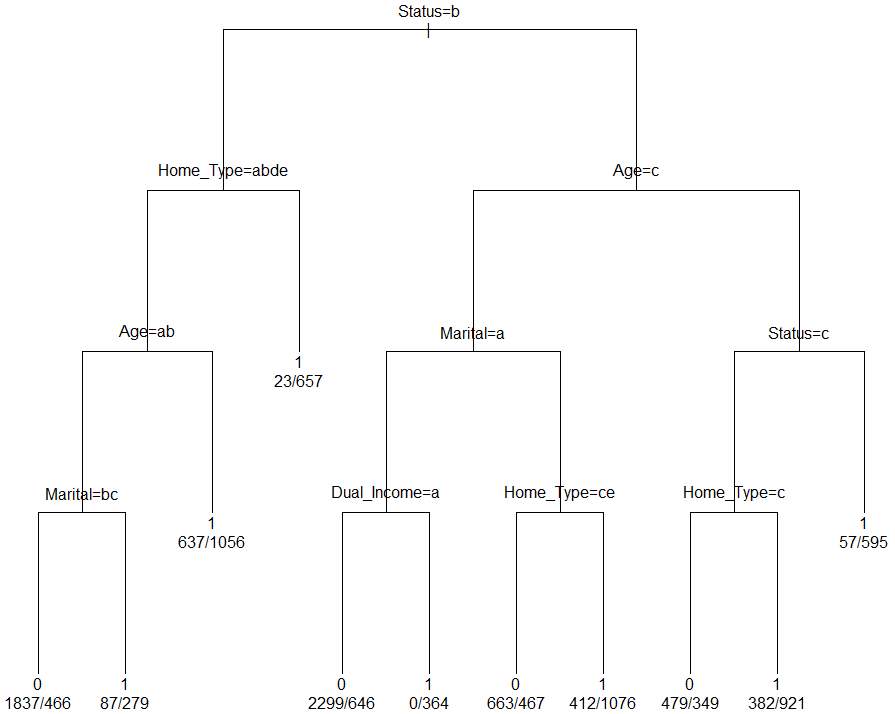
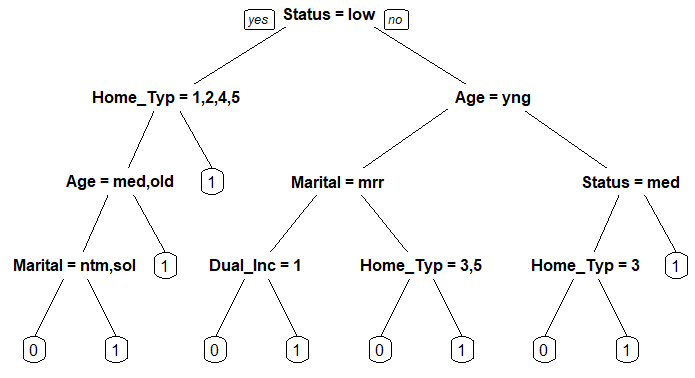
1)

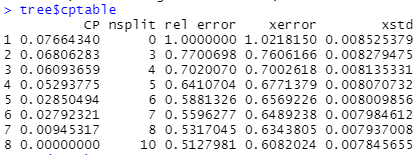
For this problem, I will be using classification tree to cluster the marketing data. To begin with, I omitted NAs in order to have more accurate data. Also, I made it in to a data frame in order to label each variable. Mostly, variables are categorized into three variables, low, med, and high. I used 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.

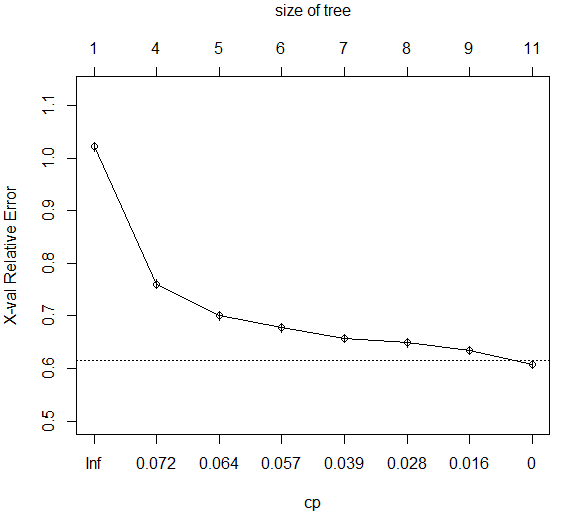
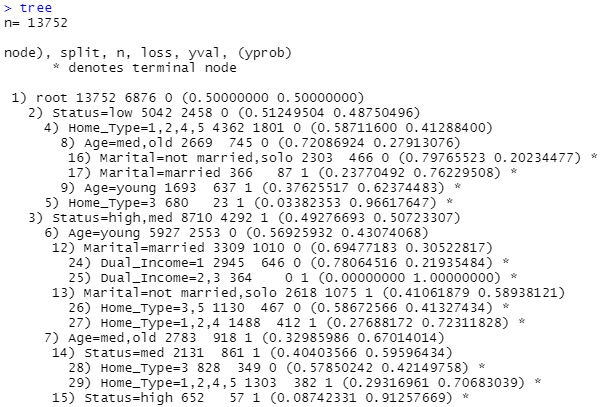
Lastly, sample zero and sample one is combined, and tree is drawn by using rpart, class method.





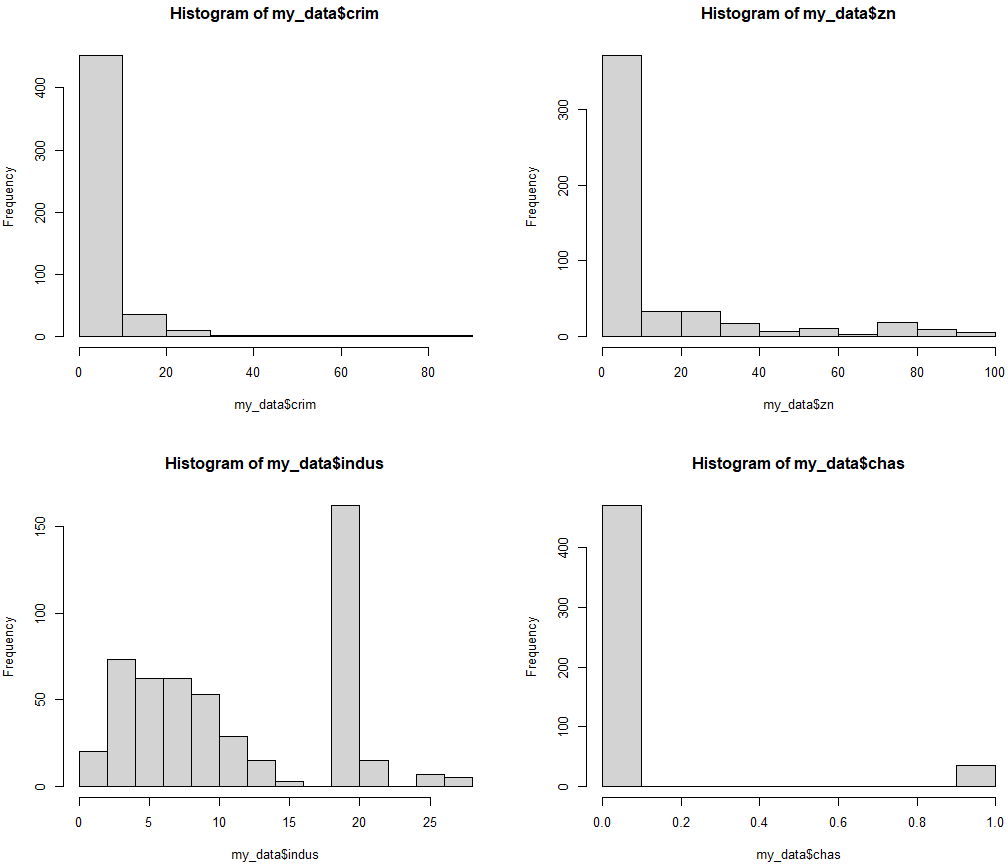


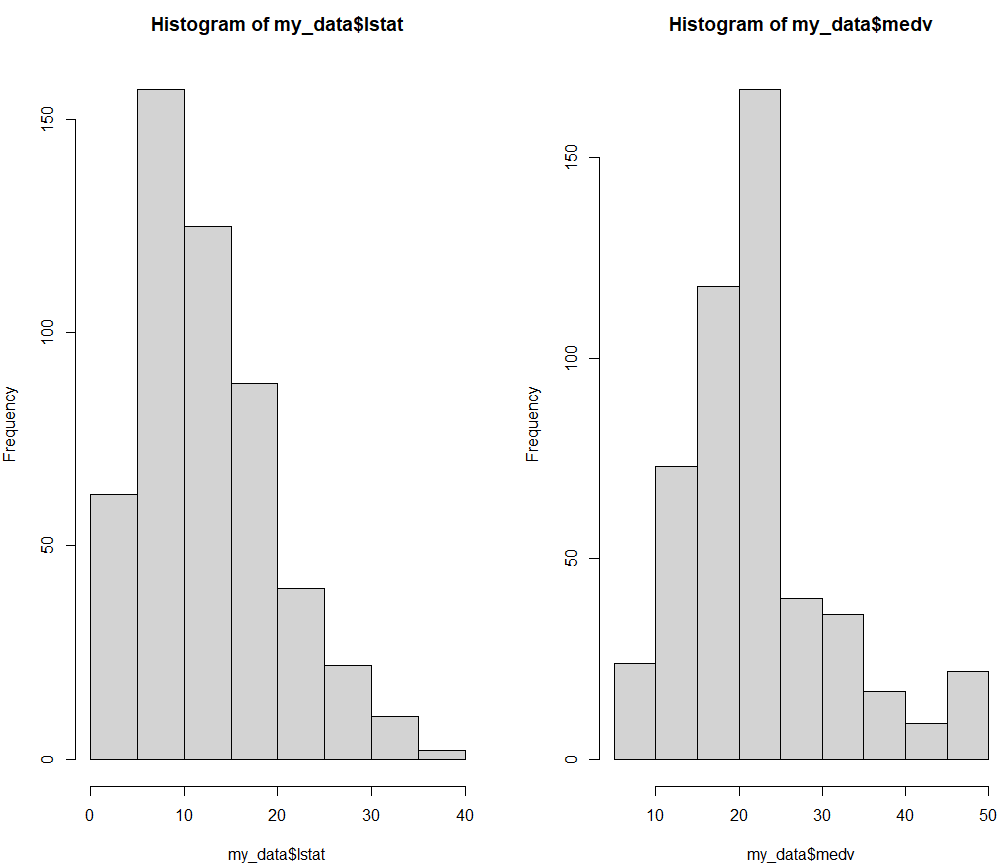
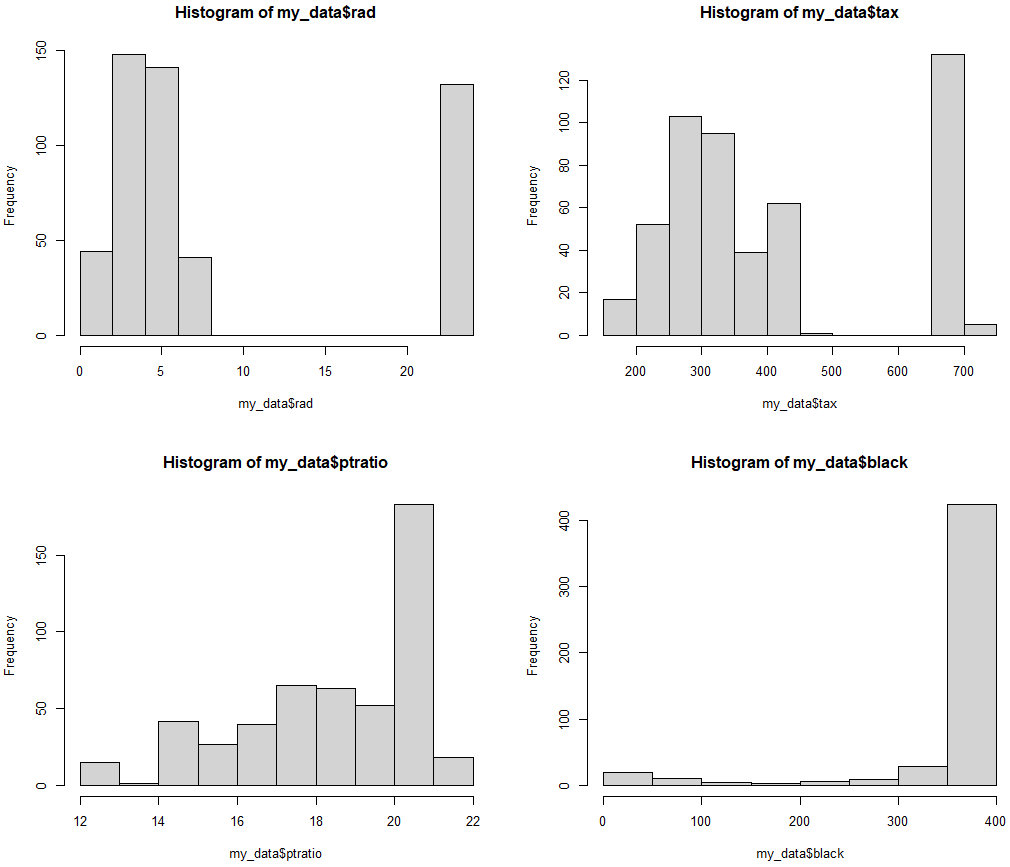
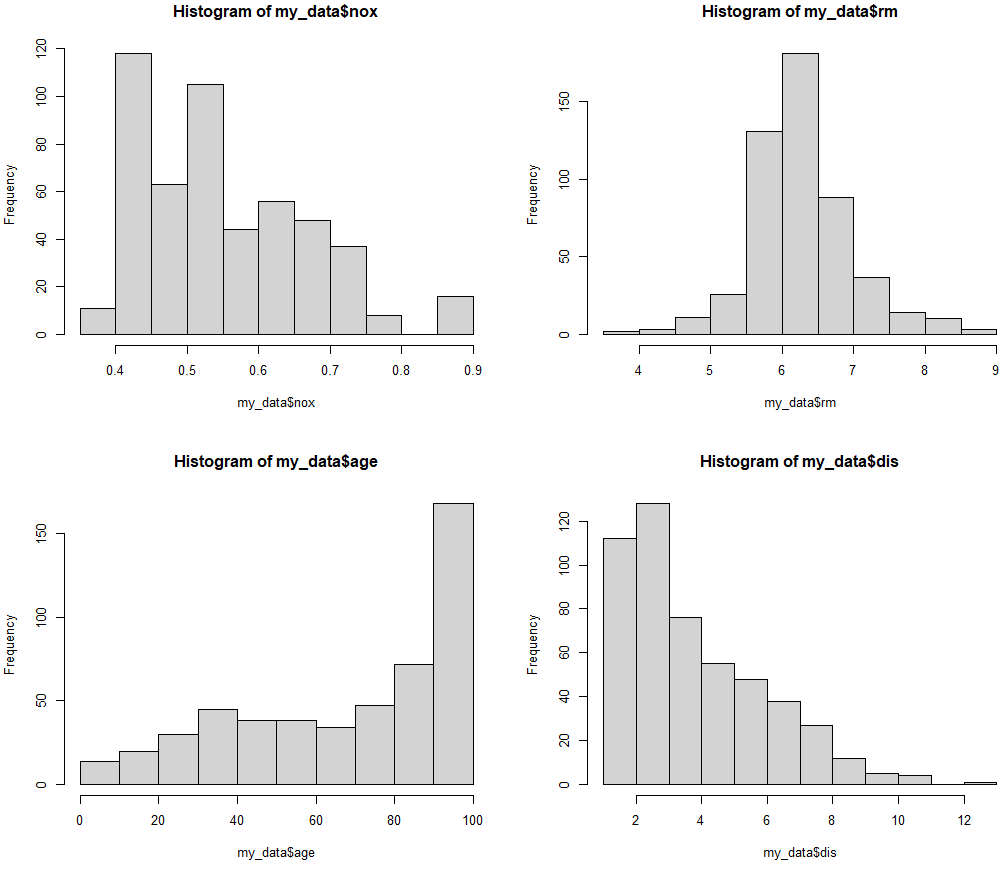
Cp becomes zero when nsplit is 8~10. Therefore, misplit for rpart is 9.



2.

a)





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"

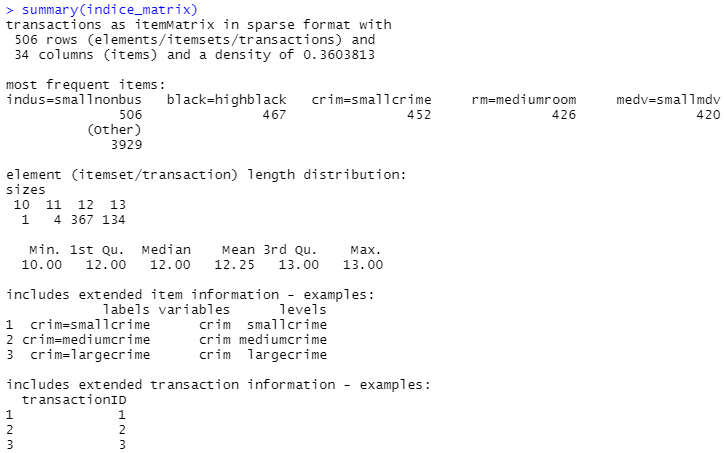
ptratio = classified into three categories: "lowtcher", "mediumtcher", "hightcher"

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

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

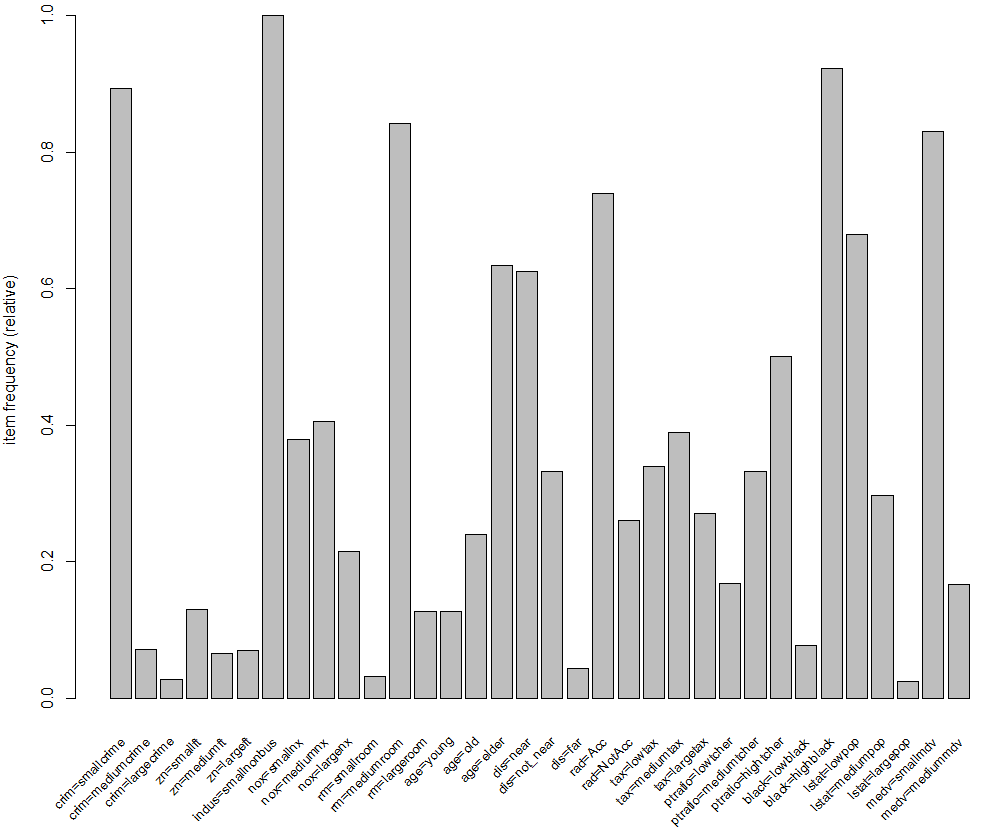
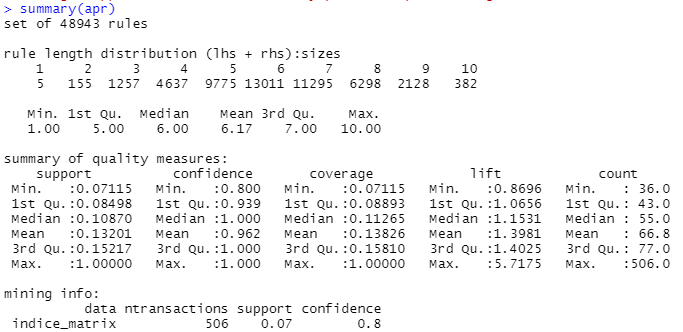
medv = classified into three categories: "smallmdv", "mediummdv", "largemdv"

Moreover, the data is changed to binary incidence matrix



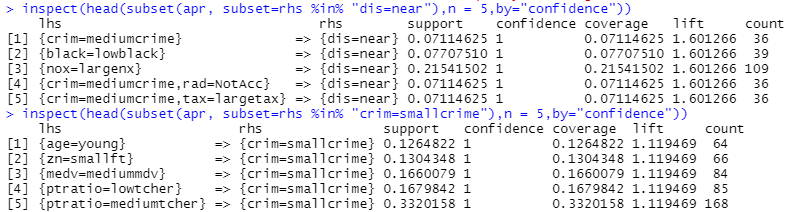
b)

ItemFrequency Plot

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)



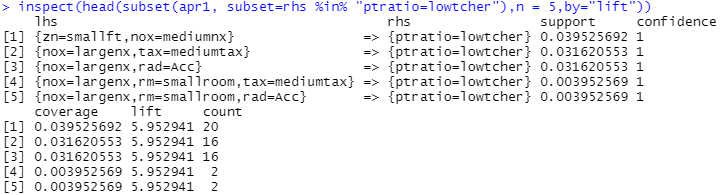
When the distance is near to Boston, crime is medium, low rate of black people, large nitrogen oxides concentration, not accessible to highways, and high tax.

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.

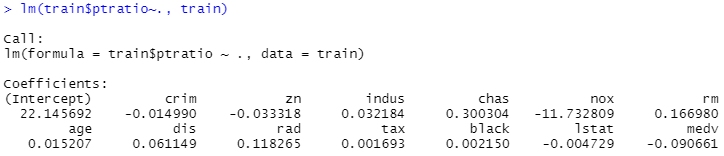
d)

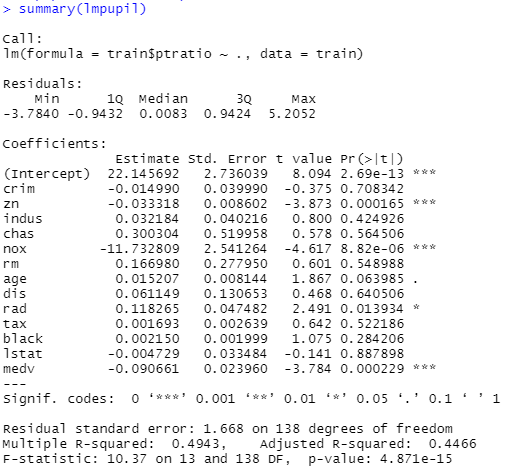
I lowered the apriori method support to 0.01, since 0 rules are shown for pupil teacher, the results are shown as follows.



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





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.