Question 1 - Consider the “College” data in the ISLR2 package:

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

In [1]:

```
#install.packages("corrplot")
#install.packages("tidyr",type="binary")
#install.packages('ISLR2')
library(ISLR2)
library(rpart)
library(rpart.plot)
library(caret)
library(dplyr)
library(tidyr)
library(corrplot)
library(arules)
```

```
Warning message:
"package 'rpart' was built under R version 3.6.3"Warning message:
"package 'rpart.plot' was built under R version 3.6.3"Warning message:
"package 'caret' was built under R version 3.6.3"Loading required package: lattice
Warning message:
"package 'lattice' was built under R version 3.6.3"Loading required package: ggplo
t2
Warning message:
"package 'ggplot2' was built under R version 3.6.3"Warning message:
"package 'dplyr' was built under R version 3.6.3"
Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

    filter, lag

The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union

Warning message:
"package 'tidyr' was built under R version 3.6.3"corrplot 0.92 loaded
Warning message:
"package 'arules' was built under R version 3.6.3"Loading required package: Matrix
Warning message:
"package 'Matrix' was built under R version 3.6.3"
Attaching package: 'Matrix'

The following objects are masked from 'package:tidyr':

    expand, pack, unpack


Attaching package: 'arules'

The following object is masked from 'package:dplyr':

    recode

The following objects are masked from 'package:base':

    abbreviate, write
```

In [2]:

```
data(College)
head(College)
dim(College)
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	[illegible]
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12
Adrian College	Yes	1428	1097	336	22	50	1036	99	1
Agnes Scott College	Yes	417	349	137	60	89	510	63	12
Alaska Pacific University	Yes	193	146	55	16	44	249	869	[illegible]
Albertson College	Yes	587	479	158	38	62	678	41	13

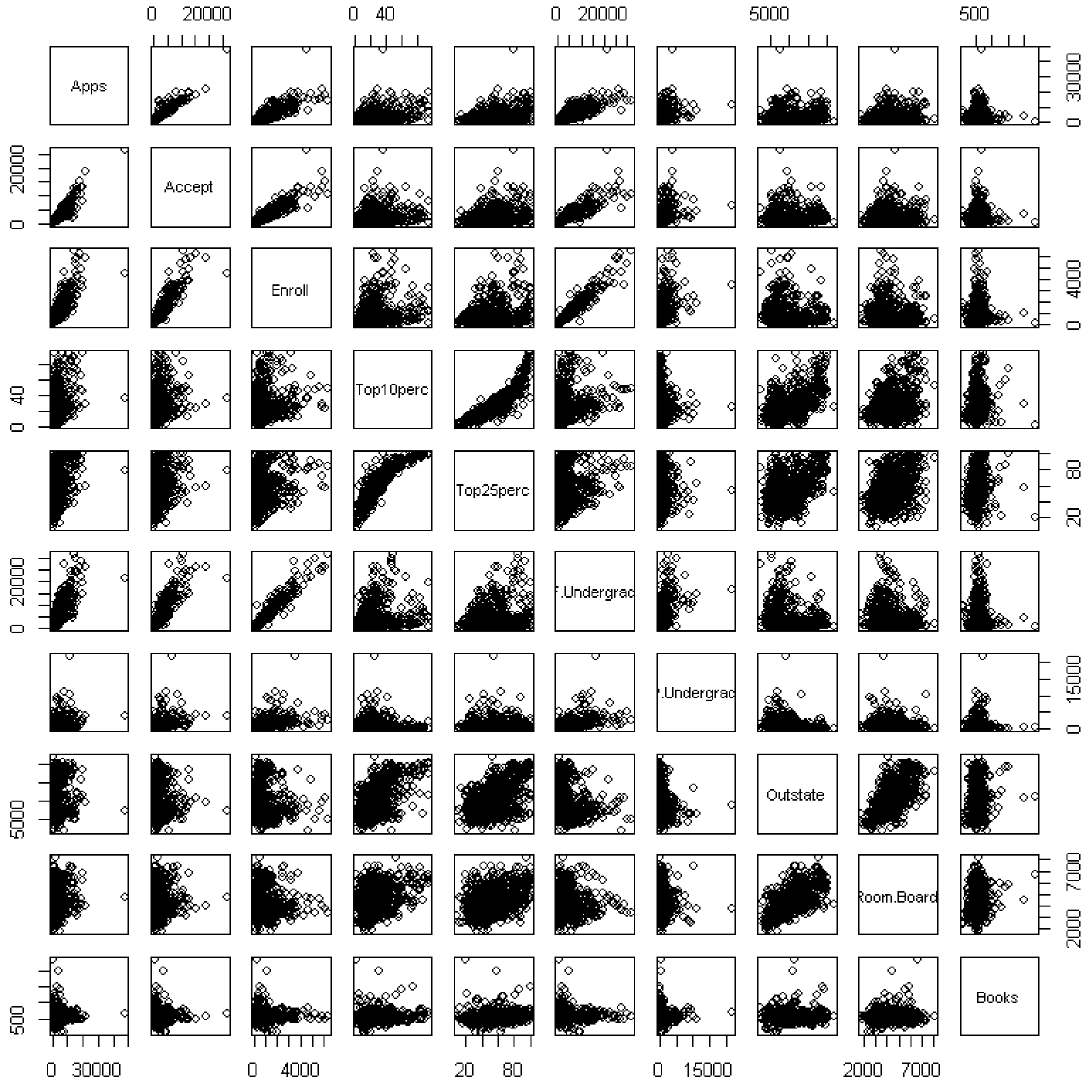
1. 777
2. 18

In [3]:

```
# here we can see that very features are highly correlated to each other
corrplot(cor(College[,2:11]), method = "number")
```

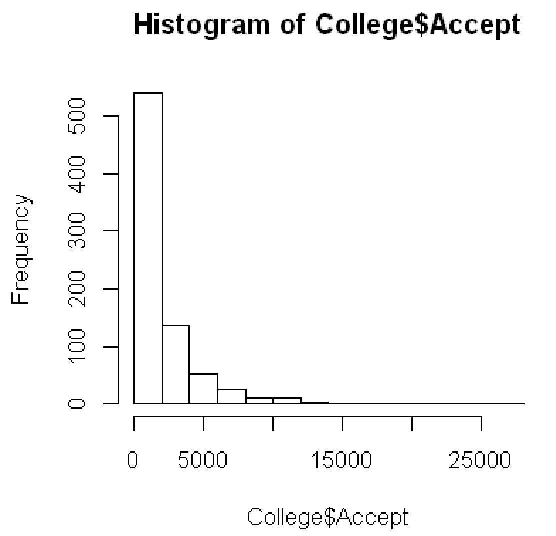
In [4]:

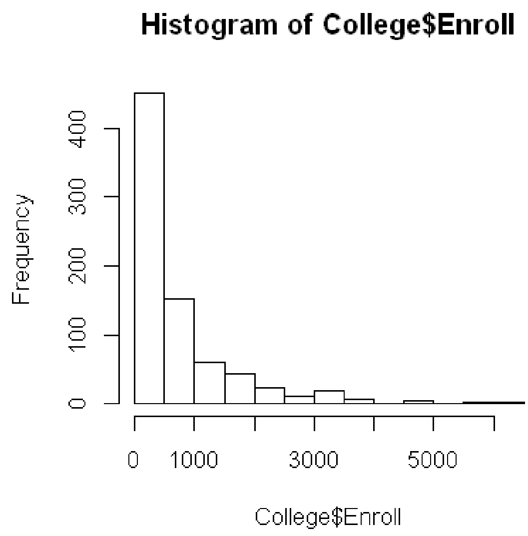
```
pairs(College[,2:11])
```

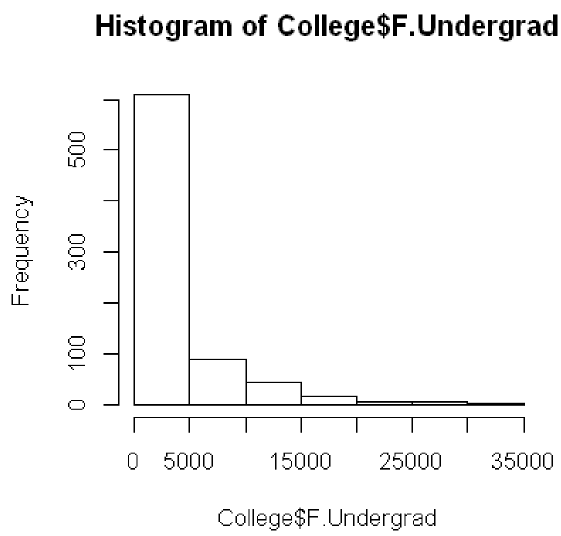


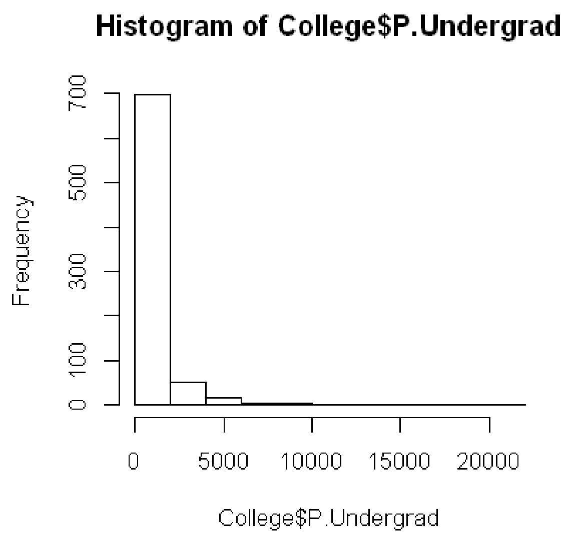
In [5]:

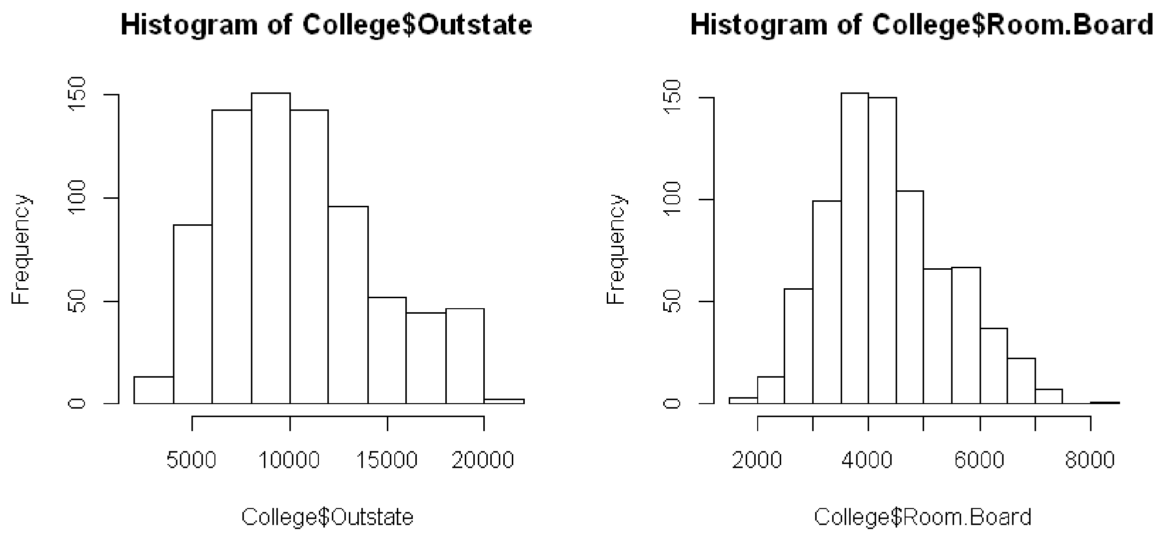
```
par(mfrow=c(2,2))
hist(College$Accept, breaks=10)
hist(College$Enroll, breaks=10)
hist(College$F.Undergrad, breaks=10)
hist(College$P.Undergrad, breaks=10)
hist(College$Outstate,breaks=10)
hist(College$Room.Board,breaks=10)

#From the histogram we can infer that the outstate and room.board requires scaling
```


Histogram of College$Accept
Frequency
0
100
200
300
400
500
0
5000
15000
25000
College$Accept


Histogram of College$Enroll
Frequency
0
100
200
300
400
0
1000
3000
5000
College$Enroll


Histogram of College$F.Undergrad
Frequency
0
100
300
500
0
5000
15000
25000
35000
College$F.Undergrad

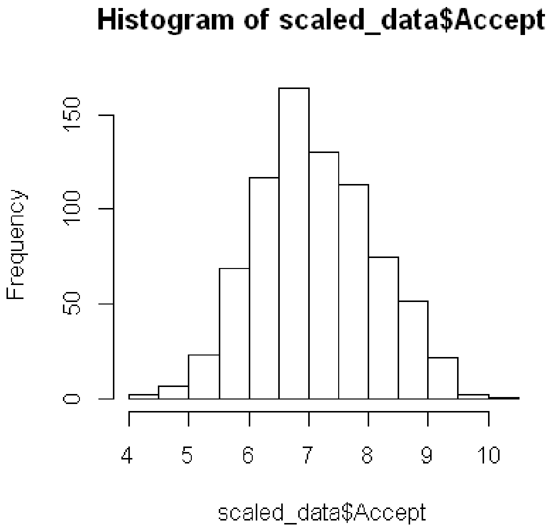

Histogram of College$P.Undergrad
Frequency
0
100
300
500
700
0
5000
10000
15000
20000
College$P.Undergrad

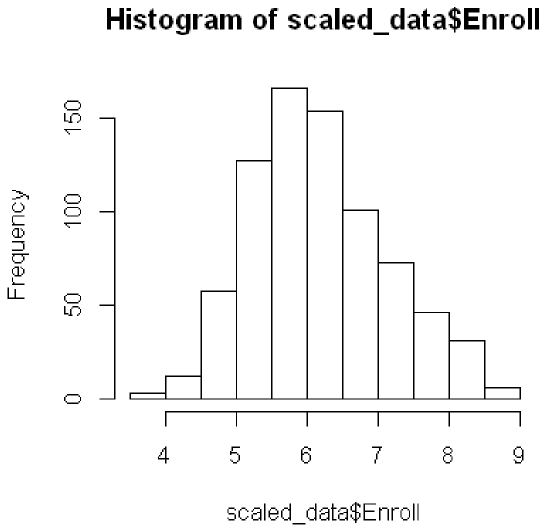


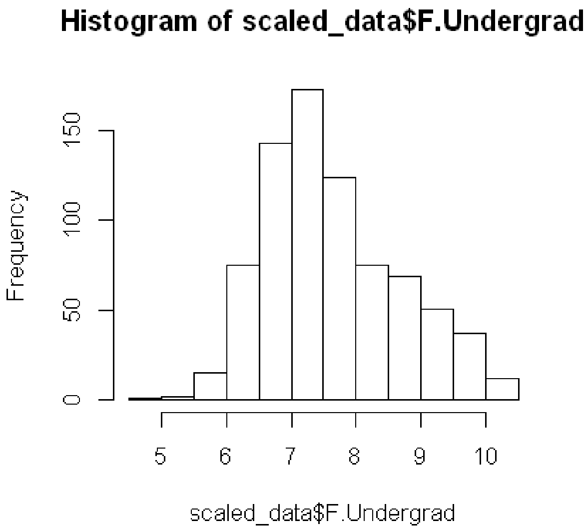
b) Scale the data appropriately (e.g., log transform) and present the visualizations in part A. Have any new relationships been revealed.

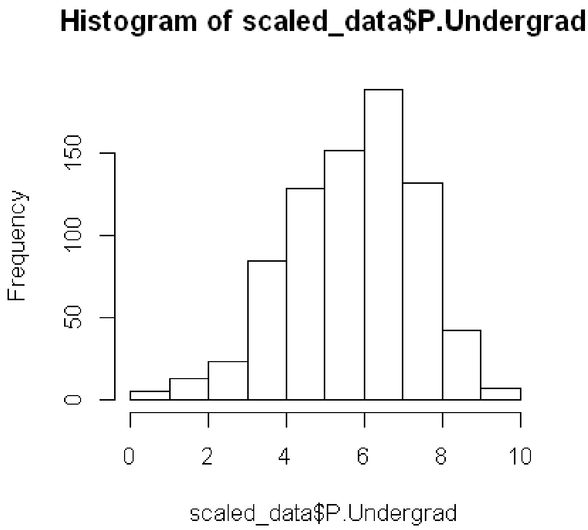
In [6]:
```
scaled_data <- College
scaled_data[, 2:18] <- log(scaled_data[, 2:18])
#scaled_data[, 2:18] <- scale(scaled_data[, 2:18])
```

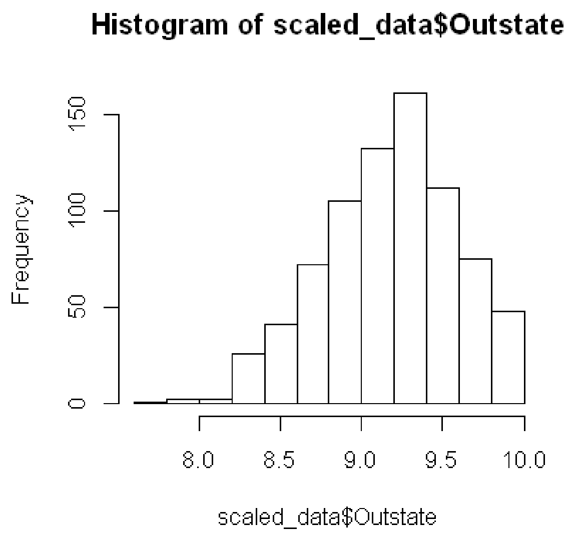
In [7]:
```
par(mfrow=c(2,2))
hist(scaled_data$Accept, breaks=10)
hist(scaled_data$Enroll, breaks=10)
hist(scaled_data$F.Undergrad, breaks=10)
hist(scaled_data$P.Undergrad, breaks=10)
hist(scaled_data$Outstate,breaks=10)
hist(scaled_data$Room.Board,breaks=10)

#So after log transformation we can see that now the new data is normally distribu
```

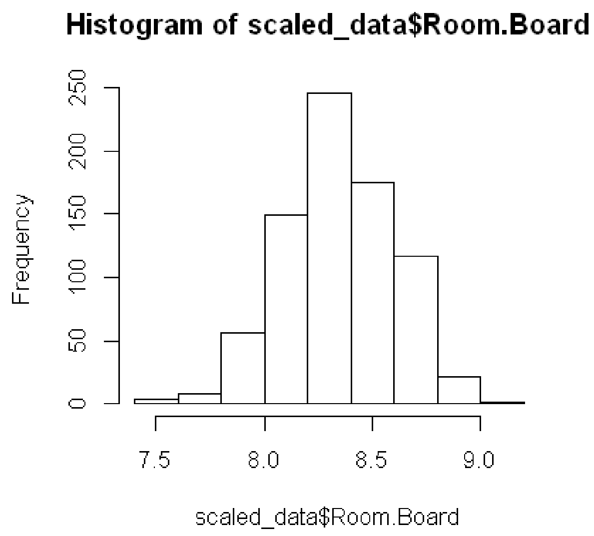

Histogram of scaled_data$Accept
Frequency
0
50
100
150
4
5
6
7
8
9
10
scaled_data$Accept


Histogram of scaled_data$Enroll
Frequency
0
50
100
150
4
5
6
7
8
9
scaled_data$Enroll


Histogram of scaled_data$F.Undergrad
Frequency
0
50
100
150
5
6
7
8
9
10
scaled_data$F.Undergrad


Histogram of scaled_data$P.Undergrad
Frequency
0
50
100
150
0
2
4
6
8
10
scaled_data$P.Undergrad

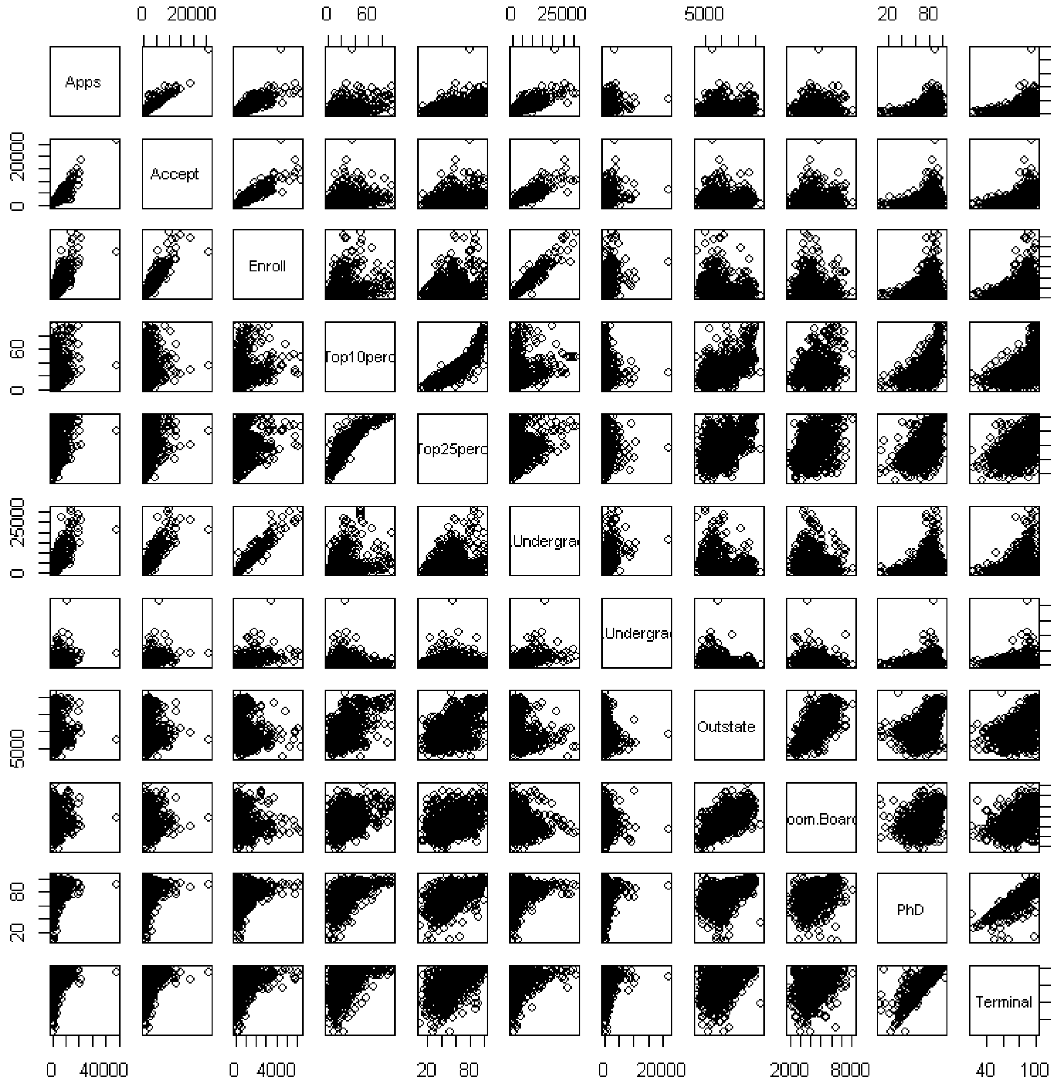




In [8]:

```
corrplot(cor(College[,2:11]), method = "number")

#After transforming the data, we can see new relations are built between the varial
```

In [9]:
```
pairs(College[,c(2:10,13,14)])

# In below graph we can see much of a linear relationship between the data after t
```



c) Subset the data into two data frames: “private” and “public”. Sort them alphabetically.

In [10]:
```
# Creating private dataframe
private_uni <- subset(College, College$Private == "Yes")
private_uni <- private_uni[order(private_uni$Private), ]
write.table(private_uni, "private_df.txt", sep = "\t", row.names = FALSE)
```

In [11]:
```
# Creating public dataframe
public_uni <- subset(College, College$Private == "No")
public_uni <- public_uni[order(public_uni$Private), ]
write.table(public_uni, "public_df.txt", sep = "\t", row.names = FALSE)
```

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”).

In [12]:
```
private_uni1 <- median(private_uni$Top25perc)
filtered_private <- subset(private_uni, private_uni$Top25perc >= private_uni1)
head(filtered_private)
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
Agnes Scott College	Yes	417	349	137	60	89	510	63	1.
Albertson College	Yes	587	479	158	38	62	678	41	1.
Albion College	Yes	1899	1720	489	37	68	1594	32	1.
Albright College	Yes	1038	839	227	30	63	973	306	1.
Alfred University	Yes	1732	1425	472	37	75	1830	110	1(
Allegheny College	Yes	2652	1900	484	44	77	1707	44	1.

In [13]:

```
public_uni1 <- median(public_uni$Top25perc)
filtered_public <- subset(public_uni, public_uni$Top25perc >= public_uni1)
head(filtered_public)
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	O
Angelo State University	No	3540	2001	1016	24	54	4190	1512	
Appalachian State University	No	7313	4664	1910	20	63	9940	1035	
Arkansas Tech University	No	1734	1729	951	12	52	3602	939	
Auburn University-Main Campus	No	7548	6791	3070	25	57	16262	1716	
Bloomsburg Univ. of Pennsylvania	No	6773	3028	1025	15	55	5847	946	
California Polytechnic-San Luis	No	7811	3817	1650	47	73	12911	1404	

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.

In [14]:

```
summary(private_uni)
summary(public_uni)
```

```
    Private          Apps           Accept          Enroll         Top10perc
 No :  0    Min.   :   81   Min.   :   72   Min.   : 35.0   Min.   : 1.00
 Yes:565    1st Qu.:  619   1st Qu.:  501   1st Qu.: 206.0   1st Qu.:17.00
            Median : 1133   Median :  859   Median : 328.0   Median :25.00
            Mean   : 1978   Mean   : 1306   Mean   : 456.9   Mean   :29.33
            3rd Qu.: 2186   3rd Qu.: 1580   3rd Qu.: 520.0   3rd Qu.:36.00
            Max.   :20192   Max.   :13007   Max.   :4615.0   Max.   :96.00
   Top25perc        F.Undergrad      P.Undergrad        Outstate
 Min.   :  9.00   Min.   :  139   Min.   :    1   Min.   : 2340
 1st Qu.: 42.00   1st Qu.:  840   1st Qu.:   63   1st Qu.: 9100
 Median : 55.00   Median : 1274   Median :  207   Median :11200
 Mean   : 56.96   Mean   : 1872   Mean   :  434   Mean   :11802
 3rd Qu.: 70.00   3rd Qu.: 2018   3rd Qu.:  541   3rd Qu.:13970
 Max.   :100.00   Max.   :27378   Max.   :10221   Max.   :21700
   Room.Board        Books           Personal          PhD
 Min.   :2370   Min.   : 250.0   Min.   : 250   Min.   :  8.00
 1st Qu.:3736   1st Qu.: 450.0   1st Qu.: 800   1st Qu.: 60.00
 Median :4400   Median : 500.0   Median :1100   Median : 73.00
 Mean   :4586   Mean   : 547.5   Mean   :1214   Mean   : 71.09
 3rd Qu.:5400   3rd Qu.: 600.0   3rd Qu.:1500   3rd Qu.: 85.00
 Max.   :8124   Max.   :2340.0   Max.   :6800   Max.   :100.00
    Terminal         S.F.Ratio       perc.alumni         Expend         Grad.Rate
 Min.   : 24.00   Min.   : 2.50   Min.   : 2.00   Min.   : 3186   Min.   : 15
 1st Qu.: 68.00   1st Qu.:11.10   1st Qu.:16.00   1st Qu.: 7477   1st Qu.: 58
 Median : 81.00   Median :12.70   Median :25.00   Median : 8954   Median : 69
 Mean   : 78.53   Mean   :12.95   Mean   :25.89   Mean   :10486   Mean   : 69
 3rd Qu.: 92.00   3rd Qu.:14.50   3rd Qu.:34.00   3rd Qu.:11625   3rd Qu.: 81
 Max.   :100.00   Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118
 Private          Apps           Accept          Enroll         Top10perc
 No :212    Min.   :  233   Min.   :  233   Min.   : 153.0   Min.   : 1.00
 Yes:  0    1st Qu.: 2191   1st Qu.: 1563   1st Qu.: 701.8   1st Qu.:12.00
            Median : 4307   Median : 2930   Median :1337.5   Median :19.00
            Mean   : 5730   Mean   : 3919   Mean   :1640.9   Mean   :22.83
            3rd Qu.: 7722   3rd Qu.: 5264   3rd Qu.:2243.8   3rd Qu.:27.50
            Max.   :48094   Max.   :26330   Max.   :6392.0   Max.   :95.00
   Top25perc       F.Undergrad      P.Undergrad        Outstate        Room.Board
 Min.   : 12.0   Min.   :  633   Min.   :    9   Min.   : 2580   Min.   :1780
 1st Qu.: 37.0   1st Qu.: 3601   1st Qu.:  600   1st Qu.: 5366   1st Qu.:3122
 Median : 51.0   Median : 6786   Median : 1375   Median : 6609   Median :3708
 Mean   : 52.7   Mean   : 8571   Mean   : 1978   Mean   : 6813   Mean   :3748
 3rd Qu.: 65.0   3rd Qu.:12507   3rd Qu.: 2495   3rd Qu.: 7844   3rd Qu.:4362
 Max.   :100.0   Max.   :31643   Max.   :21836   Max.   :15732   Max.   :6540
     Books           Personal          PhD            Terminal
 Min.   :  96.0   Min.   : 400   Min.   : 33.00   Min.   : 33.00
 1st Qu.: 500.0   1st Qu.:1200   1st Qu.: 71.00   1st Qu.: 76.00
 Median : 550.0   Median :1649   Median : 78.50   Median : 86.00
 Mean   : 554.4   Mean   :1677   Mean   : 76.83   Mean   : 82.82
 3rd Qu.: 612.0   3rd Qu.:2051   3rd Qu.: 86.00   3rd Qu.: 92.00
 Max.   :1125.0   Max.   :4288   Max.   :103.00   Max.   :100.00
   S.F.Ratio      perc.alumni        Expend         Grad.Rate
 Min.   : 6.70   Min.   : 0.00   Min.   : 3605   Min.   : 10.00
 1st Qu.:15.10   1st Qu.: 9.00   1st Qu.: 5715   1st Qu.: 46.00
 Median :17.25   Median :13.50   Median : 6716   Median : 55.00
 Mean   :17.14   Mean   :14.36   Mean   : 7458   Mean   : 56.04
 3rd Qu.:19.32   3rd Qu.:19.00   3rd Qu.: 8570   3rd Qu.: 65.00
 Max.   :28.80   Max.   :48.00   Max.   :16527   Max.   :100.00
```

In [15]:

```
filtered_private$Rate <- cut(filtered_private$Grad.Rate, c(0,58,81,118), labels=c(
head(filtered_private)
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outs
Agnes Scott College	Yes	417	349	137	60	89	510	63	1:
Albertson College	Yes	587	479	158	38	62	678	41	1:
Albion College	Yes	1899	1720	489	37	68	1594	32	1:
Albright College	Yes	1038	839	227	30	63	973	306	1!
Alfred University	Yes	1732	1425	472	37	75	1830	110	1(
Allegheny College	Yes	2652	1900	484	44	77	1707	44	1

In [16]:

```
filtered_public$Rate <- cut(filtered_public$Grad.Rate, c(0,46,65,100), labels=c('L
head(filtered_public)
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	O
Angelo State University	No	3540	2001	1016	24	54	4190	1512	
Appalachian State University	No	7313	4664	1910	20	63	9940	1035	
Arkansas Tech University	No	1734	1729	951	12	52	3602	939	
Auburn University-Main Campus	No	7548	6791	3070	25	57	16262	1716	
Bloomsburg Univ. of Pennsylvania	No	6773	3028	1025	15	55	5847	946	
California Polytechnic-San Luis	No	7811	3817	1650	47	73	12911	1404	

f) 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).

In [17]:

```
list_structure <- list(filtered_private, filtered_public)
save(list_structure, file="list_structure.RData")
```

Question 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.

In [18]:

```
# reading and checking the dimension of the data
data <- read.csv("Marketingdata.csv")
head(data)
dim(data)
```

ANNUAL_INCOME	SEX	MARITAL_STATUS	AGE	EDUCATION	OCCUPATION	YEARS_LIVED_IN_SAN
9	2	1	5	4	5	
9	1	1	5	5	5	
9	2	1	3	5	1	
1	2	5	1	2	6	
1	2	5	1	2	6	
8	1	1	6	4	8	

1. 8993
2. 14

In [19]:

```
# adding a target class =1 for training sample
train_sample<- data
train_sample$TARGET=1
dim(train_sample)
```

1. 8993
2. 15

In [20]:

```
# creating a reference sample of same size of training using permutation.
reference_sample = train_sample

for(i in 1:ncol(reference_sample)){
  reference_sample[,i] = sample(reference_sample[,i], nrow(reference_sample), repl
}

# adding a target class =0 for reference sample
reference_sample$TARGET = 0

# binding both the datasets.
combined_data = rbind(reference_sample, train_sample);

head(combined_data)
dim(combined_data)
```

ANNUAL_INCOME	SEX	MARITAL_STATUS	AGE	EDUCATION	OCCUPATION	YEARS_LIVED_IN_SAN
4	1	1	3	3	1	
1	2	2	5	4	1	
8	2	1	6	5	6	
5	2	3	3	3	5	
8	2	5	1	4	1	
2	1	5	2	3	1	

1. 17986
2. 15

In [21]:

```
#replacing na values with median of that column
head(combined_data %>% mutate(across(where(is.numeric), ~replace_na(., median(., na
```

ANNUAL_INCOME	SEX	MARITAL_STATUS	AGE	EDUCATION	OCCUPATION	YEARS_LIVED_IN_SAN
4	1	1	3	3	1	
1	2	2	5	4	1	
8	2	1	6	5	6	
5	2	3	3	3	5	
8	2	5	1	4	1	
2	1	5	2	3	1	

In [22]:

```
# Fit a classification tree to the combined data
#model.control <- rpart.control(minbucket = 2, minsplit = 100, xval=10, cp=0.02)
class_tree <- rpart(TARGET ~ ., data=combined_data, method="class")
summary(class_tree)
```

```
Call:
rpart(formula = TARGET ~ ., data = combined_data, method = "class")
  n= 17986

          CP nsplit rel error    xerror        xstd
1 0.02996775      0 1.0000000 1.0149005 0.007455632
2 0.02857778     12 0.5621039 0.6202602 0.006897922
3 0.01745802     13 0.5335261 0.5686645 0.006727156
4 0.01534527     14 0.5160681 0.5462026 0.006644488
5 0.01467808     16 0.4853775 0.5261870 0.006566349
6 0.01111976     17 0.4706994 0.5053931 0.006480535
7 0.01000000     18 0.4595797 0.4928278 0.006426311

Variable importance
        DUAL_INCOMES       MARITAL_STATUS                  AGE
                  31                   28                    9
  HOUSEHOLDER_STATUS           OCCUPATION         TYPE_OF_HOME
                   8                    8                    7
PERSONS_IN_HOUSEHOLD        ANNUAL_INCOME      PERSON_UNDER_18
                   5                    3                    1

Node number 1: 17986 observations,    complexity param=0.02996775
  predicted class=0  expected loss=0.5  P(node) =1
    class counts:  8993  8993
   probabilities: 0.500 0.500
  left son=2 (17572 obs) right son=3 (414 obs)
  Primary splits:
      ETHNICITY          < 7.5 to the left,  improve=1.5961990, (139 missing)
      AGE                < 6.5 to the left,  improve=1.2934330, (0 missing)
      OCCUPATION         < 3.5 to the left,  improve=1.1066240, (274 missing)
      EDUCATION          < 4.5 to the left,  improve=0.8662616, (163 missing)
      HOUSEHOLDER_STATUS < 2.5 to the right, improve=0.3457240, (489 missing)

Node number 2: 17572 observations,    complexity param=0.02996775
  predicted class=0  expected loss=0.4989756  P(node) =0.9769821
    class counts:  8804  8768
   probabilities: 0.501 0.499
  left son=4 (16516 obs) right son=5 (1056 obs)
  Primary splits:
      AGE                  < 6.5 to the left,  improve=2.0739470, (0 missing)
      OCCUPATION           < 3.5 to the left,  improve=1.4648770, (261 missing)
      EDUCATION            < 4.5 to the left,  improve=0.9365601, (162 missing)
      DUAL_INCOMES         < 2.5 to the left,  improve=0.5279850, (0 missing)
      PERSONS_IN_HOUSEHOLD < 3.5 to the right, improve=0.4766903, (741 missing)

Node number 3: 414 observations
  predicted class=1  expected loss=0.4565217  P(node) =0.0230179
    class counts:   189   225
   probabilities: 0.457 0.543

Node number 4: 16516 observations,    complexity param=0.02996775
  predicted class=0  expected loss=0.4970332  P(node) =0.9182698
    class counts:  8307  8209
   probabilities: 0.503 0.497
  left son=8 (1515 obs) right son=9 (15001 obs)
  Primary splits:
      OCCUPATION           < 7.5 to the right, improve=36.858670, (250 missing)
      DUAL_INCOMES         < 2.5 to the right, improve= 7.299156, (0 missing)
      PERSON_UNDER_18      < 0.5 to the left,  improve= 6.943398, (0 missing)
      HOUSEHOLDER_STATUS   < 1.5 to the left,  improve= 5.850329, (446 missing)
      PERSONS_IN_HOUSEHOLD < 2.5 to the left,  improve= 3.935403, (684 missing)

Node number 5: 1056 observations,    complexity param=0.02996775
  predicted class=1  expected loss=0.4706439  P(node) =0.05871233
```

```
      class counts:   497   559
     probabilities: 0.471 0.529 
    left son=10 (573 obs) right son=11 (483 obs)
    Primary splits:
        OCCUPATION          < 7.5 to the left,  improve=217.77410, (11 missing)
        PERSONS_IN_HOUSEHOLD < 2.5 to the right, improve=112.09110, (57 missing)
        MARITAL_STATUS       < 4.5 to the right, improve=111.12600, (16 missing)
        PERSON_UNDER_18      < 0.5 to the right, improve= 94.00331, (0 missing)
        HOUSEHOLDER_STATUS   < 1.5 to the right, improve= 89.47838, (37 missing)
    Surrogate splits:
        DUAL_INCOMES         < 2.5 to the left,  agree=0.641, adj=0.215, (11 split)
        HOUSEHOLDER_STATUS   < 1.5 to the right, agree=0.616, adj=0.161, (0 split)
        PERSON_UNDER_18      < 0.5 to the right, agree=0.590, adj=0.105, (0 split)
        MARITAL_STATUS       < 4.5 to the right, agree=0.589, adj=0.100, (0 split)
        PERSONS_IN_HOUSEHOLD < 2.5 to the right, agree=0.589, adj=0.100, (0 split)

Node number 8: 1515 observations,    complexity param=0.01745802
  predicted class=0  expected loss=0.3920792  P(node) =0.08423218
    class counts:   921   594
   probabilities: 0.608 0.392 
  left son=16 (1226 obs) right son=17 (289 obs)
  Primary splits:
      AGE             < 5.5 to the left,  improve=102.891900, (0 missing)
      OCCUPATION      < 8.5 to the left,  improve= 34.715580, (0 missing)
      DUAL_INCOMES    < 2.5 to the left,  improve= 32.600100, (0 missing)
      ANNUAL_INCOME   < 2.5 to the right, improve= 16.643610, (0 missing)
      PERSON_UNDER_18 < 0.5 to the right, improve=  7.828121, (0 missing)

Node number 9: 15001 observations,    complexity param=0.02996775
  predicted class=1  expected loss=0.4923672  P(node) =0.8340376
    class counts:  7386  7615
   probabilities: 0.492 0.508 
  left son=18 (1938 obs) right son=19 (13063 obs)
  Primary splits:
      DUAL_INCOMES         < 2.5 to the right, improve=21.852670, (0 missing)
      AGE                  < 5.5 to the right, improve=13.798330, (0 missing)
      PERSON_UNDER_18      < 0.5 to the left,  improve=11.422050, (0 missing)
      HOUSEHOLDER_STATUS   < 1.5 to the left,  improve= 9.986457, (399 missing)
      PERSONS_IN_HOUSEHOLD < 2.5 to the left,  improve= 6.364165, (615 missing)

Node number 10: 573 observations
  predicted class=0  expected loss=0.2338569  P(node) =0.03185811
    class counts:   439   134
   probabilities: 0.766 0.234 

Node number 11: 483 observations
  predicted class=1  expected loss=0.1200828  P(node) =0.02685422
    class counts:    58   425
   probabilities: 0.120 0.880 

Node number 16: 1226 observations
  predicted class=0  expected loss=0.3026101  P(node) =0.06816413
    class counts:   855   371
   probabilities: 0.697 0.303 

Node number 17: 289 observations
  predicted class=1  expected loss=0.2283737  P(node) =0.01606805
    class counts:    66   223
   probabilities: 0.228 0.772 

Node number 18: 1938 observations,    complexity param=0.02996775
  predicted class=0  expected loss=0.4375645  P(node) =0.1077505
    class counts:  1090   848
   probabilities: 0.562 0.438 
```

```
    left son=36 (608 obs) right son=37 (1330 obs)
    Primary splits:
        MARITAL_STATUS       < 2.5 to the right, improve=327.02830, (52 missing)
        AGE                  < 2.5 to the left,  improve= 83.68285, (0 missing)
        HOUSEHOLDER_STATUS   < 1.5 to the right, improve= 76.10106, (39 missing)
        PERSONS_IN_HOUSEHOLD < 1.5 to the left,  improve= 64.55316, (69 missing)
        OCCUPATION           < 4.5 to the left,  improve= 58.65023, (30 missing)
    Surrogate splits:
        AGE < 1.5 to the left,  agree=0.691, adj=0.027, (52 split)

Node number 19: 13063 observations,    complexity param=0.02996775
  predicted class=1  expected loss=0.481972  P(node) =0.7262871
    class counts:  6296  6767
   probabilities: 0.482 0.518
  left son=38 (4364 obs) right son=39 (8699 obs)
  Primary splits:
      HOUSEHOLDER_STATUS                     < 1.5 to the left,  improve=42.17405
0, (360 missing)
      MARITAL_STATUS                         < 4.5 to the left,  improve=40.64717
0, (205 missing)
      AGE                                    < 5.5 to the right, improve=33.84448
0, (0 missing)
      OCCUPATION                             < 4.5 to the right, improve= 6.48047
9, (220 missing)
      YEARS_LIVED_IN_SAN.FRAN_OAKLAND_SANJOSE < 4.5 to the right, improve= 6.18222
4, (1251 missing)
  Surrogate splits:
      ANNUAL_INCOME   < 7.5 to the right, agree=0.671, adj=0.027, (360 split)
      AGE             < 4.5 to the right, agree=0.665, adj=0.010, (0 split)
      PERSON_UNDER_18 < 7.5 to the right, agree=0.662, adj=0.001, (0 split)

Node number 36: 608 observations
  predicted class=0  expected loss=0.01315789  P(node) =0.03380407
    class counts:   600     8
   probabilities: 0.987 0.013

Node number 37: 1330 observations,    complexity param=0.01111976
  predicted class=1  expected loss=0.3684211  P(node) =0.0739464
    class counts:   490   840
   probabilities: 0.368 0.632
  left son=74 (152 obs) right son=75 (1178 obs)
  Primary splits:
      HOUSEHOLDER_STATUS   < 2.5 to the right, improve=72.19567, (25 missing)
      AGE                  < 2.5 to the left,  improve=58.91398, (0 missing)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  improve=57.18715, (40 missing)
      OCCUPATION           < 4.5 to the left,  improve=34.92526, (16 missing)
      ANNUAL_INCOME        < 1.5 to the left,  improve=30.07658, (0 missing)

Node number 38: 4364 observations,    complexity param=0.02996775
  predicted class=0  expected loss=0.4608158  P(node) =0.2426332
    class counts:  2353  2011
   probabilities: 0.539 0.461
  left son=76 (873 obs) right son=77 (3491 obs)
  Primary splits:
      TYPE_OF_HOME   < 2.5 to the right, improve=252.2238, (142 missing)
      DUAL_INCOMES   < 1.5 to the left,  improve=249.2753, (0 missing)
      ANNUAL_INCOME  < 4.5 to the left,  improve=236.7061, (0 missing)
      AGE            < 2.5 to the left,  improve=197.5281, (0 missing)
      MARITAL_STATUS < 4.5 to the right, improve=152.2977, (70 missing)
  Surrogate splits:
      PERSON_UNDER_18 < 7.5 to the right, agree=0.793, adj=0.001, (142 split)

Node number 39: 8699 observations,    complexity param=0.02996775
  predicted class=1  expected loss=0.4532705  P(node) =0.483654
```

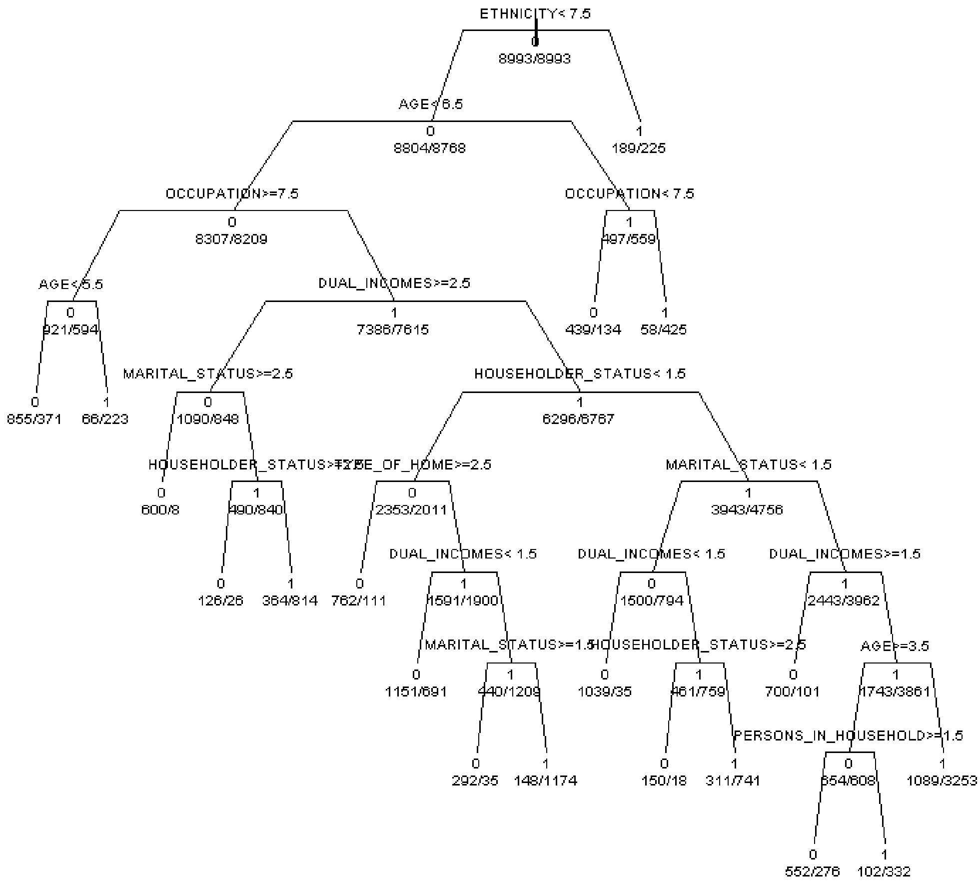
```
    class counts:  3943  4756
   probabilities: 0.453 0.547 
  left son=78 (2294 obs) right son=79 (6405 obs)
  Primary splits:
      MARITAL_STATUS < 1.5 to the left,  improve=255.05540, (135 missing)
      AGE            < 3.5 to the right, improve=203.58850, (0 missing)
      TYPE_OF_HOME   < 2.5 to the left,  improve= 97.97047, (368 missing)
      ANNUAL_INCOME  < 6.5 to the right, improve= 88.55847, (0 missing)
      DUAL_INCOMES   < 1.5 to the right, improve= 77.34054, (0 missing)
  Surrogate splits:
      DUAL_INCOMES < 1.5 to the right, agree=0.781, adj=0.175, (135 split)

Node number 74: 152 observations
  predicted class=0  expected loss=0.1710526  P(node) =0.008451017
    class counts:   126    26
   probabilities: 0.829 0.171 

Node number 75: 1178 observations
  predicted class=1  expected loss=0.3089983  P(node) =0.06549539
    class counts:   364   814
   probabilities: 0.309 0.691 

Node number 76: 873 observations
  predicted class=0  expected loss=0.1271478  P(node) =0.04853775
    class counts:   762   111
   probabilities: 0.873 0.127 

Node number 77: 3491 observations,    complexity param=0.02996775
  predicted class=1  expected loss=0.4557433  P(node) =0.1940954
    class counts:  1591  1900
   probabilities: 0.456 0.544 
  left son=154 (1842 obs) right son=155 (1649 obs)
  Primary splits:
      DUAL_INCOMES   < 1.5 to the left,  improve=223.0712, (0 missing)
      ANNUAL_INCOME  < 4.5 to the left,  improve=213.8041, (0 missing)
      AGE            < 2.5 to the left,  improve=170.9269, (0 missing)
      MARITAL_STATUS < 4.5 to the right, improve=145.3262, (54 missing)
      OCCUPATION     < 4.5 to the right, improve=126.1426, (48 missing)
  Surrogate splits:
      MARITAL_STATUS       < 1.5 to the right, agree=0.768, adj=0.508, (0 split)
      ANNUAL_INCOME        < 7.5 to the left,  agree=0.635, adj=0.227, (0 split)
      AGE                  < 3.5 to the left,  agree=0.593, adj=0.138, (0 split)
      PERSON_UNDER_18      < 0.5 to the left,  agree=0.578, adj=0.107, (0 split)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  agree=0.564, adj=0.078, (0 split)

Node number 78: 2294 observations,    complexity param=0.02996775
  predicted class=0  expected loss=0.3461203  P(node) =0.1275436
    class counts:  1500   794
   probabilities: 0.654 0.346 
  left son=156 (1074 obs) right son=157 (1220 obs)
  Primary splits:
      DUAL_INCOMES         < 1.5 to the left,  improve=397.03720, (0 missing)
      HOUSEHOLDER_STATUS   < 2.5 to the right, improve=129.33110, (86 missing)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  improve= 60.57195, (77 missing)
      ANNUAL_INCOME        < 1.5 to the left,  improve= 42.59181, (0 missing)
      TYPE_OF_HOME         < 2.5 to the left,  improve= 37.05562, (91 missing)
  Surrogate splits:
      HOUSEHOLDER_STATUS   < 2.5 to the right, agree=0.592, adj=0.129, (0 split)
      ANNUAL_INCOME        < 2.5 to the left,  agree=0.578, adj=0.099, (0 split)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  agree=0.565, adj=0.071, (0 split)
      OCCUPATION           < 4.5 to the right, agree=0.561, adj=0.063, (0 split)
      AGE                  < 1.5 to the left,  agree=0.557, adj=0.054, (0 split)

Node number 79: 6405 observations,    complexity param=0.02996775
```

```
  predicted class=1  expected loss=0.3814208  P(node) =0.3561103
    class counts:  2443  3962
   probabilities: 0.381 0.619
  left son=158 (801 obs) right son=159 (5604 obs)
  Primary splits:
      DUAL_INCOMES  < 1.5 to the right, improve=444.09180, (0 missing)
      AGE           < 3.5 to the right, improve=177.45430, (0 missing)
      ANNUAL_INCOME < 5.5 to the right, improve=120.30610, (0 missing)
      OCCUPATION    < 5.5 to the left,  improve= 79.95604, (115 missing)
      TYPE_OF_HOME  < 2.5 to the left,  improve= 57.47363, (277 missing)

Node number 154: 1842 observations
  predicted class=0  expected loss=0.3751357  P(node) =0.102413
    class counts:  1151   691
   probabilities: 0.625 0.375

Node number 155: 1649 observations,    complexity param=0.02857778
  predicted class=1  expected loss=0.2668284  P(node) =0.09168242
    class counts:   440  1209
   probabilities: 0.267 0.733
  left son=310 (327 obs) right son=311 (1322 obs)
  Primary splits:
      MARITAL_STATUS       < 1.5 to the right, improve=323.24350, (27 missing)
      ANNUAL_INCOME        < 3.5 to the left,  improve=142.72090, (0 missing)
      AGE                  < 2.5 to the left,  improve=132.34400, (0 missing)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  improve= 83.05147, (30 missing)
      OCCUPATION           < 4.5 to the right, improve= 79.55533, (22 missing)
  Surrogate splits:
      ANNUAL_INCOME        < 3.5 to the left,  agree=0.826, adj=0.127, (27 split)
      AGE                  < 2.5 to the left,  agree=0.823, adj=0.114, (0 split)
      OCCUPATION           < 5.5 to the right, agree=0.808, adj=0.040, (0 split)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  agree=0.803, adj=0.015, (0 split)

Node number 156: 1074 observations
  predicted class=0  expected loss=0.03258845  P(node) =0.05971311
    class counts:  1039    35
   probabilities: 0.967 0.033

Node number 157: 1220 observations,    complexity param=0.01467808
  predicted class=1  expected loss=0.3778689  P(node) =0.06783053
    class counts:   461   759
   probabilities: 0.378 0.622
  left son=314 (168 obs) right son=315 (1052 obs)
  Primary splits:
      HOUSEHOLDER_STATUS   < 2.5 to the right, improve=106.51050, (44 missing)
      PERSONS_IN_HOUSEHOLD < 1.5 to the left,  improve= 60.56016, (35 missing)
      AGE                  < 1.5 to the left,  improve= 34.46271, (0 missing)
      ANNUAL_INCOME        < 1.5 to the left,  improve= 29.36774, (0 missing)
      OCCUPATION           < 4.5 to the right, improve= 25.36109, (24 missing)

Node number 158: 801 observations
  predicted class=0  expected loss=0.1260924  P(node) =0.04453464
    class counts:   700   101
   probabilities: 0.874 0.126

Node number 159: 5604 observations,    complexity param=0.01534527
  predicted class=1  expected loss=0.3110278  P(node) =0.3115757
    class counts:  1743  3861
   probabilities: 0.311 0.689
  left son=318 (1262 obs) right son=319 (4342 obs)
  Primary splits:
      AGE            < 3.5 to the right, improve=139.85100, (0 missing)
      ANNUAL_INCOME  < 5.5 to the right, improve=100.19930, (0 missing)
      OCCUPATION     < 5.5 to the left,  improve= 60.80390, (104 missing)
```

```
      MARITAL_STATUS < 4.5 to the left,  improve= 34.90784, (113 missing)
      TYPE_OF_HOME   < 2.5 to the left,  improve= 33.98125, (249 missing)

Node number 310: 327 observations
  predicted class=0  expected loss=0.1070336  P(node) =0.01818081
    class counts:   292    35
   probabilities: 0.893 0.107

Node number 311: 1322 observations
  predicted class=1  expected loss=0.1119516  P(node) =0.07350161
    class counts:   148  1174
   probabilities: 0.112 0.888

Node number 314: 168 observations
  predicted class=0  expected loss=0.1071429  P(node) =0.009340598
    class counts:   150    18
   probabilities: 0.893 0.107

Node number 315: 1052 observations
  predicted class=1  expected loss=0.2956274  P(node) =0.05848994
    class counts:   311   741
   probabilities: 0.296 0.704

Node number 318: 1262 observations,    complexity param=0.01534527
  predicted class=0  expected loss=0.481775  P(node) =0.07016568
    class counts:   654   608
   probabilities: 0.518 0.482
  left son=636 (828 obs) right son=637 (434 obs)
  Primary splits:
      PERSONS_IN_HOUSEHOLD < 1.5 to the right, improve=110.17440, (68 missing)
      TYPE_OF_HOME         < 1.5 to the left,  improve= 84.11689, (59 missing)
      HOUSEHOLDER_STATUS   < 2.5 to the right, improve= 68.41283, (39 missing)
      MARITAL_STATUS       < 4.5 to the right, improve= 54.35737, (36 missing)
      PERSON_UNDER_18      < 0.5 to the right, improve= 37.18716, (0 missing)

Node number 319: 4342 observations
  predicted class=1  expected loss=0.2508061  P(node) =0.24141
    class counts:  1089  3253
   probabilities: 0.251 0.749

Node number 636: 828 observations
  predicted class=0  expected loss=0.3333333  P(node) =0.04603581
    class counts:   552   276
   probabilities: 0.667 0.333

Node number 637: 434 observations
  predicted class=1  expected loss=0.235023  P(node) =0.02412988
    class counts:   102   332
   probabilities: 0.235 0.765
```

In [23]:

```
# Plot the tree
plot(class_tree, uniform = T, compress = T, branch=0.7)
text(class_tree, cex=0.5, use.n=T, all=T)
```



In [24]:

```
predicted <- predict(class_tree, type = "prob")
df = data.frame(Actual_Target  = combined_data$TARGET, target = predicted)
df$Predicted_Target <- with(df, ifelse(target.0 > target.1, '0', '1'))
```

In [25]:

```
predicted <- predict(class_tree,combined_data[, -c(15)],type = "prob")
predicted
```

0	1
0.9674115	0.03258845
0.6666667	0.33333333
0.2956274	0.70437262
0.8739076	0.12609238
0.9868421	0.01315789
0.2508061	0.74919392
0.2956274	0.70437262
0.6973899	0.30261011
0.6973899	0.30261011
0.6973899	0.30261011
0.8739076	0.12609238
0.6973899	0.30261011
0.4565217	0.54347826
0.7661431	0.23385689
0.2508061	0.74919392
0.2508061	0.74919392
0.2508061	0.74919392
0.8739076	0.12609238
0.8739076	0.12609238
0.6973899	0.30261011
0.8928571	0.10714286
0.9868421	0.01315789
0.6248643	0.37513572
0.6666667	0.33333333
0.2508061	0.74919392
0.6973899	0.30261011
0.6973899	0.30261011
0.8739076	0.12609238
0.6973899	0.30261011
0.9868421	0.01315789
...	...
0.1119516	0.8880484
0.6248643	0.3751357
0.1119516	0.8880484
0.2508061	0.7491939
0.2508061	0.7491939

0	1
0.2508061	0.7491939
0.2508061	0.7491939
0.1200828	0.8799172
0.2508061	0.7491939
0.2508061	0.7491939
0.2508061	0.7491939
0.2508061	0.7491939
0.6248643	0.3751357
0.2508061	0.7491939
0.2508061	0.7491939
0.6248643	0.3751357
0.3089983	0.6910017
0.2508061	0.7491939
0.2508061	0.7491939
0.2508061	0.7491939
0.2508061	0.7491939
0.4565217	0.5434783
0.2508061	0.7491939
0.4565217	0.5434783
0.1119516	0.8880484
0.2508061	0.7491939
0.2508061	0.7491939
0.2508061	0.7491939
0.2956274	0.7043726
0.2508061	0.7491939

In [26]:
```
drop <- c("target.0","target.1")
head(df[ , !(names(df) %in% drop)])
```

Actual_Target	Predicted_Target
0	0
0	0
0	1
0	0
0	0
0	1

In [27]:
```
# combine predicted probabilities with original data and rename the class columns.
predicted_probability <- cbind(combined_data,predicted)
```

```
colnames(predicted_probability)[16] <- "Class_0"
colnames(predicted_probability)[17] <- "Class_1"
```

In [28]:
```
# To find just the terminal node we will remove the duplicates.
terminal_nodes <- predicted_probability %>% distinct(Class_1, .keep_all = TRUE)
terminal_nodes
```

	ANNUAL_INCOME	SEX	MARITAL_STATUS	AGE	EDUCATION	OCCUPATION	YEARS_LIVED_IN
1	4	1	1	3	3	1	
2	1	2	2	5	4	1	
3	8	2	1	6	5	6	
4	5	2	3	3	3	5	
5	8	2	5	1	4	1	
6	2	1	5	2	3	1	
8	8	2	5	2	6	8	
13	6	1	1	2	3	6	
14	7	2	2	7	3	2	
21	6	1	1	2	5	7	
23	3	2	1	2	6	1	
35	2	1	1	3	5	1	
62	8	2	2	4	3	2	
70	1	1	4	3	3	2	
79	4	2	1	4	4	1	
82	9	2	1	5	2	4	
112	4	2	5	6	5	8	
114	5	2	1	6	1	1	
145	8	2	5	7	4	9	

In [29]:
```
# Ordering the terminal nodes in descending order.
highest_node = terminal_nodes[order(-terminal_nodes$Class_1), ]
head(highest_node)
```

	ANNUAL_INCOME	SEX	MARITAL_STATUS	AGE	EDUCATION	OCCUPATION	YEARS_LIVED_IN
79	4	2	1	4	4	1	
145	8	2	5	7	4	9	
112	4	2	5	6	5	8	
62	8	2	2	4	3	2	
6	2	1	5	2	3	1	
3	8	2	1	6	5	6	

Question 3 - Consider the Boston Housing Data in the ISLR2 package. (Important – do not use data from any other packages).

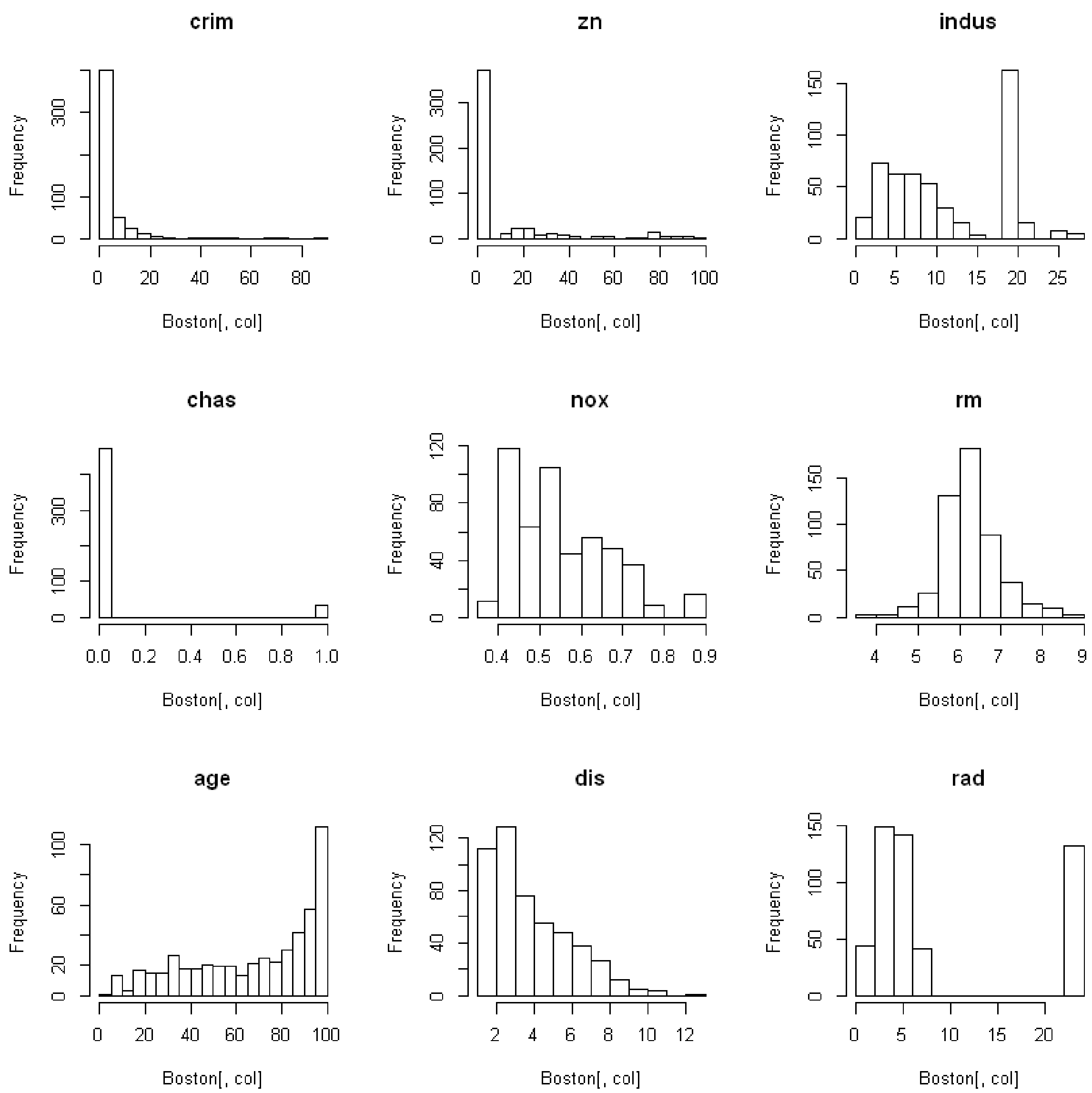
a. Visualize the data using histograms of the different variables in the data set. Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.

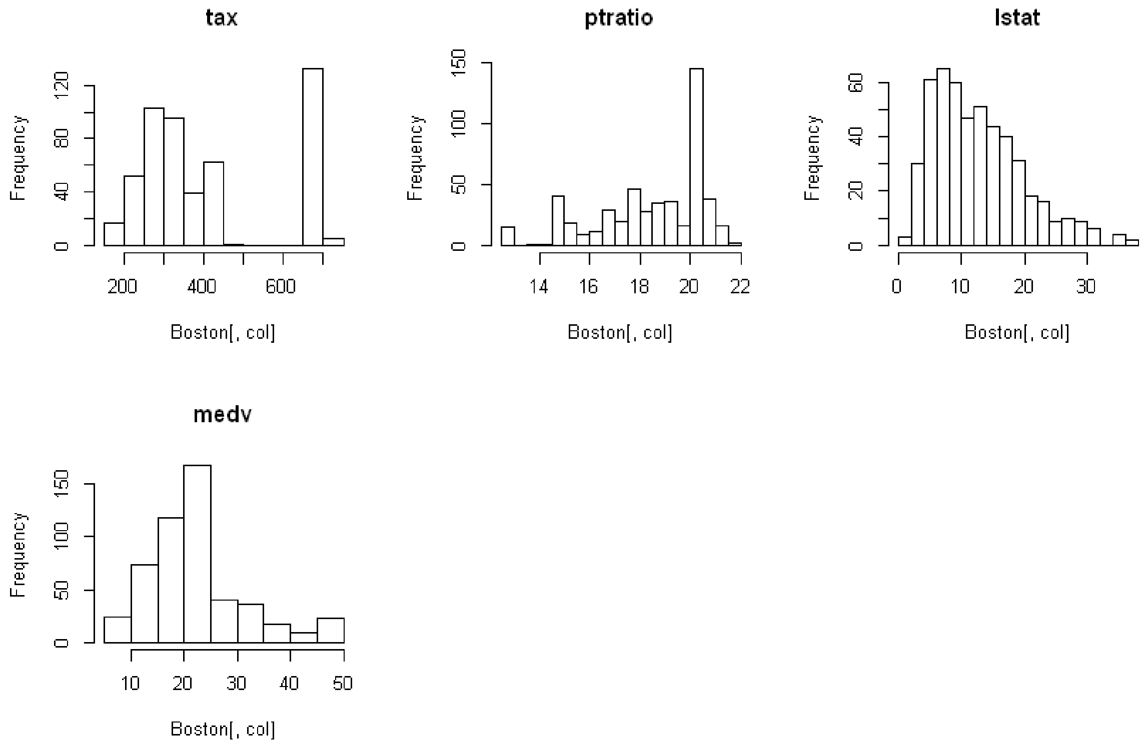
In [30]:
```
data(Boston)
head(Boston)
dim(Boston)
```

crim	zn	indus	chas	nox	rm	age	dis	rad	tax	ptratio	lstat	medv
0.00632	18	2.31	0	0.538	6.575	65.2	4.0900	1	296	15.3	4.98	24.0
0.02731	0	7.07	0	0.469	6.421	78.9	4.9671	2	242	17.8	9.14	21.6
0.02729	0	7.07	0	0.469	7.185	61.1	4.9671	2	242	17.8	4.03	34.7
0.03237	0	2.18	0	0.458	6.998	45.8	6.0622	3	222	18.7	2.94	33.4
0.06905	0	2.18	0	0.458	7.147	54.2	6.0622	3	222	18.7	5.33	36.2
0.02985	0	2.18	0	0.458	6.430	58.7	6.0622	3	222	18.7	5.21	28.7

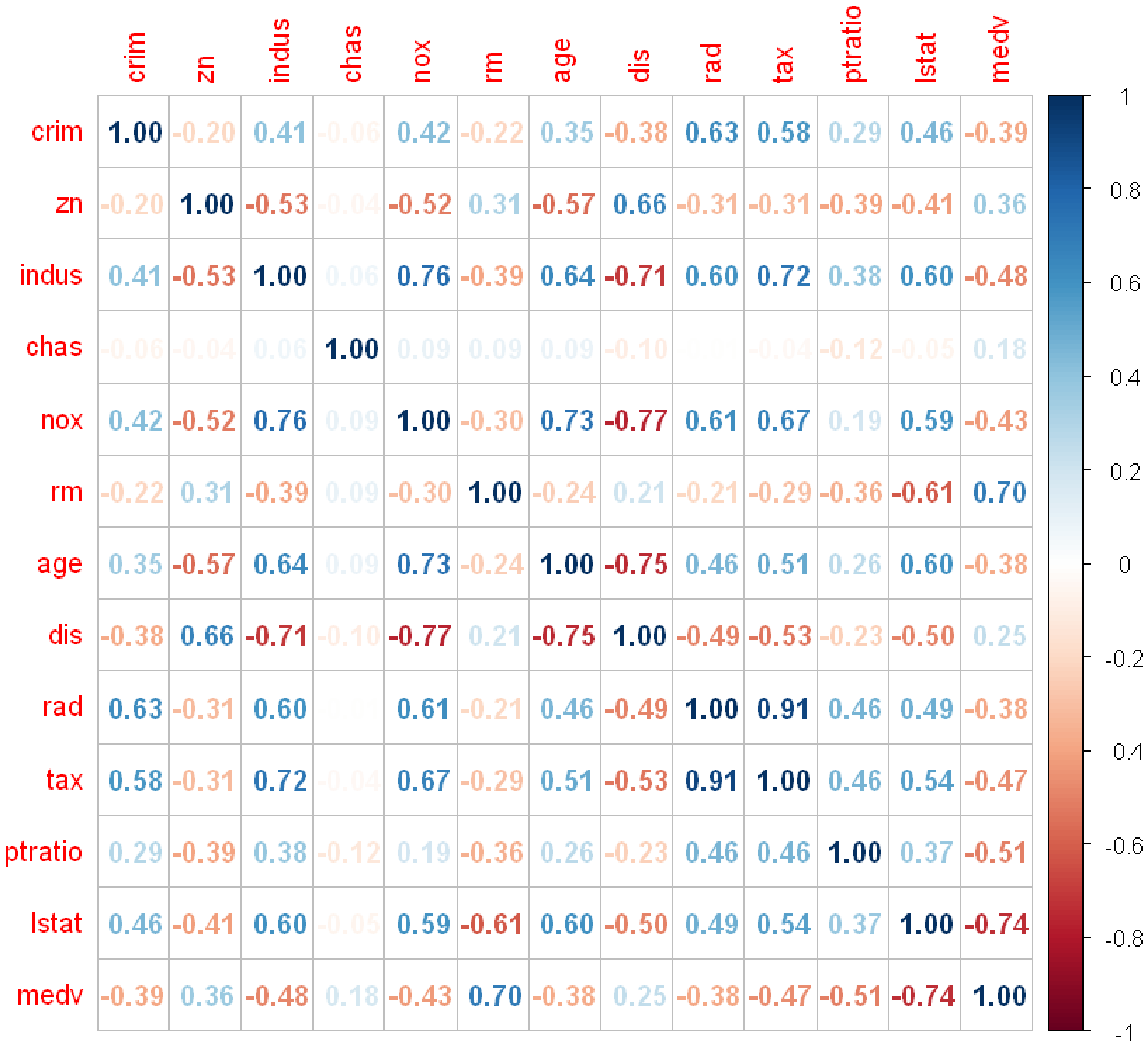
1. 506
2. 13

In [31]:
```
# Histograms of all the columns [a]
par(mfrow=c(3,3))
for (col in 1:ncol(Boston)) {
  hist(Boston[,col], main=colnames(Boston)[col], breaks=15)
}
```


crim
Frequency
0 100 300
0 20 40 60 80
Boston[, col]
zn
Frequency
0 100 200 300
0 20 40 60 80 100
Boston[, col]
indus
Frequency
0 50 100 150
0 5 10 15 20 25
Boston[, col]
chas
Frequency
0 100 300
0.0 0.2 0.4 0.6 0.8 1.0
Boston[, col]
nox
Frequency
0 40 80 120
0.4 0.5 0.6 0.7 0.8 0.9
Boston[, col]
rm
Frequency
0 50 100 150
4 5 6 7 8 9
Boston[, col]
age
Frequency
0 20 60 100
0 20 40 60 80 100
Boston[, col]
dis
Frequency
0 40 80 120
2 4 6 8 10 12
Boston[, col]
rad
Frequency
0 50 100 150
0 5 10 15 20
Boston[, col]



In [32]:
```
corrplot(cor(Boston), method = "number")
```



In [33]:
```
summary(Boston)
```

```
      crim                zn              indus           chas        
 Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
 1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
 Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
 Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
 3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
 Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
      nox               rm             age              dis        
 Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
 1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
 Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
 Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
 3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
 Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
      rad              tax           ptratio          lstat      
 Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
 1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
 Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
 Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
 3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
 Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
      medv      
 Min.   : 5.00  
 1st Qu.:17.02  
 Median :21.20  
 Mean   :22.53  
 3rd Qu.:25.00  
 Max.   :50.00  
```

In [34]:

```
#based on mean
Boston$crim <- ordered(cut(Boston$crim, c(0, 3.61, 90), labels=c('Safe', 'Unsafe')

#based on mean
Boston$zn <- ordered(cut(Boston$zn, c(0,11.36, 101), labels=c('Low', 'High')))

# based on quartile range
Boston$indus <- ordered(cut(Boston$indus,  c(0,5.19, 18.10, 30), labels=c('Low', '

# chas = 1 if tract bounds river; 0 otherwise.
Boston$chas <- ordered(cut(Boston$chas, c(0, 0.5, 1), labels=c('Unbounds', 'Tract_

# based on quartile range
Boston$nox <- ordered(cut(Boston$nox, c(0, 0.4490, 0.6240, 0.9), labels=c('Low', '

# based on quartile range
Boston$rm <- ordered(cut(Boston$rm, c(0, 5.886, 6.623, 9), labels=c('Less', 'Suffi

# based on mean
Boston$age<- ordered(cut(Boston$age, c(0, 25, 65 ,100), labels=c('Young', 'Middle-

# based on 3rd quartile range
Boston$dis <- ordered(cut(Boston$dis, c(0, 6, 12.127), labels=c('Close', 'Far')))

# Based on 1st quartile range
Boston$rad <- ordered(cut(Boston$rad, c(0, 4, 25), labels=c('Near', 'Far')))

# Based on quartile range
Boston$tax <- ordered(cut(Boston$tax, c(0, 280, 380, 712), labels=c('Low','Medium'

# Based on quartile range
Boston$ptratio <- ordered(cut(Boston$ptratio, c(0, 17.40, 20.20, 22.00), labels=c(

# Based on quartile range
Boston$lstat <- ordered(cut(Boston$lstat, c(0, 6.95, 16.95, 37.97), labels=c('Low'
```

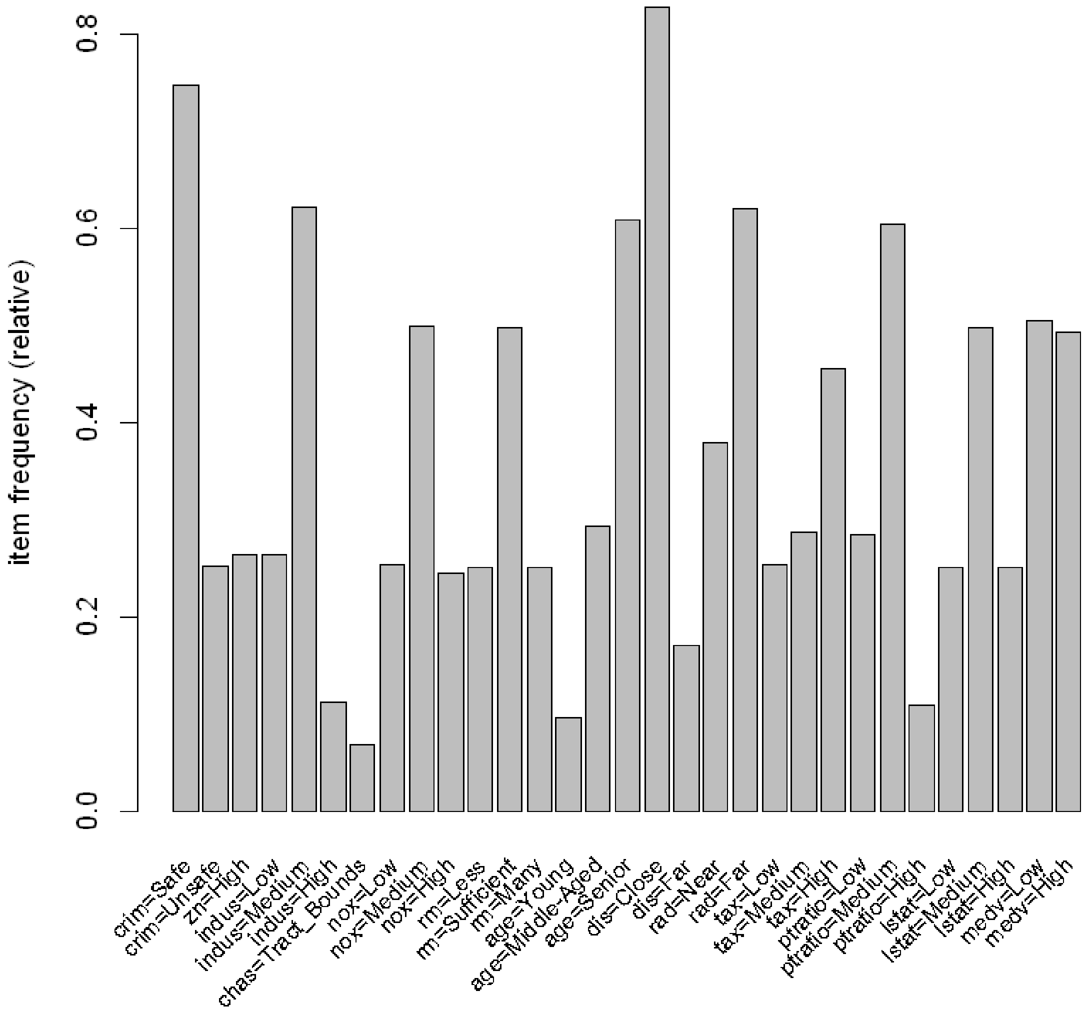
```
Boston$medv <- ordered(cut(Boston$medv, c(0, 21.20 ,  50.00  ), labels=c('Low', 'H
```

In [35]:
```
# binary incidence matrix
boston_matrix <- as(Boston, 'transactions')
summary(boston_matrix)
```

```
transactions as itemMatrix in sparse format with
 506 rows (elements/itemsets/transactions) and
 31 columns (items) and a density of 0.3656126

most frequent items:
   dis=Close     crim=Safe indus=Medium       rad=Far    age=Senior       (Other)
         419           378          315           314           308          4001

element (itemset/transaction) length distribution:
sizes
 11  12  13
344 155   7

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  11.00   11.00   11.00   11.33   12.00   13.00

includes extended item information - examples:
       labels variables levels
1   crim=Safe      crim   Safe
2 crim=Unsafe      crim Unsafe
3     zn=High        zn   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

In [36]:
```
#plot
itemFrequencyPlot(boston_matrix, support=0.03, cex.names=0.8)
```



In [37]:
```r
rules  <- apriori(boston_matrix, parameter = list(support = 0.01, confidence = 0.8
summary(rules)
sample(labels(rules), size=5)
```

```
Apriori

Parameter specification:
 confidence minval smax arem  aval originalSupport maxtime support minlen
        0.8    0.1    1 none FALSE            TRUE       5    0.01      2
 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 ...[31 item(s), 506 transaction(s)] done [0.00s].
sorting and recoding items ... [31 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 message in apriori(boston_matrix, parameter = list(support = 0.01, confide
nce = 0.8, :
"Mining stopped (maxlen reached). Only patterns up to a length of 10 returned!"
```

```
 done [0.04s].
writing ... [163362 rule(s)] done [0.06s].
creating S4 object  ... done [0.16s].
set of 163362 rules

rule length distribution (lhs + rhs):sizes
    2     3     4     5     6     7     8     9    10 
   64  1116  7587 25175 44527 44987 27371 10247  2288 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  2.000   6.000   7.000   6.586   7.000  10.000 

summary of quality measures:
    support          confidence        coverage            lift        
 Min.   :0.01186   Min.   :0.8000   Min.   :0.01186   Min.   : 0.9661  
 1st Qu.:0.01383   1st Qu.:0.9091   1st Qu.:0.01383   1st Qu.: 1.3386  
 Median :0.01779   Median :1.0000   Median :0.01976   Median : 1.7126  
 Mean   :0.02882   Mean   :0.9588   Mean   :0.03033   Mean   : 2.2255  
 3rd Qu.:0.02964   3rd Qu.:1.0000   3rd Qu.:0.03162   3rd Qu.: 3.1251  
 Max.   :0.58696   Max.   :1.0000   Max.   :0.62253   Max.   :10.3265  
     count       
 Min.   :  6.00  
 1st Qu.:  7.00  
 Median :  9.00  
 Mean   : 14.58  
 3rd Qu.: 15.00  
 Max.   :297.00  

mining info:
          data ntransactions support confidence
 boston_matrix           506    0.01        0.8
```

1. '{zn=High,indus=Low,age=Senior,lstat=Low} => {crim=Safe}'
2. '{crim=Safe,rm=Less,age=Senior,rad=Near,tax=Low,ptratio=Medium} => {dis=Close}'
3. '{crim=Safe,indus=High,rm=Sufficient,dis=Close,tax=High,ptratio=High,lstat=Medium} => {age=Senior}'
4. '{crim=Safe,nox=High,dis=Close,tax=Low,ptratio=Low,lstat=Medium} => {zn=High}'
5. '{zn=High,indus=Low,nox=Low,age=Middle-Aged,rad=Near,ptratio=Low,lstat=Low} => {rm=Many}'

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?

In [38]:

```
ruleslowCrime <- subset(rules, subset = lhs %ain% c('crim=Safe', 'tax=Low') &rhs %
summary(ruleslowCrime)
inspect(head(sort(ruleslowCrime, by='support'), n=10))


ruleslowCrime <- subset(rules, subset = lhs %ain% c('crim=Safe','dis=Close') &rhs %
summary(ruleslowCrime)
inspect(head(sort(ruleslowCrime, by='support'), n=10))
```

```
set of 1144 rules

rule length distribution (lhs + rhs):sizes
  4   5   6   7   8   9  10 
  5  53 201 346 311 170  58 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   4.00    7.00    7.00    7.44    8.00   10.00 

summary of quality measures:
    support          confidence    coverage            lift      
 Min.   :0.01186   Min.   :1    Min.   :0.01186   Min.   :1.208  
 1st Qu.:0.01186   1st Qu.:1    1st Qu.:0.01186   1st Qu.:1.208  
 Median :0.01581   Median :1    Median :0.01581   Median :1.208  
 Mean   :0.01792   Mean   :1    Mean   :0.01792   Mean   :1.208  
 3rd Qu.:0.01976   3rd Qu.:1    3rd Qu.:0.01976   3rd Qu.:1.208  
 Max.   :0.09091   Max.   :1    Max.   :0.09091   Max.   :1.208  
     count       
 Min.   : 6.000  
 1st Qu.: 6.000  
 Median : 8.000  
 Mean   : 9.066  
 3rd Qu.:10.000  
 Max.   :46.000  

mining info:
          data ntransactions support confidence
 boston_matrix           506    0.01        0.8
```

```
     lhs                   rhs             support confidence   coverage     lift cou
nt
[1]  {crim=Safe,
      age=Senior,
      tax=Low}          => {dis=Close} 0.09090909          1 0.09090909 1.207637
46
[2]  {crim=Safe,
      nox=Medium,
      tax=Low,
      lstat=Medium}     => {dis=Close} 0.07905138          1 0.07905138 1.207637
40
[3]  {crim=Safe,
      indus=Medium,
      nox=Medium,
      tax=Low}          => {dis=Close} 0.06719368          1 0.06719368 1.207637
34
[4]  {crim=Safe,
      nox=Medium,
      age=Senior,
      tax=Low}          => {dis=Close} 0.06521739          1 0.06521739 1.207637
33
[5]  {crim=Safe,
      age=Senior,
      tax=Low,
      medv=High}        => {dis=Close} 0.06521739          1 0.06521739 1.207637
33
[6]  {crim=Safe,
      nox=Medium,
      tax=Low,
      ptratio=Medium,
      lstat=Medium}     => {dis=Close} 0.06521739          1 0.06521739 1.207637
33
[7]  {crim=Safe,
      age=Senior,
      rad=Near,
      tax=Low}          => {dis=Close} 0.05928854          1 0.05928854 1.207637
30
[8]  {crim=Safe,
      nox=Medium,
      rm=Sufficient,
      tax=Low,
      lstat=Medium}     => {dis=Close} 0.05731225          1 0.05731225 1.207637
29
[9]  {crim=Safe,
      age=Senior,
      tax=Low,
      ptratio=Medium}   => {dis=Close} 0.05533597          1 0.05533597 1.207637
28
[10] {crim=Safe,
      nox=Medium,
      age=Senior,
      rad=Near,
      tax=Low}          => {dis=Close} 0.05138340          1 0.05138340 1.207637
26
```

```
set of 987 rules

rule length distribution (lhs + rhs):sizes
  5   6   7   8   9  10 
 16 112 281 325 190  63 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   5.00    7.00    8.00    7.76    9.00   10.00 

summary of quality measures:
    support           confidence        coverage             lift      
 Min.   :0.01186   Min.   :0.8000   Min.   :0.01186   Min.   :3.138  
 1st Qu.:0.01383   1st Qu.:0.8571   1st Qu.:0.01383   1st Qu.:3.362  
 Median :0.01779   Median :0.9167   Median :0.01779   Median :3.596  
 Mean   :0.02004   Mean   :0.9247   Mean   :0.02217   Mean   :3.627  
 3rd Qu.:0.02372   3rd Qu.:1.0000   3rd Qu.:0.02569   3rd Qu.:3.922  
 Max.   :0.10277   Max.   :1.0000   Max.   :0.12846   Max.   :3.922  
     count      
 Min.   : 6.00  
 1st Qu.: 7.00  
 Median : 9.00  
 Mean   :10.14  
 3rd Qu.:12.00  
 Max.   :52.00  

mining info:
          data ntransactions support confidence
 boston_matrix           506    0.01        0.8
```

```
     lhs                     rhs          support confidence   coverage     lift coun
t
[1]  {crim=Safe,
      dis=Close,
      rad=Near,
      ptratio=Medium,
      medv=High}       => {tax=Low} 0.10276680  0.8000000 0.12845850 3.137984    5
2
[2]  {crim=Safe,
      indus=Low,
      rm=Many,
      dis=Close}       => {tax=Low} 0.07312253  0.8809524 0.08300395 3.455519    3
7
[3]  {crim=Safe,
      indus=Low,
      rm=Many,
      dis=Close,
      medv=High}       => {tax=Low} 0.07312253  0.8809524 0.08300395 3.455519    3
7
[4]  {crim=Safe,
      indus=Low,
      dis=Close,
      ptratio=Medium}  => {tax=Low} 0.07114625  0.8780488 0.08102767 3.444129    3
6
[5]  {crim=Safe,
      age=Middle-Aged,
      dis=Close,
      rad=Near,
      ptratio=Medium}  => {tax=Low} 0.06521739  0.8461538 0.07707510 3.319022    3
3
[6]  {crim=Safe,
      zn=High,
      rm=Many,
      dis=Close}       => {tax=Low} 0.06126482  0.8611111 0.07114625 3.377692    3
1
[7]  {crim=Safe,
      zn=High,
      rm=Many,
      dis=Close,
      medv=High}       => {tax=Low} 0.06126482  0.8611111 0.07114625 3.377692    3
1
[8]  {crim=Safe,
      indus=Low,
      dis=Close,
      rad=Near,
      medv=High}       => {tax=Low} 0.05928854  0.8333333 0.07114625 3.268734    3
0
[9]  {crim=Safe,
      indus=Low,
      dis=Close,
      ptratio=Medium,
      medv=High}       => {tax=Low} 0.05928854  0.8571429 0.06916996 3.362126    3
0
[10] {crim=Safe,
      nox=Medium,
      dis=Close,
      rad=Near,
      ptratio=Medium,
      medv=High}       => {tax=Low} 0.05928854  0.8571429 0.06916996 3.362126    3
0
```

- Applying some association rules, we can suggest the student interested in low taxes, but wants to be in a safe area with low crime as:

1. From above we can tell that the association between safe crime area with low tax, a student can find a housing.

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?

In [39]:

```
ruleslowpt <- subset(rules, subset = rhs %in% 'ptratio=Low' & lift>2.5)
summary(ruleslowpt)
inspect(head(sort(ruleslowpt, by='support', decreasing = TRUE), n=10))
```

```
set of 4413 rules

rule length distribution (lhs + rhs):sizes
   3    4    5    6    7    8    9   10 
  10  102  476 1058 1303  951  411  102 

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  3.000   6.000   7.000   6.937   8.000  10.000 

summary of quality measures:
    support           confidence        coverage             lift      
 Min.   :0.01186   Min.   :0.800   Min.   :0.01186   Min.   :2.811  
 1st Qu.:0.01383   1st Qu.:0.875   1st Qu.:0.01383   1st Qu.:3.075  
 Median :0.01581   Median :1.000   Median :0.01779   Median :3.514  
 Mean   :0.01821   Mean   :0.942   Mean   :0.01962   Mean   :3.310  
 3rd Qu.:0.01976   3rd Qu.:1.000   3rd Qu.:0.02174   3rd Qu.:3.514  
 Max.   :0.09684   Max.   :1.000   Max.   :0.11265   Max.   :3.514  
     count       
 Min.   : 6.000  
 1st Qu.: 7.000  
 Median : 8.000  
 Mean   : 9.214  
 3rd Qu.:10.000  
 Max.   :49.000  

mining info:
          data ntransactions support confidence
 boston_matrix           506    0.01        0.8
```

```
     lhs                rhs                support confidence   coverage     lift count
[1]  {crim=Safe,
      age=Senior,
      rad=Far,
      medv=High}   => {ptratio=Low} 0.09683794  0.8596491 0.11264822 3.020712    49
[2]  {crim=Safe,
      age=Senior,
      dis=Close,
      rad=Far,
      medv=High}   => {ptratio=Low} 0.09683794  0.8750000 0.11067194 3.074653    49
[3]  {crim=Safe,
      rm=Many,
      dis=Close,
      rad=Far}     => {ptratio=Low} 0.07509881  0.8260870 0.09090909 2.902778    38
[4]  {crim=Safe,
      rm=Many,
      dis=Close,
      rad=Far,
      medv=High}   => {ptratio=Low} 0.07509881  0.8260870 0.09090909 2.902778    38
[5]  {crim=Safe,
      nox=Medium,
      age=Senior,
      rad=Far,
      medv=High}   => {ptratio=Low} 0.07509881  0.8444444 0.08893281 2.967284    38
[6]  {crim=Safe,
      nox=Medium,
      age=Senior,
      dis=Close,
      rad=Far,
      medv=High}   => {ptratio=Low} 0.07509881  0.8636364 0.08695652 3.034722    38
[7]  {crim=Safe,
      rm=Many,
      rad=Far,
      lstat=Low}   => {ptratio=Low} 0.06521739  0.8048780 0.08102767 2.828252    33
[8]  {crim=Safe,
      rm=Many,
      rad=Far,
      lstat=Low,
      medv=High}   => {ptratio=Low} 0.06521739  0.8048780 0.08102767 2.828252    33
[9]  {crim=Safe,
      dis=Close,
      rad=Far,
      lstat=Low}   => {ptratio=Low} 0.06324111  0.8000000 0.07905138 2.811111    32
[10] {crim=Safe,
      dis=Close,
      rad=Far,
      lstat=Low,
      medv=High}   => {ptratio=Low} 0.06324111  0.8000000 0.07905138 2.811111    32
```

- From above, we can tell that for schools in areas with crime safe areas with high median home values, are most likely to have low PTRatio.

Extra Credit: Use a regression model to solve part d. Are you results comparable? Which provides an easier interpretation? When would regression be preferred, and when would association models be preferred?

In [40]:
```
data(Boston)
```

```r
model_lm <- lm(ptratio~., data=Boston)
summary(model_lm)
```

```
Call:
lm(formula = ptratio ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-4.2228 -1.0341 -0.0015  0.9260  4.8646 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.571e+01  1.263e+00  20.357  < 2e-16 ***
crim        -1.766e-02  1.083e-02  -1.632  0.10339    
zn          -2.499e-02  4.415e-03  -5.660 2.57e-08 ***
indus        5.633e-02  2.001e-02   2.815  0.00507 ** 
chas        -2.697e-01  2.851e-01  -0.946  0.34469    
nox         -1.066e+01  1.186e+00  -8.989  < 2e-16 ***
rm          -1.118e-01  1.464e-01  -0.764  0.44527    
age          7.725e-03  4.312e-03   1.792  0.07382 .  
dis         -1.855e-02  6.896e-02  -0.269  0.78806    
rad          1.145e-01  2.151e-02   5.322 1.56e-07 ***
tax          6.951e-04  1.247e-03   0.557  0.57748    
lstat       -4.034e-02  1.822e-02  -2.214  0.02730 *  
medv        -9.873e-02  1.392e-02  -7.091 4.63e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.557 on 493 degrees of freedom
Multiple R-squared:  0.495,	Adjusted R-squared:  0.4828 
F-statistic: 40.28 on 12 and 493 DF,  p-value: < 2.2e-16
```

When we want to identify patterns or relation between two or more variables we use association rules and we need to understand thar relationship we use regression model.

For example in our case we were interested to know if the family moving to a certain area has low teacher-pupil ratio and so we used associantion rule. However, if we wanted to understand this relation we can use regression.

For me association rules provide easy way of interpretation. As per my understanding if once we identify the relation or pattern it would be easier to apply any regression model if required.

References :

- a. https://datascience.stackexchange.com/questions/106369/print-histogram-for-each-of-the-columns-in-my-table-with-one-single-command
- b. https://www.statology.org/train-test-split-r/