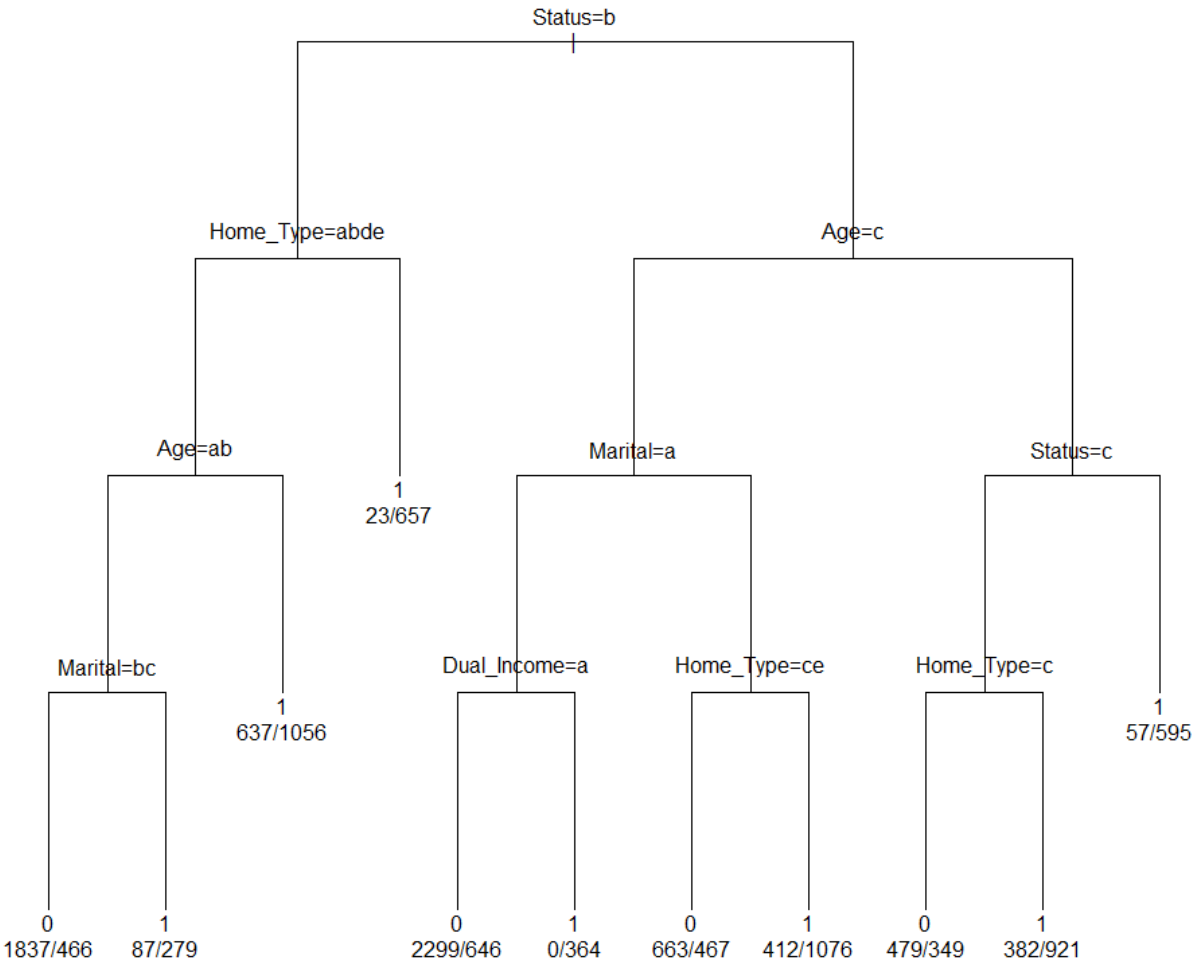
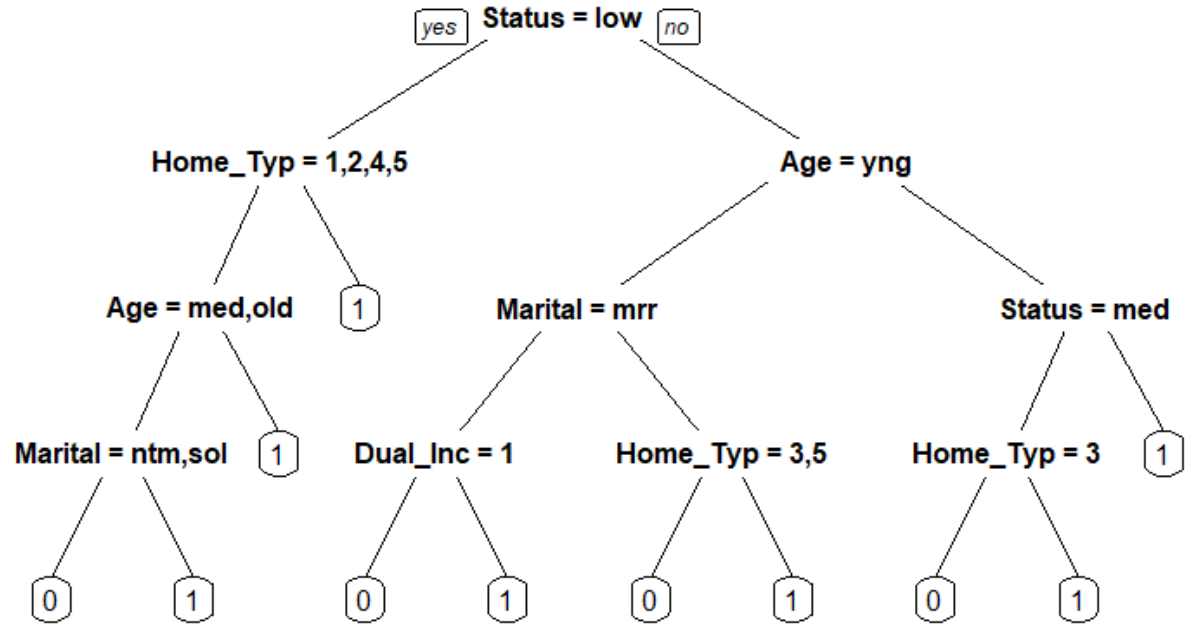
1)

For this problem, I will be using classification tree to cluster the marketing data. To begin with, I omitted NAs in order to have more accurate data. Also, I made it in to a data frame in order to label each variable. Mostly, variables are categorized into three variables, low, med, and high. I used histogram as a reference to categorize the data. I ignored the data by using NULL, as I thought that they were irrelevant for our analysis. For the categorized data, I added a class column to mark that it is sample zero.

For sample one, I duplicated my_marketing data, and replaced each variable as True, so that it can be differentiated with sample zero.

Lastly, sample zero and sample one is combined, and tree is drawn by using rpart, class method.

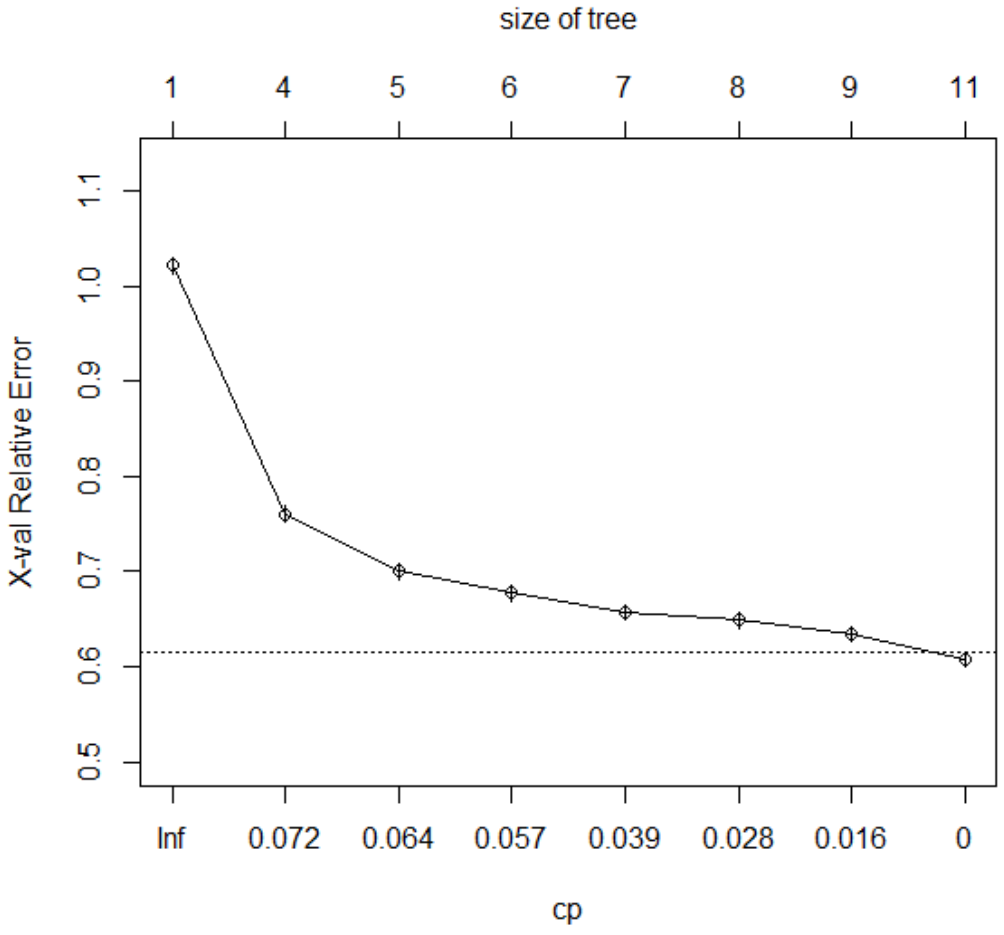





```
> tree$cptable
          CP nsplit rel error    xerror        xstd
1 0.07664340      0 1.0000000 1.0218150 0.008525379
2 0.06806283      3 0.7700698 0.7606166 0.008279475
3 0.06093659      4 0.7020070 0.7002618 0.008135331
4 0.05293775      5 0.6410704 0.6771379 0.008070732
5 0.02850494      6 0.5881326 0.6569226 0.008009856
6 0.02792321      7 0.5596277 0.6489238 0.007984612
7 0.00945317      8 0.5317045 0.6343805 0.007937008
8 0.00000000     10 0.5127981 0.6082024 0.007845655
```

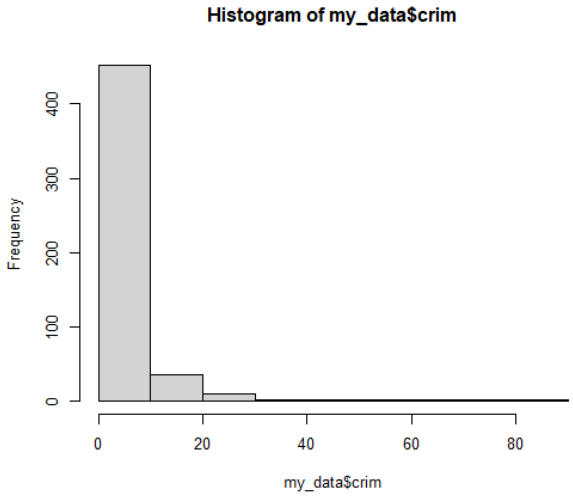
Cp becomes zero when nsplit is 8~10. Therefore, misplit for rpart is 9.

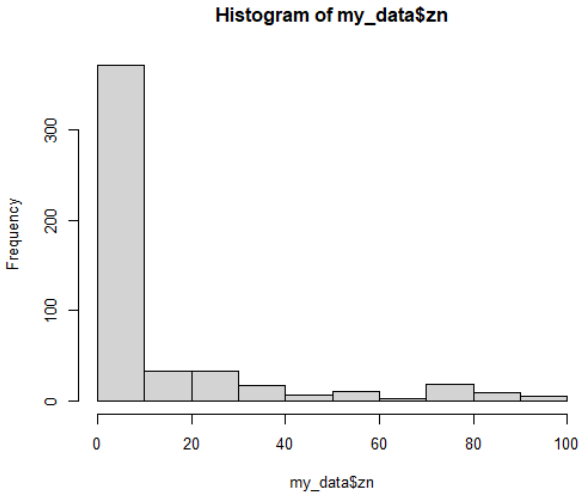
```
> tree
n= 13752

node), split, n, loss, yval, (yprob)
      * denotes terminal node

 1) root 13752 6876 0 (0.50000000 0.50000000)
   2) Status=low 5042 2458 0 (0.51249504 0.48750496)
     4) Home_Type=1,2,4,5 4362 1801 0 (0.58711600 0.41288400)
       8) Age=med,old 2669  745 0 (0.72086924 0.27913076)
        16) Marital=not married,solo 2303  466 0 (0.79765523 0.20234477) *
        17) Marital=married 366   87 1 (0.23770492 0.76229508) *
       9) Age=young 1693  637 1 (0.37625517 0.62374483) *
     5) Home_Type=3 680   23 1 (0.03382353 0.96617647) *
   3) Status=high,med 8710 4292 1 (0.49276693 0.50723307)
     6) Age=young 5927 2553 0 (0.56925932 0.43074068)
      12) Marital=married 3309 1010 0 (0.69477183 0.30522817)
        24) Dual_Income=1 2945  646 0 (0.78064516 0.21935484) *
        25) Dual_Income=2,3 364    0 1 (0.00000000 1.00000000) *
      13) Marital=not married,solo 2618 1075 1 (0.41061879 0.58938121)
        26) Home_Type=3,5 1130  467 0 (0.58672566 0.41327434) *
        27) Home_Type=1,2,4 1488  412 1 (0.27688172 0.72311828) *
     7) Age=med,old 2783  918 1 (0.32985986 0.67014014)
      14) Status=med 2131  861 1 (0.40403566 0.59596434)
        28) Home_Type=3 828  349 0 (0.57850242 0.42149758) *
        29) Home_Type=1,2,4,5 1303  382 1 (0.29316961 0.70683039) *
      15) Status=high 652   57 1 (0.08742331 0.91257669) *
```

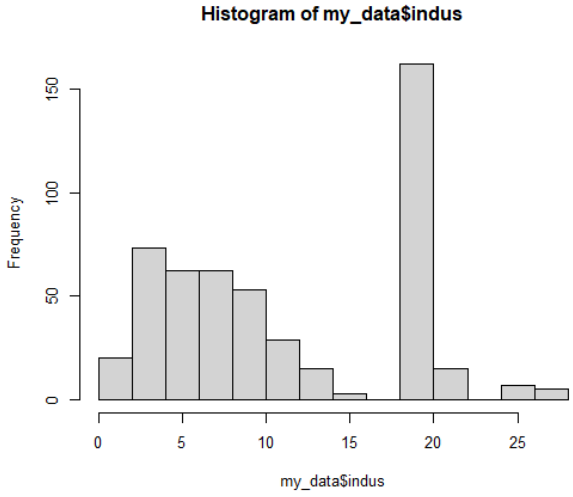


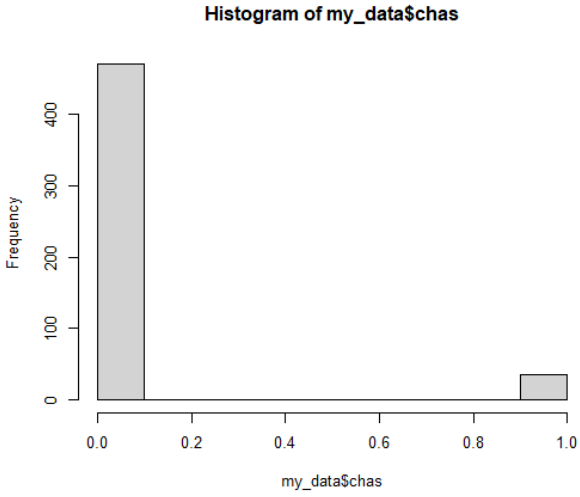
2.

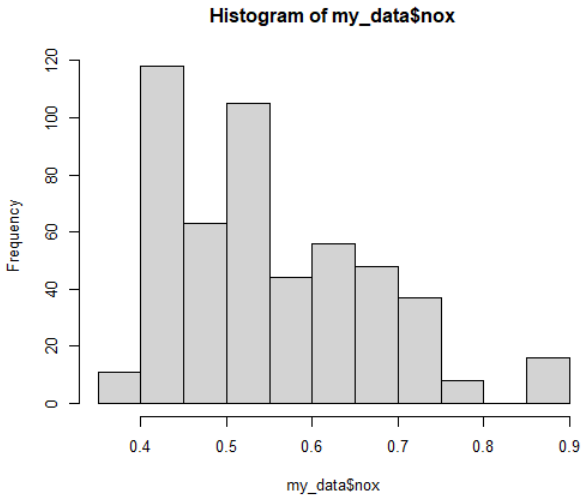
a)

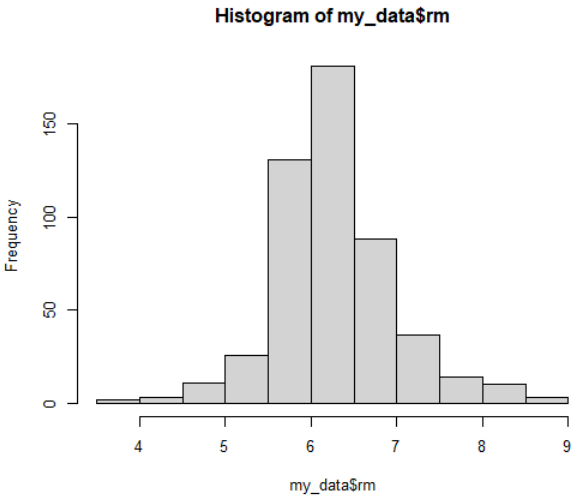

Histogram of my_data$crim
Frequency
0
100
200
300
400
0
20
40
60
80
my_data$crim

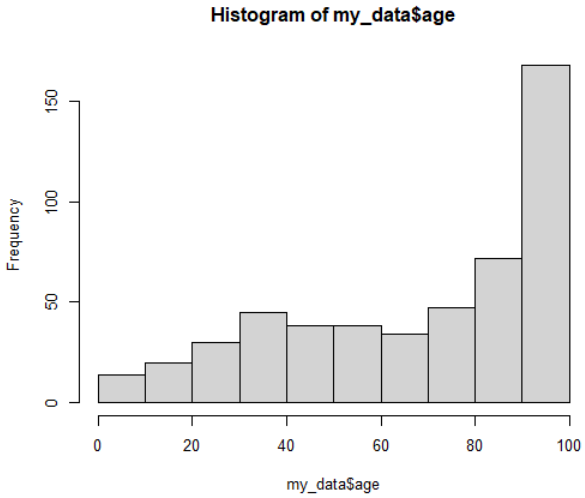

Histogram of my_data$zn
Frequency
0
100
200
300
0
20
40
60
80
100
my_data$zn

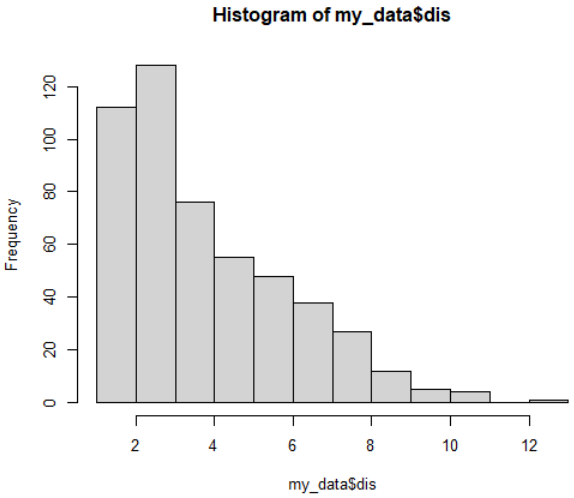

Histogram of my_data$indus
Frequency
0
50
100
150
0
5
10
15
20
25
my_data$indus

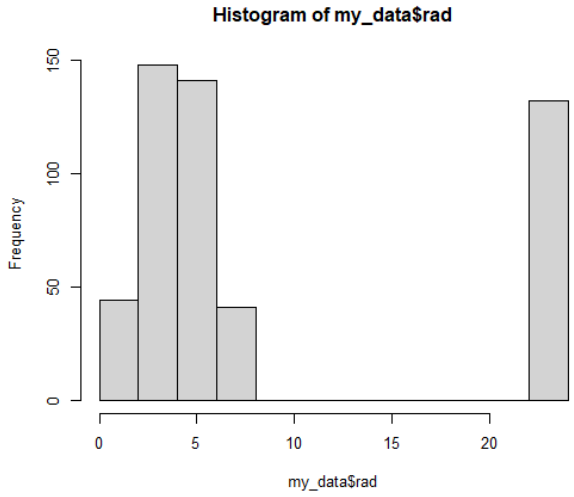

Histogram of my_data$chas
Frequency
0
100
200
300
400
0.0
0.2
0.4
0.6
0.8
1.0
my_data$chas

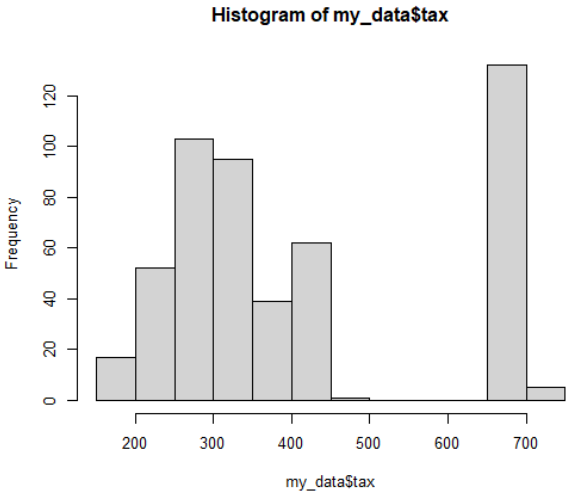

Histogram of my_data$nox
Frequency
0
20
40
60
80
100
120
0.4
0.5
0.6
0.7
0.8
0.9
my_data$nox

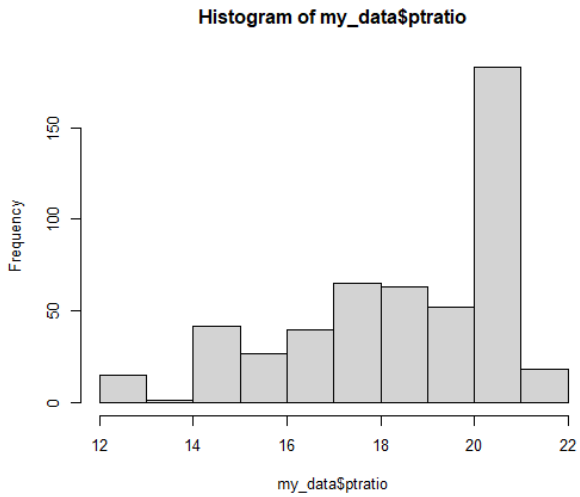

Histogram of my_data$rm
Frequency
0
50
100
150
4
5
6
7
8
9
my_data$rm

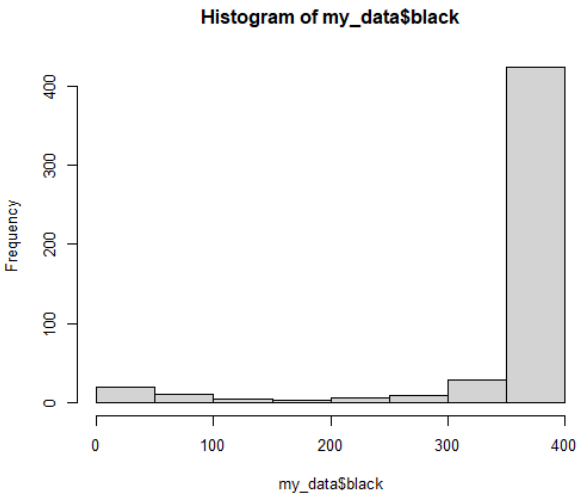

Histogram of my_data$age
Frequency
0
50
100
150
0
20
40
60
80
100
my_data$age

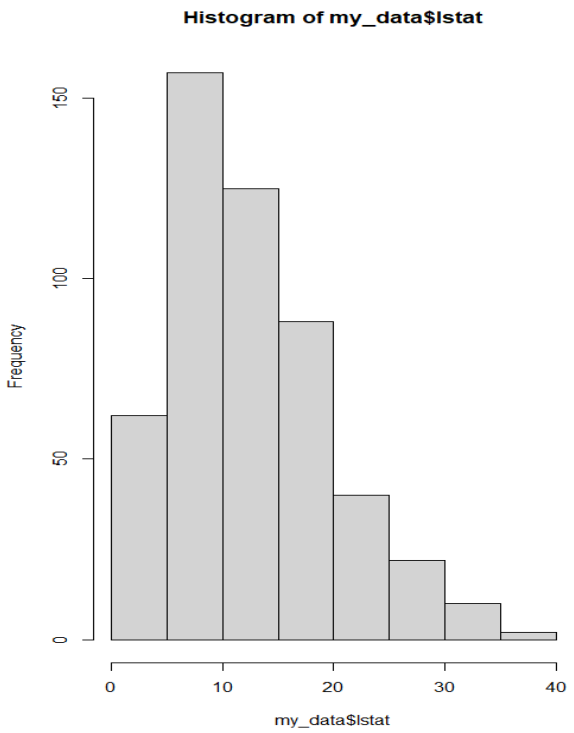

Histogram of my_data$dis
Frequency
0
20
40
60
80
100
120
2
4
6
8
10
12
my_data$dis

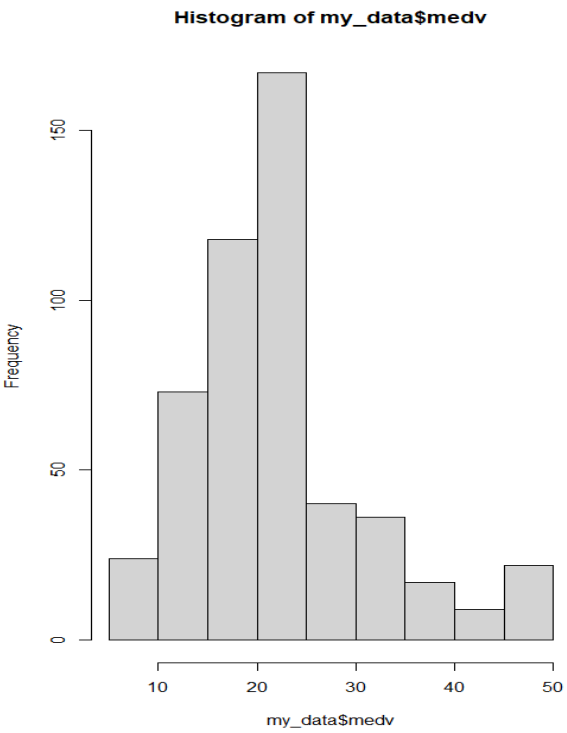

Histogram of my_data$rad
Frequency
0
50
100
150
0
5
10
15
20
my_data$rad


Histogram of my_data$tax
Frequency
0
20
40
60
80
100
120
200
300
400
500
600
700
my_data$tax


Histogram of my_data$ptratio
Frequency
0
50
100
150
12
14
16
18
20
22
my_data$ptratio


Histogram of my_data$black
Frequency
0
100
200
300
400
0
100
200
300
400
my_data$black


Histogram of my_data$lstat
Frequency
0
50
100
150
0
10
20
30
40
my_data$lstat


Histogram of my_data$medv
Frequency
0
50
100
150
10
20
30
40
50
my_data$medv

Most of the variables are classified into three variables, small medium, large.

Crim = classified into three categories: "smallcrime", "mediumcrime", "largecrime"

Zn = classified into three categories: "smallft", "mediumft", "largeft"

Indus = classified into three categories: "smallnonbus", "mediumnonbus", "largenonbus"

Nox = classified into three categories: "smallnx", "mediumnx", "largenx"

Rm = classified into three categories: "smallroom", "mediumroom", "largeroom"

age = classified into three categories: "young", "old", "elder"

dis = classified into three categories: "near", "not_near", "far"

rad = classified into three categories: "Acc", "NotAcc"

tax = classified into three categories: "lowtax", "mediumtax", "largetax"

ptratio = classified into three categories: "lowtcher", "mediumtcher", "hightcher"

black = classified into two categories: "lowblack", "highblack"

lstat = classified into three categories: "lowpop", "mediumpop", "largepop"

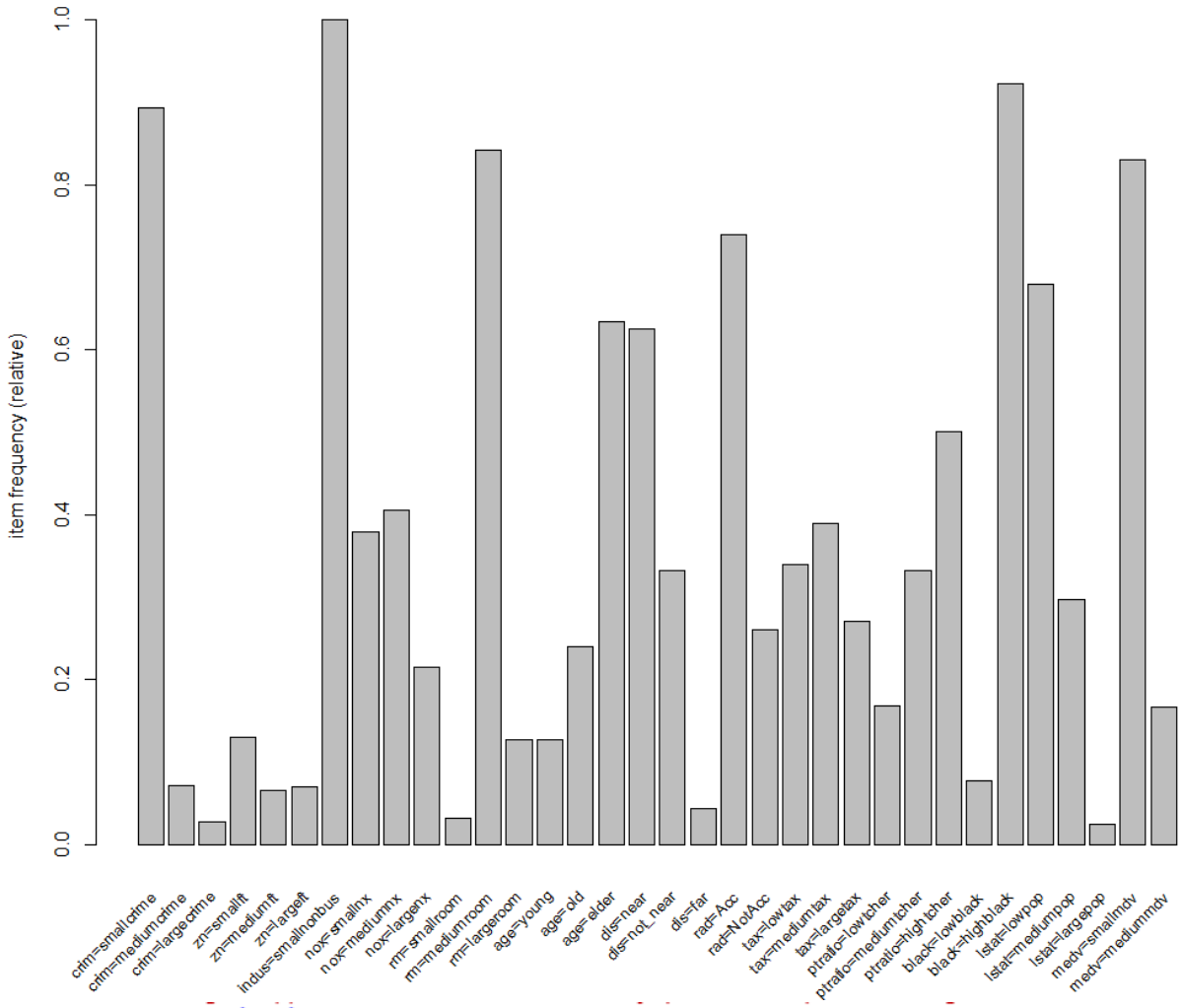
medv = classified into three categories: "smallmdv", "mediummdv", "largemdv"

Moreover, the data is changed to binary incidence matrix

```
> summary(indice_matrix)
transactions as itemMatrix in sparse format with
 506 rows (elements/itemsets/transactions) and
 34 columns (items) and a density of 0.3603813

most frequent items:
indus=smallnonbus   black=highblack   crim=smallcrime     rm=mediumroom     medv=smallmdv
              506               467               452               426               420
          (Other)
             3929

element (itemset/transaction) length distribution:
sizes
 10  11  12  13
  1   4 367 134

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  10.00   12.00   12.00   12.25   13.00   13.00

includes extended item information - examples:
            labels variables      levels
1  crim=smallcrime      crim  smallcrime
2 crim=mediumcrime      crim mediumcrime
3  crim=largecrime      crim  largecrime

includes extended transaction information - examples:
  transactionID
1             1
2             2
3             3
```

b)

ItemFrequency Plot




```
> summary(apr)
set of 48943 rules

rule length distribution (lhs + rhs):sizes
    1     2     3     4     5     6     7     8     9    10
    5   155  1257  4637  9775 13011 11295  6298  2128   382

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   1.00    5.00    6.00    6.17    7.00   10.00

summary of quality measures:
    support         confidence        coverage            lift            count
 Min.   :0.07115   Min.   :0.800   Min.   :0.07115   Min.   :0.8696   Min.   : 36.0
 1st Qu.:0.08498   1st Qu.:0.939   1st Qu.:0.08893   1st Qu.:1.0656   1st Qu.: 43.0
 Median :0.10870   Median :1.000   Median :0.11265   Median :1.1531   Median : 55.0
 Mean   :0.13201   Mean   :0.962   Mean   :0.13826   Mean   :1.3981   Mean   : 66.8
 3rd Qu.:0.15217   3rd Qu.:1.000   3rd Qu.:0.15810   3rd Qu.:1.4025   3rd Qu.: 77.0
 Max.   :1.00000   Max.   :1.000   Max.   :1.00000   Max.   :5.7175   Max.   :506.0

mining info:
          data ntransactions support confidence
 indice_matrix           506    0.07        0.8
```

The apriori method is applied, set of 48943 rules are made. The support is 0.07, and confidence is 0.8 in this data.

c)

```
> inspect(head(subset(apr, subset=rhs %in% "dis=near"),n = 5,by="confidence"))
    lhs                                rhs        support    confidence coverage   lift     count
[1] {crim=mediumcrime}              => {dis=near} 0.07114625 1          0.07114625 1.601266  36
[2] {black=lowblack}                => {dis=near} 0.07707510 1          0.07707510 1.601266  39
[3] {nox=largenx}                   => {dis=near} 0.21541502 1          0.21541502 1.601266 109
[4] {crim=mediumcrime,rad=NotAcc}   => {dis=near} 0.07114625 1          0.07114625 1.601266  36
[5] {crim=mediumcrime,tax=largetax} => {dis=near} 0.07114625 1          0.07114625 1.601266  36
> inspect(head(subset(apr, subset=rhs %in% "crim=smallcrime"),n = 5,by="confidence"))
    lhs                     rhs                 support   confidence coverage  lift     count
[1] {age=young}          => {crim=smallcrime} 0.1264822 1          0.1264822 1.119469  64
[2] {zn=smallft}         => {crim=smallcrime} 0.1304348 1          0.1304348 1.119469  66
[3] {medv=mediummdv}     => {crim=smallcrime} 0.1660079 1          0.1660079 1.119469  84
[4] {ptratio=lowtcher}   => {crim=smallcrime} 0.1679842 1          0.1679842 1.119469  85
[5] {ptratio=mediumtcher} => {crim=smallcrime} 0.3320158 1          0.3320158 1.119469 168
```

When the distance is near to Boston, crime is medium, low rate of black people, large nitrogen oxides concentration, not accessible to highways, and high tax.

When there is low crime in Boston, young people are located, residential zone is low, median value is medium, pupil teacher ratio is low.

As a result, in order to live near Boston with low crime rate, the place should have low black people rate, high tax, residential zone should be low, young people should be located, and pupil teacher ratio should be low.

d)

I lowered the apriori method support to 0.01, since 0 rules are shown for pupil teacher, the results are shown as follows.

```
> inspect(head(subset(apr1, subset=rhs %in% "ptratio=lowtcher"),n = 5,by="lift"))
    lhs                                   rhs                  support     confidence
[1] {zn=smallft,nox=mediumnx}          => {ptratio=lowtcher} 0.039525692 1
[2] {nox=largenx,tax=mediumtax}        => {ptratio=lowtcher} 0.031620553 1
[3] {nox=largenx,rad=Acc}              => {ptratio=lowtcher} 0.031620553 1
[4] {nox=largenx,rm=smallroom,tax=mediumtax} => {ptratio=lowtcher} 0.003952569 1
[5] {nox=largenx,rm=smallroom,rad=Acc} => {ptratio=lowtcher} 0.003952569 1
    coverage    lift     count
[1] 0.039525692 5.952941 20
[2] 0.031620553 5.952941 16
[3] 0.031620553 5.952941 16
[4] 0.003952569 5.952941  2
[5] 0.003952569 5.952941  2
```

When ptratio is low, proportion of residential land is low, nitrogen oxides concentration is medium, average number of rooms is small, medium tax rate, radial highways are accessible.

Regression model

```
> lm(train$ptratio~., train)

Call:
lm(formula = train$ptratio ~ ., data = train)

Coefficients:
(Intercept)         crim           zn        indus         chas          nox           rm
  22.145692    -0.014990    -0.033318     0.032184     0.300304   -11.732809     0.166980
        age          dis          rad          tax        black        lstat         medv
   0.015207     0.061149     0.118265     0.001693     0.002150    -0.004729    -0.090661
```

```
> summary(lmpupil)

Call:
lm(formula = train$ptratio ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max
-3.7840 -0.9432  0.0083  0.9424  5.2052

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  22.145692   2.736039   8.094 2.69e-13 ***
crim         -0.014990   0.039990  -0.375 0.708342
zn           -0.033318   0.008602  -3.873 0.000165 ***
indus         0.032184   0.040216   0.800 0.424926
chas          0.300304   0.519958   0.578 0.564506
nox         -11.732809   2.541264  -4.617 8.82e-06 ***
rm            0.166980   0.277950   0.601 0.548988
age           0.015207   0.008144   1.867 0.063985 .
dis           0.061149   0.130653   0.468 0.640506
rad           0.118265   0.047482   2.491 0.013934 *
tax           0.001693   0.002639   0.642 0.522186
black         0.002150   0.001999   1.075 0.284206
lstat        -0.004729   0.033484  -0.141 0.887898
medv         -0.090661   0.023960  -3.784 0.000229 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.668 on 138 degrees of freedom
Multiple R-squared:  0.4943,    Adjusted R-squared:  0.4466
F-statistic: 10.37 on 13 and 138 DF,  p-value: 4.871e-15
```

By looking at the data, positive relationship between ptratio is indus, chas, rm, age, dis, rad, tax, and black. Compared to association rules, if we want to have specific interpretation, association rules should be preferred. We can set various standards such as support, and confidence. Moreover, association rules show lift, which helps us to have specific interpretation.