Homework 1

Rachael Hageman Blair

2023-02-22

Loading Packages:

```
library(ISLR2)
library(arules)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'arules'
```

```
## The following objects are masked from 'package:base':
##
##     abbreviate, write
```

```
library(moments)
library(ggplot2)
library(reshape2)
library(rpart)
library(caret)
```

```
## Loading required package: lattice
```

```
library (rpart.plot)
```

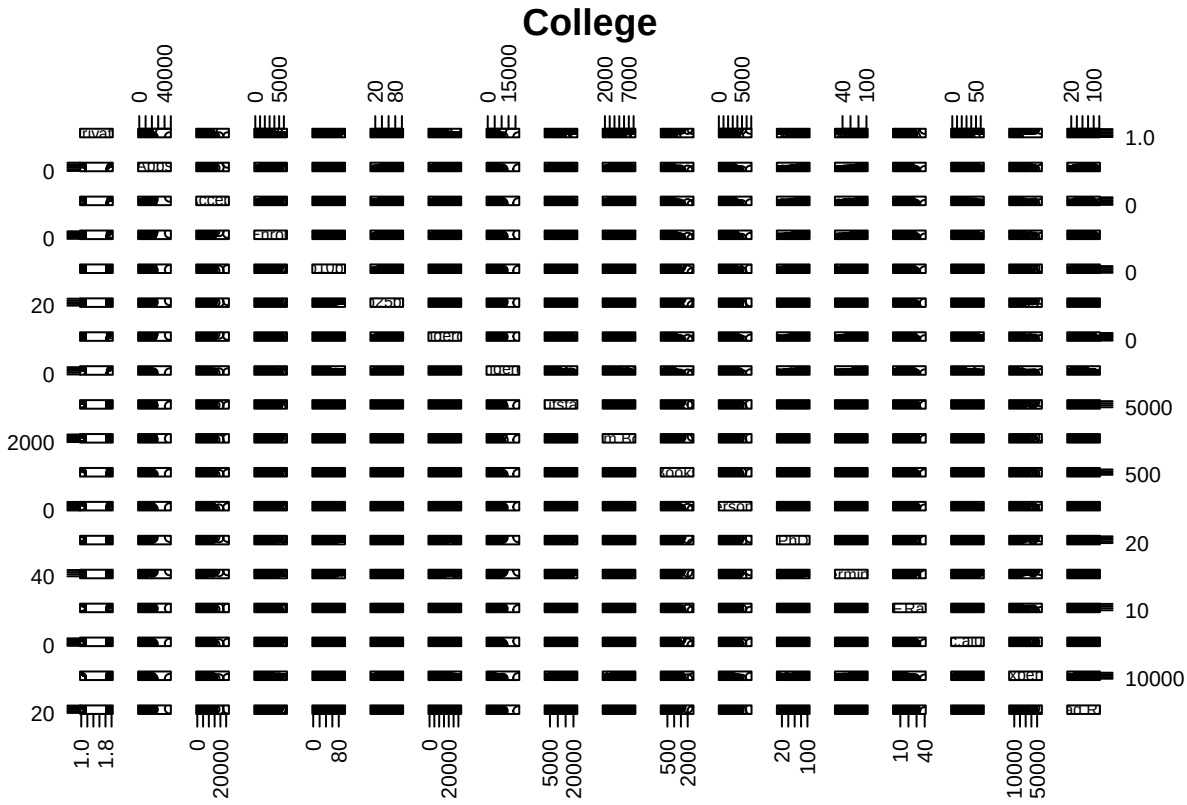
Question 1:

####Consider the "College" data in the ISLR2 package: ####> library(ISLR2) ####> data(College) ####> head(College)

####a) Present some visualizations of this data such as pair plots and histograms? Do you think any scaling or transformation is required?

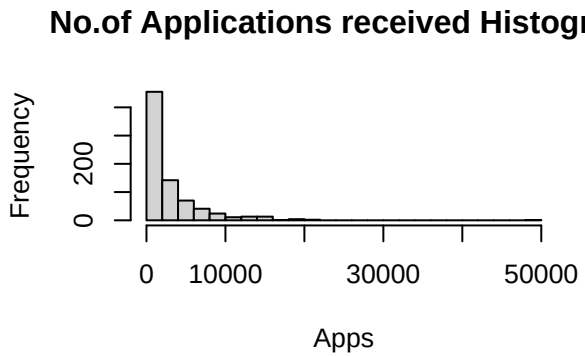
plotting a pair plot

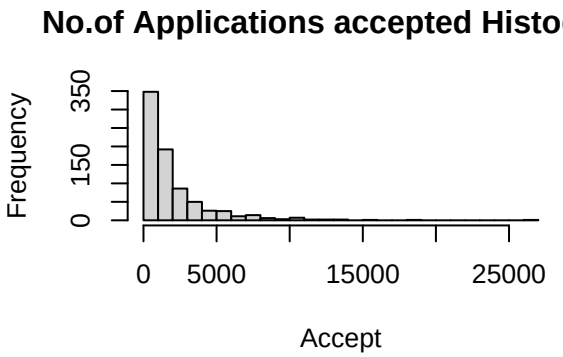
```
pairs(College, las=2, main= "College")
```

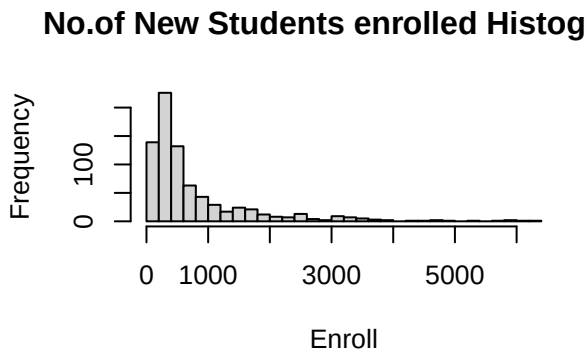


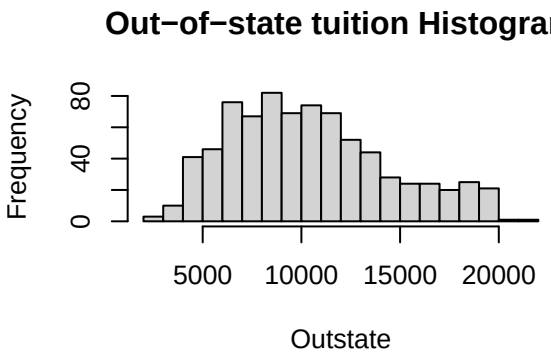
Plotting Histograms

```
par(mfrow=c(2,2))
hist(College[,2], breaks = 25, main = "No.of Applications received Histogram", xlab = "Apps")
hist(College[,3], breaks = 25, main = "No.of Applications accepted Histogram", xlab = "Accept")
hist(College[,4], breaks = 25, main = "No.of New Students enrolled Histogram", xlab = "Enroll")
hist(College[,9], breaks = 25, main = "Out-of-state tuition Histogram", xlab = "Outstate")
```









Check whether scaling required or not

```
sapply(College[, -1], skewness)
```

```
##        Apps      Accept      Enroll   Top10perc   Top25perc F.Undergrad
##   3.7165574   3.4111259   2.6852679   1.4104871   0.2588394   2.6054157
## P.Undergrad    Outstate  Room.Board       Books    Personal         PhD
##   5.6813582   0.5082943   0.4764335   3.4782933   1.7391308  -0.7666864
##    Terminal   S.F.Ratio perc.alumni      Expend   Grad.Rate
##  -0.8149652   0.6661462   0.6057190   3.4526399  -0.1135575
```

As shown above some of the variables have right-skewness greater than 0.5, so they need to be scaled using log transformation. So, that the distribution will be more symmetrical and it is easy to identify patterns and also will reduce the effect of outliers.

####b) Scale the data appropriately (e.g., log transform) and present the visualizations in part A. Have any new relationships been revealed.

Applying log transformation to the data

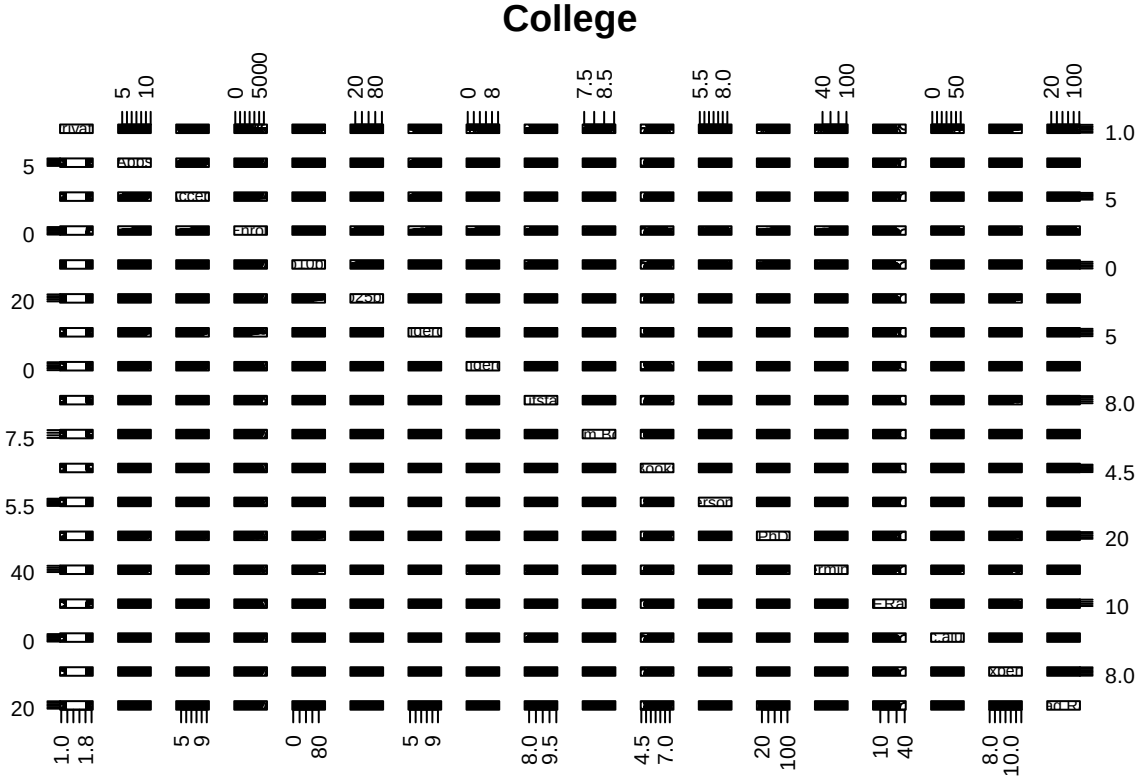
```
# Applying log transformation to the data
College <- College
College[, c(2:3,7:12,17)] <- log(College[,c(2:3,7:12,17)])
head(College)
```

```
##                              Private     Apps   Accept Enroll Top10perc
## Abilene Christian University     Yes 7.414573 7.116394    721        23
## Adelphi University               Yes 7.689829 7.562162    512        16
## Adrian College                   Yes 7.264030 7.000334    336        22
## Agnes Scott College              Yes 6.033086 5.855072    137        60
```

```
## Alaska Pacific University        Yes 5.262690 4.983607     55        16
## Albertson College                Yes 6.375025 6.171701    158        38
##                               Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University         52    7.967280    6.285998 8.914626
## Adelphi University                   29    7.894691    7.112327 9.415727
## Adrian College                       50    6.943122    4.595120 9.328123
## Agnes Scott College                  89    6.234411    4.143135 9.469623
## Alaska Pacific University            44    5.517453    6.767343 8.930626
## Albertson College                    62    6.519147    3.713572 9.510445
##                               Room.Board    Books Personal PhD Terminal
## Abilene Christian University   8.101678 6.109248 7.696213  70       78
## Adelphi University             8.771835 6.620073 7.313220  29       30
## Adrian College                 8.229511 5.991465 7.060476  53       66
## Agnes Scott College            8.603371 6.109248 6.774224  92       97
## Alaska Pacific University      8.323608 6.684612 7.313220  76       72
## Albertson College              8.112228 6.214608 6.514713  67       73
##                               S.F.Ratio perc.alumni   Expend Grad.Rate
## Abilene Christian University      18.1           12 8.859505        60
## Adelphi University                12.2           16 9.261699        56
## Adrian College                    12.9           30 9.075093        54
## Agnes Scott College                7.7           37 9.853036        59
## Alaska Pacific University         11.9            2 9.298534        15
## Albertson College                  9.4           11 9.182661        55
```

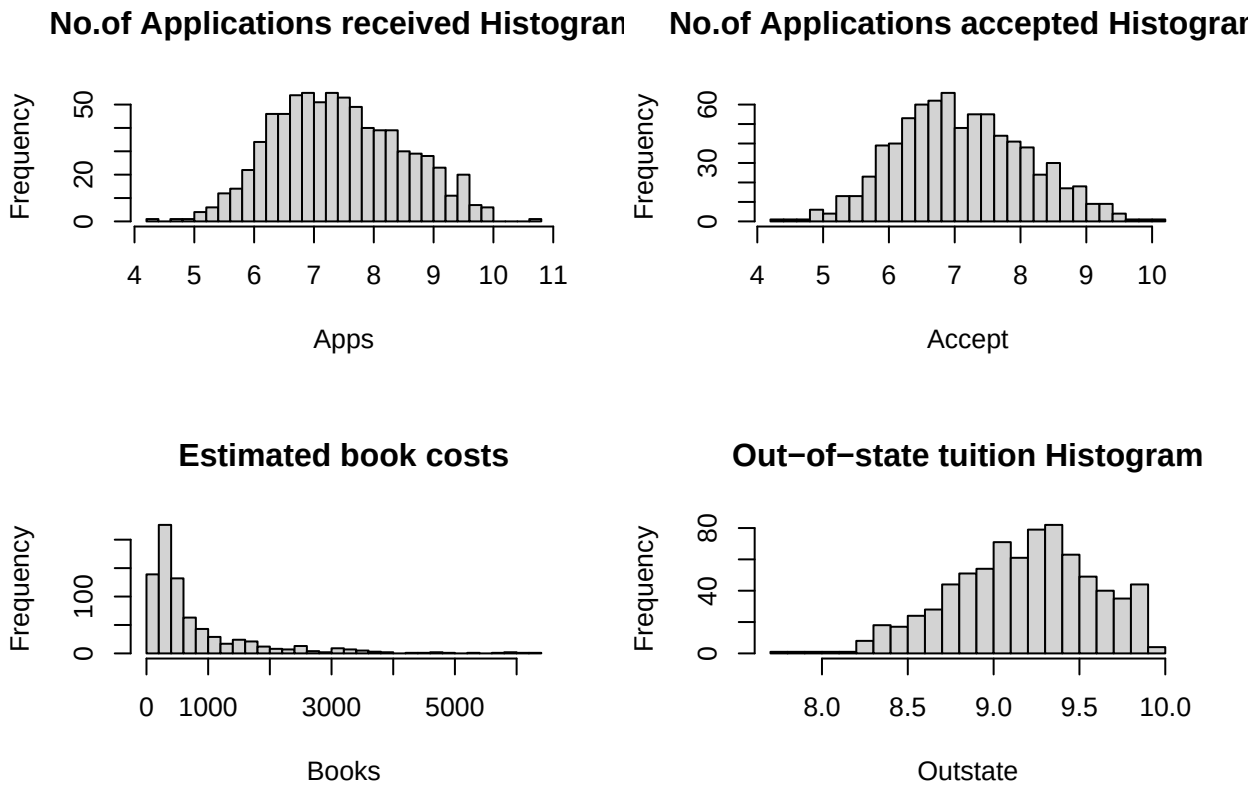
plotting a pair to the transformed data

```
pairs(College, las=2, main= "College")
```



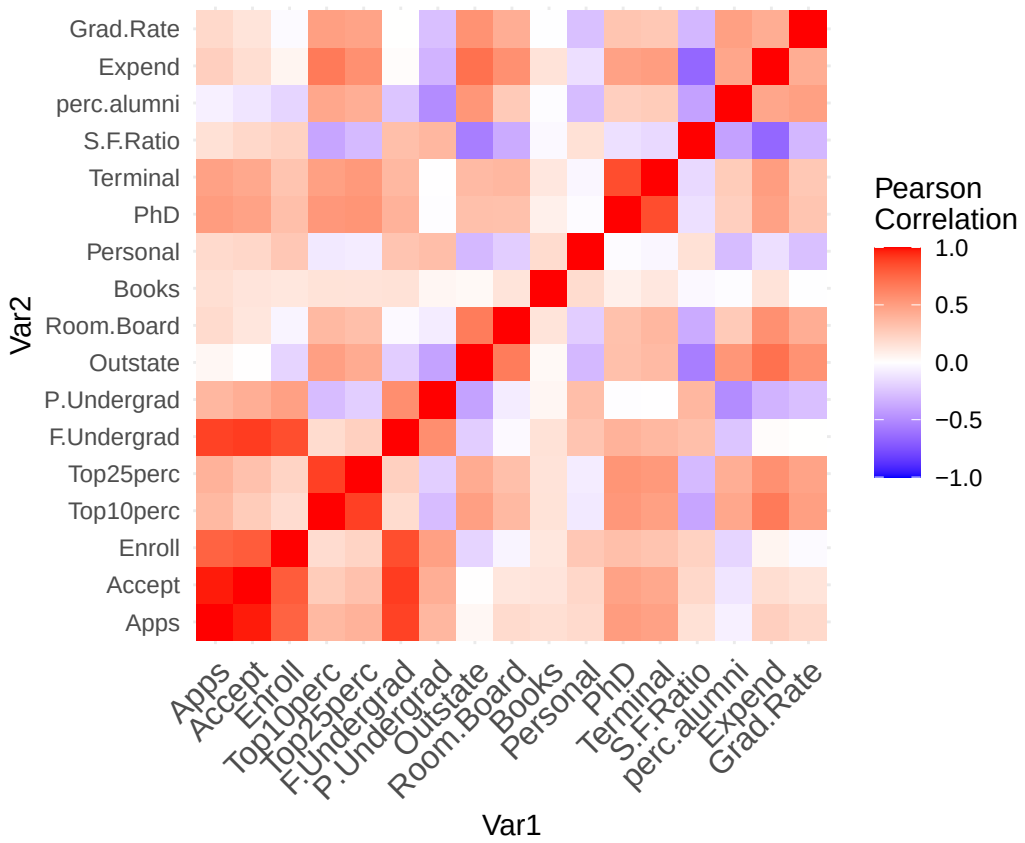
plotting Histograms to the transformed data.

```
par(mfrow=c(2,2))
hist(College[,2], breaks = 25, main = "No.of Applications received Histogram", xlab = "Apps")
hist(College[,3], breaks = 25, main = "No.of Applications accepted Histogram", xlab = "Accept")
hist(College[,4], breaks = 25, main = "Estimated book costs", xlab = "Books")
hist(College[,9], breaks = 25, main = "Out-of-state tuition Histogram", xlab = "Outstate")
```



Plotting a co-relation matrix

```
corr_matrix <- cor(College[,-1])
melted_corr_matrix <- melt(corr_matrix)
ggplot(data = melted_corr_matrix, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile() +
  scale_fill_gradient2(low = "blue", mid = "white", high = "red",
                       midpoint = 0, limit = c(-1,1), space = "Lab",
                       name="Pearson\nCorrelation") +
  theme_minimal() +
  theme(axis.text.x = element_text(angle = 45, vjust = 1,
                                   size = 12, hjust = 1)) +
  coord_fixed()
```



After applying the log transformation, we can see that the variables are more normally distributed.

Additionally, in the scatter plot matrix, the relationship between "Out-state" and "Private" is less clear.

Number of new students Enrolled and Number of applications accepted are linearly related with Number of applications Received.

Number of full-time Undergraduate students has linear relationship with Number of applications Received, accepted and new students enrolled.

Student-Faculty Ratio and Out-of-state tution are non linearly related. Instructional expenditure per student and out-of-state tution are linearly related

c) Subset the data into two data frames: "private" and "public". Sort them alphabetically. Save them as tab delimited txt files. Be sure these are the only two objects saved in that file. Submit it with your assignment (only on ublearns). Forming the data frames public and private

```
private <- College[College$Private == "Yes",]
public <- College[College$Private == "No",]
```

Sorting the data frames alphabetically by using the row names.

```
private <- private[order(rownames(private)), ]
public <- public[order(rownames(public)), ]
```

```
head(private)
```

```
##                              Private     Apps   Accept Enroll Top10perc
## Abilene Christian University     Yes 7.414573 7.116394    721        23
## Adelphi University               Yes 7.689829 7.562162    512        16
```

```
## Adrian College                   Yes 7.264030 7.000334    336        22
## Agnes Scott College              Yes 6.033086 5.855072    137        60
## Alaska Pacific University        Yes 5.262690 4.983607     55        16
## Albertson College                Yes 6.375025 6.171701    158        38
##                                Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University          52    7.967280    6.285998 8.914626
## Adelphi University                    29    7.894691    7.112327 9.415727
## Adrian College                        50    6.943122    4.595120 9.328123
## Agnes Scott College                   89    6.234411    4.143135 9.469623
## Alaska Pacific University             44    5.517453    6.767343 8.930626
## Albertson College                     62    6.519147    3.713572 9.510445
##                                Room.Board    Books Personal PhD Terminal
## Abilene Christian University     8.101678 6.109248 7.696213  70       78
## Adelphi University               8.771835 6.620073 7.313220  29       30
## Adrian College                   8.229511 5.991465 7.060476  53       66
## Agnes Scott College              8.603371 6.109248 6.774224  92       97
## Alaska Pacific University        8.323608 6.684612 7.313220  76       72
## Albertson College                8.112228 6.214608 6.514713  67       73
##                                S.F.Ratio perc.alumni   Expend Grad.Rate
## Abilene Christian University        18.1          12 8.859505        60
## Adelphi University                  12.2          16 9.261699        56
## Adrian College                      12.9          30 9.075093        54
## Agnes Scott College                  7.7          37 9.853036        59
## Alaska Pacific University           11.9           2 9.298534        15
## Albertson College                    9.4          11 9.182661        55
```

```
head(public)
```

```
##                                   Private     Apps   Accept Enroll Top10perc
## Angelo State University                No 8.171882 7.601402   1016        24
## Appalachian State University           No 8.897409 8.447629   1910        20
## Arizona State University Main campus   No 9.457903 9.240676   3761        24
## Arkansas Tech University               No 7.458186 7.455298    951        12
## Auburn University-Main Campus          No 8.929038 8.823353   3070        25
## Bemidji State University               No 7.096721 6.776507    546        12
##                                   Top25perc F.Undergrad P.Undergrad Outstate
## Angelo State University                  54    8.340456    7.321189 8.542861
## Appalachian State University             63    9.204322    6.942157 8.825560
## Arizona State University Main campus     49   10.025395    8.933928 8.913819
## Arkansas Tech University                 52    8.189245    6.844815 8.149024
## Auburn University-Main Campus            57    9.696586    7.447751 8.748305
## Bemidji State University                 36    8.241703    6.714171 8.395026
##                                   Room.Board    Books Personal PhD Terminal
## Angelo State University             8.186464 6.214608 7.600902  60       62
## Appalachian State University        7.839919 4.564348 7.600902  83       96
## Arizona State University Main campus 8.486734 6.551080 7.649693  88       93
## Arkansas Tech University            7.882315 6.109248 6.907755  57       60
## Auburn University-Main Campus       8.277158 6.396930 7.553811  85       91
## Bemidji State University            7.901007 6.492240 7.495542  57       62
##                                   S.F.Ratio perc.alumni   Expend Grad.Rate
## Angelo State University                23.1           5 8.296547        34
## Appalachian State University           18.3          14 8.674880        70
## Arizona State University Main campus   18.9           5 8.434246        48
## Arkansas Tech University               19.6           5 8.463581        48
## Auburn University-Main Campus          16.7          18 8.801168        69
```

```
## Bemidji State University                   19.6          16 8.230044        46
```

Save the public and private data frames as tab delimited *txt files

```
write.table(public, "public_colleges.txt", sep = "\t", row.names = TRUE,col.names = colnames(College))
write.table(private, "private_colleges.txt", sep = "\t", row.names = TRUE,col.names = colnames(College))
```

```
#No.of Top 25%  scored High shool students that joined in each public university
public$Top25HS <- public$Enroll * (public$Top25perc / 100)
#Print median of Top 25 High school students
median(public$Top25HS)
```

d) Within each new data frame from part C, eliminate Universities that have less than the median number of HS students admitted from the top 25% of the class("Top25perc").

```
## [1] 631.54
```

```
# Within each data frame, eliminate public universities with less than median Top25perc High school stu
public <- public[public$Top25HS >= median(public$Top25HS),]
```

```
#No.of Top 25%  scored Highshool students that joined in each public university
private$Top25HS <- private$Enroll * (private$Top25perc / 100)
#print median of Top 25 High school students
median(private$Top25HS)
```

```
## [1] 179.2
```

```
# Within each data frame, eliminate public universities with less than median Top25perc High school stu
private <- private[private$Top25HS >= median(private$Top25HS),]
```

```
head(public)
```

```
##                                     Private     Apps   Accept Enroll
## Appalachian State University             No 8.897409 8.447629   1910
## Arizona State University Main campus     No 9.457903 9.240676   3761
## Auburn University-Main Campus            No 8.929038 8.823353   3070
## Bowling Green State University           No 9.132487 8.900140   3076
## California Polytechnic-San Luis          No 8.963288 8.247220   1650
## California State University at Fresno    No 8.420682 8.099858   1483
##                                     Top10perc Top25perc F.Undergrad
## Appalachian State University               20        63    9.204322
## Arizona State University Main campus       24        49   10.025395
## Auburn University-Main Campus              25        57    9.696586
## Bowling Green State University             14        45    9.525078
## California Polytechnic-San Luis            47        73    9.465835
## California State University at Fresno       5        60    9.510000
##                                     P.Undergrad Outstate Room.Board    Books
## Appalachian State University           6.942157 8.825560   7.839919 4.564348
## Arizona State University Main campus   8.933928 8.913819   8.486734 6.551080
## Auburn University-Main Campus          7.447751 8.748305   8.277158 6.396930
## Bowling Green State University         7.100852 8.916238   8.117312 6.396930
## California Polytechnic-San Luis        7.247081 8.906529   8.492286 6.416732
## California State University at Fresno  7.134094 8.949755   8.382061 6.396930
##                                     Personal PhD Terminal S.F.Ratio
## Appalachian State University        7.600902  83       96      18.3
## Arizona State University Main campus 7.649693  88       93      18.9
```

```
## Auburn University-Main Campus         7.553811  85       91      16.7
## Bowling Green State University        7.438384  81       89      21.1
## California Polytechnic-San Luis       7.645398  72       81      19.8
## California State University at Fresno 7.563201  90       90      21.2
##                                       perc.alumni   Expend Grad.Rate Top25HS
## Appalachian State University                   14 8.674880        70 1203.30
## Arizona State University Main campus            5 8.434246        48 1842.89
## Auburn University-Main Campus                  18 8.801168        69 1749.90
## Bowling Green State University                 14 8.841882        67 1384.20
## California Polytechnic-San Luis                13 9.042277        59 1204.50
## California State University at Fresno           8 8.891236        61  889.80
```

```
head(private)
```

```
##                                     Private     Apps   Accept Enroll
## Abilene Christian University            Yes 7.414573 7.116394    721
## Albion College                          Yes 7.549083 7.450080    489
## Alfred University                       Yes 7.457032 7.261927    472
## Allegheny College                       Yes 7.883069 7.549609    484
## Allentown Coll. of St. Francis de Sales Yes 7.072422 6.659294    290
## Alma College                            Yes 7.144407 6.984716    385
##                                     Top10perc Top25perc F.Undergrad
## Abilene Christian University               23        52    7.967280
## Albion College                             37        68    7.374002
## Alfred University                          37        75    7.512071
## Allegheny College                          44        77    7.442493
## Allentown Coll. of St. Francis de Sales    38        64    7.029973
## Alma College                               44        73    7.174724
##                                     P.Undergrad Outstate Room.Board
## Abilene Christian University           6.285998 8.914626   8.101678
## Albion College                         3.465736 9.537339   8.481773
## Alfred University                      4.700480 9.714021   8.595265
## Allegheny College                      3.784190 9.745663   8.398410
## Allentown Coll. of St. Francis de Sales 6.458338 9.178850   8.473241
## Alma College                           3.332205 9.439227   8.423322
##                                        Books Personal PhD Terminal
## Abilene Christian University        6.109248 7.696213  70       78
## Albion College                      6.109248 6.745236  89      100
## Alfred University                   6.214608 6.396930  82       88
## Allegheny College                   5.991465 6.396930  73       91
## Allentown Coll. of St. Francis de Sales 6.396930 6.907755  60       84
## Alma College                        5.991465 5.991465  79       87
##                                     S.F.Ratio perc.alumni   Expend
## Abilene Christian University             18.1          12 8.859505
## Albion College                           13.7          37 9.348971
## Alfred University                        11.3          31 9.299450
## Allegheny College                         9.9          41 9.368284
## Allentown Coll. of St. Francis de Sales  13.3          21 8.979669
## Alma College                             15.3          32 9.138307
##                                     Grad.Rate Top25HS
## Abilene Christian University               60  374.92
## Albion College                             73  332.52
## Alfred University                          73  354.00
## Allegheny College                          76  372.68
## Allentown Coll. of St. Francis de Sales    74  185.60
```

```
## Alma College                                    68  281.05
```

```
# Determine cut-off points using quantile
private.quantiles <- quantile(private$Grad.Rate, probs = c(0, 0.5, 0.75, 1))
public.quantiles <- quantile(public$Grad.Rate, probs = c(0, 0.5, 0.75, 1))
private.quantiles
```

e) Create a new variable that categorizes graduation rate into "High", "Medium" and "Low", use a histogram or quantiles to determine how to create this variable. Append this variable to your "private" and "public" datasets.

```
##   0%  50%  75% 100%
##   18   77   87  118
```

```
public.quantiles
```

```
##   0%  50%  75% 100%
##   10   58   68   98
```

```
# Create new variable using cut
private$Grad_Rate_Category <- cut(private$Grad.Rate, breaks = private.quantiles,
                                  labels = c("Low", "Medium", "High"), include.lowest = TRUE)
public$Grad_Rate_Category <- cut(public$Grad.Rate, breaks = public.quantiles,
                                 labels = c("Low", "Medium", "High"), include.lowest = TRUE)
```

```
head(private)
```

```
##                                         Private     Apps   Accept Enroll
## Abilene Christian University               Yes 7.414573 7.116394    721
## Albion College                             Yes 7.549083 7.450080    489
## Alfred University                          Yes 7.457032 7.261927    472
## Allegheny College                          Yes 7.883069 7.549609    484
## Allentown Coll. of St. Francis de Sales    Yes 7.072422 6.659294    290
## Alma College                               Yes 7.144407 6.984716    385
##                                         Top10perc Top25perc F.Undergrad
## Abilene Christian University                   23        52    7.967280
## Albion College                                 37        68    7.374002
## Alfred University                              37        75    7.512071
## Allegheny College                              44        77    7.442493
## Allentown Coll. of St. Francis de Sales        38        64    7.029973
## Alma College                                   44        73    7.174724
##                                         P.Undergrad Outstate Room.Board
## Abilene Christian University               6.285998 8.914626   8.101678
## Albion College                             3.465736 9.537339   8.481773
## Alfred University                          4.700480 9.714021   8.595265
## Allegheny College                          3.784190 9.745663   8.398410
## Allentown Coll. of St. Francis de Sales    6.458338 9.178850   8.473241
## Alma College                               3.332205 9.439227   8.423322
##                                            Books Personal PhD Terminal
## Abilene Christian University            6.109248 7.696213  70       78
## Albion College                          6.109248 6.745236  89      100
## Alfred University                       6.214608 6.396930  82       88
## Allegheny College                       5.991465 6.396930  73       91
## Allentown Coll. of St. Francis de Sales 6.396930 6.907755  60       84
## Alma College                            5.991465 5.991465  79       87
```

```
##                                     S.F.Ratio perc.alumni   Expend
## Abilene Christian University             18.1          12 8.859505
## Albion College                           13.7          37 9.348971
## Alfred University                        11.3          31 9.299450
## Allegheny College                         9.9          41 9.368284
## Allentown Coll. of St. Francis de Sales  13.3          21 8.979669
## Alma College                             15.3          32 9.138307
##                                     Grad.Rate Top25HS Grad_Rate_Category
## Abilene Christian University               60  374.92                Low
## Albion College                             73  332.52                Low
## Alfred University                          73  354.00                Low
## Allegheny College                          76  372.68                Low
## Allentown Coll. of St. Francis de Sales    74  185.60                Low
## Alma College                               68  281.05                Low
```

```
head(public)
```

```
##                                    Private     Apps   Accept Enroll
## Appalachian State University            No 8.897409 8.447629   1910
## Arizona State University Main campus    No 9.457903 9.240676   3761
## Auburn University-Main Campus           No 8.929038 8.823353   3070
## Bowling Green State University          No 9.132487 8.900140   3076
## California Polytechnic-San Luis         No 8.963288 8.247220   1650
## California State University at Fresno   No 8.420682 8.099858   1483
##                                    Top10perc Top25perc F.Undergrad
## Appalachian State University              20        63    9.204322
## Arizona State University Main campus      24        49   10.025395
## Auburn University-Main Campus             25        57    9.696586
## Bowling Green State University            14        45    9.525078
## California Polytechnic-San Luis           47        73    9.465835
## California State University at Fresno      5        60    9.510000
##                                    P.Undergrad Outstate Room.Board    Books
## Appalachian State University          6.942157 8.825560   7.839919 4.564348
## Arizona State University Main campus  8.933928 8.913819   8.486734 6.551080
## Auburn University-Main Campus         7.447751 8.748305   8.277158 6.396930
## Bowling Green State University        7.100852 8.916238   8.117312 6.396930
## California Polytechnic-San Luis       7.247081 8.906529   8.492286 6.416732
## California State University at Fresno 7.134094 8.949755   8.382061 6.396930
##                                    Personal PhD Terminal S.F.Ratio
## Appalachian State University       7.600902  83       96      18.3
## Arizona State University Main campus 7.649693  88       93      18.9
## Auburn University-Main Campus      7.553811  85       91      16.7
## Bowling Green State University     7.438384  81       89      21.1
## California Polytechnic-San Luis    7.645398  72       81      19.8
## California State University at Fresno 7.563201  90       90      21.2
##                                    perc.alumni   Expend Grad.Rate Top25HS
## Appalachian State University                14 8.674880        70 1203.30
## Arizona State University Main campus         5 8.434246        48 1842.89
## Auburn University-Main Campus               18 8.801168        69 1749.90
## Bowling Green State University              14 8.841882        67 1384.20
## California Polytechnic-San Luis             13 9.042277        59 1204.50
## California State University at Fresno        8 8.891236        61  889.80
##                                    Grad_Rate_Category
## Appalachian State University                     High
## Arizona State University Main campus              Low
```

```
## Auburn University-Main Campus                        High
## Bowling Green State University                     Medium
## California Polytechnic-San Luis                    Medium
## California State University at Fresno              Medium
```

Create a "list structure" that contains your two datasets and save this to an .RData file. Make sure that your file contains only the list structure. Submit this with your homework (only on ublearns). Creating a list structure and copying public and private data frames to list

```
private_public_data <- list()
private_public_data[[1]] <- public
private_public_data[[2]] <- private
```

```
save(private_public_data,file='private_public_data.RData')
```

Question 2:

2) You are going to derive generalized association rules to the marketing data from your book ESL. This data is in the available on UB learns. Specifically, generate a reference sample of the same size of the training set. This can be done in a couple of ways, e.g., (i) sample uniformly for each variable, or (ii) by randomly permuting the values within each variable independently. Build a classification tree to the training sample (class 1) and the reference sample (class 0) and describe the terminal nodes having highest estimated class 1 probability. Compare the results to the results near Table 14.1 (ESL), which were derived using PRIM.

Imported Marketing Data through CSV into R

```
Marketing <- read.csv("Marketing.csv")
head(Marketing)
```

```
##   Income Sex Marital_Status Age Education Occupation Years_In_Bay_Area
## 1      9   2              1   5         4          5                 5
## 2      9   1              1   5         5          5                 5
## 3      9   2              1   3         5          1                 5
## 4      1   2              5   1         2          6                 5
## 5      1   2              5   1         2          6                 3
## 6      8   1              1   6         4          8                 5
##   Dual_Incomes Number_In_Household Number_Of_Children Householder_Status
## 1            3                   3                  0                 1
## 2            3                   5                  2                 1
## 3            2                   3                  1                 2
## 4            1                   4                  2                 3
## 5            1                   4                  2                 3
## 6            3                   2                  0                 1
##   Type_Of_Home Ethnic_Classification Language_In_Home
## 1            1                     7               NA
## 2            1                     7                1
## 3            3                     7                1
## 4            1                     7                1
## 5            1                     7                1
## 6            1                     7                1
```

Removing NA values

```
Marketing <- na.omit(Marketing)
```

Creating the reference sample by permuting or sampling the values within each variable independently.

```
set.seed(123)
ref_sample = Marketing
for(i in 1:ncol(ref_sample)){
  ref_sample[,i] = sample(ref_sample[,i], nrow(ref_sample), replace = T)
}
dim(ref_sample)
```

```
## [1] 6876   14
```

Assigning class variables to the sample observations both train and reference (binary)

```
ref_sample$class <- 0
Marketing$class <- 1
```

Combine the two samples and store in data_all variable

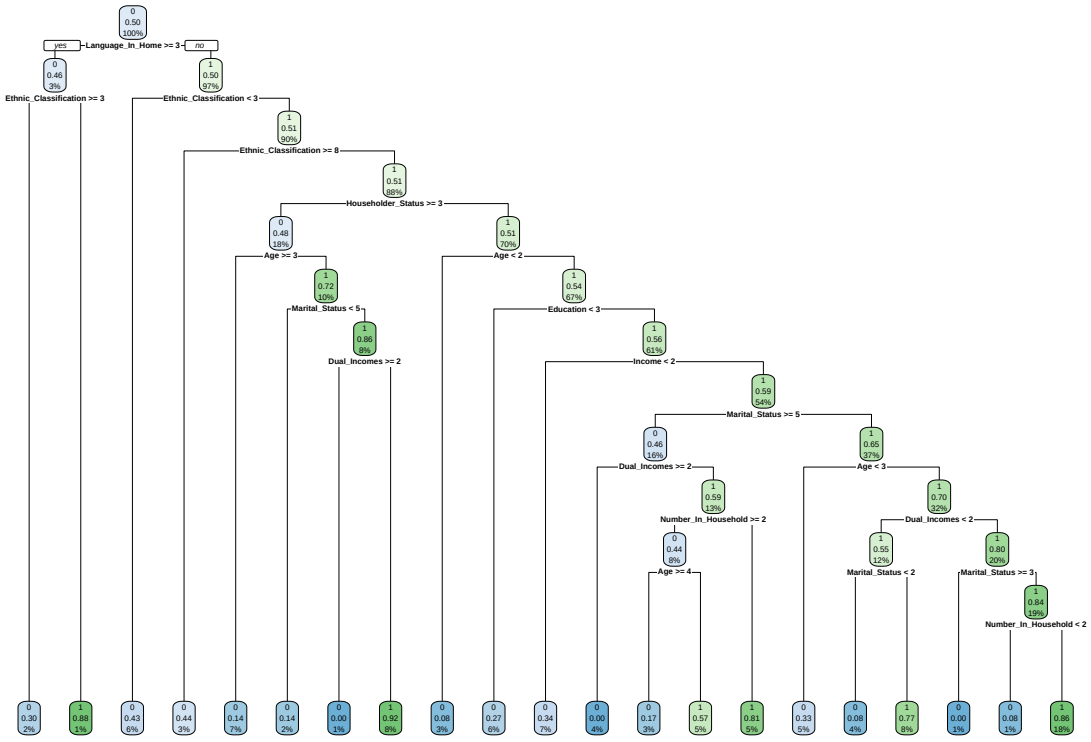
```
data_all <- rbind(Marketing, ref_sample)
```

Fit the Tree model

```
tree <- rpart(class ~ ., data = data_all, method = "class")
```

Plot the tree

```
rpart.plot(tree)
```




```
#summary(tree)
```

From the above plot it is evident that the highest estimated class 1 probability is 0.84.

Obersavations from the tree comparing to ESL PRIM data.

If Number of Household <= 8 and Number of children <=5 and Language_in_home>=2 => Ethnic classification>=7 with a value of 0.7 and confidence 5%

If Dual_income < 2, Marital_status >= 5, Household status < 3, Ethnic classification < 3,Age < 3 => languages in home <3 with a value of 0.73 and confidence 7%

If Number_of_Children<=5,Language_In_Home <=2, Income>=2, Marital status>=5 => Education <3 with a value of 0.70 and confidence 8%

Question 3:

Consider the Boston Housing Data in the ISLR2 package. (Important – do not use data from any other packages).

a) Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.
Visualizing the data using histograms.

```
par(mfrow=c(3,2))
#x11()
hist(Boston$crim, breaks = 4, main = "per capita crime rate by town.", xlab = "crim")

#x11()
hist(Boston$zn, breaks = 3, main = "proportion of residential land zoned for lots over 25,000 sq.ft.", x

#x11()
hist(Boston$indus, breaks = 3, main = "proportion of non-retail business acres per town.", xlab = "indus

#x11()
hist(Boston$age, breaks = 3, main = "proportion of owner-occupied units built prior to 1940.", xlab = "a

#x11()
hist(Boston$tax, breaks = 3, main = "full-value property-tax rate per $10,000.", xlab = "tax")
```

#Copy Boston data to Boston2 for later use.

```
Boston2<-Boston
```

Categorize the variables using cut function

```
Boston$crim <- ordered(cut(Boston$crim, breaks = c(-Inf, 1, 5, Inf), labels = c("Low", "Medium", "High"))
Boston$zn <- cut(Boston$zn, breaks = c(-Inf, 10, 20, Inf), labels = c("Low", "Medium", "High"))
Boston$indus <- cut(Boston$indus, breaks = c(-Inf, 15, 25, Inf), labels = c("Low", "Medium", "High"))
Boston$chas <- factor(Boston$chas)
Boston$nox <- cut(Boston$nox, breaks = c(-Inf, 0.5, 0.7, Inf), labels = c("Low", "Medium", "High"))
Boston$rm <- cut(Boston$rm, breaks = c(-Inf, 6, 7, Inf), labels = c("Low", "Medium", "High"))
Boston$age <- ordered(cut(Boston$age, breaks = c(-Inf, 70, 85, Inf), labels = c("Low", "Medium", "High"))
Boston$dis <- cut(Boston$dis, breaks = c(-Inf, 3, 5, Inf), labels = c("short", "Medium", "Long"))
Boston$rad <- factor(Boston$rad)
Boston$tax <- cut(Boston$tax, breaks = c(-Inf, 300, 500, Inf), labels = c("Low", "Medium", "High"))
Boston$ptratio <- cut(Boston$ptratio, breaks = c(-Inf, 17, 20, Inf), labels = c("Low", "Medium", "High"))
Boston$lstat <- cut(Boston$lstat, breaks = c(-Inf, 10, 15, Inf), labels = c("Low", "Medium", "High"))
Boston$medv <- cut(Boston$medv, breaks = c(-Inf, 25, Inf), labels = c("low", "high"))
```

I have used different break points for each variable, based on the distribution in histograms.

I have also looked at the data and used relevant domain knowledge to categorize each variable, so some variables like tax, dis, age.

As the CRIM and AGE are ordered variables general cut function doesn't preserve the ordering. So, I have used order function to preserve the order of variables.

For the un ordered categorical variables like CHAS and RAD we leave them as un ordered factors only as they are already categorized.

Convert to a binary incidence matrix

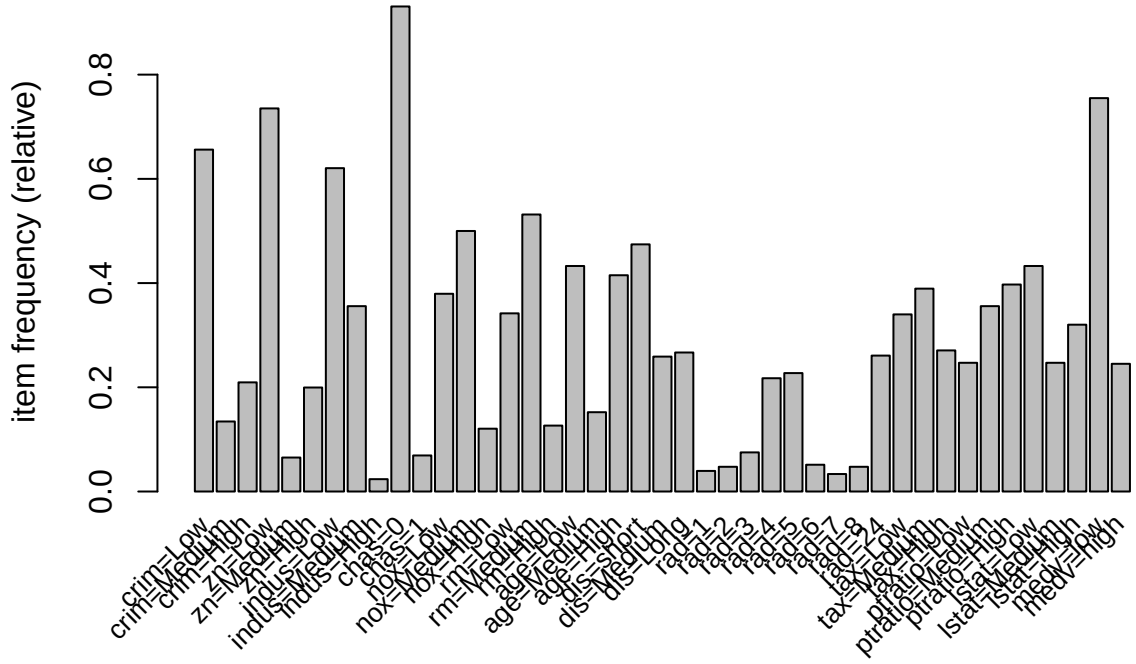
```
Boston <- as(Boston, "transactions")
summary(Boston)
```

```
## transactions as itemMatrix in sparse format with
##  506 rows (elements/itemsets/transactions) and
##  43 columns (items) and a density of 0.3023256
##
## most frequent items:
##    chas=0  medv=low    zn=Low  crim=Low indus=Low   (Other)
##       471       382       372       332       314      4707
##
## element (itemset/transaction) length distribution:
## sizes
##  13
## 506
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      13      13      13      13      13      13
##
## includes extended item information - examples:
##        labels variables levels
## 1    crim=Low      crim    Low
## 2 crim=Medium      crim Medium
## 3   crim=High      crim   High
##
## includes extended transaction information - examples:
```

```
##   transactionID
## 1             1
## 2             2
## 3             3
```

b) Visualize the data using the itemFrequencyPlot in the "arules" package. Apply the apriori algorithm (Do not forget to specify parameters in your write up). Plot the item Frequency plot

```
#x11()
itemFrequencyPlot(Boston, support = 0.01, cex.names = 0.8)
```



Apply the Apriori algorithm

```
rules  <- apriori(Boston, parameter = list(support = 0.01, confidence = 0.7))
```

```
## Apriori
##
## Parameter specification:
##  confidence minval smax arem  aval originalSupport maxtime support minlen
##         0.7    0.1    1 none FALSE            TRUE       5    0.01      1
##  maxlen target  ext
##      10  rules TRUE
##
## Algorithmic control:
##  filter tree heap memopt load sort verbose
##     0.1 TRUE TRUE  FALSE TRUE    2    TRUE
##
## Absolute minimum support count: 5
```

```
##
## set item appearances ...[0 item(s)] done [0.00s].
## set transactions ...[43 item(s), 506 transaction(s)] done [0.00s].
## sorting and recoding items ... [43 item(s)] done [0.00s].
## creating transaction tree ... done [0.00s].
## checking subsets of size 1 2 3 4 5 6 7 8 9 10
```

```
## Warning in apriori(Boston, parameter = list(support = 0.01, confidence = 0.7)):
## Mining stopped (maxlen reached). Only patterns up to a length of 10 returned!
```

```
##  done [0.05s].
## writing ... [747660 rule(s)] done [0.10s].
## creating S4 object  ... done [0.36s].
```

```
summary(rules)
```

```
## set of 747660 rules
##
## rule length distribution (lhs + rhs):sizes
##      1      2      3      4      5      6      7      8      9     10
##      3    226   3786  24465  80637 157672 196423 161878  89576  32994
##
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   6.000   7.000   7.042   8.000  10.000
##
## summary of quality measures:
##     support          confidence        coverage            lift
##  Min.   :0.01186   Min.   :0.7000   Min.   :0.01186   Min.   : 0.752
##  1st Qu.:0.01383   1st Qu.:0.8889   1st Qu.:0.01383   1st Qu.: 1.360
##  Median :0.01779   Median :1.0000   Median :0.01976   Median : 2.000
##  Mean   :0.02977   Mean   :0.9427   Mean   :0.03208   Mean   : 2.353
##  3rd Qu.:0.02964   3rd Qu.:1.0000   3rd Qu.:0.03360   3rd Qu.: 2.635
##  Max.   :0.93083   Max.   :1.0000   Max.   :1.00000   Max.   :42.167
##      count
##  Min.   :  6.00
##  1st Qu.:  7.00
##  Median :  9.00
##  Mean   : 15.06
##  3rd Qu.: 15.00
##  Max.   :471.00
##
## mining info:
##    data ntransactions support confidence
##  Boston           506    0.01        0.7
##                                                                                call
##  apriori(data = Boston, parameter = list(support = 0.01, confidence = 0.7))
```

I have considered minimum support value of 0.01, so we get the most of the association rules(1-5%) without missing many.

And confidence which can be from 50% to 80% in general. so I choose 0.7 to get good rules.

```
rulestax_and_crime_low <- subset(rules, subset = lhs %in% c("crim=Low") & rhs %in% c("tax=Low") & lift >
inspect(head(sort(rulestax_and_crime_low , by = "confidence"), n = 6))
```

c) A student is interested low taxes, but wants to be in a safe aera with low crime. What can you advise on this matter through the mining of association rules?

```
##     lhs                                          rhs       support    confidence
## [1] {crim=Low, indus=High, rad=2}          => {tax=Low} 0.01383399 1
## [2] {crim=Low, indus=High, ptratio=Medium} => {tax=Low} 0.01383399 1
## [3] {crim=Low, dis=Medium, rad=2}          => {tax=Low} 0.02173913 1
## [4] {crim=Low, rm=Low, rad=2}              => {tax=Low} 0.01185771 1
## [5] {crim=Low, age=High, rad=2}            => {tax=Low} 0.01185771 1
## [6] {crim=Low, dis=short, rad=2}           => {tax=Low} 0.01383399 1
##     coverage   lift    count
## [1] 0.01383399 2.94186  7
## [2] 0.01383399 2.94186  7
## [3] 0.02173913 2.94186 11
## [4] 0.01185771 2.94186  6
## [5] 0.01185771 2.94186  6
## [6] 0.01383399 2.94186  7
```

```
inspect(head(sort(rulestax_and_crime_low , by = "lift"), n = 6))
```

```
##     lhs                                          rhs       support    confidence
## [1] {crim=Low, indus=High, rad=2}          => {tax=Low} 0.01383399 1
## [2] {crim=Low, indus=High, ptratio=Medium} => {tax=Low} 0.01383399 1
## [3] {crim=Low, dis=Medium, rad=2}          => {tax=Low} 0.02173913 1
## [4] {crim=Low, rm=Low, rad=2}              => {tax=Low} 0.01185771 1
## [5] {crim=Low, age=High, rad=2}            => {tax=Low} 0.01185771 1
## [6] {crim=Low, dis=short, rad=2}           => {tax=Low} 0.01383399 1
##     coverage   lift    count
## [1] 0.01383399 2.94186  7
## [2] 0.01383399 2.94186  7
## [3] 0.02173913 2.94186 11
## [4] 0.01185771 2.94186  6
## [5] 0.01185771 2.94186  6
## [6] 0.01383399 2.94186  7
```

If we need crime rate low area which also has low tax

Rule 1: {crim=Low, indus=High, rad=2}=>{tax=Low}, From rule 1 we can say that he need to choose a industrial area and also an area far from radial highways

Rule 2: {crim=Low, indus=High, ptratio=Medium} =>{tax=Low}, From rule 2 we say that he need to choose industrial area and also an area with less pupil-teacher ratio.(generally we have more taxes in the area of less pupil-teacher ratio)

Rule 3: {crim=Low, dis=Medium, rad=2}=>{tax=Low}, From rule 3 we can say that he little bit near to Boston employment centers.

Finally from the rules we can say that if we want to choose a area with low crime and tax he need to choose an industrial area instead of residential area and also an area with less accessibility to radial highways and area with schools having medium pupil-teacher ratio.

```
low_pupil_teacher_ratio <- subset(rules, subset = rhs %in% "ptratio=Low" & lift>1.2)
inspect(head(sort(low_pupil_teacher_ratio , by = "confidence"), n = 6))
```

d) A family is moving to the area, and has made schooling a priority. They want schools with low pupil-teacher ratios. What can you advise on this matter through the mining of association rules?

```
##     lhs                    rhs           support    confidence coverage   lift
## [1] {zn=High, rad=6}   => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [2] {dis=Long, rad=6}  => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [3] {rad=6, tax=Low}   => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [4] {nox=Low, rad=6}   => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [5] {rad=6, lstat=Low} => {ptratio=Low} 0.01185771 1          0.01185771 4.048
## [6] {zn=Medium, rad=5} => {ptratio=Low} 0.04545455 1          0.04545455 4.048
##     count
## [1]  9
## [2]  9
## [3]  9
## [4]  9
## [5]  6
## [6] 23
```

```
inspect(head(sort(low_pupil_teacher_ratio , by = "lift"), n = 6))
```

```
##     lhs                    rhs           support    confidence coverage   lift
## [1] {zn=High, rad=6}   => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [2] {dis=Long, rad=6}  => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [3] {rad=6, tax=Low}   => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [4] {nox=Low, rad=6}   => {ptratio=Low} 0.01778656 1          0.01778656 4.048
## [5] {rad=6, lstat=Low} => {ptratio=Low} 0.01185771 1          0.01185771 4.048
## [6] {zn=Medium, rad=5} => {ptratio=Low} 0.04545455 1          0.04545455 4.048
##     count
## [1]  9
## [2]  9
## [3]  9
## [4]  9
## [5]  6
## [6] 23
```

If we need low pupil-teacher ratios.

Rule1:{zn=High, rad=6}=> {ptratio=Low} They need to choose an area with good proportion of residential land area (>25000 sq.ft).

Rule2:{dis=Long, rad=6}=> {ptratio=Low} They need to choose an area which is far from the five Boston employment centers.

Rule3:{nox=Low, rad=6}=>{ptratio=Low} They need to choose an area with low nitrogen oxides concentration.

Rule4:{rad=6, lstat=Low}=> {ptratio=Low} They need to choose area which has less percent lower status population.

Finally if they need low pupil-teacher ratio they need to choose an area with good proportion of residential land area, far from five Boston employment centers, low nitrogen oxides concentration and less percent of lower status population.

Extra credit

```
model.control <- rpart.control(minsplit = 5, xval = 10, cp = 0)
pttree<- rpart(ptratio~., data = Boston2, method = "class", control = model.control)
```

```
#x11()
rpart.plot(pttree)
```

```
## Warning: All boxes will be white (the box.palette argument will be ignored) because
```

```
## the number of classes in the response 46 is greater than length(box.palette) 6.
## To silence this warning use box.palette=0 or trace=-1.

## Warning: labs do not fit even at cex 0.15, there may be some overplotting
```