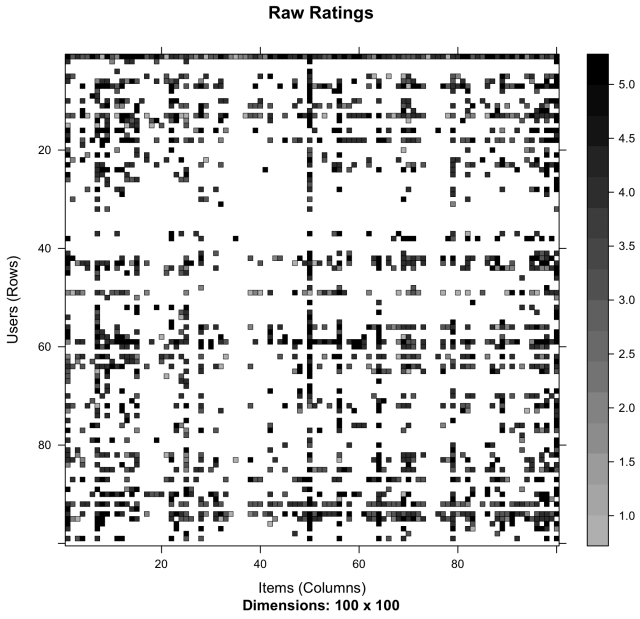
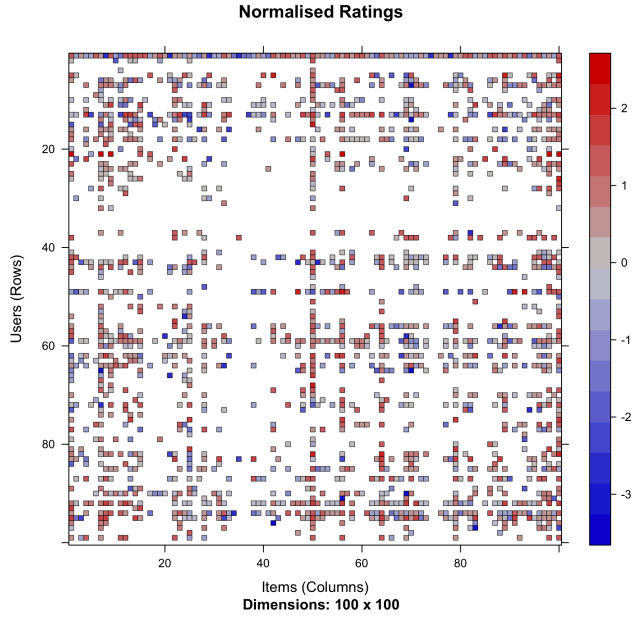
EAS 507
HW -1

By - Mohit Tripathi

Ans-1

Plots of Raw ratings and Normalized Ratings





- 943 users and 1664 movies
- Method used is “UBCF” and similarity function is “Cosine”
- 50 nearest neighbors are used to do the prediction as taking all users will be computationally very difficult.
- Predicted ratings for first 3 users using Rating Matrix

b.)

- Good rating is 3

- Average RMSE 1.16

- Error Values for 5 cross folds

```
[[1]]
       RMSE      MSE       MAE
res 1.13223 1.281945 0.8953529

[[2]]
      RMSE      MSE       MAE
res 1.2373 1.530912 0.9792443

[[3]]
       RMSE      MSE       MAE
res 1.14351 1.307615 0.9045241

[[4]]
        RMSE      MSE       MAE
res 1.181802 1.396655 0.9393876

[[5]]
        RMSE      MSE       MAE
res 1.192379 1.421768 0.9427041
```

Ans-2

a) **UBCF using Pearson Correlation (Normalized Matrix)**

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]
[1,]	0.000000	1.000000	2.00	-1.000000	-2.000000	5.324102
[2,]	0.000000	4.080462	-1.00	3.844854	1.000000	0.000000
[3,]	2.124195	0.750000	1.75	-1.250000	-1.250000	2.753220
[4,]	2.200000	-0.800000	-1.80	1.200000	4.444854	-0.800000
[5,]	-1.600000	2.875376	0.40	-0.600000	-0.600000	2.400000

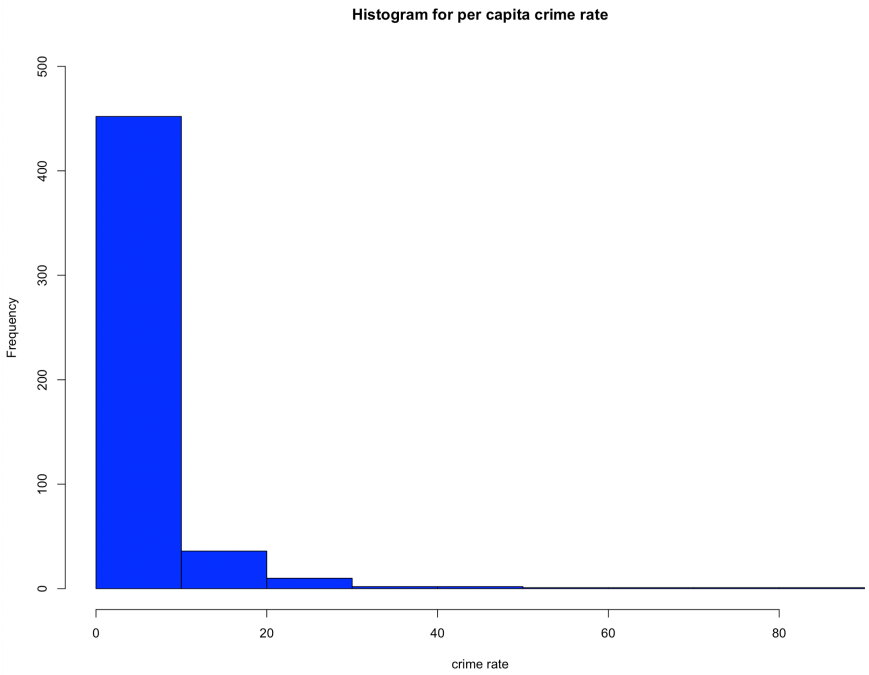
b) **IBCF using Cosine Similarity (Normalized Matrix)**

	1	2	3	4	5	6
[1,]	0.000000	1.000000	2.00	-1.00000	-2.000000	5.062993
[2,]	0.000000	3.999355	-1.00	3.98319	1.000000	0.000000
[3,]	2.328387	0.750000	1.75	-1.25000	-1.250000	2.255409
[4,]	2.200000	-0.800000	-1.80	1.20000	4.454013	-0.800000
[5,]	-1.600000	2.734110	0.40	-0.60000	-0.600000	2.400000

Ans-3

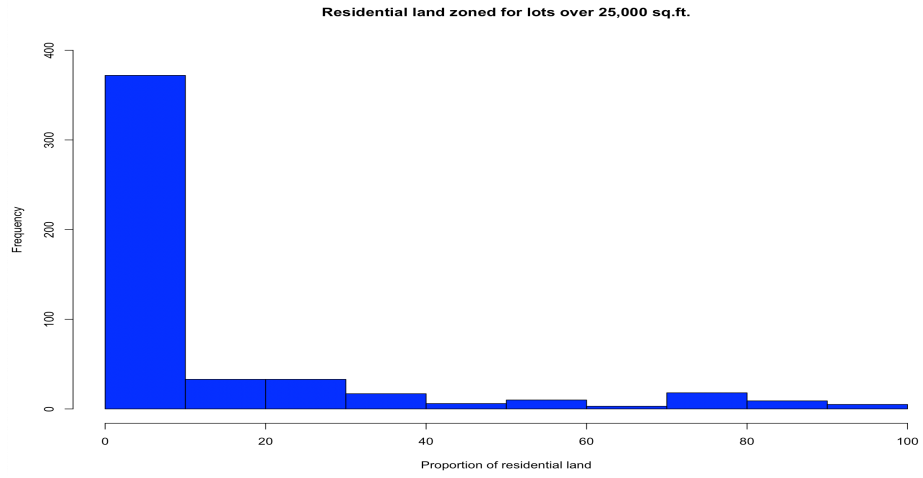
(a) Histograms

Crime Rate



- Crime rate is divided into 3 categories as low rate (0-10), medium (10-30) and high rate (30-90). Generally, if the crime rate per capita is less than 10, we categorize it as a city with low crime rate.

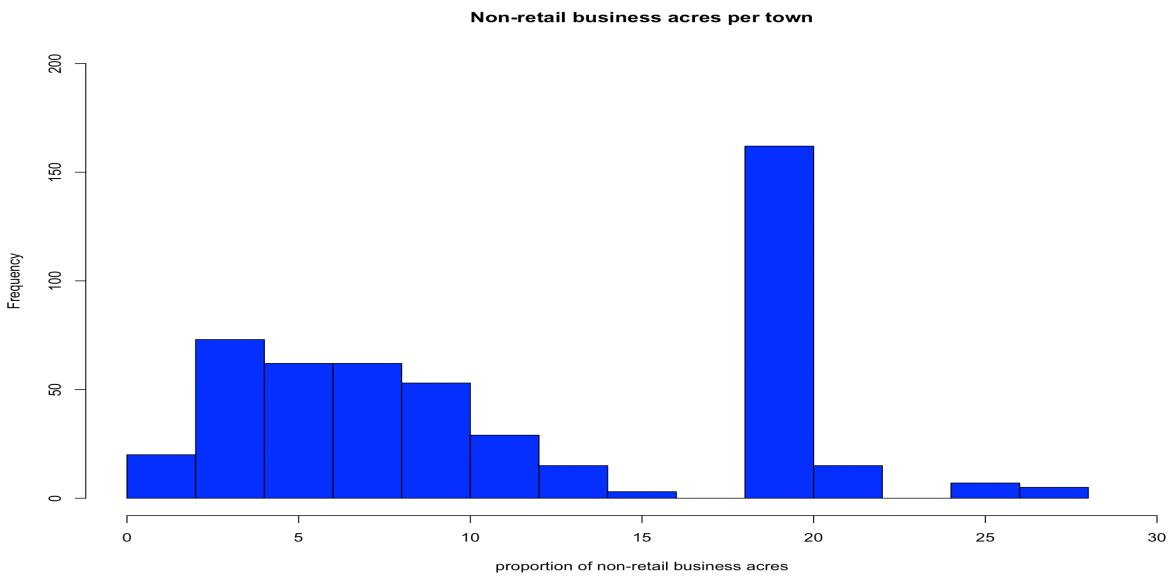
ZN



- Categorized it as small land (0-25), medium land (25-50) and large land (50-100)

Indus

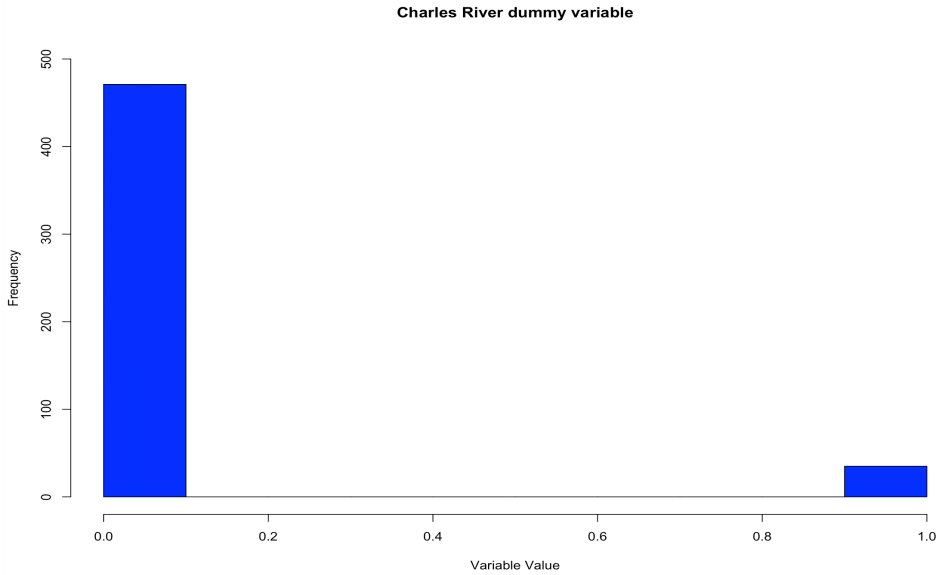
(proportion of non-retail business acres per town)



- Categorized it as low business (0-8), medium business (8-18) and large business (18-28)

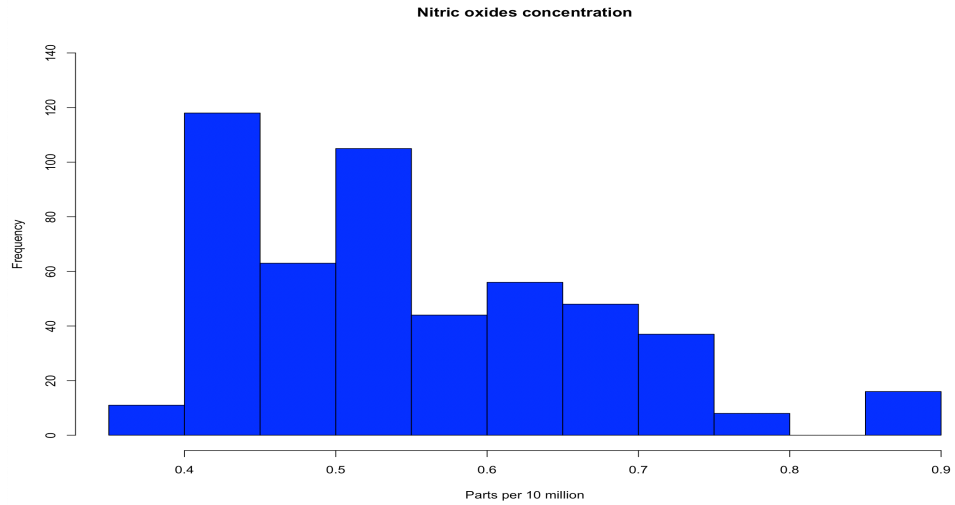
CHAS

Charles River dummy variable



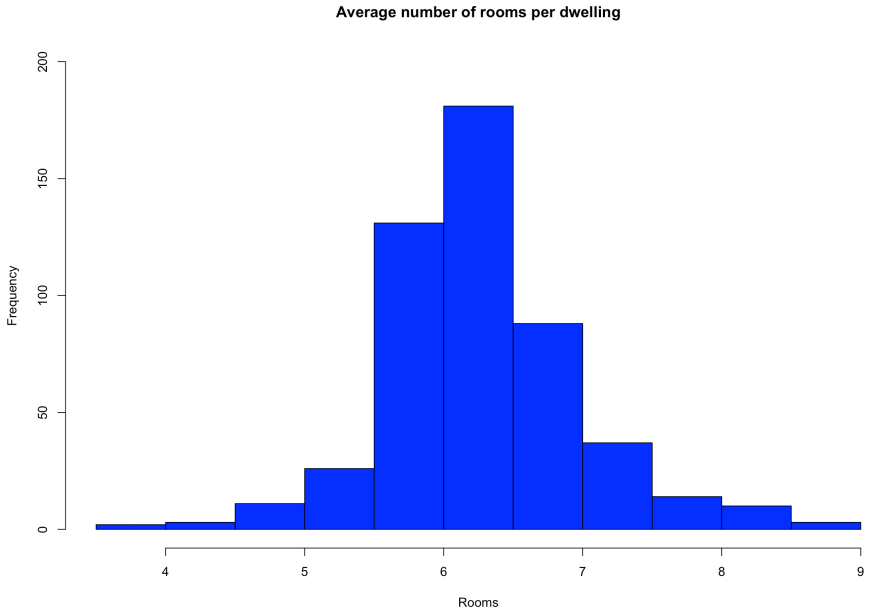
- Categorized 0 as “No” and 1 as “Yes”

NOX
Nitric oxides concentration



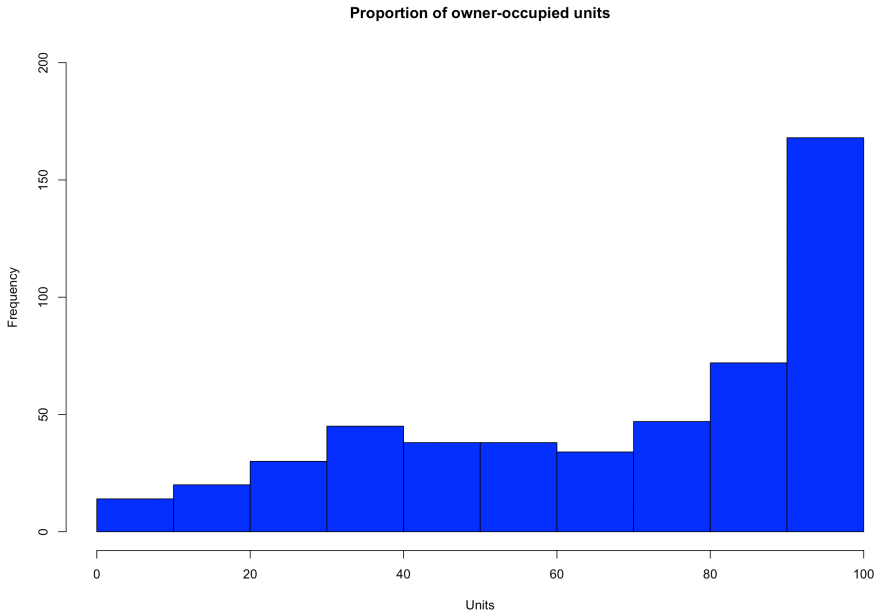
- Categorized it as low NOX rate (0.36,0.56), medium NOX rate (0.56,.68) and high NOX rate (.68,1)

RM
Average number of rooms per dwelling



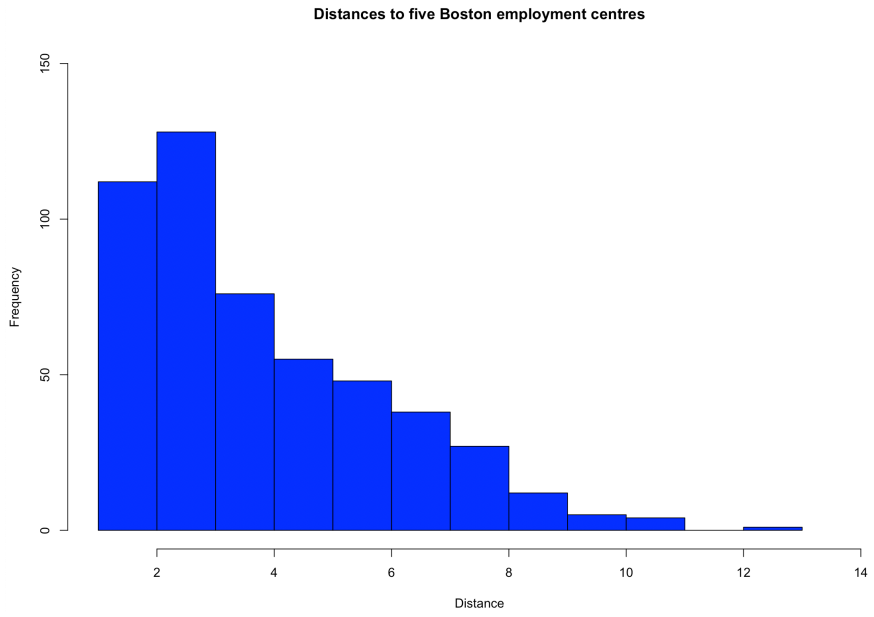
- Categorized it as less rooms (3.5,5.9), medium rooms (5.9,6.9) and high rooms (6.9,9)

AGE
Proportion of owner-occupied units



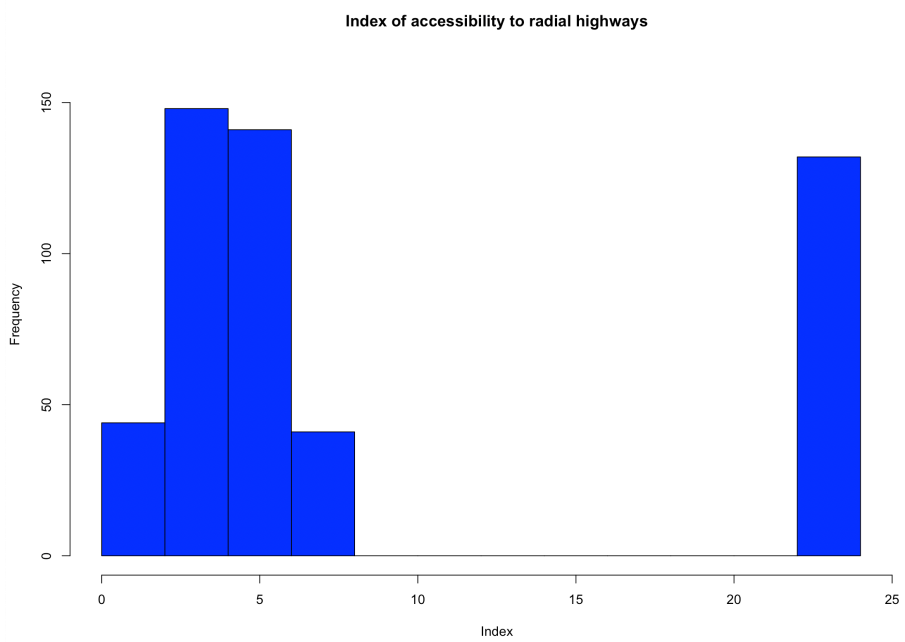
- Categorized it as low units (0,35), medium units (35,70) and high units (70,100)

DIS
Distances to five Boston employment centers



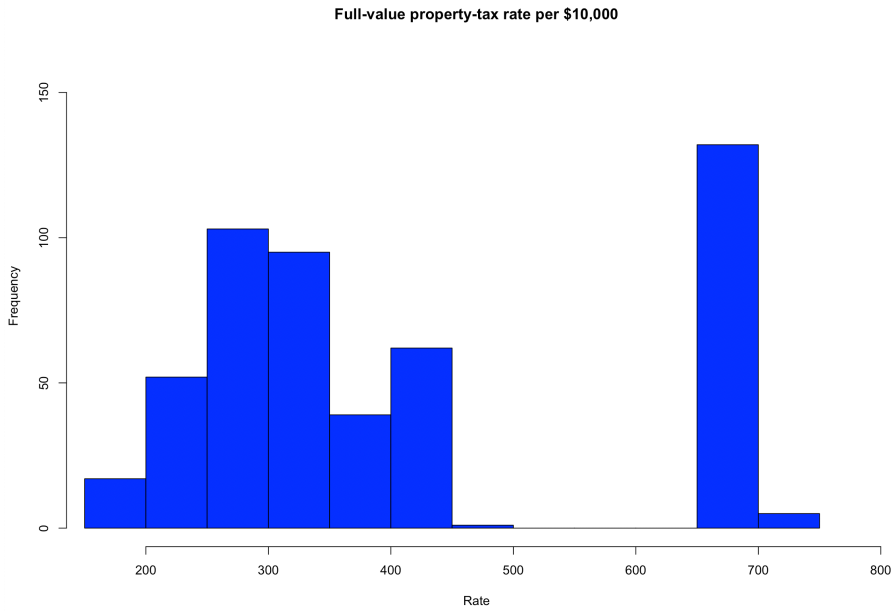
- Categorized it as less distance (1,3), medium distance (3,5.5) and high distance (5.5,13)

RAD
Index of accessibility to radial highways



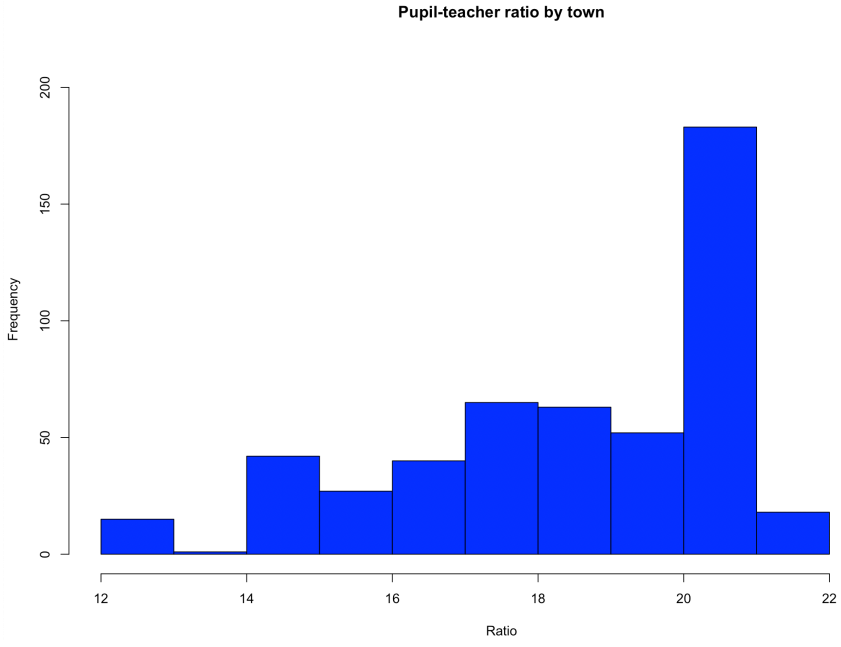
- Categorized it as low index (0,4.5), medium index (4.5,9) and high index (9,25)

TAX
Full-value property-tax rate per $10,000



- Categorized it as low tax (180,300), medium tax (300,500) and high tax (500,750)

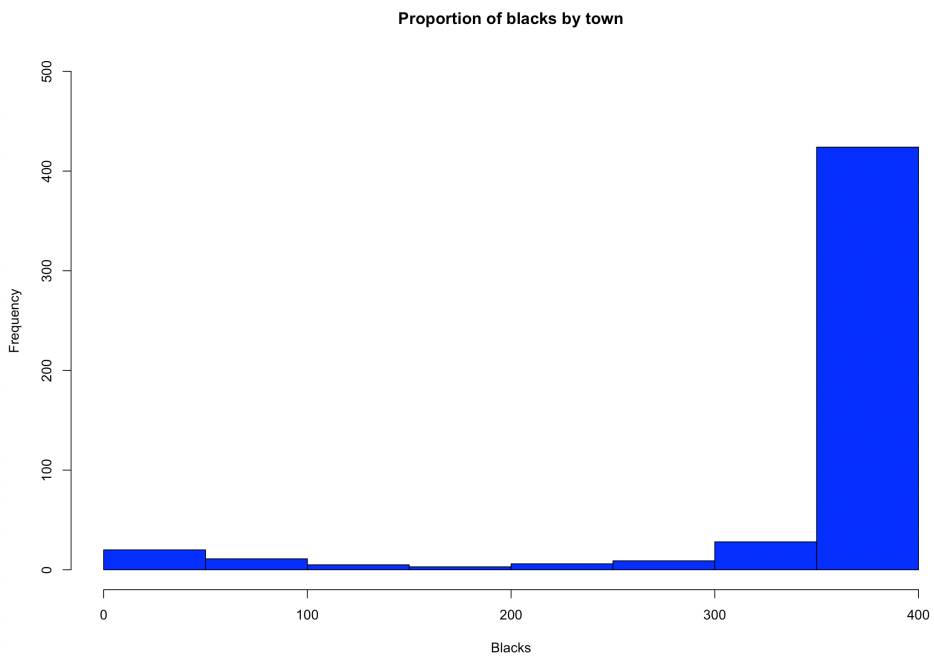
PTRATIO
Pupil-teacher ratio by town



- Categorized it as low ptratio (12,16), medium ptratio (16,20) and high ptratio (20,23)

BLACK

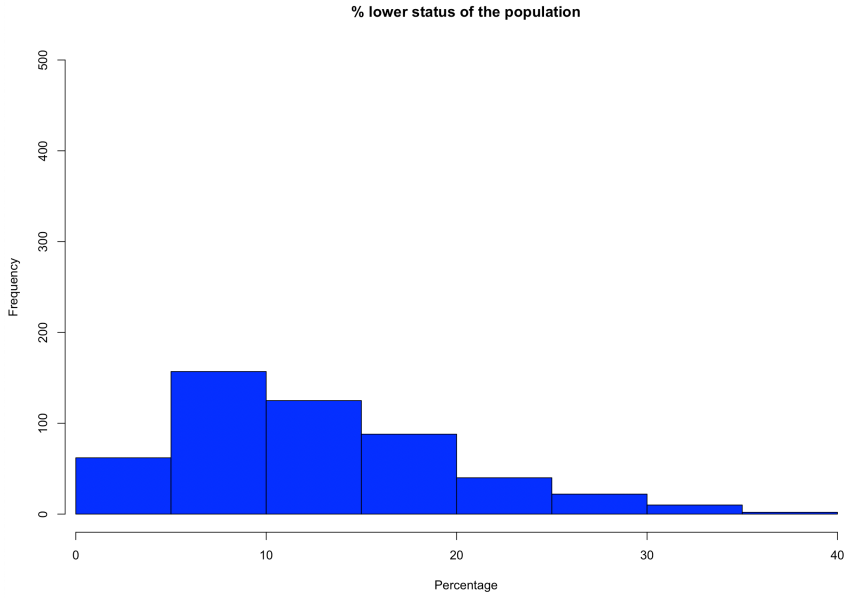
1000(Bk - 0.63)^2 , Bk is Proportion of blacks by town



- Categorized it as low Bk (0,100), medium Bk (100,350) and high Bk (300,400)

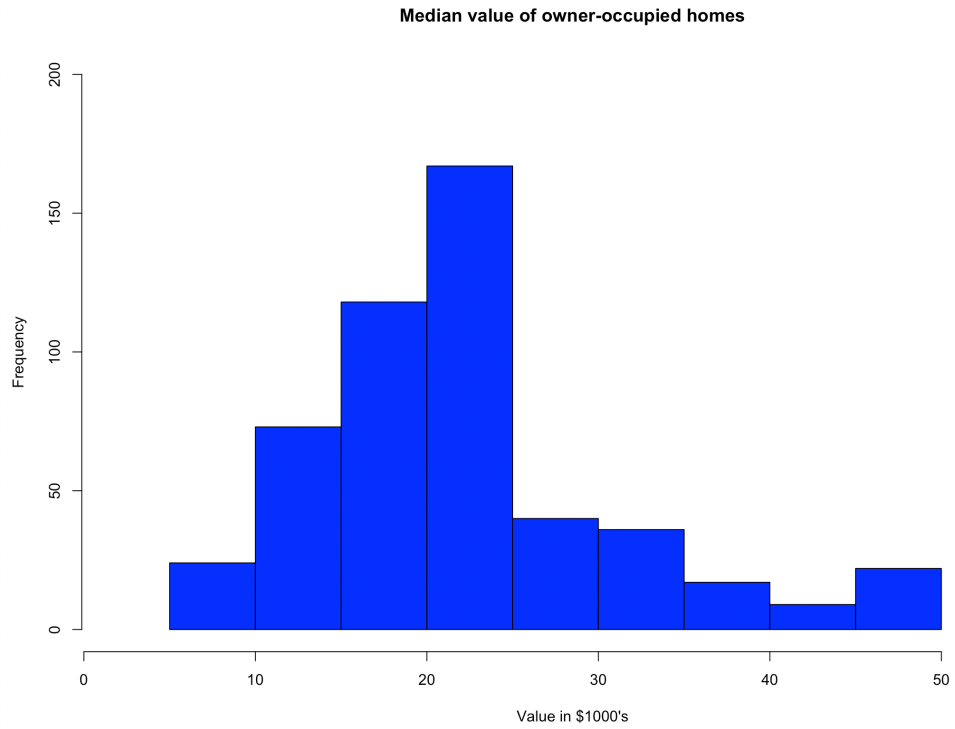
LSTAT

% lower status of the population



- Categorized it as low lstat (0,10), medium lstat (10,17) and high lstat (17,40)

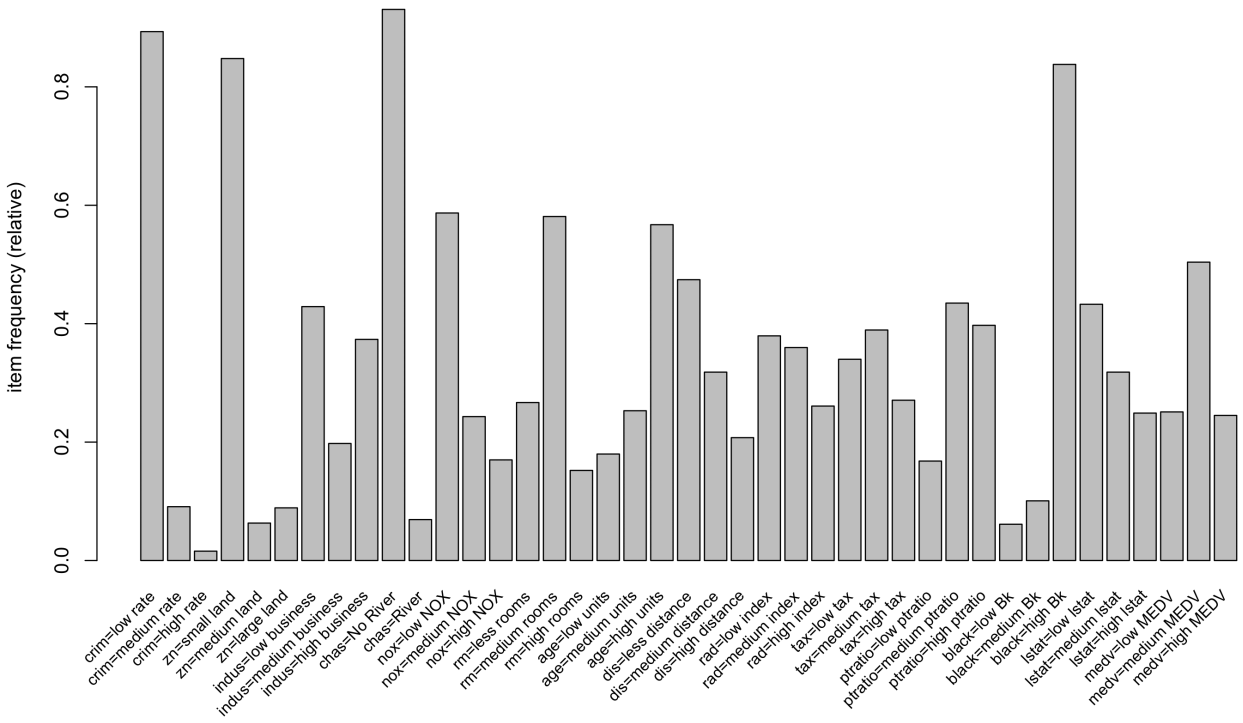
MEDV
Median value of owner-occupied homes



- Categorized it as low MEDV (4,17), medium MEDV (17,25) and high MEDV (25,51)

For most of the variables, categorization was done by checking the Quantile Values using Summary command.

(b) Visualize the data



- Charles River dummy variable(Chas) with value 0 (No River) has the highest frequency(relative)
- Support is .01
- Confidence is 0.6
- Maxlen (Maximum number of items per item set) is 15

(c) Low crime and as close as to city (dis = "less distance"), there are 30186 such rules. Lift is greater than .5.

- After inspecting the subset of rules, we can infer that if a student wants to live in an area with low crime rate and close to city, then he has to stay near Charles river.
- if a student wants to live in an area with low crime rate, then the area will be having low pupil ratio by town.

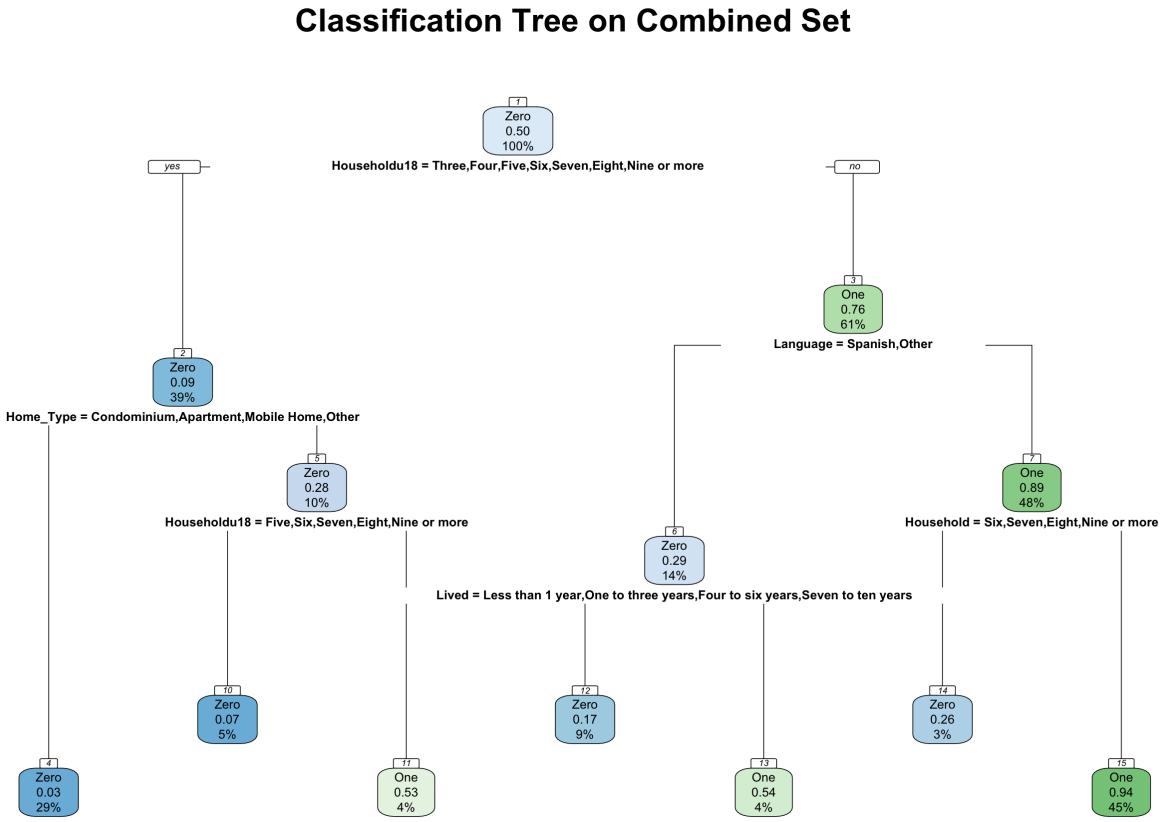
(d) Lift is greater than .5

- Since the family is looking for schools with low pupil-teacher ratio, they have to stay in an area with high NOX (Nitrogen Oxides Concentration) and medium RAD (index of accessibility to radial highways).
- They may stay in an area with high NOX (Nitrogen Oxides Concentration) and medium TAX value (Full Value property tax rate per $10,000)

(e)

- The results are approximately similar with linear regression. Significant variables (by checking p value) includes NOX and RAD with 3 stars. However, Tax is not significant variable to predict PTratio as per lm model.
- Association rules provides much easier interpretation over linear regression.
- Linear regression would be preferred when the target variable is given (supervised problem) and association rules are preferred when the data is not labelled. For higher interpretability, Association rules are better. However, if accuracy is the sole criteria and in we have target variable defined, linear regression is a better choice.

Ans 4.)



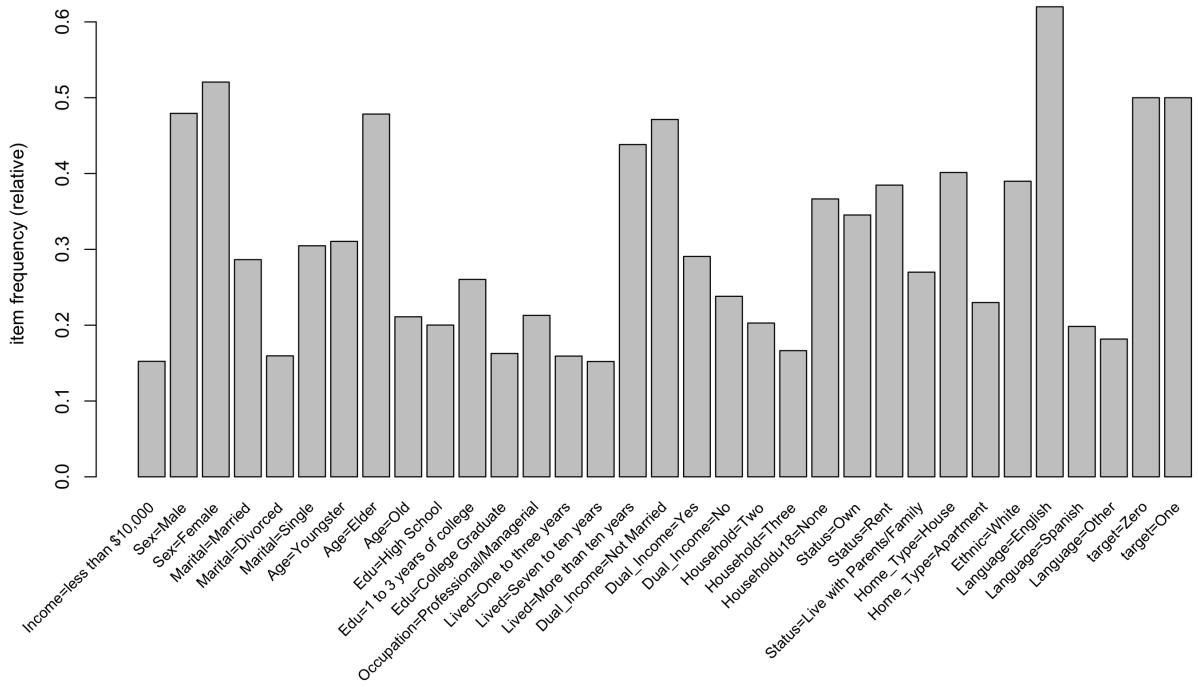
- Important variable after checking the summary of model generated using Rpart are mentioned below

  Variable importance

  Householdu18 Language Household Home_Type Lived Ethnic

- Node 15 is the terminal node having the highest class 1 probabilty (.94%)
- Node number: 15
  root
  Householdu18= None, One, Two (No or 1 or two PERSONS IN HOUSEHOLD UNDER 18)
  Language=English
  Household=One, Two, Three, Four, Five (1 or 2 or 3 or 4 or 5 PERSONS IN YOUR HOUSEHOLD)

Using Apriori Algo (Item Frequency Plot)



- Language English has the highest frequency(relative)

- Target Class 1 has Household18 value as None as per 1st rule
- Target Class 1 has Language as English for 2nd rule.

We can say that the result using apriori algo is relatable with the important variables we got from classification tree.