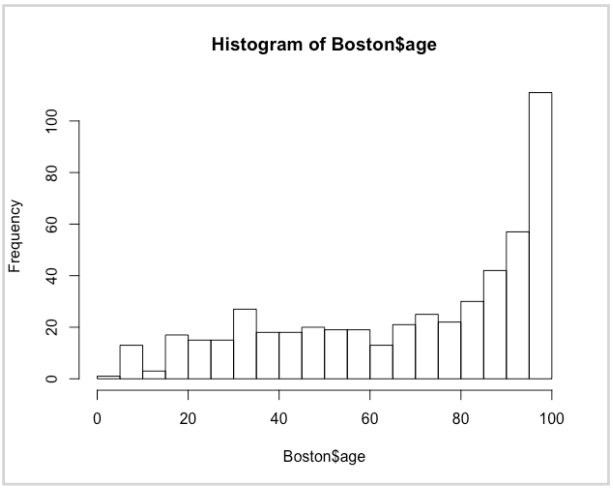
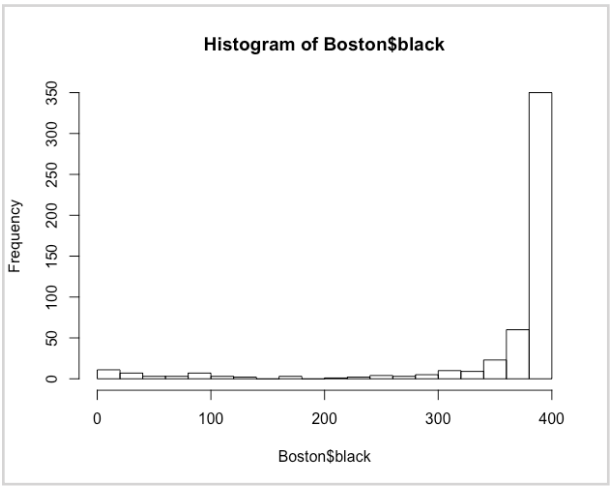
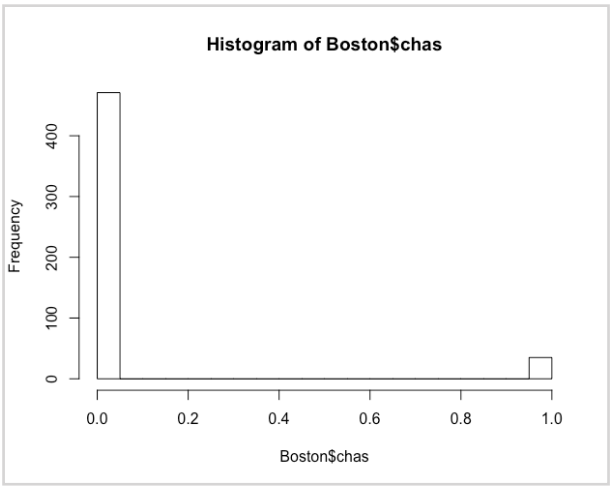
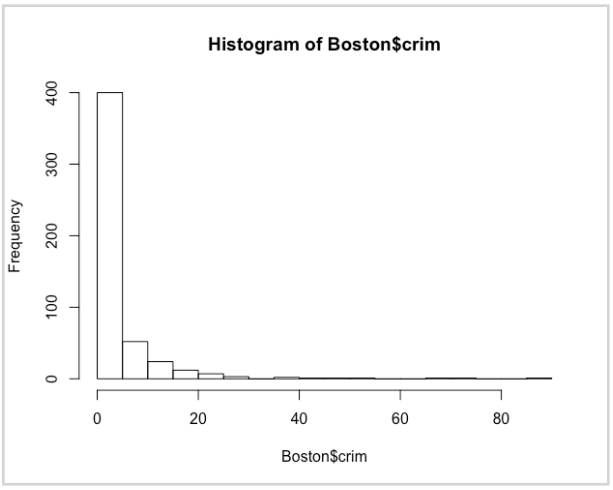
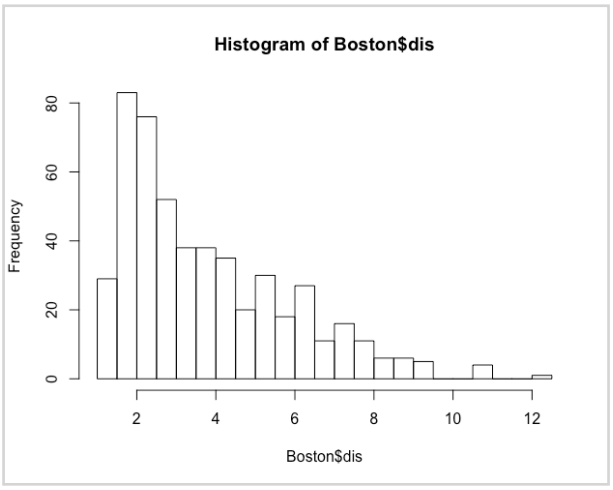
2. (a) Visualize the data using histograms of the different variables in the data set.

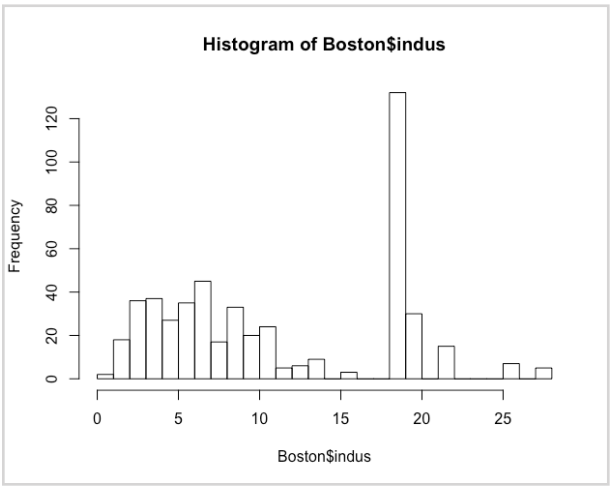


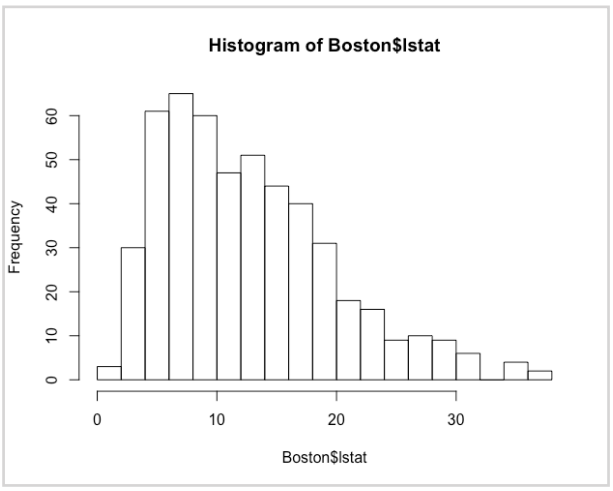


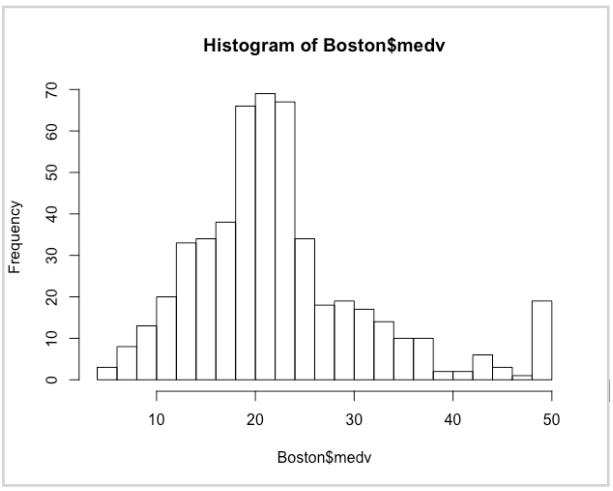


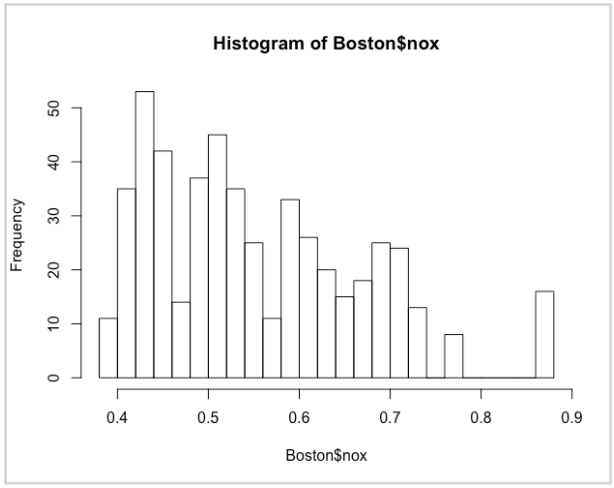


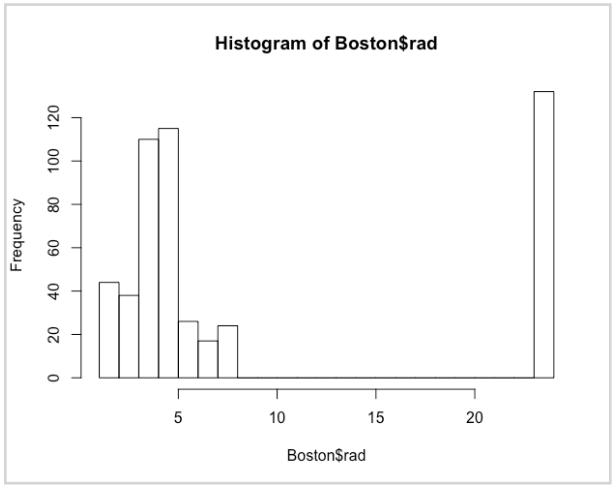


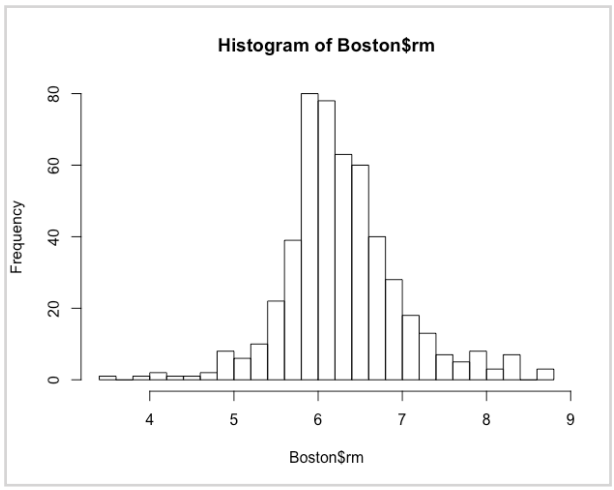


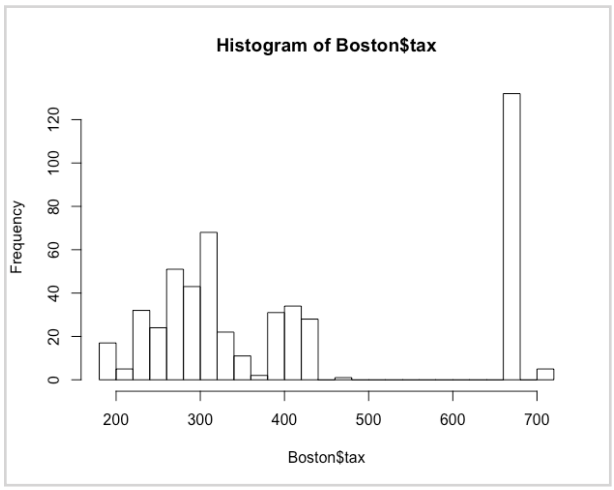


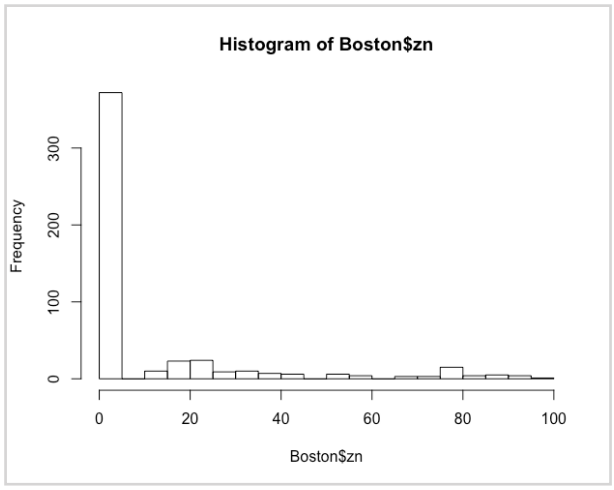












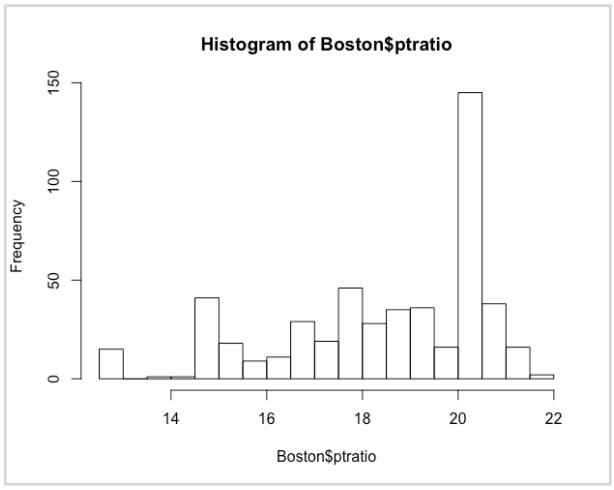


Fig 1~14. Histograms of the different variables of Boston

Transform the data into a binary incidence matrix, and justify the choices you make in grouping categories.

I used quartile 0%, 50%,100% to classify the data to two parts for each variable first, then itemFrequencyPlot will display un-recognized plot(all bars are equal), so I re-classified the group. The following is the rule I made for grouping categories:

1. CRIM: Almost 90% of the data is less than 5, so I cut the point in 0, 3, 5 and 100.
    - Grouping:0~3(Low), 3~5(Mid) and 5~100 (High)
    - Quantile: 0% 50% 75% 80% 90% 100%
      Val : 0.006320 0.256510 3.677083 5.581070 10.753000 88.976200

2. ZN: Most of the data is 0, so there is a category "Not-above" for it.
    - Grouping:-1~12.5(Not-above), 12.5~60(small), 60~80(Mid) and 80-100 (Large)

3. Indus: Separate the range into 3 part.
    - Grouping:0~10(Low), 10~20(Mid) and 20~28 (High)

4. CHAS: I eliminate the variable, because there is only two types of value.

5. NOX: Separate the range into 3 part.
    - Grouping:0.3~0.5(Low), 0.5~0.8(Mid) and 0.8~0.9 (High)

6. RM: Separate the range into 3 part.
    - Grouping:0~5(Small), 5~7(Mid) and 5~9 (Large)

7. AGE: Separate the range into 3 part.
    - Grouping:0~40(Youth), 40~80(Senior) and 80~101 (Elderly)

8. DIS: Separate the range into 3 part.
    - Grouping:0~4(Low), 4~28(Mid) and 8~13 (High)

9. RAD: Eliminate variable due to polarize.

10. TAX: Separate the range into 3 part.
    - Grouping:180~300(Low), 300~500(Mid) and 500~720 (High)

11. PTRATIO: Separate the range into 3 part.
    - Grouping:12~16(Low), 16~20(Mid) and 20~23 (High)

12. BLACK: Eliminate variable due to highly concentrative.

13. LSTAT: Separate the range into 3 part.
    - Grouping:1~10(Low), 10~20(Mid) and 20~38 (High)

14. MEDV: Separate the range into 3 part.
    - Grouping:0~20(Low), 20~40(Mid) and 40~51 (High)

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

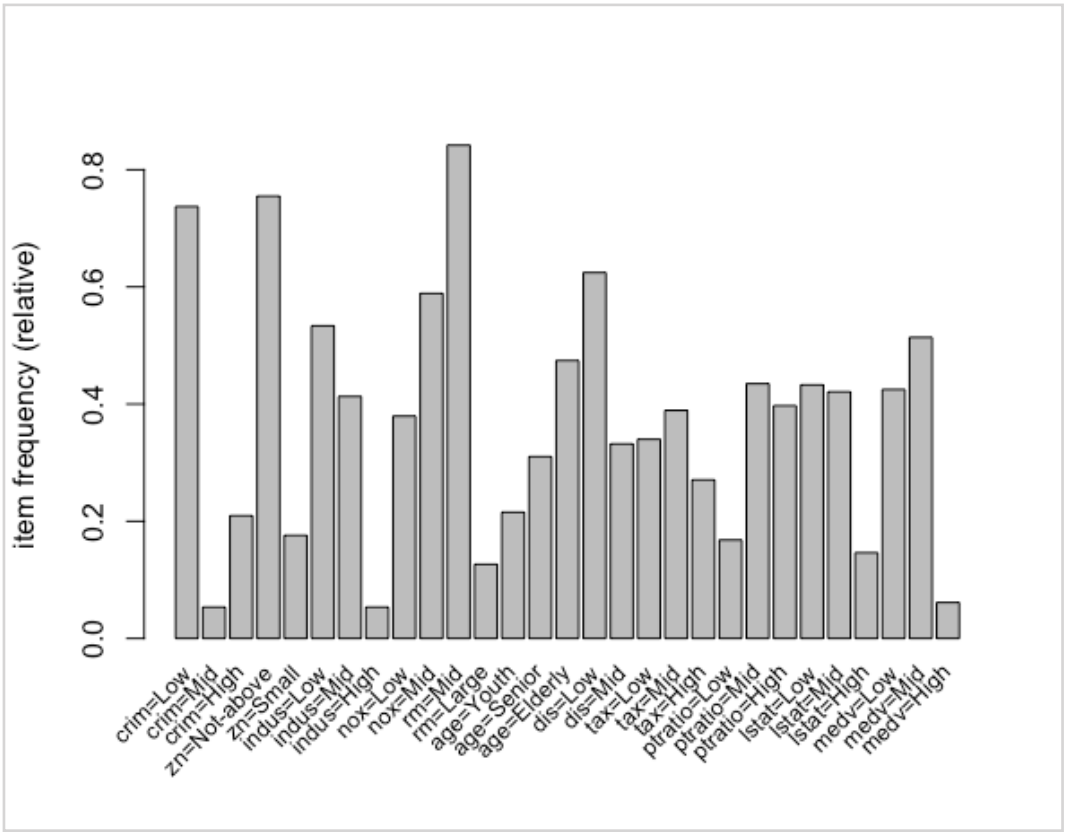


Fig 15. itemFrequencyPlot with support 0.05

I generated itemFrequencyPlot with the support = 0.2 and 0.1, but only several bars would be displayed, so I chose to use support 0.05.

```
summary of quality measures:
    support          confidence          lift             count
 Min.   :0.01186   Min.   :0.8000   Min.   : 0.9502   Min.   :  6.00
 1st Qu.:0.01581   1st Qu.:0.9091   1st Qu.: 1.3566   1st Qu.:  8.00
 Median :0.02174   Median :1.0000   Median : 1.6980   Median : 11.00
 Mean   :0.03844   Mean   :0.9578   Mean   : 2.0019   Mean   : 19.45
 3rd Qu.:0.03953   3rd Qu.:1.0000   3rd Qu.: 2.3105   3rd Qu.: 20.00
 Max.   :0.84190   Max.   :1.0000   Max.   :31.6250   Max.   :426.00
```

Fig 16. Summary of quality measures with Support = 0.01, Confidence = 0.8

For the Apriori, lift is 0.71 with support = 0.01, confidence = 0.6; lift is 0.71 with support = 0.01, confidence = 0.6, after several tests, I found that the highest lift is 0.95 when support > 0.01 and confidence > 0.8.

(c) A student is interested is a low crime area, but wants to be as close to the city as possible (as measured by "dis"). What can you advise on this matter through the mining of association rules?

	lhs		rhs	support	confidence	lift	count
[1]	{indus=High,dis=Low}	=>	{crim=Low}	0.05335968	1	1.356568	27
[2]	{zn=Small,dis=Low}	=>	{crim=Low}	0.04150198	1	1.356568	21
[3]	{age=Youth,dis=Low}	=>	{crim=Low}	0.02371542	1	1.356568	12
[4]	{dis=Low,tax=Low}	=>	{crim=Low}	0.14426877	1	1.356568	73
[5]	{nox=Low,dis=Low}	=>	{crim=Low}	0.07707510	1	1.356568	39
[6]	{dis=Low,ptratio=Mid}	=>	{crim=Low}	0.19565217	1	1.356568	99
[7]	{indus=Low,dis=Low}	=>	{crim=Low}	0.19565217	1	1.356568	99
[8]	{indus=High,dis=Low,lstat=High}	=>	{crim=Low}	0.01383399	1	1.356568	7
[9]	{indus=High,dis=Low,tax=Low}	=>	{crim=Low}	0.01383399	1	1.356568	7
[10]	{indus=High,dis=Low,tax=Mid}	=>	{crim=Low}	0.02964427	1	1.356568	15

Fig 17. Low dis and crime associations table (sorted by lift)

Fig 17. shows the top ten rules with low crime and low dis. I would advise the student to give priority to the high indus and low dis areas, because these areas have the highest probability to have low crime.

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

	lhs		rhs	support	confidence	lift	count
[1]	{nox=High}	=>	{ptratio=Low}	0.03162055	1	5.952941	16
[2]	{nox=High,lstat=High}	=>	{ptratio=Low}	0.01383399	1	5.952941	7
[3]	{nox=High,tax=Mid}	=>	{ptratio=Low}	0.03162055	1	5.952941	16
[4]	{indus=Mid,nox=High}	=>	{ptratio=Low}	0.03162055	1	5.952941	16
[5]	{nox=High,lstat=Mid}	=>	{ptratio=Low}	0.01581028	1	5.952941	8
[6]	{nox=High,medv=Low}	=>	{ptratio=Low}	0.02766798	1	5.952941	14
[7]	{nox=High,age=Elderly}	=>	{ptratio=Low}	0.03162055	1	5.952941	16
[8]	{nox=High,dis=Low}	=>	{ptratio=Low}	0.03162055	1	5.952941	16
[9]	{crim=Low,nox=High}	=>	{ptratio=Low}	0.02569170	1	5.952941	13
[10]	{zn=Not-above,nox=High}	=>	{ptratio=Low}	0.03162055	1	5.952941	16

Fig 18. Low ptratio associations table (sorted by lift)

Fig 18. shows the top ten rules in low ptratio. It is better to choose the high nox areas, because these areas have the highest probability to have low ptratio.

(e) 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?

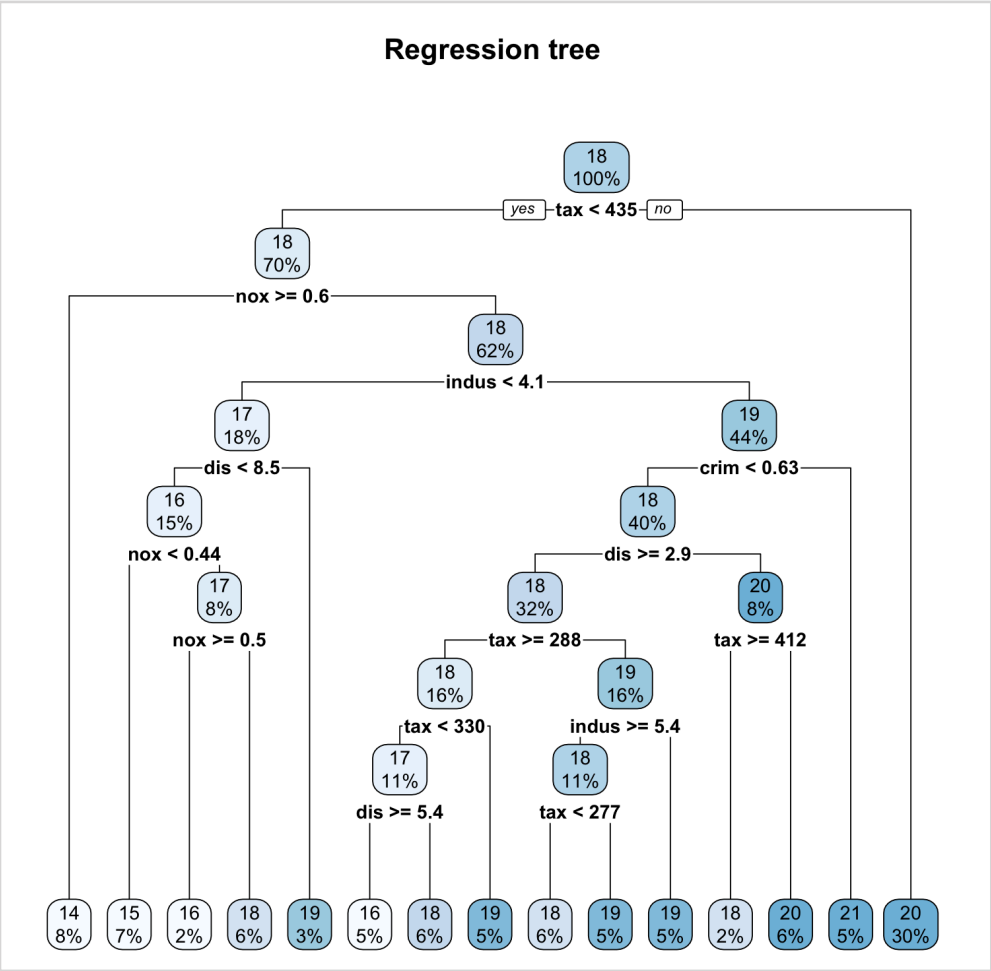


Fig 19. Regression tree with ptratio

Regression tree can still show the relationship between ptratio and other variables, but the data of ptratio has been separated into many terminal nodes, it makes the result be more complicated than in association model.
I believe that regression tree is more suitable for data set with binary variables, because there are only two branches of each node, the results can be easily analyzed; association models are more suitable for judging the relationship of multiple subset combinations.

3. (Modified Exercise 14.4 in ESL) Cluster the demographic data (>data(marketing in ESL package)) of Table 14.1 using a classification tree. Specifically, generate a reference sample the same size as the training set, by randomly permuting the values within each feature. 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.

A. I created a random reference sample, and added a column class as 0, also added a column as 1 into training sample (original data).
B. Before generating classification tree, I set rpart fit with cp = 0 to show all the result. I used the default value of minsplit because the tree is already huge.
C. For pruning the classification tree, I set cp to the minimum value of xerror in the data.
D. The followings are the full tree and pruned tree:

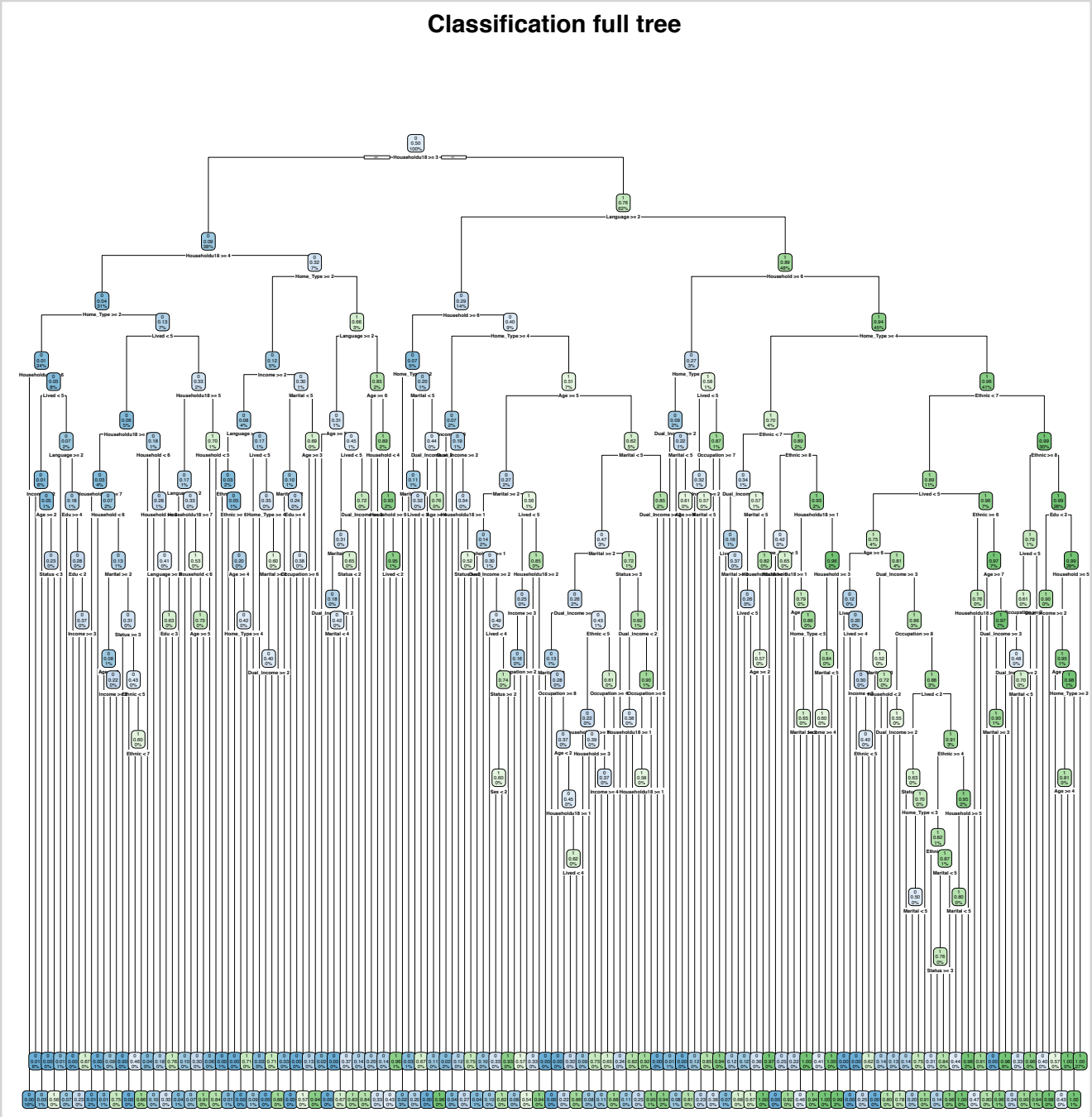


Fig 20. Classification full tree

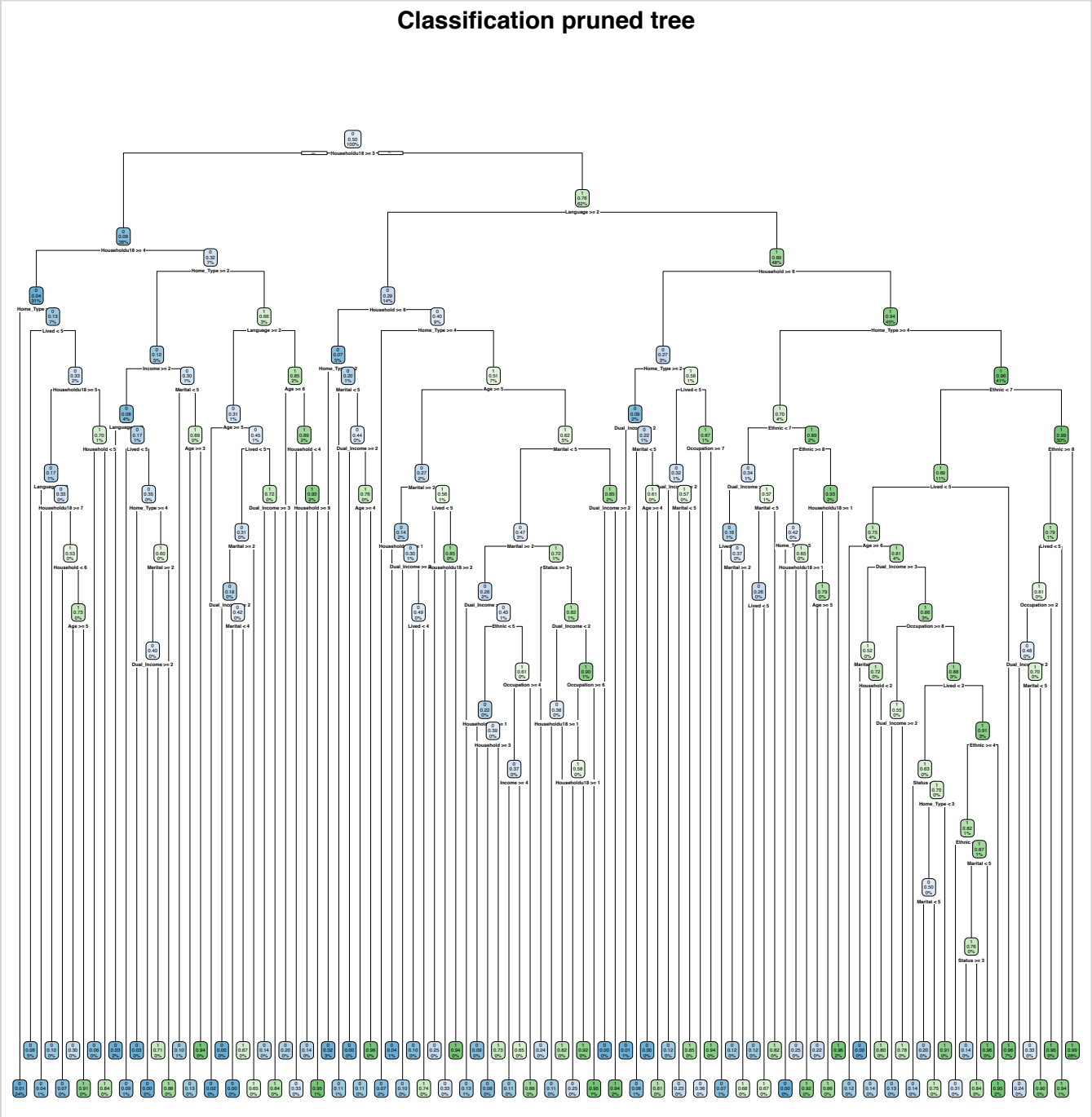


Fig 21. Classification pruned tree

E. Result: According to the confusion matrix (Fig 22), this model have high predictive power, it have 0.967 accuracy to distinguish class 1. Also, the accuracy of pruned tree is 0.965, it is slightly smaller than full tree, so pruning seems useless in this model.

	predict_class	
	0	1
0	8711	282
1	310	8683

Fig 22. Confusion matrix (full tree)

	predict_class	
	0	1
0	8690	303
1	325	8668

Fig 23. Confusion matrix (pruned tree)