Case\_Study\_1740256

Jeevan-1740256

24 January 2019

Analyse the churn in Telecom’s dataset

setwd("C:/Users/Jeevan/Desktop/Christ University/R Studio") #setting the directory  
datas = read.csv('bigml\_59c28831336c6604c800002a.csv') #importing the dataset  
head(datas,5) #1st 5 rows of the dataset

## state account.length area.code phone.number international.plan  
## 1 KS 128 415 382-4657 no  
## 2 OH 107 415 371-7191 no  
## 3 NJ 137 415 358-1921 no  
## 4 OH 84 408 375-9999 yes  
## 5 OK 75 415 330-6626 yes  
## voice.mail.plan number.vmail.messages total.day.minutes total.day.calls  
## 1 yes 25 265.1 110  
## 2 yes 26 161.6 123  
## 3 no 0 243.4 114  
## 4 no 0 299.4 71  
## 5 no 0 166.7 113  
## total.day.charge total.eve.minutes total.eve.calls total.eve.charge  
## 1 45.07 197.4 99 16.78  
## 2 27.47 195.5 103 16.62  
## 3 41.38 121.2 110 10.30  
## 4 50.90 61.9 88 5.26  
## 5 28.34 148.3 122 12.61  
## total.night.minutes total.night.calls total.night.charge  
## 1 244.7 91 11.01  
## 2 254.4 103 11.45  
## 3 162.6 104 7.32  
## 4 196.9 89 8.86  
## 5 186.9 121 8.41  
## total.intl.minutes total.intl.calls total.intl.charge  
## 1 10.0 3 2.70  
## 2 13.7 3 3.70  
## 3 12.2 5 3.29  
## 4 6.6 7 1.78  
## 5 10.1 3 2.73  
## customer.service.calls churn  
## 1 1 FALSE  
## 2 1 FALSE  
## 3 0 FALSE  
## 4 2 FALSE  
## 5 3 FALSE

View(datas) #viewing the whole dataset  
summary(datas)

## state account.length area.code phone.number   
## WV : 106 Min. : 1.0 Min. :408.0 327-1058: 1   
## MN : 84 1st Qu.: 74.0 1st Qu.:408.0 327-1319: 1   
## NY : 83 Median :101.0 Median :415.0 327-3053: 1   
## AL : 80 Mean :101.1 Mean :437.2 327-3587: 1   
## OH : 78 3rd Qu.:127.0 3rd Qu.:510.0 327-3850: 1   
## OR : 78 Max. :243.0 Max. :510.0 327-3954: 1   
## (Other):2824 (Other) :3327   
## international.plan voice.mail.plan number.vmail.messages  
## no :3010 no :2411 Min. : 0.000   
## yes: 323 yes: 922 1st Qu.: 0.000   
## Median : 0.000   
## Mean : 8.099   
## 3rd Qu.:20.000   
## Max. :51.000   
##   
## total.day.minutes total.day.calls total.day.charge total.eve.minutes  
## Min. : 0.0 Min. : 0.0 Min. : 0.00 Min. : 0.0   
## 1st Qu.:143.7 1st Qu.: 87.0 1st Qu.:24.43 1st Qu.:166.6   
## Median :179.4 Median :101.0 Median :30.50 Median :201.4   
## Mean :179.8 Mean :100.4 Mean :30.56 Mean :201.0   
## 3rd Qu.:216.4 3rd Qu.:114.0 3rd Qu.:36.79 3rd Qu.:235.3   
## Max. :350.8 Max. :165.0 Max. :59.64 Max. :363.7   
##   
## total.eve.calls total.eve.charge total.night.minutes total.night.calls  
## Min. : 0.0 Min. : 0.00 Min. : 23.2 Min. : 33.0   
## 1st Qu.: 87.0 1st Qu.:14.16 1st Qu.:167.0 1st Qu.: 87.0   
## Median :100.0 Median :17.12 Median :201.2 Median :100.0   
## Mean :100.1 Mean :17.08 Mean :200.9 Mean :100.1   
## 3rd Qu.:114.0 3rd Qu.:20.00 3rd Qu.:235.3 3rd Qu.:113.0   
## Max. :170.0 Max. :30.91 Max. :395.0 Max. :175.0   
##   
## total.night.charge total.intl.minutes total.intl.calls total.intl.charge  
## Min. : 1.040 Min. : 0.00 Min. : 0.000 Min. :0.000   
## 1st Qu.: 7.520 1st Qu.: 8.50 1st Qu.: 3.000 1st Qu.:2.300   
## Median : 9.050 Median :10.30 Median : 4.000 Median :2.780   
## Mean : 9.039 Mean :10.24 Mean : 4.479 Mean :2.765   
## 3rd Qu.:10.590 3rd Qu.:12.10 3rd Qu.: 6.000 3rd Qu.:3.270   
## Max. :17.770 Max. :20.00 Max. :20.000 Max. :5.400   
##   
## customer.service.calls churn   
## Min. :0.000 Mode :logical   
## 1st Qu.:1.000 FALSE:2850   
## Median :1.000 TRUE :483   
## Mean :1.563   
## 3rd Qu.:2.000   
## Max. :9.000   
##

names(datas) #knowing the names of the columns in the sheet

## [1] "state" "account.length"   
## [3] "area.code" "phone.number"   
## [5] "international.plan" "voice.mail.plan"   
## [7] "number.vmail.messages" "total.day.minutes"   
## [9] "total.day.calls" "total.day.charge"   
## [11] "total.eve.minutes" "total.eve.calls"   
## [13] "total.eve.charge" "total.night.minutes"   
## [15] "total.night.calls" "total.night.charge"   
## [17] "total.intl.minutes" "total.intl.calls"   
## [19] "total.intl.charge" "customer.service.calls"  
## [21] "churn"

mean.totaldaycalls = mean(datas$total.day.calls,na.rm = TRUE) #getting the mean of total day calls  
mean.totaldaycalls

## [1] 100.4356

plot(datas$total.day.calls,datas$total.eve.calls,main = "Total Day Calls vs Total Evening Calls",xlab = "Total Day calls",ylab = "Total Evening Calls",pch=20,data = datas) #plotting total day calls with total evening calls

## Warning in plot.window(...): "data" is not a graphical parameter

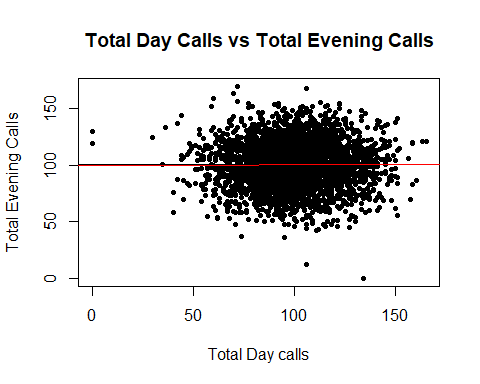
## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not  
## a graphical parameter  
  
## Warning in axis(side = side, at = at, labels = labels, ...): "data" is not  
## a graphical parameter

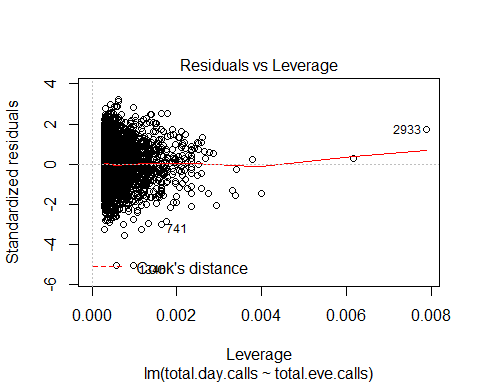
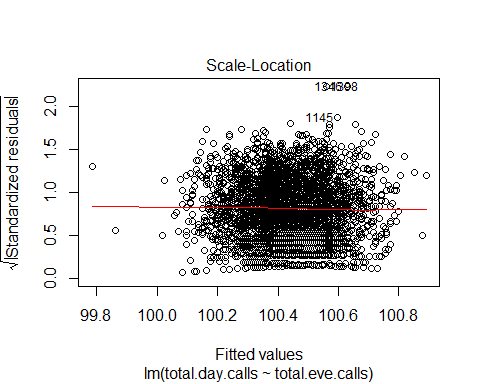
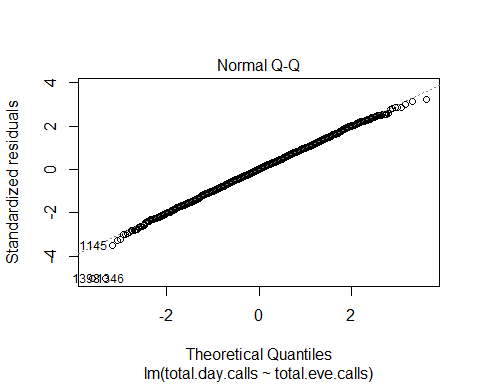
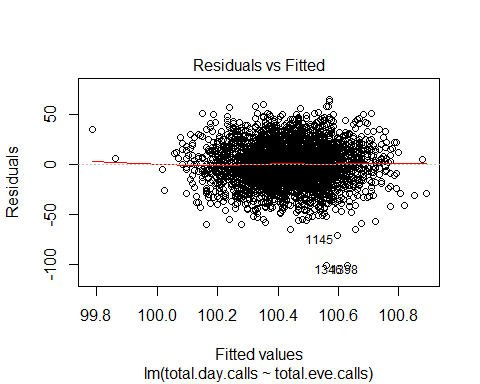
## Warning in box(...): "data" is not a graphical parameter

## Warning in title(...): "data" is not a graphical parameter

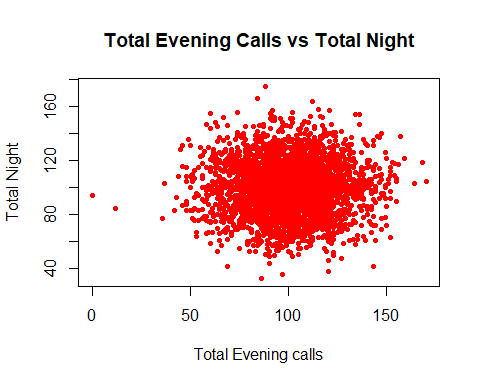
abline(h=mean.totaldaycalls) #adding the mean to the graph  
model1=lm(total.day.calls~total.eve.calls,data=datas) #getting a linear model  
abline(model1,col="red") #Fitting the linear model to the graph



plot(model1)



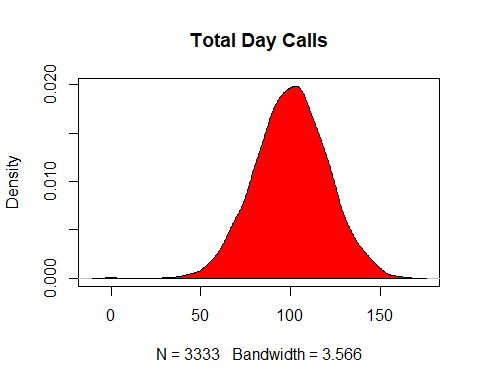
plot(datas$total.eve.calls,datas$total.night.calls,main = "Total Evening Calls vs Total Night",xlab = "Total Evening calls",ylab = "Total Night",pch=20,col='red')



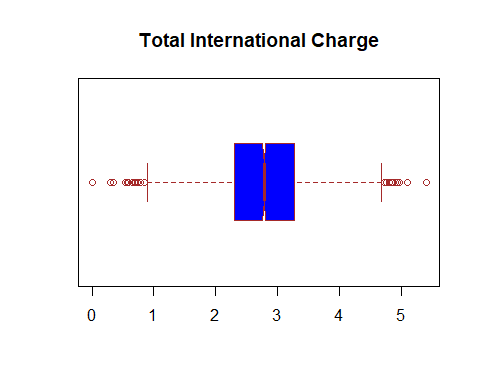
d=density(datas$total.day.calls) #Kernel density plot  
d

##   
## Call:  
## density.default(x = datas$total.day.calls)  
##   
## Data: datas$total.day.calls (3333 obs.); Bandwidth 'bw' = 3.566  
##   
## x y   
## Min. :-10.7 Min. :1.500e-08   
## 1st Qu.: 35.9 1st Qu.:5.925e-05   
## Median : 82.5 Median :1.511e-03   
## Mean : 82.5 Mean :5.360e-03   
## 3rd Qu.:129.1 3rd Qu.:1.003e-02   
## Max. :175.7 Max. :1.982e-02

plot(d,main="Total Day Calls")  
polygon(d,col="red",border="black")



boxplot(datas$total.intl.charge,main = "Total International Charge",col="blue",border="brown",horizontal = TRUE,notch = TRUE)



datas$churn = factor(datas$churn, #converting textual values to numerical values  
 levels = c('TRUE','FALSE'),  
 labels = c(1,0))  
datas$voice.mail.plan = factor(datas$voice.mail.plan,  
 levels = c('yes','no'),  
 labels = c(1,0))  
datas$international.plan = factor(datas$international.plan,  
 levels = c('yes','no'),  
 labels = c(1,0))  
datas$total.day.calls = ifelse(is.na(datas$total.day.calls),  
 ave(datas$total.day.calls,FUN = function(x) mean(x,na.rm = TRUE)),  
 datas$total.day.calls)  
datas$total.day.minutes = ifelse(is.na(datas$total.day.minutes),  
 ave(datas$total.day.minutes,FUN = function(x) mean(x,na.rm = TRUE)),  
 datas$total.day.minutes)

In this case study, the excel sheet of churning the Telecom’s dataset has been imported into RStudio. To have a rough overview of how the data is, the first 5 rows of the sheet has been coded. In the summary of the sheet - we can see minimum value, maximum value, 1st Quartile & 3rd Quartile deviation. The mean and the median for each column are also shown in the summary. The code with the names of the sheet tells us the different columns that are ther. We are finding the mean of total day calls to replace all the values that are not applicable. We are then plotting the graph between Total Day calls & Total evening to see whether they are related. Using the concept of linear regression, we are fitting a linear model to the graph of total day calls. A kernel density plot is shown of the total day calls which tells us that the graph is normally distributed. To replace the textual data - to make TRUE = 1 & FALSE = 0, we are using the factors function and labelling it with 1 & 0. This is the case for the churn column, voice mail plan and international plan. The total day calls and total day minutes might have nil values and we are replacing that with the mean of the corresponding column.