R Code for Customer Churn Project

This document contains all the relevant R code utilized in the cleaning of the churn dataset, which was subsequently used for the data analysis and machine learning for the customer churn project. Please consult the full Project Report document for related notes, explanations, analysis and results.

churn <- read.csv("churn.csv",  
header = T,  
stringsAsFactors = F,  
na.strings = c("","NA"))

#changing names  
names(churn) <- c('state', 'acclength', 'areacode', 'phone', 'interplan',  
'vmailplan', 'vmailmess', 'daymin', 'daycalls', 'daycharge', 'evemins', 'evecalls',  
'evecharge', 'nightmins', 'nightcalls', 'nightcharge', 'intlmins', 'intlcalls', 'intlcharge', 'custservcalls', 'churn')

#replacing all True and False with yes and no  
churn[churn == "True."] <-"yes"  
churn[churn == "False."] <-"no"

length(which(churn$churn == "yes"))

## [1] 483

length(which(churn$churn == "no"))

## [1] 2850

#structure of the dataset

str(churn)

## 'data.frame': 3333 obs. of 21 variables:  
## $ state : chr "KS" "OH" "NJ" "OH" ...  
## $ acclength : int 128 107 137 84 75 118 121 147 117 141 ...  
## $ areacode : int 415 415 415 408 415 510 510 415 408 415 ...  
## $ phone : chr "382-4657" "371-7191" "358-1921" "375-9999" ...  
## $ interplan : chr "no" "no" "no" "yes" ...  
## $ vmailplan : chr "yes" "yes" "no" "no" ...  
## $ vmailmess : int 25 26 0 0 0 0 24 0 0 37 ...  
## $ daymin : num 265 162 243 299 167 ...  
## $ daycalls : int 110 123 114 71 113 98 88 79 97 84 ...  
## $ daycharge : num 45.1 27.5 41.4 50.9 28.3 ...  
## $ evemins : num 197.4 195.5 121.2 61.9 148.3 ...  
## $ evecalls : int 99 103 110 88 122 101 108 94 80 111 ...  
## $ evecharge : num 16.78 16.62 10.3 5.26 12.61 ...  
## $ nightmins : num 245 254 163 197 187 ...  
## $ nightcalls : int 91 103 104 89 121 118 118 96 90 97 ...  
## $ nightcharge : num 11.01 11.45 7.32 8.86 8.41 ...  
## $ intlmins : num 10 13.7 12.2 6.6 10.1 6.3 7.5 7.1 8.7 11.2 ...  
## $ intlcalls : int 3 3 5 7 3 6 7 6 4 5 ...  
## $ intlcharge : num 2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...  
## $ custservcalls: int 1 1 0 2 3 0 3 0 1 0 ...  
## $ churn : chr "no" "no" "no" "no" ...

# Here we can see that there is a significant positive correlation between  
#nightcharge and nightmins , daycharge and daymin , evecharge and evemins, intlmins and intlcharge. This means as one variable increases, the other increases as well. We can remove nightmins, daymin, evemins,and intlmins accordingly.   
corr2 <- cor(churn[,7:20])  
corr2

## vmailmess daymin daycalls daycharge  
## vmailmess 1.0000000000 0.0007782741 -0.009548068 0.0007755235  
## daymin 0.0007782741 1.0000000000 0.006750414 0.9999999522  
## daycalls -0.0095480677 0.0067504139 1.000000000 0.0067529620  
## daycharge 0.0007755235 0.9999999522 0.006752962 1.0000000000  
## evemins 0.0175620343 0.0070425110 -0.021451408 0.0070496072  
## evecalls -0.0058643513 0.0157689932 0.006462114 0.0157692822  
## evecharge 0.0175777801 0.0070290353 -0.021449263 0.0070361315  
## nightmins 0.0076811359 0.0043233666 0.022937845 0.0043238794  
## nightcalls 0.0071230629 0.0229724555 -0.019556965 0.0229724195  
## nightcharge 0.0076632904 0.0043003570 0.022926638 0.0043008608  
## intlmins 0.0028561959 -0.0101545856 0.021564794 -0.0101568616  
## intlcalls 0.0139573387 0.0080333570 0.004574268 0.0080315720  
## intlcharge 0.0028836579 -0.0100919742 0.021666095 -0.0100942572  
## custservcalls -0.0132625831 -0.0134231864 -0.018941930 -0.0134269694  
## evemins evecalls evecharge nightmins nightcalls  
## vmailmess 0.017562034 -0.005864351 0.017577780 0.007681136 0.0071230629  
## daymin 0.007042511 0.015768993 0.007029035 0.004323367 0.0229724555  
## daycalls -0.021451408 0.006462114 -0.021449263 0.022937845 -0.0195569654  
## daycharge 0.007049607 0.015769282 0.007036131 0.004323879 0.0229724195  
## evemins 1.000000000 -0.011430108 0.999999776 -0.012583678 0.0075856431  
## evecalls -0.011430108 1.000000000 -0.011422894 -0.002092768 0.0077097055  
## evecharge 0.999999776 -0.011422894 1.000000000 -0.012592020 0.0075958430  
## nightmins -0.012583678 -0.002092768 -0.012592020 1.000000000 0.0112038563  
## nightcalls 0.007585643 0.007709706 0.007595843 0.011203856 1.0000000000  
## nightcharge -0.012592806 -0.002055984 -0.012601142 0.999999215 0.0111878197  
## intlmins -0.011034714 0.008702881 -0.011042582 -0.015207297 -0.0136049964  
## intlcalls 0.002541292 0.017433692 0.002541458 -0.012353432 0.0003045795  
## intlcharge -0.011066621 0.008673858 -0.011074499 -0.015179849 -0.0136301696  
## custservcalls -0.012984553 0.002422575 -0.012987407 -0.009287613 -0.0128019273  
## nightcharge intlmins intlcalls intlcharge  
## vmailmess 0.007663290 0.002856196 0.0139573387 0.002883658  
## daymin 0.004300357 -0.010154586 0.0080333570 -0.010091974  
## daycalls 0.022926638 0.021564794 0.0045742682 0.021666095  
## daycharge 0.004300861 -0.010156862 0.0080315720 -0.010094257  
## evemins -0.012592806 -0.011034714 0.0025412917 -0.011066621  
## evecalls -0.002055984 0.008702881 0.0174336921 0.008673858  
## evecharge -0.012601142 -0.011042582 0.0025414580 -0.011074499  
## nightmins 0.999999215 -0.015207297 -0.0123534324 -0.015179849  
## nightcalls 0.011187820 -0.013604996 0.0003045795 -0.013630170  
## nightcharge 1.000000000 -0.015213526 -0.0123292150 -0.015186139  
## intlmins -0.015213526 1.000000000 0.0323038841 0.999992742  
## intlcalls -0.012329215 0.032303884 1.0000000000 0.032372145  
## intlcharge -0.015186139 0.999992742 0.0323721453 1.000000000  
## custservcalls -0.009276954 -0.009639680 -0.0175605992 -0.009674732  
## custservcalls  
## vmailmess -0.013262583  
## daymin -0.013423186  
## daycalls -0.018941930  
## daycharge -0.013426969  
## evemins -0.012984553  
## evecalls 0.002422575  
## evecharge -0.012987407  
## nightmins -0.009287613  
## nightcalls -0.012801927  
## nightcharge -0.009276954  
## intlmins -0.009639680  
## intlcalls -0.017560599  
## intlcharge -0.009674732  
## custservcalls 1.000000000

corr3 <- round(corr2, 1)  
ggcorrplot(corr3, method = "circle")

#Drop columns with significant correlation to daycharge  
churn$phone <- NULL  
churn$daymin <- NULL  
churn$evemins <- NULL  
churn$intlmins <- NULL  
churn$nightmins <- NULL

#Find max, min, mean and standard deviation of attributes.  
#From this statistical summary of our data set we can observe some very relatively high Maximum values and relatively low minimum values. This is an indication that the data has outliers.   
#  
summary(churn)

## state acclength areacode interplan   
## Length:3333 Min. : 1.0 Min. :408.0 Length:3333   
## Class :character 1st Qu.: 74.0 1st Qu.:408.0 Class :character   
## Mode :character Median :101.0 Median :415.0 Mode :character   
## Mean :101.1 Mean :437.2   
## 3rd Qu.:127.0 3rd Qu.:510.0   
## Max. :243.0 Max. :510.0   
## vmailplan vmailmess daycalls daycharge   
## Length:3333 Min. : 0.000 Min. : 0.0 Min. : 0.00   
## Class :character 1st Qu.: 0.000 1st Qu.: 87.0 1st Qu.:24.43   
## Mode :character Median : 0.000 Median :101.0 Median :30.50   
## Mean : 8.099 Mean :100.4 Mean :30.56   
## 3rd Qu.:20.000 3rd Qu.:114.0 3rd Qu.:36.79   
## Max. :51.000 Max. :165.0 Max. :59.64   
## evecalls evecharge nightcalls nightcharge   
## Min. : 0.0 Min. : 0.00 Min. : 33.0 Min. : 1.040   
## 1st Qu.: 87.0 1st Qu.:14.16 1st Qu.: 87.0 1st Qu.: 7.520   
## Median :100.0 Median :17.12 Median :100.0 Median : 9.050   
## Mean :100.1 Mean :17.08 Mean :100.1 Mean : 9.039   
## 3rd Qu.:114.0 3rd Qu.:20.00 3rd Qu.:113.0 3rd Qu.:10.590   
## Max. :170.0 Max. :30.91 Max. :175.0 Max. :17.770   
## intlcalls intlcharge custservcalls churn   
## Min. : 0.000 Min. :0.000 Min. :0.000 Length:3333   
## 1st Qu.: 3.000 1st Qu.:2.300 1st Qu.:1.000 Class :character   
## Median : 4.000 Median :2.780 Median :1.000 Mode :character   
## Mean : 4.479 Mean :2.765 Mean :1.563   
## 3rd Qu.: 6.000 3rd Qu.:3.270 3rd Qu.:2.000   
## Max. :20.000 Max. :5.400 Max. :9.000

sd(churn$acclength)

## [1] 39.82211

sd(churn$vmailmess)

## [1] 13.68837

sd(churn$daycalls)

## [1] 20.06908

sd(churn$daycharge)

## [1] 9.259435

sd(churn$evecalls)

## [1] 19.92263

sd(churn$evecharge)

## [1] 4.310668

sd(churn$nightcalls)

## [1] 19.56861

sd(churn$nightcharge)

## [1] 2.275873

sd(churn$intlcalls)

## [1] 2.461214

sd(churn$intlcharge)

## [1] 0.7537726

sd(churn$custservcalls)

## [1] 1.315491

#Here we can examine the structure of data again. The number of rows is still (3333) but the columns have reduced to 16.

str(churn)

## 'data.frame': 3333 obs. of 16 variables:  
## $ state : chr "KS" "OH" "NJ" "OH" ...  
## $ acclength : int 128 107 137 84 75 118 121 147 117 141 ...  
## $ areacode : int 415 415 415 408 415 510 510 415 408 415 ...  
## $ interplan : chr "no" "no" "no" "yes" ...  
## $ vmailplan : chr "yes" "yes" "no" "no" ...  
## $ vmailmess : int 25 26 0 0 0 0 24 0 0 37 ...  
## $ daycalls : int 110 123 114 71 113 98 88 79 97 84 ...  
## $ daycharge : num 45.1 27.5 41.4 50.9 28.3 ...  
## $ evecalls : int 99 103 110 88 122 101 108 94 80 111 ...  
## $ evecharge : num 16.78 16.62 10.3 5.26 12.61 ...  
## $ nightcalls : int 91 103 104 89 121 118 118 96 90 97 ...  
## $ nightcharge : num 11.01 11.45 7.32 8.86 8.41 ...  
## $ intlcalls : int 3 3 5 7 3 6 7 6 4 5 ...  
## $ intlcharge : num 2.7 3.7 3.29 1.78 2.73 1.7 2.03 1.92 2.35 3.02 ...  
## $ custservcalls: int 1 1 0 2 3 0 3 0 1 0 ...  
## $ churn : chr "no" "no" "no" "no" ...

head(churn)

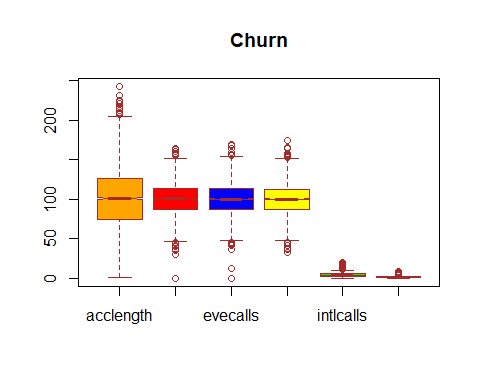
## state acclength areacode interplan vmailplan vmailmess daycalls daycharge  
## 1 KS 128 415 no yes 25 110 45.07  
## 2 OH 107 415 no yes 26 123 27.47  
## 3 NJ 137 415 no no 0 114 41.38  
## 4 OH 84 408 yes no 0 71 50.90  
## 5 OK 75 415 yes no 0 113 28.34  
## 6 AL 118 510 yes no 0 98 37.98  
## evecalls evecharge nightcalls nightcharge intlcalls intlcharge custservcalls  
## 1 99 16.78 91 11.01 3 2.70 1  
## 2 103 16.62 103 11.45 3 3.70 1  
## 3 110 10.30 104 7.32 5 3.29 0  
## 4 88 5.26 89 8.86 7 1.78 2  
## 5 122 12.61 121 8.41 3 2.73 3  
## 6 101 18.75 118 9.18 6 1.70 0  
## churn  
## 1 no  
## 2 no  
## 3 no  
## 4 no  
## 5 no  
## 6 no

#Below, I run a check for null values in the dataset but it returns zero, meaning the dataset contains no null values.  
sum(is.na(churn))

## [1] 0

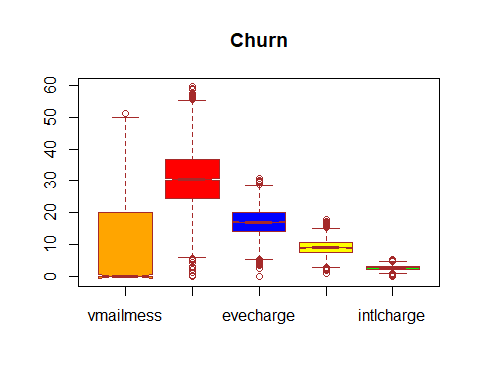
#Next, we run a check for outliers using boxplots. We can see from the results that the data contains outliers. All outliers will be removed from the data accordingly.  
boxplot(churn$acclength, churn$daycalls, churn$evecalls,  
 churn$nightcalls, churn$intlcalls, churn$custservcalls, main = "Churn",  
col = c("orange","red", "blue", "yellow", "green", "pink"),  
border = "brown",  
names = c("acclength", "daycalls", "evecalls",   
 "nightcharge", "intlcalls", "custservcalls" ),  
horizontal = F,  
notch = TRUE)

## Warning in bxp(list(stats = structure(c(1, 74, 101, 127, 205, 47, 87, 101, :  
## some notches went outside hinges ('box'): maybe set notch=FALSE



boxplot(churn$vmailmess, churn$daycharge, churn$evecharge,  
 churn$nightcharge, churn$intlcharge, main = "Churn",  
col = c("orange","red", "blue", "yellow", "green", "pink"),  
border = "brown",  
names = c("vmailmess", "daycharge","evecharge",  
 "nightcharge", "intlcharge"),  
horizontal = F,  
notch = TRUE)

## Warning in bxp(list(stats = structure(c(0, 0, 0, 20, 50, 5.97, 24.43, 30.5, :  
## some notches went outside hinges ('box'): maybe set notch=FALSE

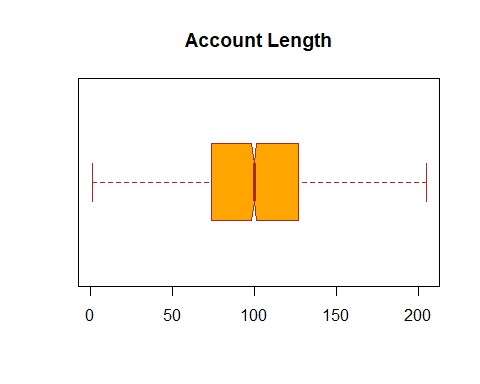


#Removing outliers  
#First, we identify the outliers:  
boxplot(churn$acclength, plot=FALSE)$out

## [1] 208 215 209 224 243 217 210 212 232 225 225 224 212 210 217 209 221 209

#Then save the outliers in a vector:  
outliers <- boxplot(churn$acclength, plot=FALSE)$out  
#This vector will be used to exclude outliers from the dataset. The which() function tells us the rows in which the outliers exist, after which the identified rows will be removed.

churn2<-churn  
#out\_ind <- which(churn$acclength %in% c(outliers))  
#out\_ind  
churn2<- churn2[-which(churn2$acclength %in% outliers),]  
boxplot(churn2$acclength, main = "Account Length",  
col = "orange",  
border = "brown",  
horizontal = TRUE,  
notch = TRUE)

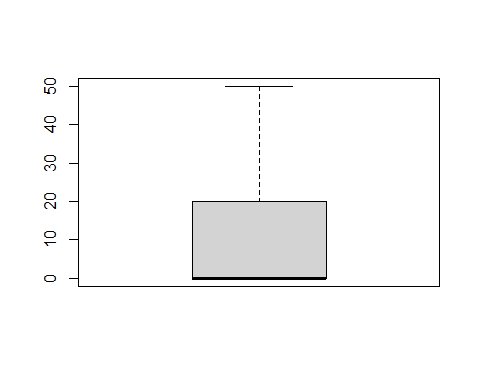


#We proceed to remove all outliers from the other columns.  
boxplot(churn$vmailmess, plot=FALSE)$out

## [1] 51

#Then save the outliers in a vector:  
outliers1 <- boxplot(churn$vmailmess, plot=FALSE)$out

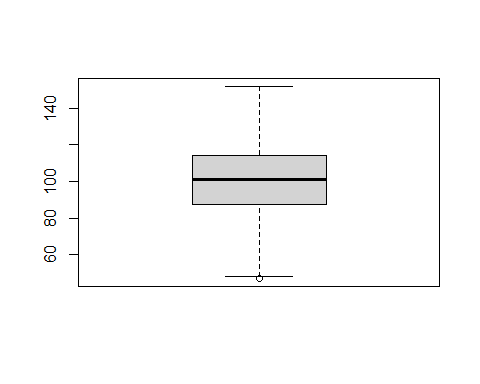
churn3<-churn2  
churn3<- churn3[-which(churn3$vmailmess %in% outliers1),]  
boxplot(churn3$vmailmess)



boxplot(churn3$daycalls, plot=FALSE)$out

## [1] 158 163 36 40 158 165 30 42 0 45 0 45 160 156 35 42 158 157 45  
## [20] 44 44 44 40

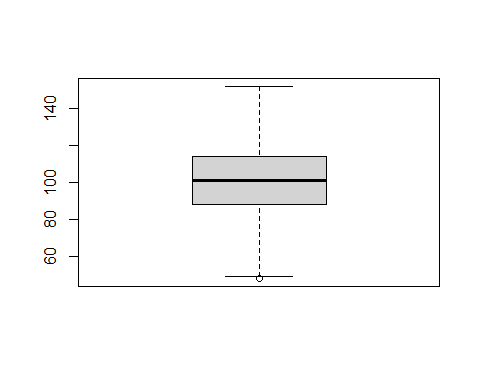
outliers2 <- boxplot(churn3$daycalls, plot=FALSE)$out  
  
churn4<-churn3  
churn4<- churn4[-which(churn4$daycalls %in% outliers2),]  
boxplot(churn4$daycalls)



boxplot(churn4$daycalls, plot=FALSE)$out

## [1] 47 47

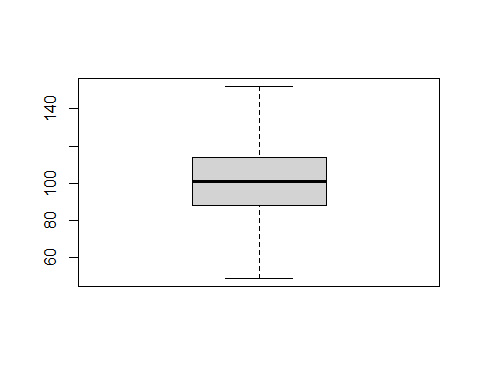
outliers3 <- boxplot(churn4$daycalls, plot=FALSE)$out  
churn5<-churn4  
churn5<- churn5[-which(churn5$daycalls %in% outliers3),]  
boxplot(churn5$daycalls)



boxplot(churn5$daycalls, plot=FALSE)$out

## [1] 48 48 48

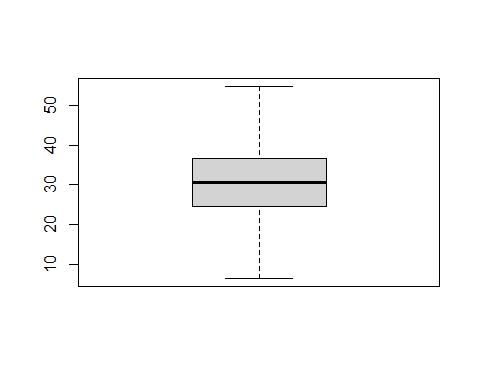
outliers4 <- boxplot(churn5$daycalls, plot=FALSE)$out  
churn6<-churn5  
churn6<- churn6[-which(churn6$daycalls %in% outliers4),]  
boxplot(churn6$daycalls)



boxplot(churn6$daycharge, plot=FALSE)$out

## [1] 56.59 57.36 59.64 57.04 5.25 5.78 56.83 58.96 2.13 4.40 5.97 3.32  
## [13] 56.07 1.34 55.78 4.59 2.99 55.47 58.70 0.44 1.33 3.21 5.08

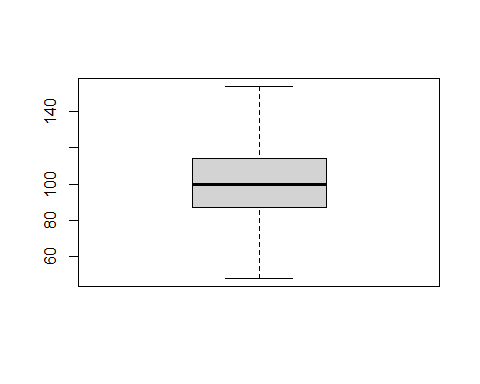
outliers5 <- boxplot(churn6$daycharge, plot=FALSE)$out  
churn7<-churn6  
churn7<- churn7[-which(churn7$daycharge %in% outliers5),]  
boxplot(churn7$daycharge)



boxplot(churn7$evecalls, plot=FALSE)$out

## [1] 164 46 168 42 37 12 157 155 45 36 156 46 44 155 46 43 0 155 159  
## [20] 170

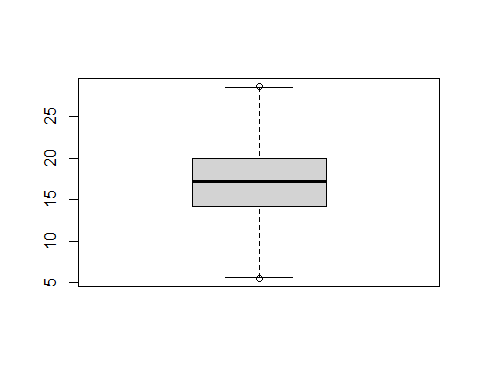
outliers6 <- boxplot(churn7$evecalls, plot=FALSE)$out  
churn8<-churn7  
churn8<- churn8[-which(churn8$evecalls %in% outliers6),]  
boxplot(churn8$evecalls)



boxplot(churn8$evecharge, plot=FALSE)$out

## [1] 5.26 29.62 29.89 2.65 5.47 29.79 3.59 29.52 5.01 3.73 4.50 3.61  
## [13] 5.17 4.98 4.76 4.09 5.10 29.83 4.18 28.89 30.75 30.11 30.91 29.01

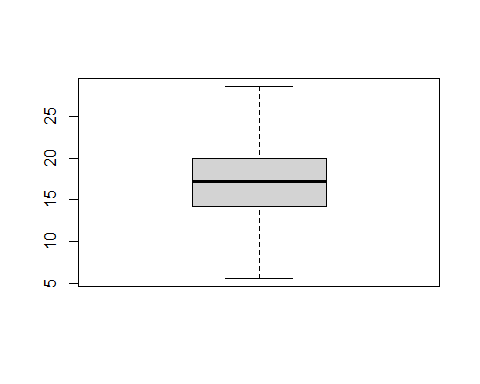
outliers7 <- boxplot(churn8$evecharge, plot=FALSE)$out  
churn9<-churn8  
churn9<- churn9[-which(churn9$evecharge %in% outliers7),]  
boxplot(churn9$evecharge)



boxplot(churn9$evecharge, plot=FALSE)$out

## [1] 28.65 5.54

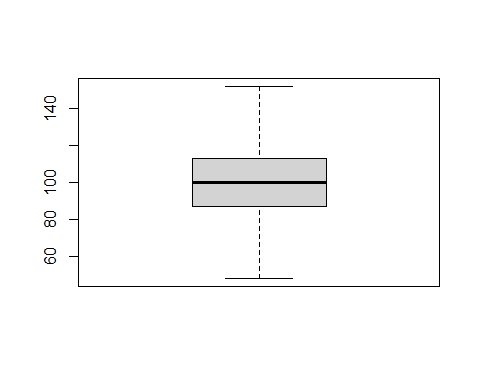
outliers8 <- boxplot(churn9$evecharge, plot=FALSE)$out  
churn10<-churn9  
churn10<- churn10[-which(churn10$evecharge %in% outliers8),]  
boxplot(churn10$evecharge)



boxplot(churn10$nightcalls, plot=FALSE)$out

## [1] 42 44 42 153 175 154 158 155 157 157 154 153 166 33 155 38 36 156 164  
## [20] 153

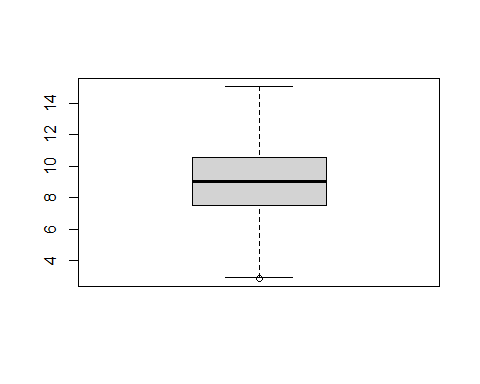
outliers9 <- boxplot(churn10$nightcalls, plot=FALSE)$out  
churn11<-churn10  
churn11<- churn11[-which(churn11$nightcalls %in% outliers9),]  
boxplot(churn11$nightcalls)



boxplot(churn11$nightcharge, plot=FALSE)$out

## [1] 2.59 15.97 15.71 15.56 2.03 15.43 16.39 2.85 2.45 2.25 1.97 15.74  
## [13] 15.86 1.04 2.86 17.19 16.99 16.55 2.55 2.43 15.49 17.77 15.76 2.25  
## [25] 2.40 15.85 16.42 2.76 2.13

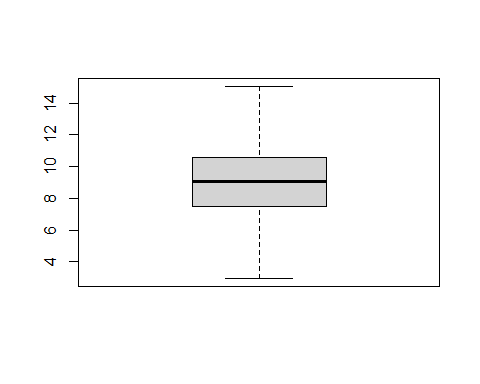
outliers10 <- boxplot(churn11$nightcharge, plot=FALSE)$out  
churn12<-churn11  
churn12<- churn12[-which(churn12$nightcharge %in% outliers10),]  
boxplot(churn12$nightcharge)



boxplot(churn12$nightcharge, plot=FALSE)$out

## [1] 2.89

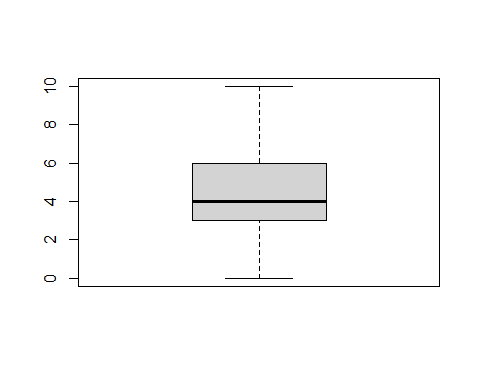
outliers11 <- boxplot(churn12$nightcharge, plot=FALSE)$out  
churn13<-churn12  
churn13<- churn13[-which(churn13$nightcharge %in% outliers11),]  
boxplot(churn13$nightcharge)



boxplot(churn13$intlcalls, plot=FALSE)$out

## [1] 19 15 11 12 13 11 12 11 13 12 11 11 18 11 12 13 12 12 11 15 13 15 11 11 14  
## [26] 13 11 13 13 11 14 15 18 12 13 11 14 11 12 14 15 12 11 16 11 11 11 11 15 14  
## [51] 11 11 12 13 11 11 16 13 11 13 11 15 11 12 13 18 12 12 12 11 13 11 13 14 20  
## [76] 17

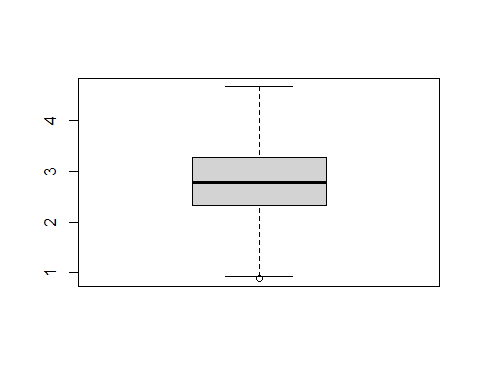
outliers12 <- boxplot(churn13$intlcalls, plot=FALSE)$out  
churn14<-churn13  
churn14<- churn14[-which(churn14$intlcalls %in% outliers12),]  
boxplot(churn14$intlcalls)



boxplot(churn14$intlcharge, plot=FALSE)$out

## [1] 5.40 0.00 4.75 0.73 5.10 0.00 4.86 0.54 0.00 4.73 4.73 4.91 0.00 0.00 0.35  
## [16] 0.00 0.00 0.00 0.59 4.86 0.00 4.83 0.00 4.97 0.54 4.81 0.78 0.84 4.75 0.70  
## [31] 0.00 0.00 4.91 0.00 4.86 0.30 0.00 4.94 0.00 0.00 0.57 0.78 4.73 0.57 0.00  
## [46] 0.00 4.81

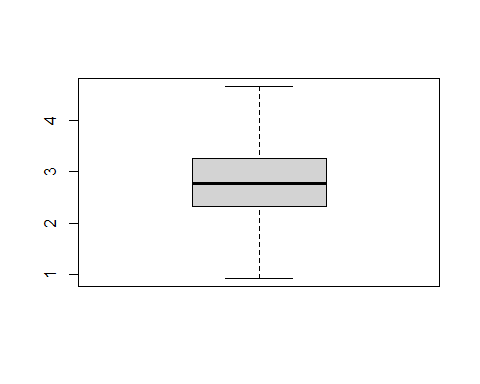
outliers13 <- boxplot(churn14$intlcharge, plot=FALSE)$out  
churn15<-churn14  
churn15<- churn15[-which(churn15$intlcharge %in% outliers13),]  
boxplot(churn15$intlcharge)



boxplot(churn15$intlcharge, plot=FALSE)$out

## [1] 0.89 0.89

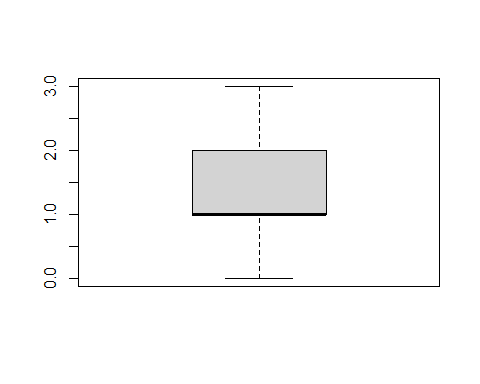
outliers14 <- boxplot(churn15$intlcharge, plot=FALSE)$out  
churn16<-churn15  
churn16<- churn16[-which(churn16$intlcharge %in% outliers14),]  
boxplot(churn16$intlcharge)



boxplot(churn16$custservcalls, plot=FALSE)$out

## [1] 4 4 5 5 5 4 4 4 4 4 4 4 4 4 4 5 4 5 4 4 5 4 4 4 4 5 4 4 7 4 4 4 4 4 5 4 4  
## [38] 4 4 4 5 4 7 4 9 5 4 4 4 4 5 5 6 4 6 5 5 5 6 5 4 4 5 4 4 7 4 6 4 4 4 6 4 4  
## [75] 5 4 4 4 4 4 5 6 5 4 4 4 5 4 4 4 4 5 5 4 4 4 4 4 5 4 6 4 4 4 4 4 4 4 4 6 4  
## [112] 4 4 4 8 4 4 5 4 4 4 6 5 5 7 4 4 5 4 4 5 4 4 5 7 4 4 5 7 4 4 4 4 8 6 4 4 5  
## [149] 5 5 4 5 4 4 4 4 4 4 4 4 4 4 5 6 4 5 4 4 5 5 4 6 4 4 4 9 6 4 5 5 4 6 4 4 5  
## [186] 4 4 4 5 5 6 4 4 4 4 5 4 4 4 4 5 6 4 4 5 4 4 4 4 4 4 4 5 7 6 5 6 7 5 5 4 6  
## [223] 4 4 4 4 5 6 7 4 4 4 5 5 5 4 4 5 6 5 5 4 4 4 4 4 4 5

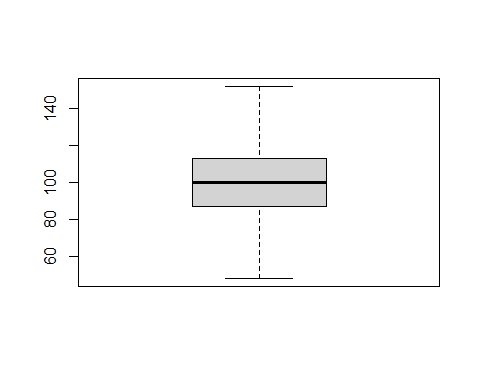
outliers15 <- boxplot(churn16$custservcalls, plot=FALSE)$out  
churn17<-churn16  
churn17<- churn17[-which(churn17$custservcalls %in% outliers15),]  
boxplot(churn17$custservcalls)



boxplot(churn17$evecalls, plot=FALSE)$out

## [1] 153 154 154

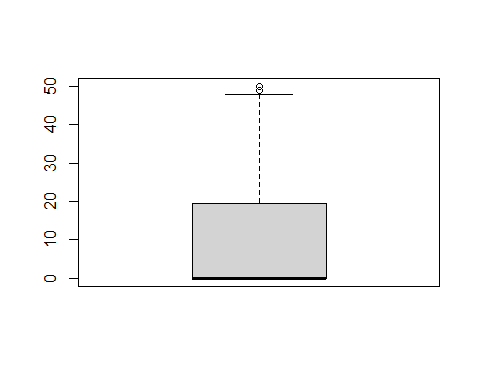
outliers16 <- boxplot(churn17$evecalls, plot=FALSE)$out  
churn18<-churn17  
churn18<- churn18[-which(churn18$evecalls %in% outliers16),]  
boxplot(churn18$evecalls)



boxplot(churn18$vmailmess, plot=FALSE)$out

## [1] 49 50 50

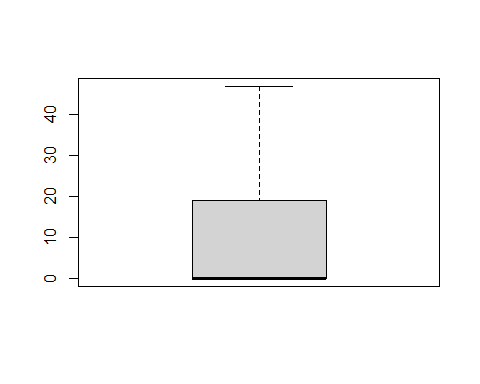
outliers17 <- boxplot(churn18$vmailmess, plot=FALSE)$out  
churn19<-churn18  
churn19<- churn19[-which(churn19$vmailmess %in% outliers17),]  
boxplot(churn18$vmailmess)



boxplot(churn19$vmailmess, plot=FALSE)$out

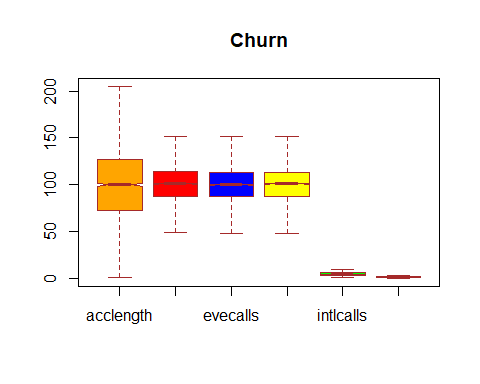
## [1] 48 48

outliers18 <- boxplot(churn19$vmailmess, plot=FALSE)$out  
churn20<-churn19  
churn20<- churn20[-which(churn20$vmailmess %in% outliers18),]  
boxplot(churn20$vmailmess)



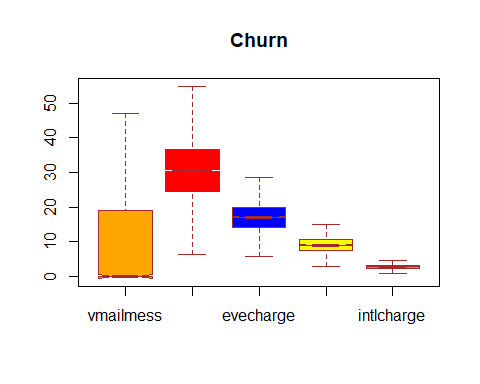
#here we can visualize the cleaned data without outliers   
boxplot(churn20$acclength, churn20$daycalls, churn20$evecalls,  
 churn20$nightcalls, churn20$intlcalls, churn20$custservcalls, main = "Churn",  
col = c("orange","red", "blue", "yellow", "green", "pink"),  
border = "brown",  
names = c("acclength", "daycalls", "evecalls",   
 "nightcharge", "intlcalls", "custservcalls" ),  
horizontal = F,  
notch = TRUE)

## Warning in bxp(list(stats = structure(c(1, 73, 100, 127, 205, 49, 87, 101, :  
## some notches went outside hinges ('box'): maybe set notch=FALSE



boxplot(churn20$vmailmess, churn20$daycharge, churn20$evecharge,  
 churn20$nightcharge, churn20$intlcharge, main = "Churn",  
col = c("orange","red", "blue", "yellow", "green", "pink"),  
border = "brown",  
names = c("vmailmess", "daycharge","evecharge",  
 "nightcharge", "intlcharge"),  
horizontal = F,  
notch = TRUE)

## Warning in bxp(list(stats = structure(c(0, 0, 0, 19, 47, 6.41, 24.48, 30.56, :  
## some notches went outside hinges ('box'): maybe set notch=FALSE



length(which(churn20$churn == "yes"))

## [1] 302

length(which(churn20$churn == "no"))

## [1] 2484

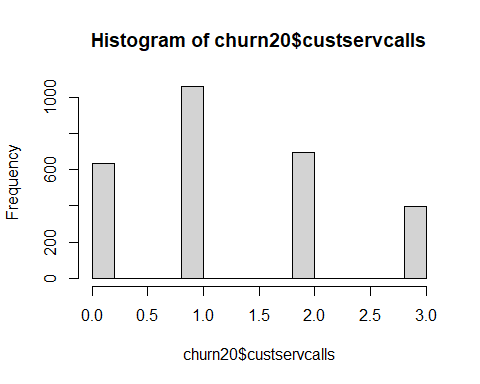
summary(churn20)

## state acclength areacode interplan   
## Length:2786 Min. : 1.0 Min. :408.0 Length:2786   
## Class :character 1st Qu.: 73.0 1st Qu.:408.0 Class :character   
## Mode :character Median :100.0 Median :415.0 Mode :character   
## Mean :100.4 Mean :437.4   
## 3rd Qu.:127.0 3rd Qu.:510.0   
## Max. :205.0 Max. :510.0   
## vmailplan vmailmess daycalls daycharge   
## Length:2786 Min. : 0.000 Min. : 49.0 Min. : 6.41   
## Class :character 1st Qu.: 0.000 1st Qu.: 87.0 1st Qu.:24.48   
## Mode :character Median : 0.000 Median :101.0 Median :30.56   
## Mean : 8.061 Mean :100.6 Mean :30.61   
## 3rd Qu.:19.000 3rd Qu.:114.0 3rd Qu.:36.73   
## Max. :47.000 Max. :152.0 Max. :54.83   
## evecalls evecharge nightcalls nightcharge   
## Min. : 48.00 Min. : 5.61 Min. : 48 Min. : 2.960   
## 1st Qu.: 87.00 1st Qu.:14.21 1st Qu.: 87 1st Qu.: 7.540   
## Median :100.00 Median :17.14 Median :101 Median : 9.050   
## Mean : 99.97 Mean :17.12 Mean :100 Mean : 9.053   
## 3rd Qu.:113.00 3rd Qu.:20.00 3rd Qu.:113 3rd Qu.:10.610   
## Max. :152.00 Max. :28.56 Max. :152 Max. :15.060   
## intlcalls intlcharge custservcalls churn   
## Min. : 1.000 Min. :0.920 Min. :0.000 Length:2786   
## 1st Qu.: 3.000 1st Qu.:2.320 1st Qu.:1.000 Class :character   
## Median : 4.000 Median :2.780 Median :1.000 Mode :character   
## Mean : 4.315 Mean :2.785 Mean :1.307   
## 3rd Qu.: 6.000 3rd Qu.:3.270 3rd Qu.:2.000   
## Max. :10.000 Max. :4.670 Max. :3.000

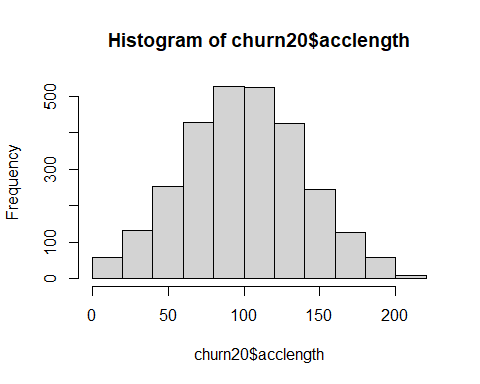
str(churn20)

## 'data.frame': 2786 obs. of 16 variables:  
## $ state : chr "KS" "OH" "NJ" "OK" ...  
## $ acclength : int 128 107 137 75 118 147 141 74 168 95 ...  
## $ areacode : int 415 415 415 415 510 415 415 415 408 510 ...  
## $ interplan : chr "no" "no" "no" "yes" ...  
## $ vmailplan : chr "yes" "yes" "no" "no" ...  
## $ vmailmess : int 25 26 0 0 0 0 37 0 0 0 ...  
## $ daycalls : int 110 123 114 113 98 79 84 127 96 88 ...  
## $ daycharge : num 45.1 27.5 41.4 28.3 38 ...  
## $ evecalls : int 99 103 110 122 101 94 111 148 71 75 ...  
## $ evecharge : num 16.8 16.6 10.3 12.6 18.8 ...  
## $ nightcalls : int 91 103 104 121 118 96 97 94 128 115 ...  
## $ nightcharge : num 11.01 11.45 7.32 8.41 9.18 ...  
## $ intlcalls : int 3 3 5 3 6 6 5 5 2 5 ...  
## $ intlcharge : num 2.7 3.7 3.29 2.73 1.7 1.92 3.02 2.46 3.02 3.32 ...  
## $ custservcalls: int 1 1 0 3 0 0 0 0 1 3 ...  
## $ churn : chr "no" "no" "no" "no" ...

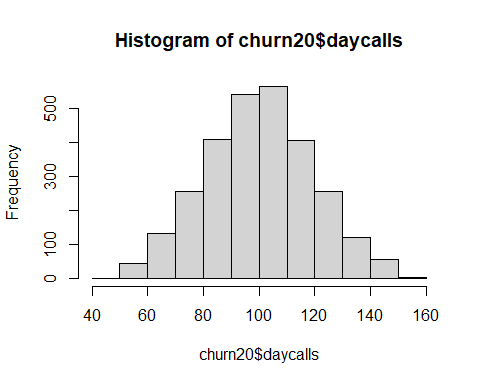
hist(churn20$custservcalls)



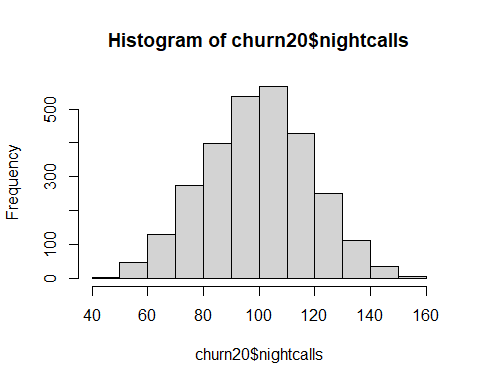
hist(churn20$acclength)



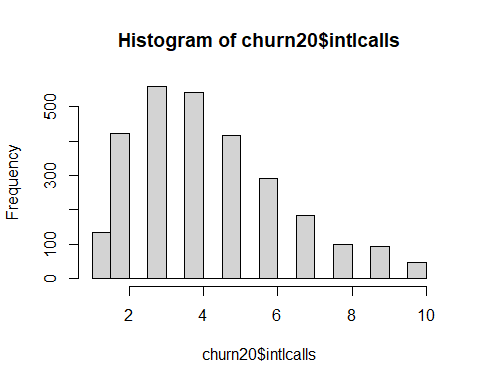
hist(churn20$daycalls)



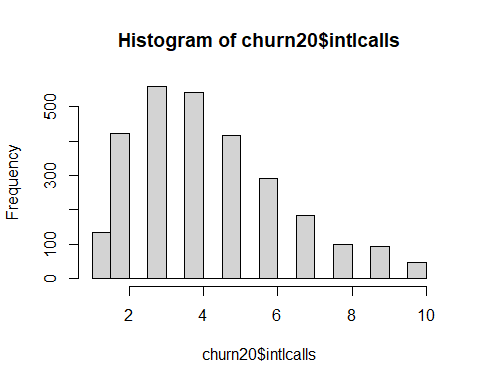
hist(churn20$nightcalls)



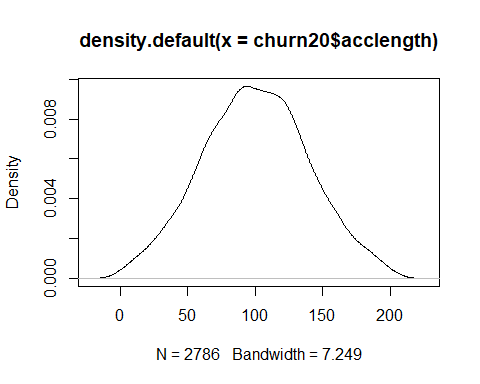
hist(churn20$intlcalls)



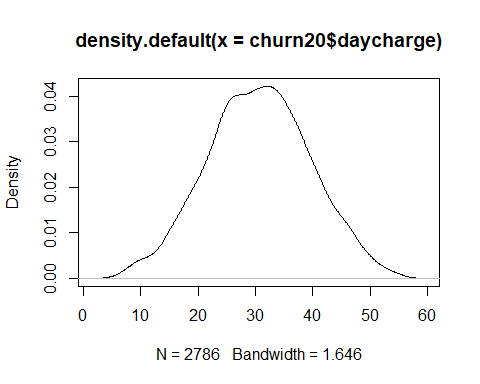
hist(churn20$intlcalls)



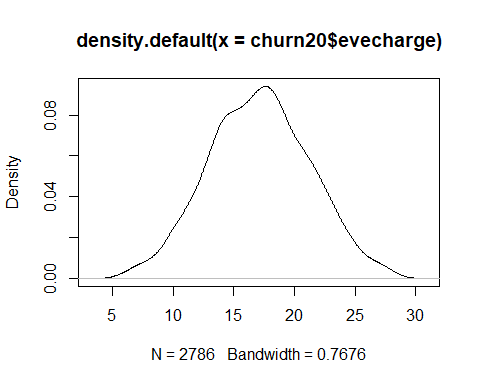
plot(density(churn20$acclength))



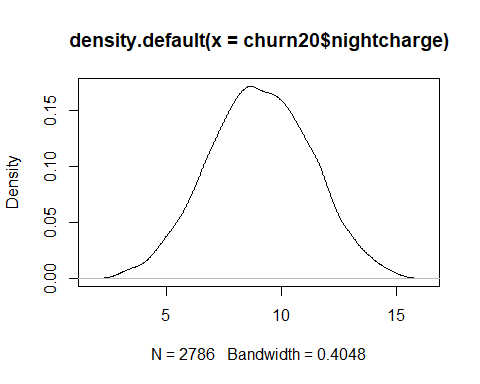
plot(density(churn20$daycharge))



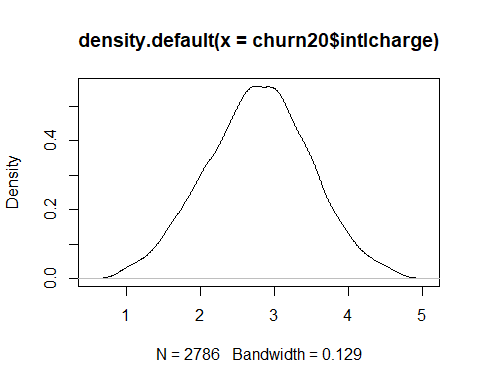
plot(density(churn20$evecharge))



plot(density(churn20$nightcharge))



plot(density(churn20$intlcharge))



write.table(churn20, file ="D:/SIGLI/Project files/churnclean2.csv", row.names = F, sep =",")