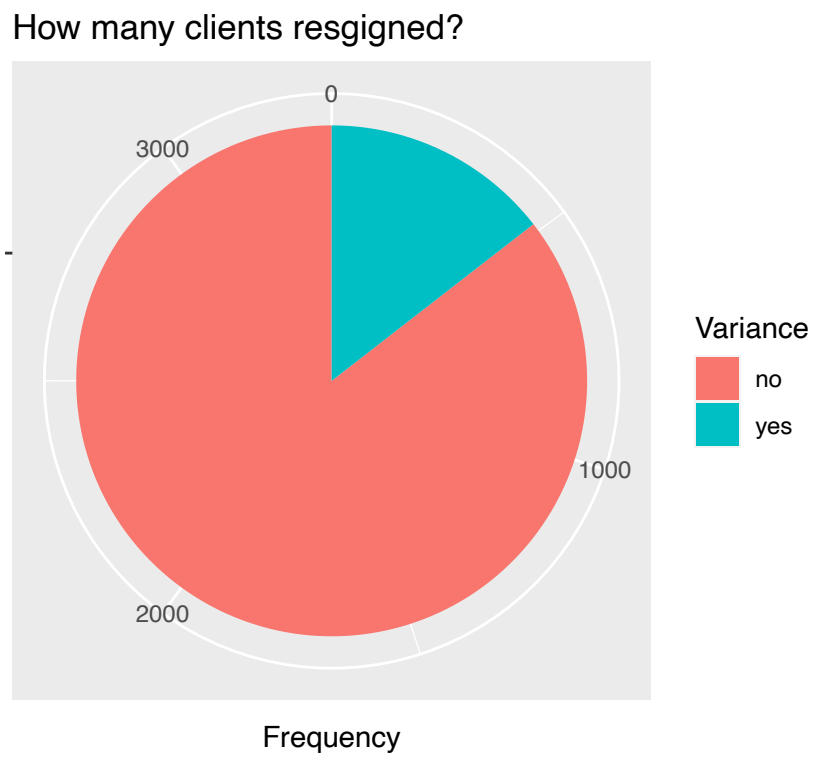
```
plot_matrix1<-which(output>0.95,arr.ind=TRUE)
plot_matrix2<-matrix(nrow=4,ncol=2)
n<-length(plot_matrix1[,1])/2
for(i in 1:n)
  plot_matrix2[i,]<-plot_matrix1[2*i,]
plot_matrix2 #coordinates matrix of correlated variables
```

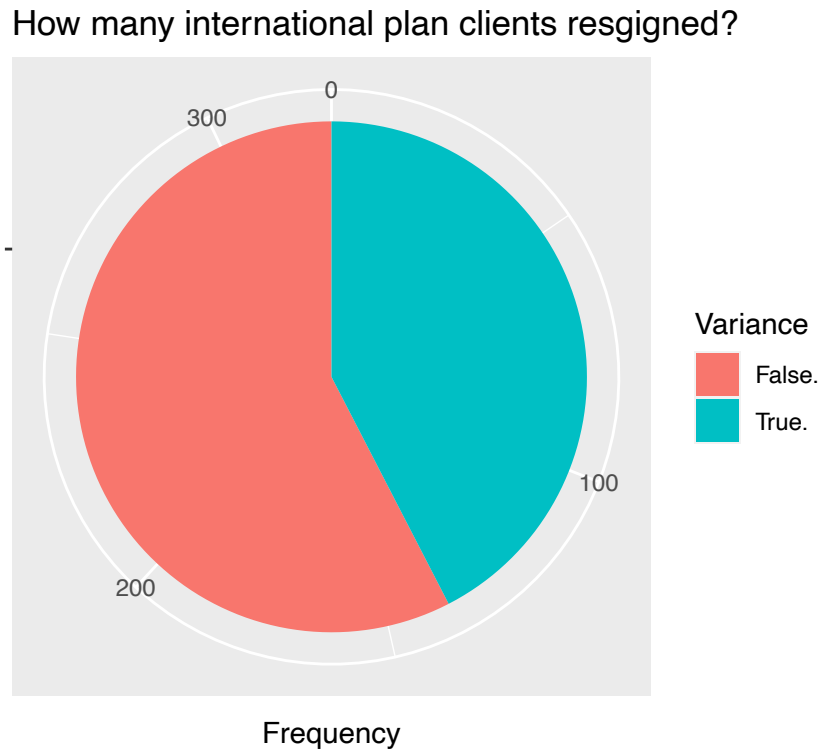
```
##      [,1] [,2]
## [1,]    4    6
## [2,]    7    9
## [3,]   10   12
## [4,]   13   15
```

From the matrix above, we can easily say which variables are correlated.
To draw important conclusions let's create plots for loyal and former clients.

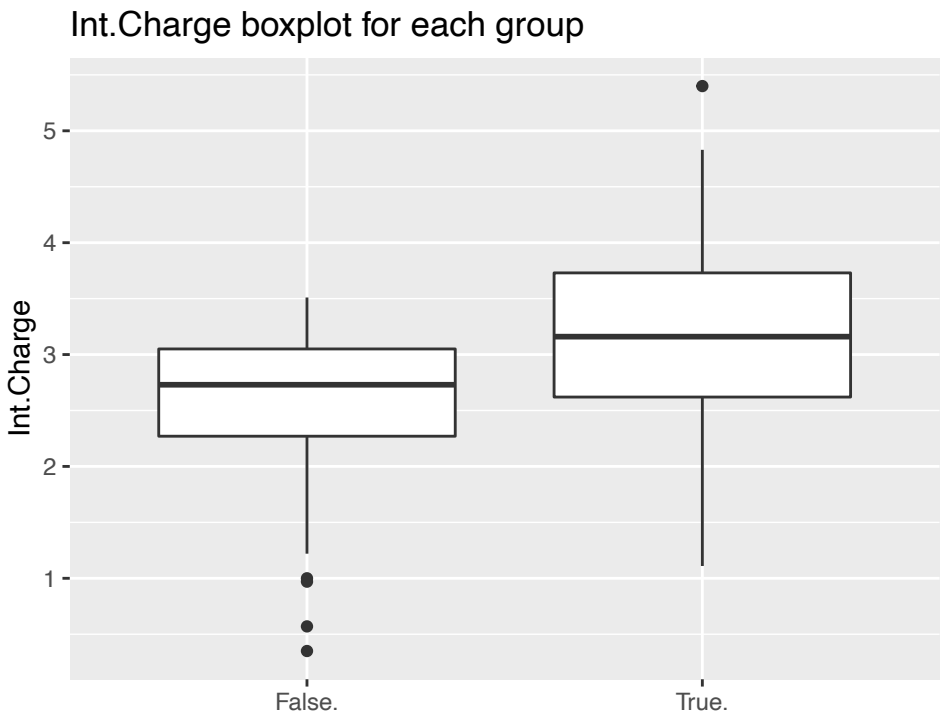
```
dane.lojalni <- data.frame(subset(daneInt, Churn.=="False.")) #subsetting data
dane.odeszli <- subset(daneInt, Churn.=="True.")
Churn_data<-data.frame(table(dane$Churn.))
Frequency1<-Churn_data$Freq
Variance1<-Churn_data$Var1
ggplot(Churn_data, aes(x="", y=Frequency1, fill=Variance)) +
  geom_bar(stat="identity", width=1) +
```

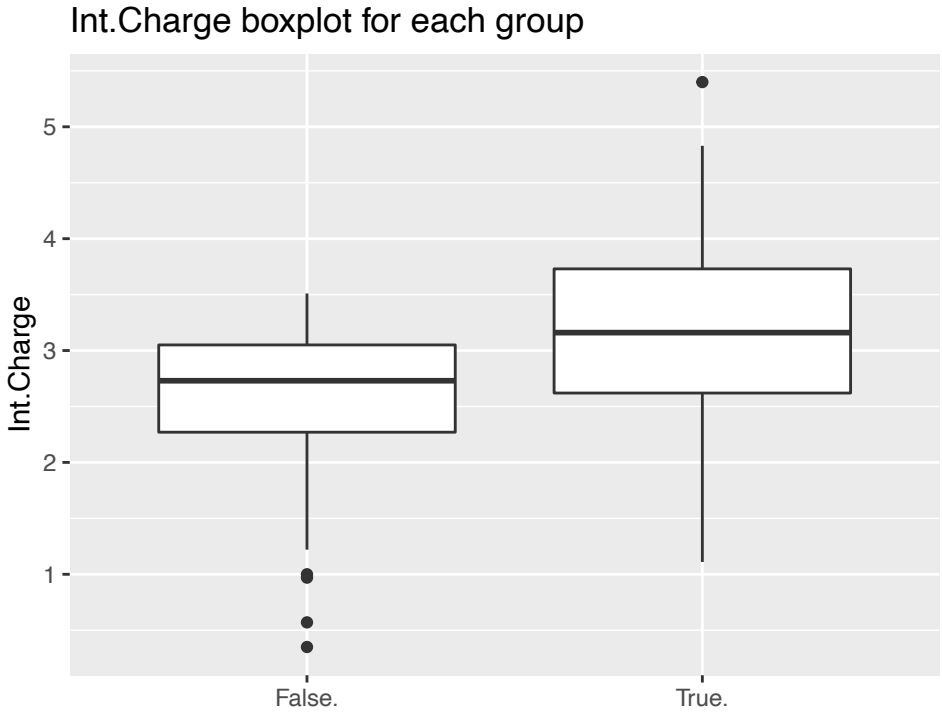
```
  coord_polar("y", start=0)+
  labs(x="", y="Frequency",fill="Variance",title = "How many clients resgigned?")
```



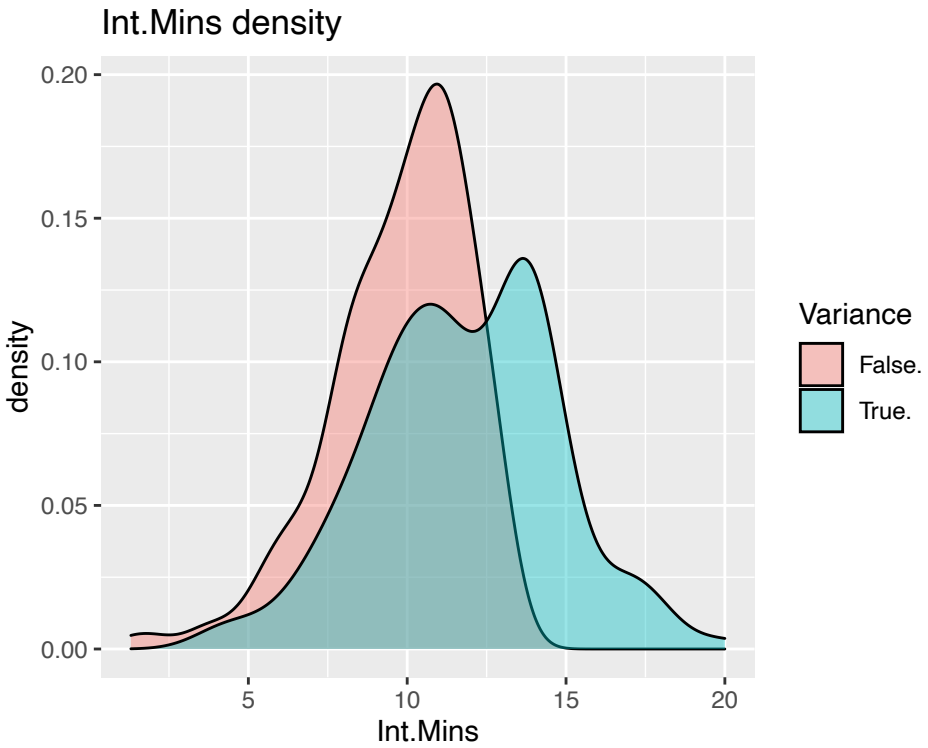


As we can see many international plan clients resigned compared to all clients.

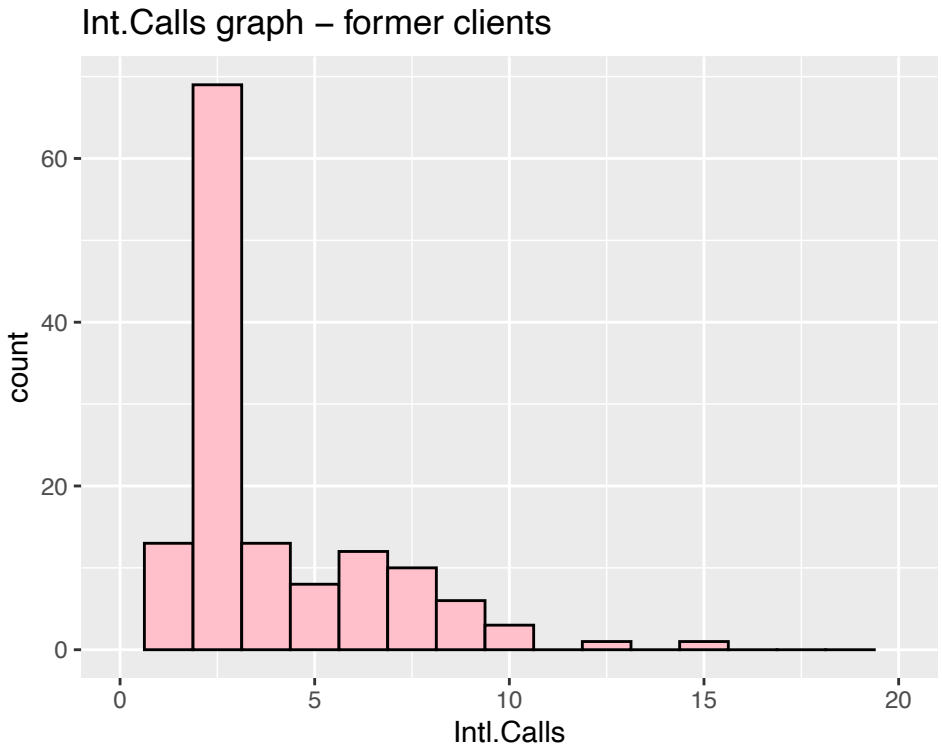


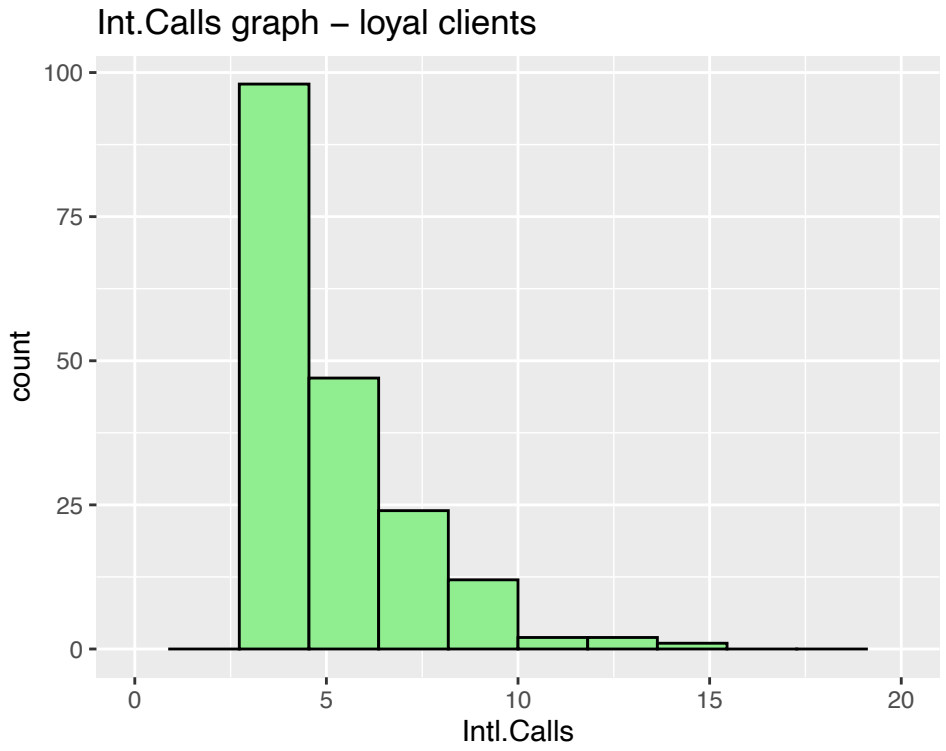


Clients who resigned, paid more for services

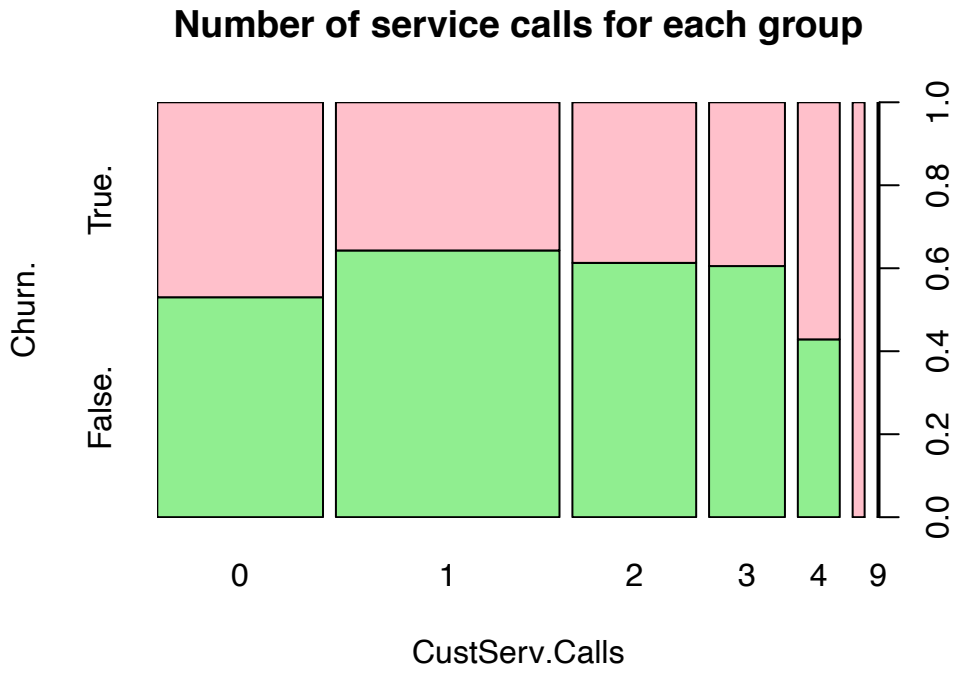


Clients who resigned, had longer talks

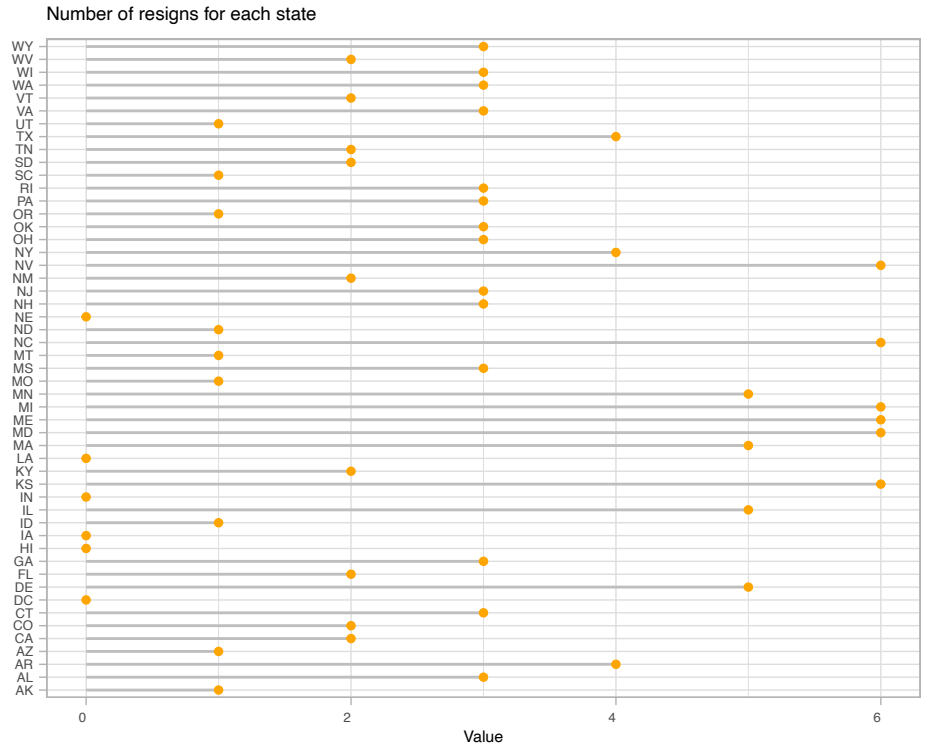




Many clients who resigned, didn't even use the service



Clients who resigned, called customer service more often. Let's see if resignation is linked to state of living:



As we can see there are states where more clients resigned.

4 Summary

To summarise, I think that the most important reasons why some clients resigned are:

- Bad cellphone infrastructure in some states
- Too high price for services
- Unprofitable plan
- Not very helpful customer service
- Mismatched plan (many clients didn't even call internationaly)

References

[1] Stackoverflow ,https://stackoverflow.com

[2] Data-visualisation site, https://www.data-to-viz.com

[3] Data mining course site, http://prac.im.pwr.wroc.pl/~zagdan/polish_ver/ED2020/index.html