Statistical Data Mining Project 1

Question 1

In this problem we had to consider the following as a "utility matrix"

	a	b	c	d	e	f	g	h
A	4	5		5	1		3	2
B		3	4	3	1	2	1	
C	2		1	3		4	5	3

Part A

In this problem we had to treat the utility matrix as Boolean as shown below:

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    1    0    1    1    0    1    1
[2,]    0    1    1    1    1    1    1    0
[3,]    1    0    1    1    0    1    1    1
```

We had to compute the Jaccard distance between users A (v1), B (v2) and C (v3) as shown below:

Metric: 'jaccard"; comparing; 3 vectors

```
    v1  v2  v3
v1 0.0 0.5 0.5
v2 0.5 0.0 0.5
v3 0.5 0.5 0.0
```

We had to compute the cosine between users as shown below

Metric: 'cosine"; comparing; 3 vectors

```
          v1        v2        v3
v1 1.0000000 0.6666667 0.6666667
v2 0.6666667 1.0000000 0.6666667
v3 0.6666667 0.6666667 1.0000000
```

Part B

In this problem we had to use a different discretization: treat ratings 3, 4, 5 as 1, and ratings 1, 2, and blank as 0.

```
     [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8]
[1,]    1    1    0    1    0    0    1    0
[2,]    0    1    1    1    0    0    0    0
[3,]    0    0    0    1    0    1    1    1
```

We had to compute the Jaccard distance between users as shown below:

```
          v1        v2        v3
v1 0.0000000 0.6000000 0.6666667
v2 0.6000000 0.0000000 0.8333333
v3 0.6666667 0.8333333 0.0000000
```

We had to compute the cosine between users as shown below:

Metric: 'cosine"; comparing; 3 vectors

```
          v1        v2        v3
v1 1.0000000 0.5773503 0.5000000
v2 0.5773503 1.0000000 0.2886751
v3 0.5000000 0.2886751 1.0000000
```

Comparing part a and b, we can see that discretizing the matrix gives better results than using the matrix as Boolean. The Boolean matrix made the Jaccard distance and cosine similarity equal between A and B, A and C, B and C users. However, when we discretized the matrix, the jaccard and cosine similarity were different between the users. Using the Jaccard distance we can see that B and C distance is the greatest, then A and C and then A and B. Using the cosine rule we can see that A and B are the greatest distance then A and C and then the B and C user pair.

Part C

In this problem we had to normalize the matrix by subtracting from each nonblank entry the average value for its user.

```
          [,1]      [,2]      [,3]      [,4]      [,5]       [,6]       [,7]      [,8]
[1,]  0.6666667 1.6666667  0.000000 1.6666667 -2.333333  0.0000000 -0.3333333 -1.333333
[2,]  0.0000000 0.6666667  1.666667 0.6666667 -1.333333 -0.3333333 -1.3333333  0.000000
[3,] -1.0000000 0.0000000 -2.000000 0.0000000  0.000000  1.0000000  2.0000000  0.000000
```

Using this matrix we had to compute the cosine similarity between each pair of users, as shown below.

Metric: 'cosine"; comparing; 3 vectors

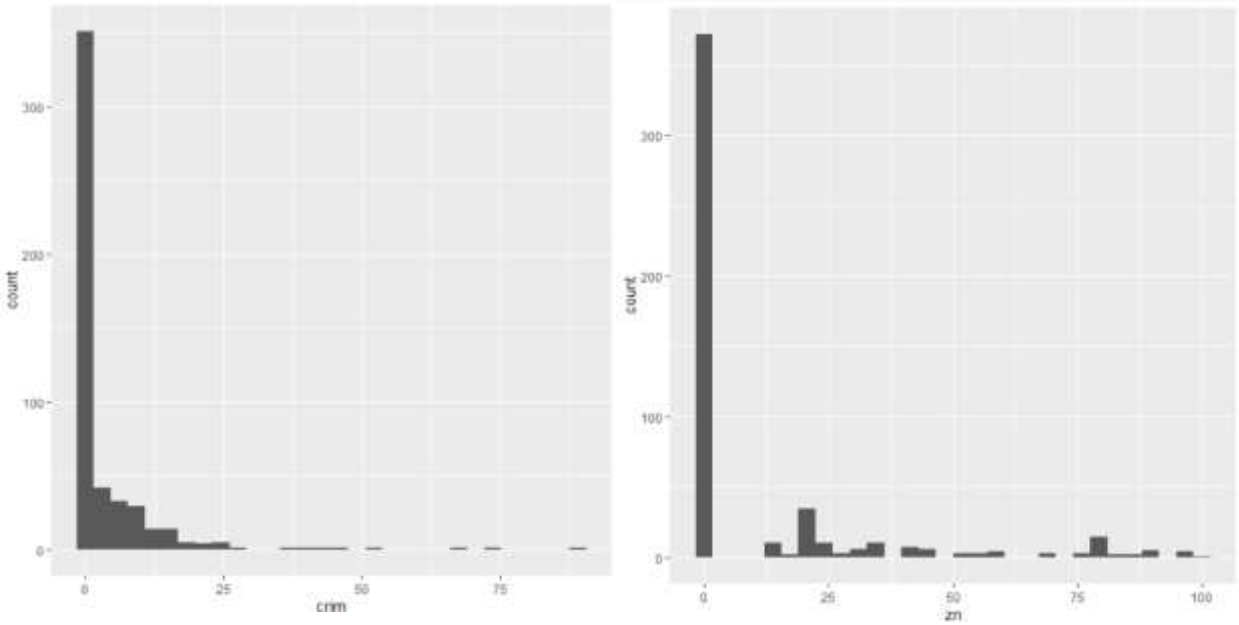
```
           v1         v2         v3
v1  1.0000000  0.5843065 -0.1154701
v2  0.5843065  1.0000000 -0.7395740
v3 -0.1154701 -0.7395740  1.0000000
```

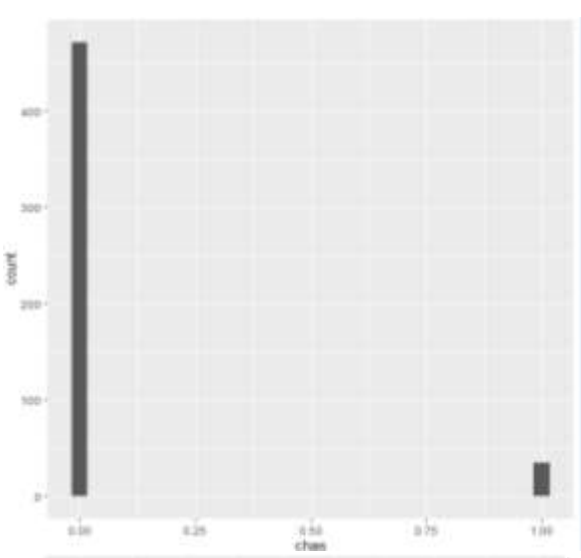
From the cosine similarity we can see that the distance between A and B is less than the distance between A and C, which is less than the distance between B and C.

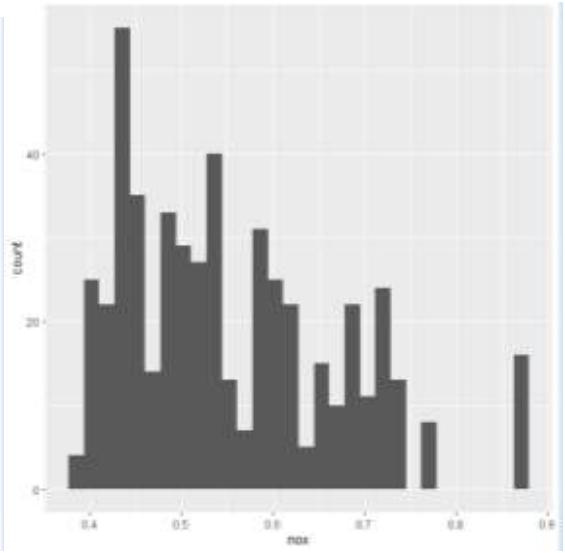
Question 2

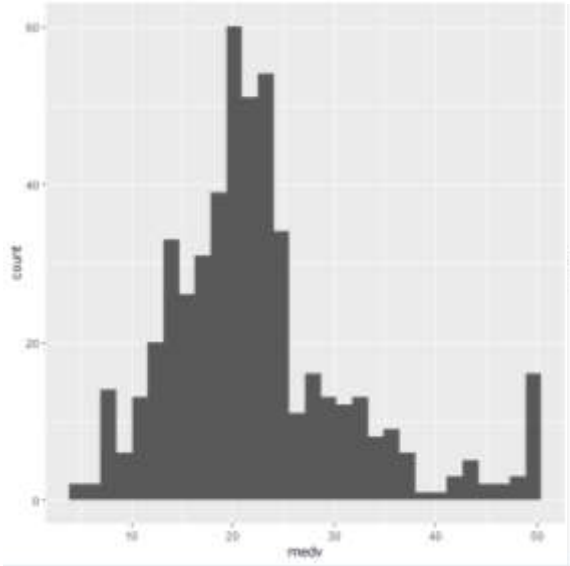
Part 2a)

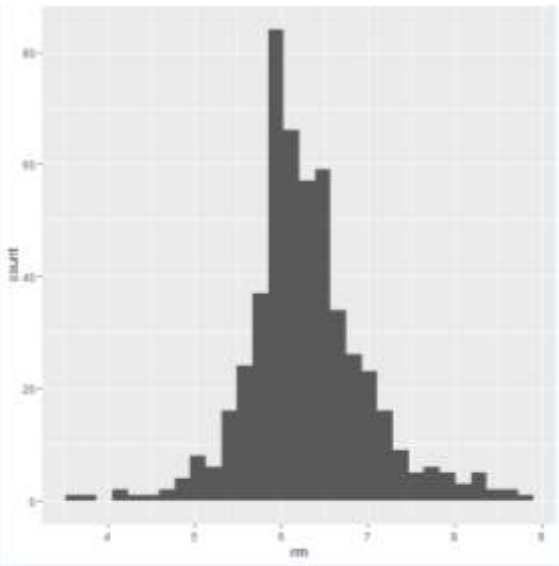
In this problem we had to consider the Boston Housing Data, which was accessed in the MASS package. The Boston data frame has 506 rows and 14 columns. We had to visualize the data using histograms of the different variables in the data set.

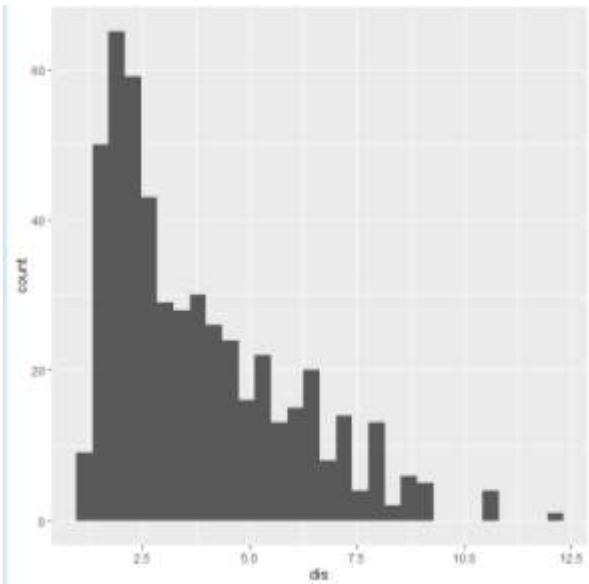

300
200
100
0
count
0
25
50
75
crim
300
200
100
0
count
0
25
50
75
100
zn

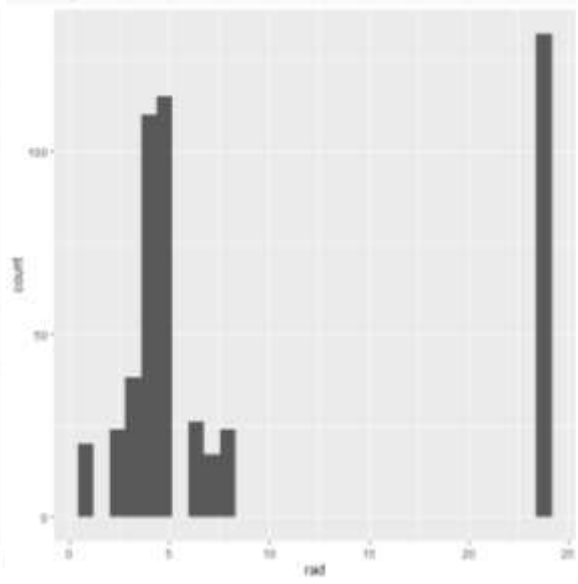

count
chas

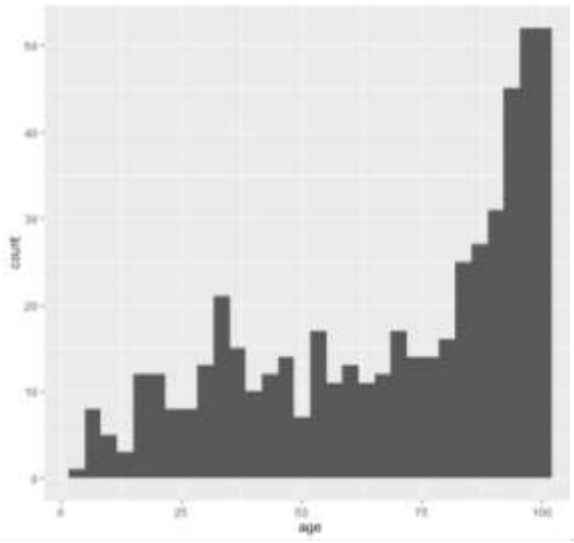

count
nox

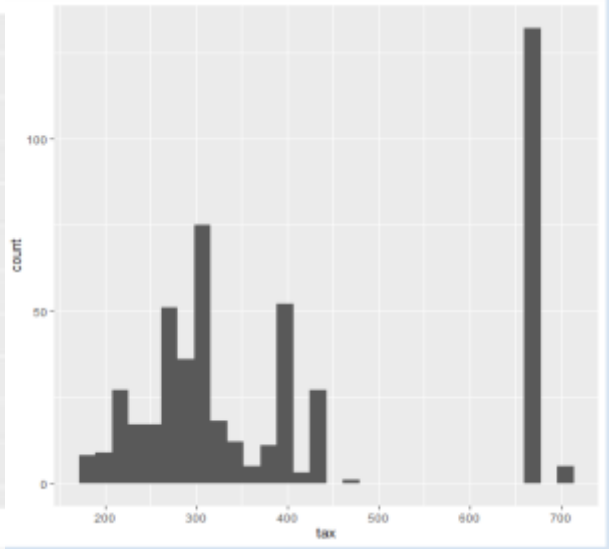

count
medv

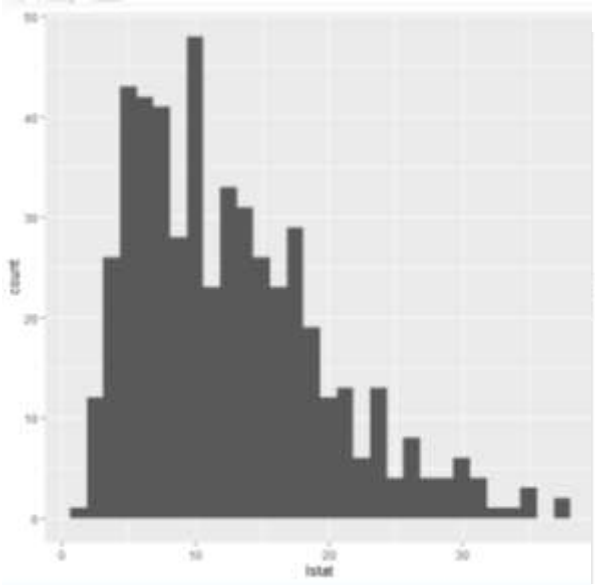

count
rm

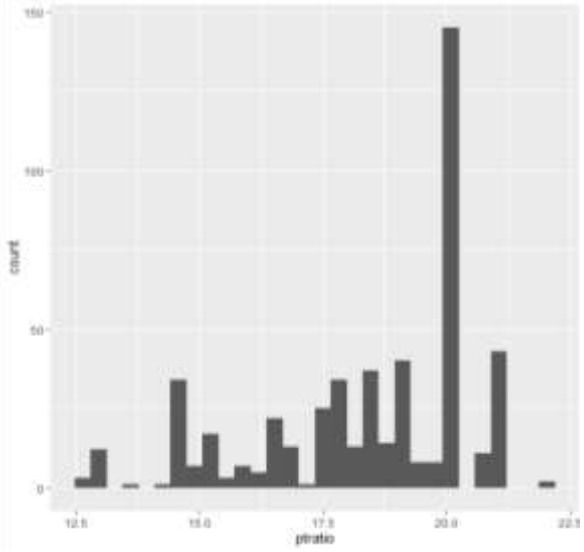

count
dis

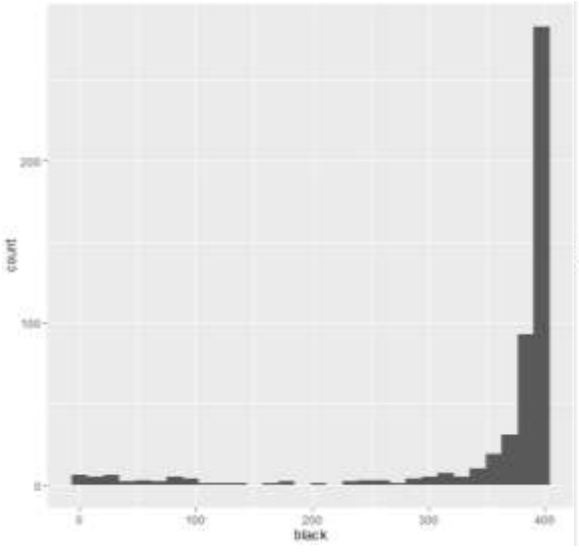

count
rad

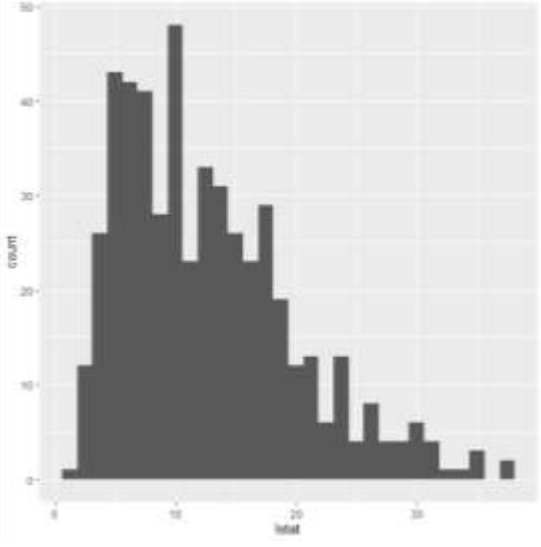

count
age


count
100
50
0
200
300
400
500
600
700
tax


count
lstat


count
150
100
50
0
12.5
15.0
17.5
20.0
22.5
ptratio


count
200
100
0
0
100
200
300
400
black


count
50
40
30
20
10
0
0
10
20
30
lstat

We had to later transform the data into a binary incidence matrix, and by categorizing the variables in grouping categories. From the above histograms we can see that chas is the only variable which is already discretized into 0 and 1, thus we kept it the same. However, the other variables have a continuous output and thus we had to discretize them into a categorical output.

Discretizing the variables was performed by determining the minimum, mean and maximum output of the variables. Thus, the variables where discretized into the categories. For example, the nox (nitric oxide concentration in parts per million) variable was discretized based on minimum= 0 mean=0.5 and maximum=0.8 nox values as shown in figure 1. This provided us with two class categories “Low nox” and “High nox” as shown below in figure 2. The discretizing also ensured that the data was distrubted fairly equally into each cateogry so that all the nox categories survive in the rules and are not dropped out.

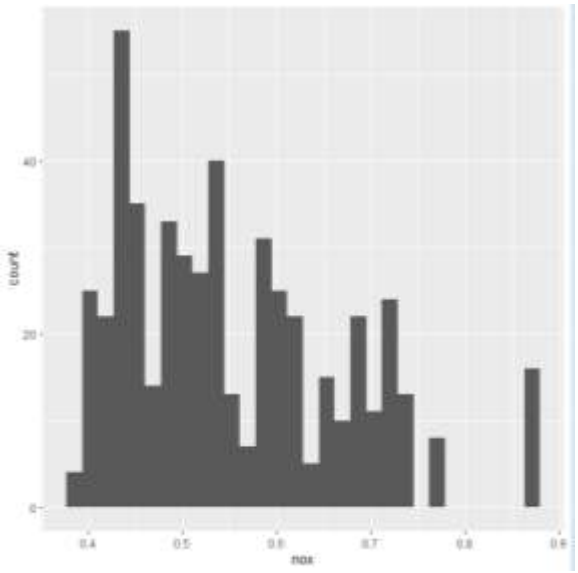


Figure 1: A histogram of the nox continuous variable

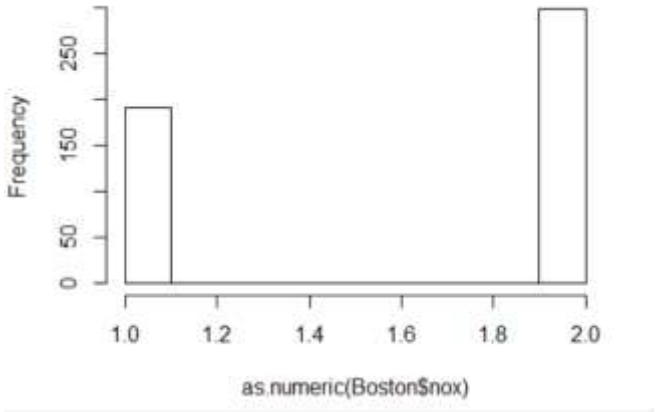


Figure 2: A histogram of the nox discretized variable with 2 categories of “Low nox” and “High nox”

Another example of the variable which was discretized was crime rate. As we can see from figure 3, crim (crime per capita crime rate per town) is a continous varibale and it had to discretized. The variable was discrtiezed using minium=0, median=0.2 and maxium=90 value for crime. This helped to discretize data and produced the catgeroies "Low crime " and "High crime" as shown in figure 3. The discretizing also ensured that the data was distrubted fairly equally into each crime cateogry so that all the crime categories survived in the rules and were not dropped out.

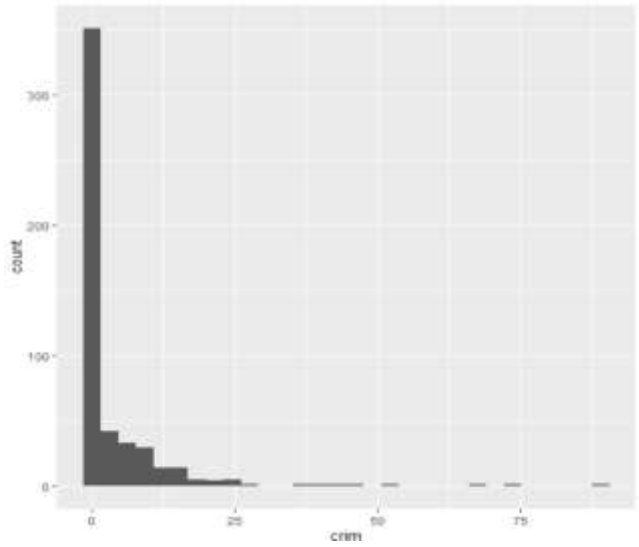


Figure 3: A histogram of the crim continuous variable

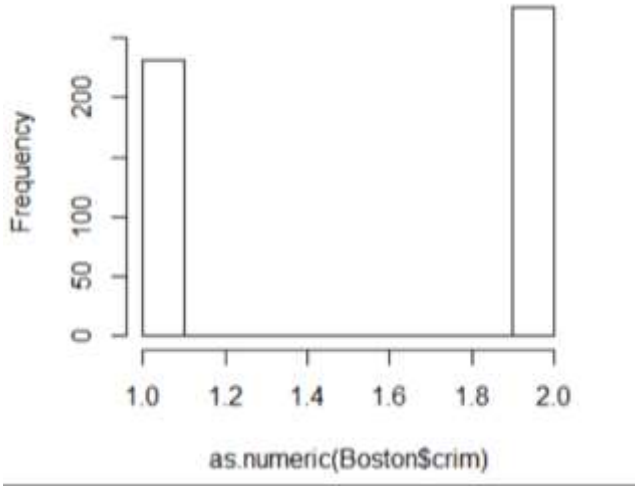


Figure 4: A histogram of the crim discretized variable with 3 categories "Low crime " and "High crime"

Part 2b)

In this problem we had visualize the data using the item Frequency Plot in the “arules” package and we had to apply the apriori algorithm. Based on the algorithm there was a total of 35018 rules involving the 14 predictors with a support of 5% as the parameters. We can see the relative frequency in the association rules from the relative frequency plot in figure 5. From the item Frequency Plot, we can see that the 14 predictors survived and were not dropped out and thus the frequency and rare categories were considered given the 5% support.

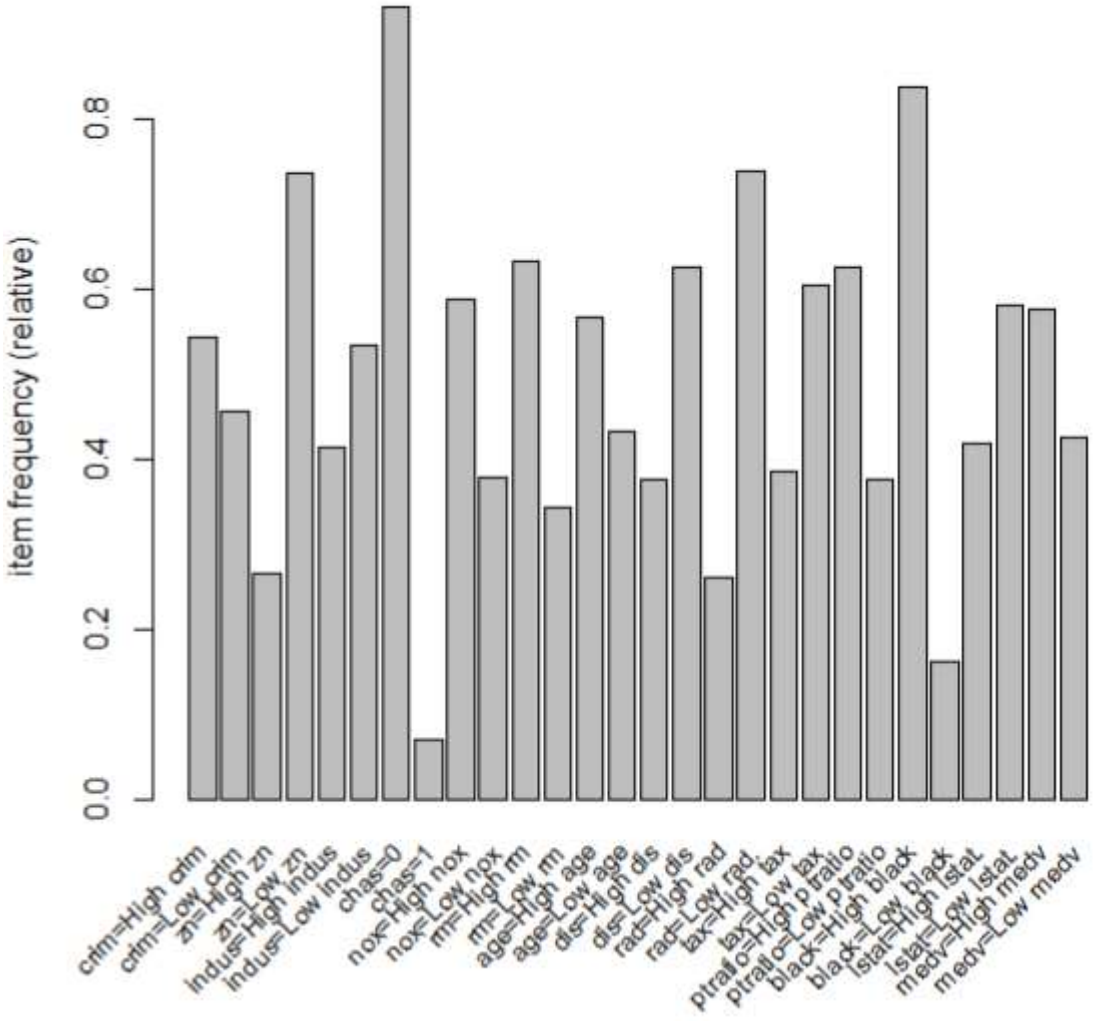


Figure 5: Item Frequency plot for the association rules of the Boston data

Part 2c)

In this problem we had to advise a student who is interested in a low crime area as close to the city as possible (as measured by "dis). We had to advise him/her on this matter through the mining of association rules, shown below.

```
    lhs                    rhs              support confidence     lift count
[1] {crim=Low crim,
     zn=Low zn,
     nox=High nox,
     age=High age}      => {dis=Low dis} 0.06916996          1 1.601266    35
[2] {crim=Low crim,
     zn=Low zn,
     nox=High nox,
     age=High age,
     rad=Low rad.}      => {dis=Low dis} 0.06916996          1 1.601266    35
[3] {crim=Low crim,
     zn=Low zn,
     nox=High nox,
     age=High age,
     black=High black} => {dis=Low dis} 0.06324111          1 1.601266    32
[4] {crim=Low crim,
     zn=Low zn,
     chas=0,
     nox=High nox,
     age=High age}      => {dis=Low dis} 0.06521739          1 1.601266    33
[5] {crim=Low crim,
     zn=Low zn,
     nox=High nox,
     age=High age,
     rad=Low rad.,
     black=High black} => {dis=Low dis} 0.06324111          1 1.601266    32
[6] {crim=Low crim,
     zn=Low zn,
     chas=0,
     nox=High nox,
     age=High age,
     rad=Low rad.}      => {dis=Low dis} 0.06521739          1 1.601266    33
```

From the association rules above, we can see relationship 1 and 2 has the highest support and counts. A support of 0.069 implies that low crim, low zn, high nox, high age and low dis appear together 6.9% in the Boston area. A confidence of 1 implies that low crim, low zn, high nox, high age were found, 100% of the time there was low dis in the area. A lift of 1.601, which is >1, indicates that the rule is better than predicting the result than guessing. It shows that the item sets are likely to be together than independent, as lift measures how independent the item set are. To live in area where there is low dis, i.e. less distance to the Boston employment centers, and low crime, then it would be recommended to live in an area where there is a high proportion of owner occupied units prior to 1940s and the area would be accessible to radial highways, low nitric oxides, and have a low proportion of residential land.

Part 2d)

In this problem we had to advise a family is moving to the area, and has made schooling a priority and they want schools with low pupil-teacher ratios. We had to advise them on this matter through the mining of association rules.

```
    lhs                                                                rhs                     support    confidence lift     count
[1] {indus=High indus,rad=Low rad.,tax=High tax}                    => {ptratio=Low ptratio} 0.07707510 1          2.663158 39
[2] {crim=High crim,indus=High indus,rad=Low rad.,tax=High tax}     => {ptratio=Low ptratio} 0.06126482 1          2.663158 31
[3] {indus=High indus,age=High age,rad=Low rad.,tax=High tax}       => {ptratio=Low ptratio} 0.07509881 1          2.663158 38
[4] {indus=High indus,dis=Low dis,rad=Low rad.,tax=High tax}        => {ptratio=Low ptratio} 0.07707510 1          2.663158 39
[5] {zn=Low zn,indus=High indus,rad=Low rad.,tax=High tax}          => {ptratio=Low ptratio} 0.07707510 1          2.663158 39
[6] {indus=High indus,chas=0,rad=Low rad.,tax=High tax}             => {ptratio=Low ptratio} 0.06324111 1          2.663158 32
> |
```

From the association rules above, we can see relationship 1, 4, and 5 have the highest support and counts. A support of 0.077 implies that high indus, low rad, high tax, and low zn appear together 7.7% in the Boston area. A confidence of 1 implies that when high indus, low rad, high tax, and low zn were found, 100% of the time there was low ptratio in the area. A lift of 2.66, which is >1, indicates that the rule is better than predicting the result than guessing. It shows that the item sets are likely to be together than independent, as lift measures how independent the item set are. To live in area where there is low ptratio, i.e. low pupil to teacher ratio, it would be recommended to live in an area where there is high tax, areas with less access to radial highways, low proportion of residential land, high proportion of non-retail businesses.

Part 2e)

In this problem we had to use a regression model to solve part d i.e. predict the ptratio. We used the p-values for the regression coefficient from the summary of the model to tell us the important feature.

Table 1: Summary of the regression model fitted on the Boston data set

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

Residuals:
    Min      1Q  Median      3Q     Max
-4.1190 -1.0126 -0.0060  0.8961  4.8945

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.484e+01  1.352e+00  18.379  < 2e-16 ***
crim        -1.578e-02  1.085e-02  -1.454  0.14661
zn          -2.473e-02  4.408e-03  -5.611 3.35e-08 ***
indus        5.722e-02  1.997e-02   2.865  0.00434 **
chas        -2.824e-01  2.846e-01  -0.992  0.32152
nox         -1.050e+01  1.187e+00  -8.848  < 2e-16 ***
rm          -7.076e-02  1.479e-01  -0.478  0.63255
age          7.198e-03  4.313e-03   1.669  0.09577 .
dis         -2.187e-02  6.883e-02  -0.318  0.75084
rad          1.177e-01  2.154e-02   5.465 7.35e-08 ***
tax          6.983e-04  1.244e-03   0.561  0.57491
black        1.573e-03  8.873e-04   1.773  0.07692 .
lstat       -3.770e-02  1.824e-02  -2.067  0.03929 *
medv        -1.021e-01  1.402e-02  -7.283 1.31e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.554 on 492 degrees of freedom
Multiple R-squared:  0.4982,    Adjusted R-squared:  0.485
F-statistic: 37.58 on 13 and 492 DF,  p-value: < 2.2e-16
```

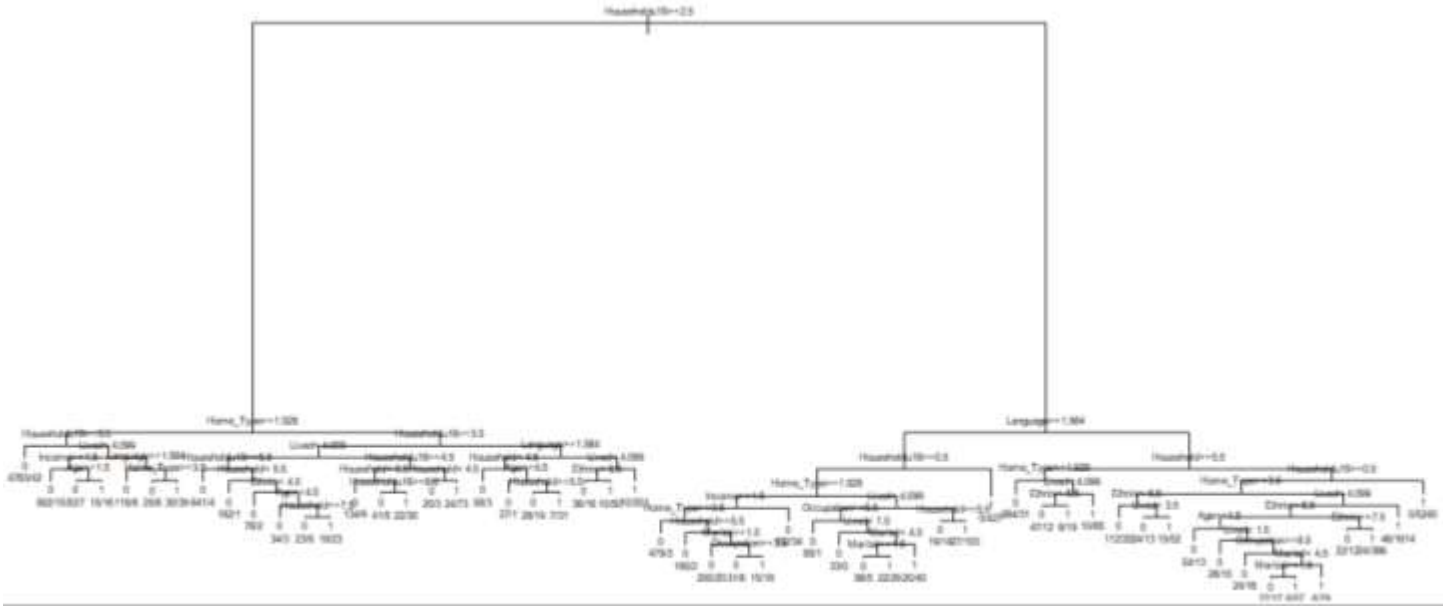
To determine the predictors which appear to have a significant relationship to the response ptratio. The p-value of each feature under the coefficients was used to check for statistical significance. The three significance stars in the last column indicate $p<0.001$, which means the feature is very significant (as the more the stars, the more significant the predictor) and based on the table zn, nox,rad and medv are the significant predictors. Thus, the results of the association rules are comparable, because the regression follows are similar trend with zn, nox,rad and medv being significant predictors /features for ptratio.

Regression model is easier to interpret using the regression coefficient p-values, as sometimes association rules results in a very large number of rules, leaving the user with the task to go through all the rules and discover interesting ones. As selecting and going manually through large sets of rules is time consuming and effortful.

Both regression and association rules are used for supervised learning, but association rules can also be used in unsupervised learning and both are used to for prediction. Regression is based using the p-values of the regression coefficients for hypothesis testing while association rules are based on support, lift and confidence. Regression is preferred when the features are continuous and can handle non-categorical input data. Association rules find associations between specific values of categorical variables in large data sets. The apriori algorithm used in the above problem is be preferred for massive database, where we have big N and big p, there are many advantages, such as it is very computationally efficient but requires to take data and create a binary incidence matrix, and but we have to make decisions about the variables, if we have continuous have to bin them into intervals, if have categorical data we have to spread them out or lump them into categories. If computation is not an issue, it imposes thresholds as it is arbitrary, we cannot threshold using cross-validation where we select the threshold parameter and complexity parameter. If we have continuous data we don't want to discretize it, as we lose information. If we have categories on the low level of frequency we don't want to threshold them in a way to get rid of them and we don't want to merge them, thus association rules is preferred not to hand non-categorical input data.

Question 3

In this problem relates to general association rules. We had to cluster the demographic data of Table 14.1 using a classification tree. We had use the marketing data set from the ElemStatlearn package. We had to generate a reference sample the same size as the marketing dataset which will be the training set, by randomly permuting the values within each feature independently i.e. the class labels to break the relationship between the variables. We had to create a response variable for the training set and the reference set, where g(x)=1 and $g(x)_0$=0, respectively. We had to build a classification tree to the training sample (class 1) and the reference sample (class 0).



From the tree we can see that the householdu18, language, household, home-type and lived ar e the main primary splits for this tree. The terminal nodes which have the highest estimated cla ss 1 probability are Household>=1.5 which had a probability of 0.95 and appears in sample 13%, followed by age>=1.5 as it has a probability of 93.9%, this node has 31 class counts, as well as Occupation< 6.5 which has a probability of 0.90 and appears in sample 2% as we can see from t he tree above and the summary of the nodes.