Case study on data of telecom industry

Jeevan Koshy, Parth Takkar, Anushka Gupta

24th January 2019

**OBJECTIVES**

-To obtain and import a dataset in RStudio.

-To find the descriptive statistics for various fields in the data obtained regarding the telecom industry including the total number of calls during different times of the day, the charges and the total number of minutes spent on call.

-To find if the charges and the minutes spent on the call have any correlation.

-To check whether there is any significant difference in the evening call total charges and the night call total charges.

-To check whether high cost leads to higher requirement of customer service.

-To visualize the data using various graphs.

**PROCEDURE**

Importing the dataset regarding the telecom industry in RStudio.

setwd("C:/Users/Jeevan/Desktop/Christ University/R Studio") #setting the directory  
datafile=read.csv('bigml\_59c28831336c6604c800002a.csv') #importing the dataset  
print("day minutes")

## [1] "day minutes"

summary(datafile$`total day minutes`)

## Length Class Mode   
## 0 NULL NULL

print("eve minutes")

## [1] "eve minutes"

summary(datafile$`total eve minutes`)

## Length Class Mode   
## 0 NULL NULL

print("night minutes")

## [1] "night minutes"

summary(datafile$`total night minutes`)

## Length Class Mode   
## 0 NULL NULL

print("Sample of size 25")

## [1] "Sample of size 25"

sample1<-sample(1:nrow(datafile),25)  
df1<-data.frame(datafile[sample1,])  
df1

## state account.length area.code phone.number international.plan  
## 569 IN 108 510 329-1955 no  
## 3203 WA 143 510 340-4989 no  
## 1809 NH 83 415 415-6145 no  
## 325 VA 129 408 384-2632 no  
## 1359 ND 51 510 337-3740 no  
## 1781 AL 68 510 344-4970 no  
## 428 NH 67 415 355-1113 no  
## 1581 WI 120 415 414-2905 no  
## 2538 NC 122 415 396-8662 no  
## 3087 NY 54 510 390-6932 yes  
## 3000 KY 76 415 407-8575 no  
## 2368 AK 74 415 336-6533 no  
## 2975 UT 201 510 373-8900 no  
## 604 MI 53 415 346-5707 no  
## 1114 IA 152 415 387-6716 no  
## 1534 WY 127 510 400-2181 yes  
## 972 TN 59 415 399-5564 no  
## 2370 NY 112 415 391-1737 no  
## 786 PA 69 415 390-5686 no  
## 3201 CT 100 510 416-1536 yes  
## 2621 TN 115 415 374-6525 no  
## 2362 MA 66 415 416-7393 no  
## 3125 ND 75 408 396-4171 no  
## 2442 HI 111 408 401-6671 no  
## 1381 WI 54 415 364-8981 no  
## voice.mail.plan number.vmail.messages total.day.minutes  
## 569 no 0 293.0  
## 3203 no 0 160.4  
## 1809 no 0 231.3  
## 325 no 0 207.0  
## 1359 no 0 227.2  
## 1781 no 0 157.3  
## 428 yes 40 104.9  
## 1581 yes 29 244.3  
## 2538 no 0 215.6  
## 3087 no 0 236.3  
## 3000 no 0 204.0  
## 2368 no 0 262.3  
## 2975 no 0 212.7  
## 604 no 0 57.5  
## 1114 no 0 206.3  
## 1534 no 0 242.2  
## 972 no 0 160.9  
## 2370 no 0 174.3  
## 786 no 0 228.2  
## 3201 no 0 107.2  
## 2621 no 0 206.2  
## 2362 no 0 116.4  
## 3125 yes 24 225.5  
## 2442 yes 13 193.1  
## 1381 no 0 116.8  
## total.day.calls total.day.charge total.eve.minutes total.eve.calls  
## 569 88 49.81 160.6 101  
## 3203 120 27.27 285.9 104  
## 1809 100 39.32 210.4 84  
## 325 91 35.19 154.9 121  
## 1359 89 38.62 194.4 106  
## 1781 83 26.74 220.9 85  
## 428 65 17.83 216.3 93  
## 1581 140 41.53 322.3 89  
## 2538 86 36.65 167.8 59  
## 3087 91 40.17 152.8 130  
## 3000 69 34.68 225.1 110  
## 2368 114 44.59 198.9 96  
## 2975 72 36.16 225.2 90  
## 604 95 9.78 265.5 131  
## 1114 98 35.07 292.8 82  
## 1534 102 41.17 226.1 80  
## 972 95 27.35 251.2 65  
## 2370 123 29.63 140.2 124  
## 786 70 38.79 263.7 80  
## 3201 98 18.22 86.8 122  
## 2621 113 35.05 176.4 102  
## 2362 98 19.79 95.6 74  
## 3125 119 38.34 182.0 108  
## 2442 104 32.83 111.6 98  
## 1381 119 19.86 123.1 123  
## total.eve.charge total.night.minutes total.night.calls  
## 569 13.65 143.9 87  
## 3203 24.30 182.5 85  
## 1809 17.88 217.4 106  
## 325 13.17 245.1 112  
## 1359 16.52 243.4 126  
## 1781 18.78 218.9 129  
## 428 18.39 217.4 128  
## 1581 27.40 166.8 83  
## 2538 14.26 207.0 67  
## 3087 12.99 160.3 98  
## 3000 19.13 240.3 85  
## 2368 16.91 165.9 90  
## 2975 19.14 195.1 99  
## 604 22.57 244.3 128  
## 1114 24.89 43.7 121  
## 1534 19.22 252.0 96  
## 972 21.35 273.4 97  
## 2370 11.92 215.4 89  
## 786 22.41 142.6 60  
## 3201 7.38 156.2 117  
## 2621 14.99 297.1 119  
## 2362 8.13 181.5 94  
## 3125 15.47 270.9 106  
## 2442 9.49 227.4 94  
## 1381 10.46 217.5 101  
## total.night.charge total.intl.minutes total.intl.calls  
## 569 6.48 10.0 6  
## 3203 8.21 6.9 4  
## 1809 9.78 12.4 2  
## 325 11.03 13.4 5  
## 1359 10.95 14.9 2  
## 1781 9.85 12.0 7  
## 428 9.78 9.6 9  
## 1581 7.51 10.6 6  
## 2538 9.32 6.4 8  
## 3087 7.21 11.2 8  
## 3000 10.81 9.6 5  
## 2368 7.47 6.6 5  
## 2975 8.78 7.0 6  
## 604 10.99 11.6 6  
## 1114 1.97 10.6 4  
## 1534 11.34 13.9 5  
## 972 12.30 5.0 5  
## 2370 9.69 9.0 6  
## 786 6.42 10.7 5  
## 3201 7.03 9.7 4  
## 2621 13.37 11.0 7  
## 2362 8.17 10.5 3  
## 3125 12.19 9.4 2  
## 2442 10.23 12.1 4  
## 1381 9.79 12.0 2  
## total.intl.charge customer.service.calls churn  
## 569 2.70 2 FALSE  
## 3203 1.86 3 FALSE  
## 1809 3.35 3 FALSE  
## 325 3.62 3 FALSE  
## 1359 4.02 0 FALSE  
## 1781 3.24 1 FALSE  
## 428 2.59 1 FALSE  
## 1581 2.86 0 FALSE  
## 2538 1.73 3 FALSE  
## 3087 3.02 3 FALSE  
## 3000 2.59 1 FALSE  
## 2368 1.78 3 FALSE  
## 2975 1.89 1 FALSE  
## 604 3.13 3 FALSE  
## 1114 2.86 1 FALSE  
## 1534 3.75 2 TRUE  
## 972 1.35 3 FALSE  
## 2370 2.43 4 TRUE  
## 786 2.89 3 FALSE  
## 3201 2.62 1 FALSE  
## 2621 2.97 1 FALSE  
## 2362 2.84 3 FALSE  
## 3125 2.54 3 FALSE  
## 2442 3.27 1 FALSE  
## 1381 3.24 1 FALSE

print("Correlation of minutes and charges")

## [1] "Correlation of minutes and charges"

cor.test(datafile$total.night.minutes,datafile$total.night.charge,method="pearson")

##   
## Pearson's product-moment correlation  
##   
## data: datafile$total.night.minutes and datafile$total.night.charge  
## t = 46058, df = 3331, p-value < 2.2e-16  
## alternative hypothesis: true correlation is not equal to 0  
## 95 percent confidence interval:  
## 0.9999992 0.9999993  
## sample estimates:  
## cor   
## 0.9999992

print("Variance of evening and night charges")

## [1] "Variance of evening and night charges"

var(datafile$total.eve.charge)

## [1] 18.58186

var(datafile$total.night.charge)

## [1] 5.179597

print("T test")

## [1] "T test"

t.test(df1$total.eve.charge,df1$total.night.charge,var.equal = FALSE)

##   
## Welch Two Sample t-test  
##   
## data: df1$total.eve.charge and df1$total.night.charge  
## t = 6.5293, df = 33.527, p-value = 1.904e-07  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 5.236846 9.973554  
## sample estimates:  
## mean of x mean of y   
## 16.8320 9.2268

chisq.test(df1$total.day.charge,df1$customer.service.calls)

## Warning in chisq.test(df1$total.day.charge, df1$customer.service.calls):  
## Chi-squared approximation may be incorrect

##   
## Pearson's Chi-squared test  
##   
## data: df1$total.day.charge and df1$customer.service.calls  
## X-squared = 100, df = 96, p-value = 0.3697

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

## [1] 100.4356

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

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

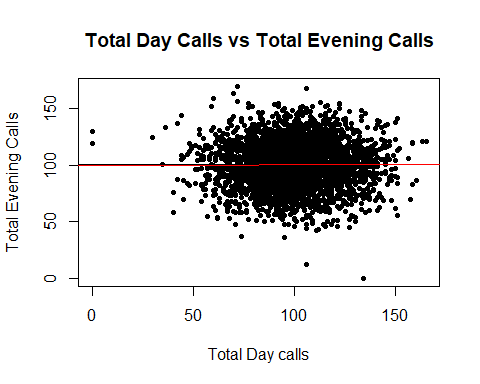
## Warning in plot.xy(xy, type, ...): "data" is not a graphical parameter

## 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

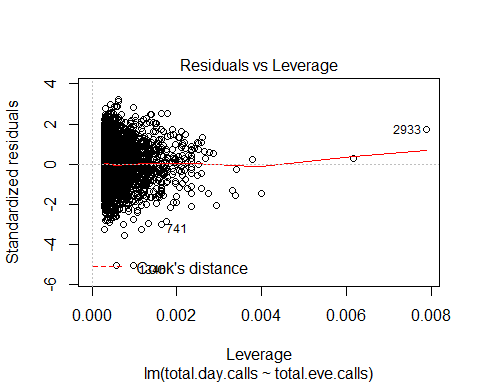
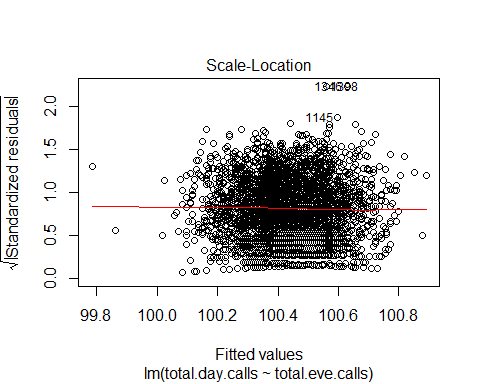
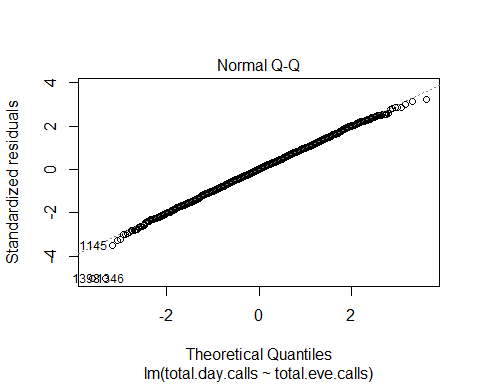
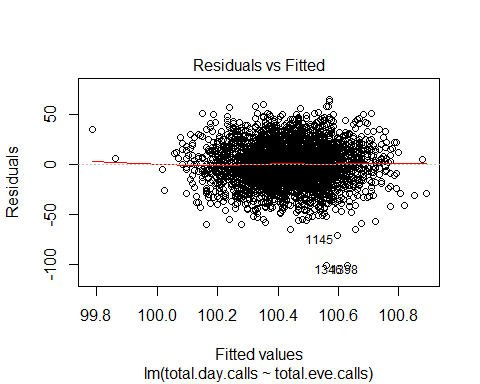
## Warning in box(...): "data" is not a graphical parameter

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

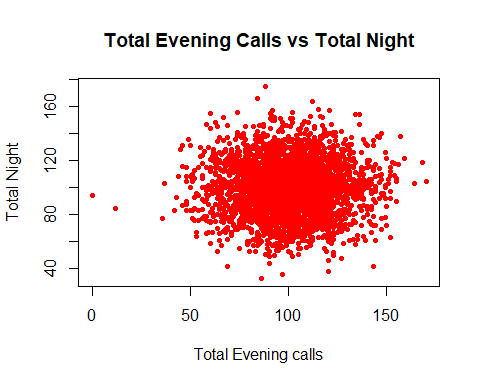
abline(h=mean.totaldaycalls) #adding the mean to the graph  
model1=lm(total.day.calls~total.eve.calls,data=datafile) #getting a linear model  
abline(model1,col="red") #Fitting the linear model to the graph



plot(model1)



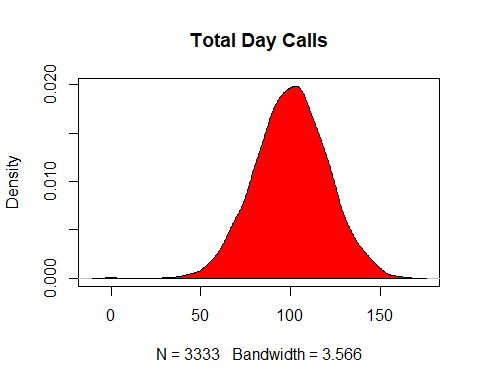
plot(datafile$total.eve.calls,datafile$total.night.calls,main = "Total Evening Calls vs Total Night",xlab = "Total Evening calls",ylab = "Total Night",pch=20,col='red')



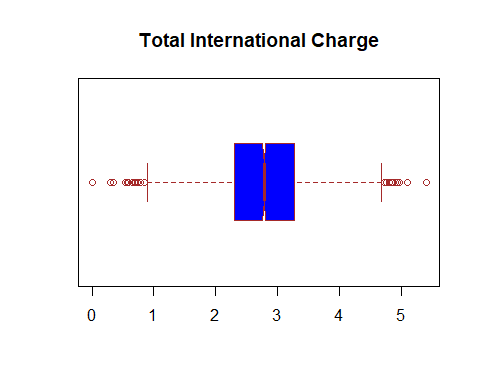
d=density(datafile$total.day.calls) #Kernel density plot  
d

##   
## Call:  
## density.default(x = datafile$total.day.calls)  
##   
## Data: datafile$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(datafile$total.intl.charge,main = "Total International Charge",col="blue",border="brown",horizontal = TRUE,notch = TRUE)



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

**INTERPRETATION**

DESCRIPTIVE STATISTICS

From the descriptive statistics thus obtained for the total minutes for calls during the day, evening and night, we see that the mean of total evening minutes is the maximum. Also, the value of mean of total night minutes is very close to the value of the highest total minutes. Hence, we compare if there is a significant difference in the means of total evening and night minutes.

In order to compare the means, we use the total evening and night charges since the call minutes and charges are directly proportional to the charges. This can be verified by evaluating the correlation between the two.

CORRELATION

We see that the Karl Pearson’s coefficient of correlation has a value of 0.9999 which is positive hence directly proportional and is approximately equal to 1. Thus, we extract a sample from the dataset and apply a test of significance in order to understand if there is a significant difference in the means of total evening and night charges.

T-TEST

Since the sample taken is a small sample, we apply the t-test. We see that the variances of the two columns considered are very different, hence we need to use t-test for unequal variances.

H0: There is no significant difference between the means of total evening and night charges.   
H1: There is a significant difference between the means of total evening and night charges.

From the results so obtained, we see that the p-value=1.078e-09 which is less than the level of significance i.e. 0.05 due to which we reject the null hypothesis. Hence, we accept the alternative hypothesis that there is a significant difference between the means of the charges.

CHI-SQUARE TESTS

H0: The means of the day charges and the customer service calls are independent of each other.

H1: The means of the day charges and the customer service calls are not independent of each other.

From the result obtained, we see that the p-value=0.3697 which is greater than the level of significance hence we accept the null hypothesis that they are independent. Therefore, the assumption that higher costs will lead to more complaints consequently more customer service calls is wrong.

VISUALISATION USING GRAPHS

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.