Question-1

The problem is to design a recommendation system and analyze the performance of our own system by k-fold cross validation.

The given dataset is the MovieLense which has some NA values we need to predict those values by converting the data frame into sparse matrix. In order to find the NA values, we will be using recommendation system.

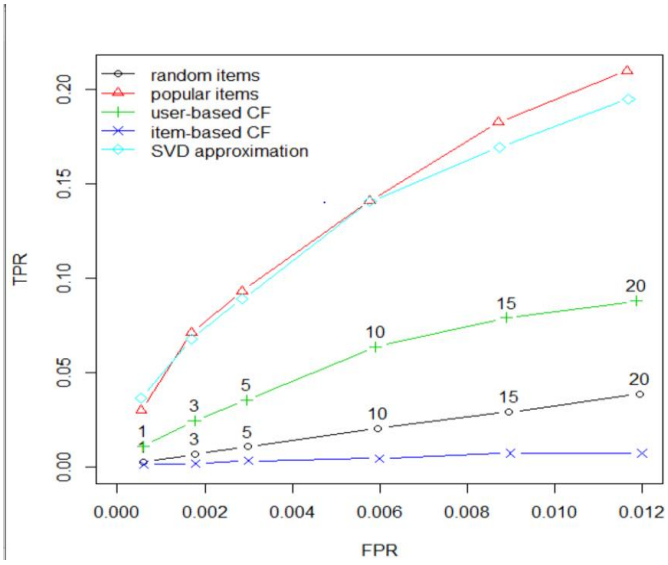
When we look into the matrix the NA values are because of some of the user didn't rate the movie so, we need to predict the rating of unknown values by comparing with movie rated by the users with similar taste.

To predict the ratings of the unspecified data for a user we will using top 10 similar user who has already seen the movie. If the top 10 didn't seen the particular movie, then those values are neglected. These things are taken care by the library 'recommender lab'. The predicted values are converted into matrix format and stored in our local file.

The next part of the question is to analyze the performance of our own recommendation system using different cross fold validation. In our case we had split the data into 50 % of training data and remaining as testing data. We had also considered the ratings which are 4 and above 4 as good rating. To analyze the performance of the recommendation we will be using 'evaluate' function. We had applied different algorithm like SVM, POPULAR , RANDOM , and UBCF to evaluate the system. We had also taken the different K-values like 1,3,5,10,15,20 for performing cross fold validation.

Then we had plotted the graphs of True positive rate and False positive rate and also plotting the graph of precision vs recall to check the value of 'k' for which the Cross Validation would work perfectly.

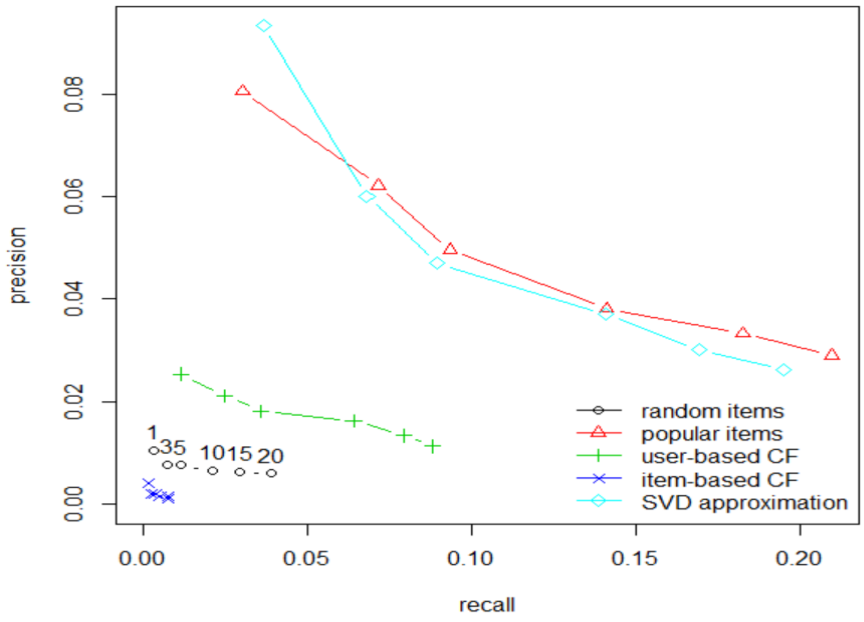
Below is the picture of graph between True positive rate and False positive rate



When we look into the graph we can see the performance of different algorithm for different cross validation set. The random items method and item-based method works very bad because the false positive rate is very high correspondingly True positive rate is very low. User based approach works better than the popular and random methods. The SVD approximation and popular item methods works far better than other methods. Because of the reason that SVN able to handle the sparse matrix better than other methods

Thus, looking at the above plot we can state that k value i.e. 5 would be ideal to check the accuracy of the recommender system created. The main reason behind choosing this value is that there should be a balance between false positive rate and True positive rate. When we see the graph the true positive rate increases along with false positive rate so in order to choose the tradeoff between TP and FP we had taken K=5

Below is the picture of graph between Recall and Precision

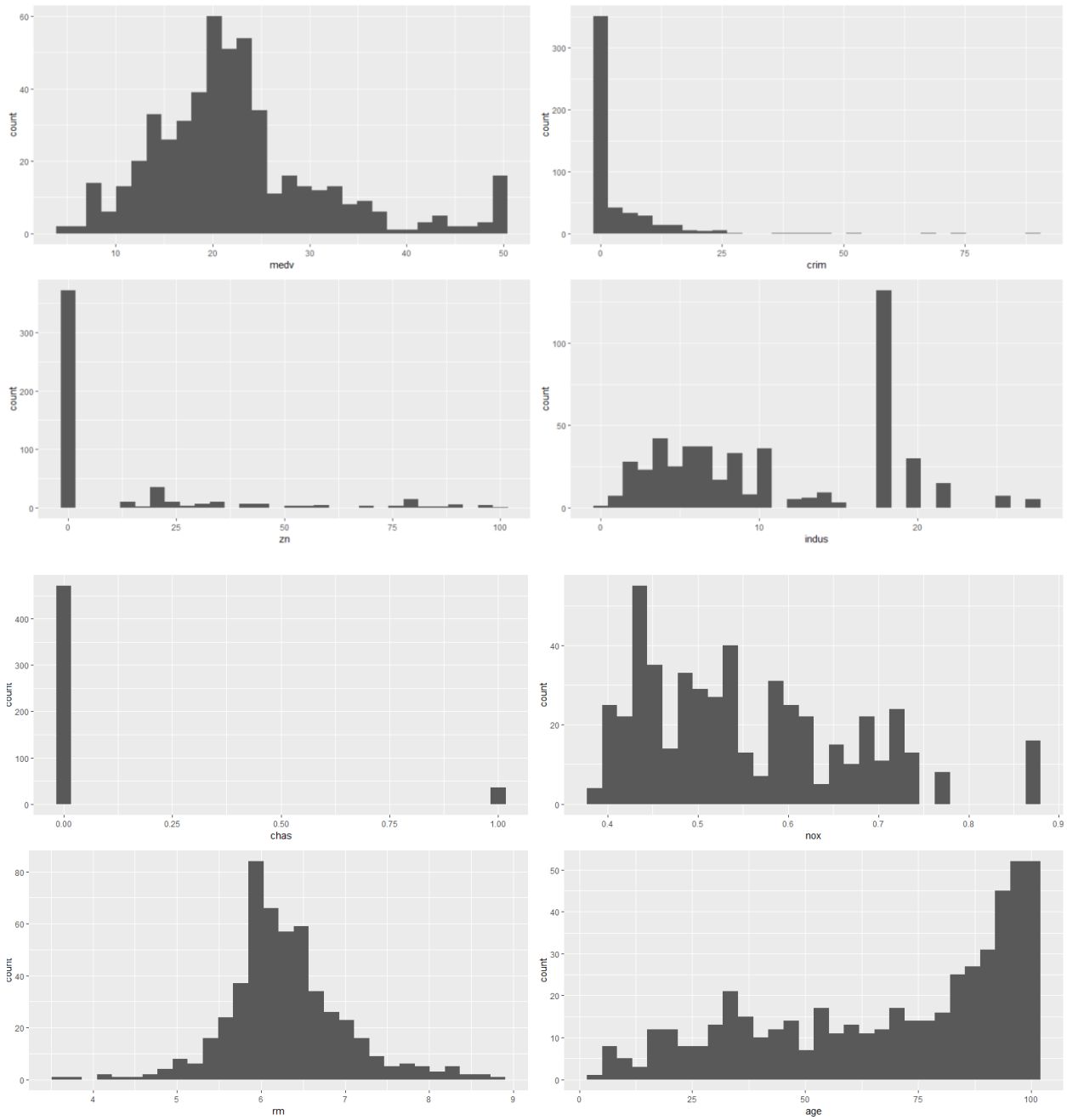


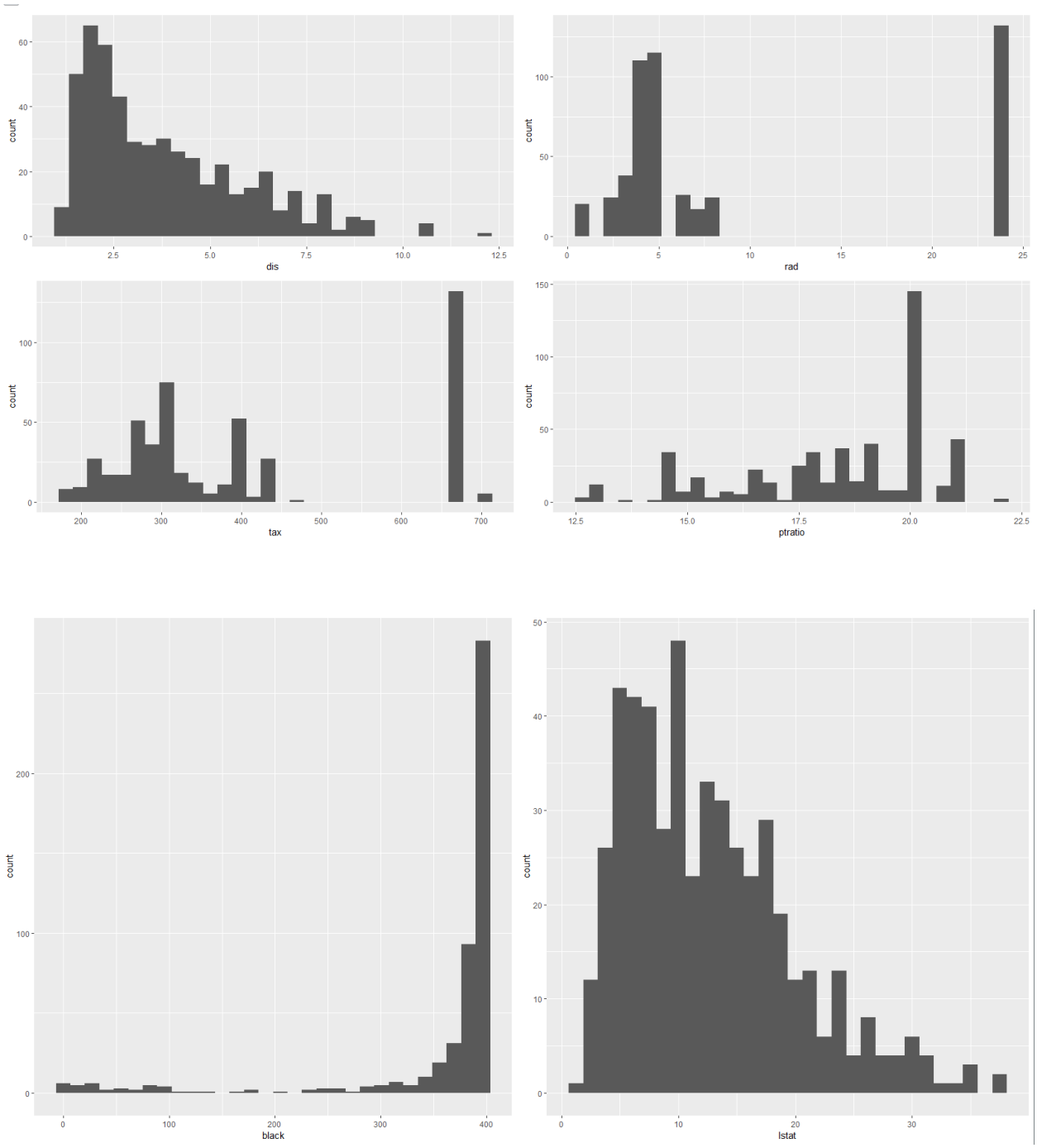
Thus, looking at the above plot we can state that k value i.e. 10 would be ideal to check the accuracy of the recommender system created. The main reason behind choosing this value is that there should be a balance between Recall and Precision. One more reason for choosing this is because there will be an elbow in the graph so K= 10 will be the optimal values.

By considering both the graphs k= 5 would be better considering all the factors like Recall, Precision, True Positive rate and False Positive rate.

Question-3

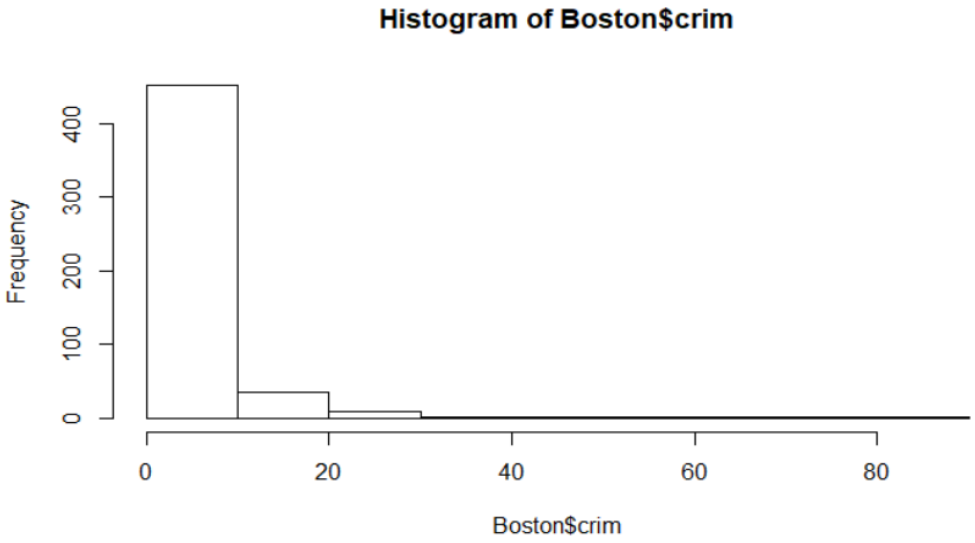
The problem is to analyze the Boston Housing Data and provide the suggestions or set of rules for the given conditions.



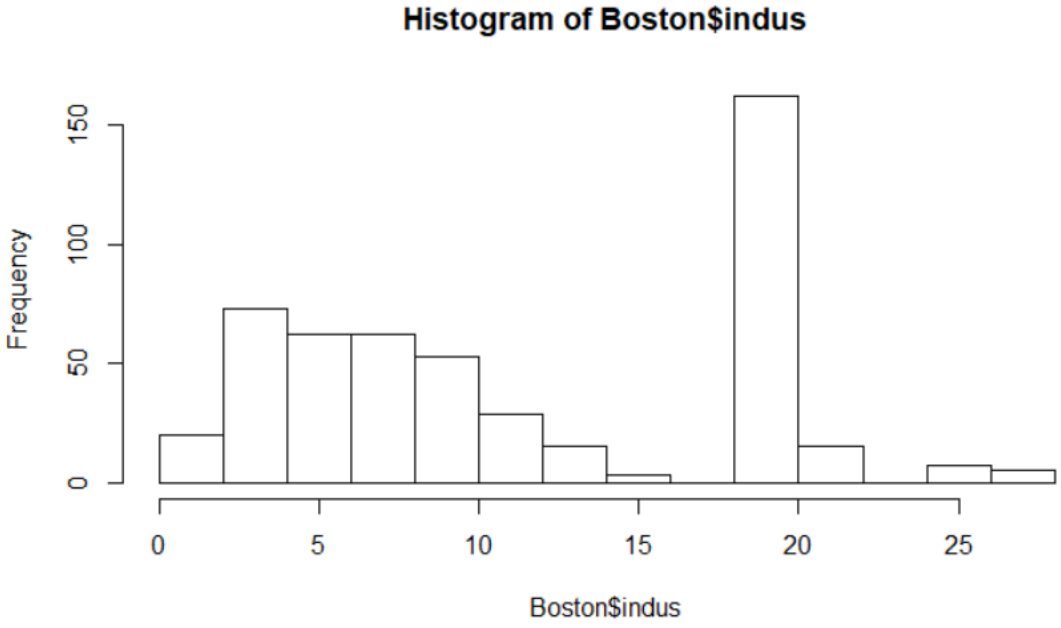


A) The task is to plot the histogram for each of the feature in the data frame and convert the values in the feature into binary incidence matrix to apply the association rules. The conversion is based upon the

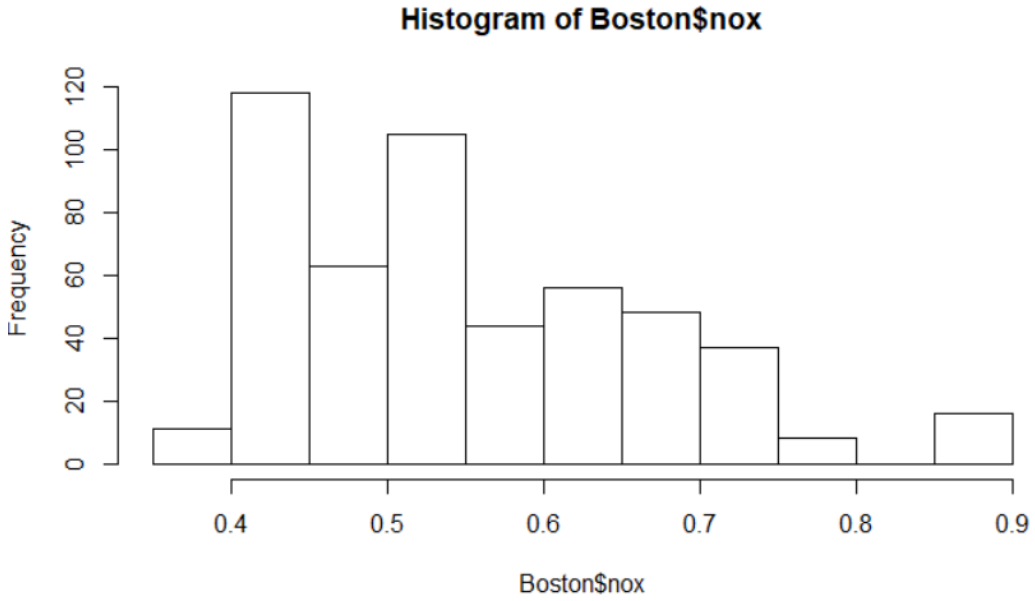
distribution of the values.



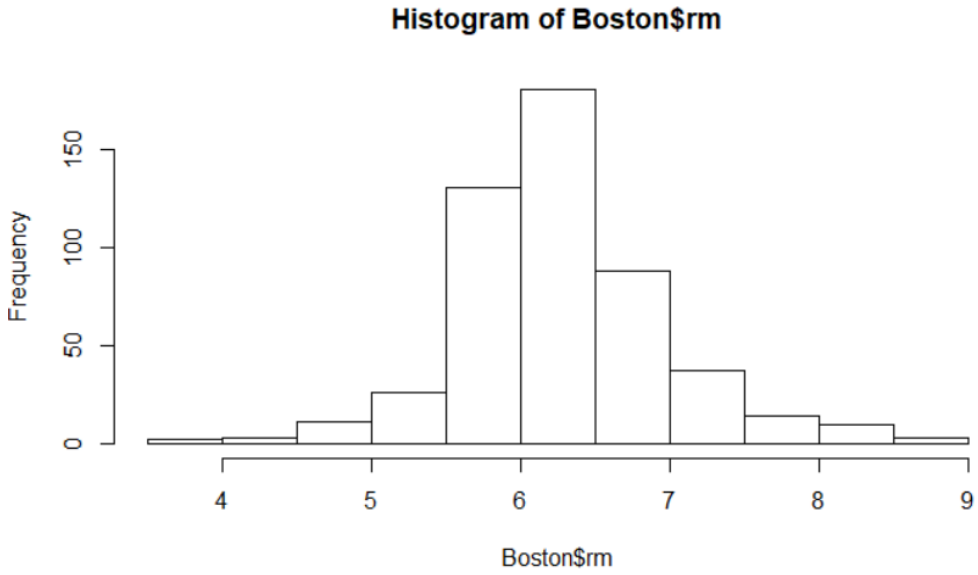
When we look into the histogram the most of the values ranges from 0 to 10 so I had taken the split from 0 to 5 as ‘low’ and remaining values as ‘high’.



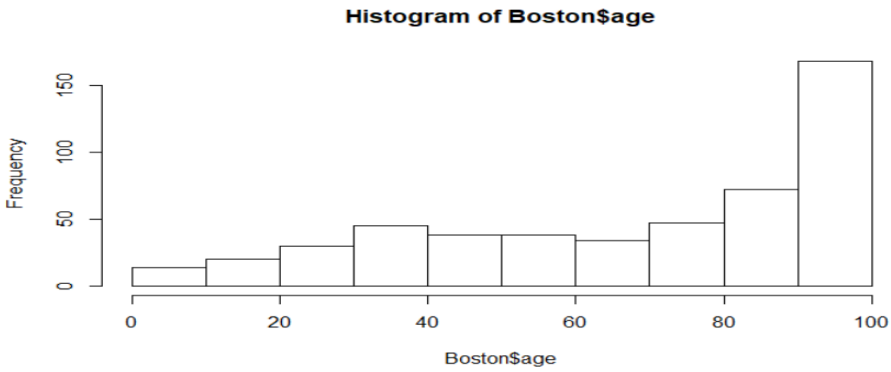
When we look into the histogram we can divide the distribution into two. The value from 0 to 15 is considered ‘low’ and the values from above 15 is considered as the other class.



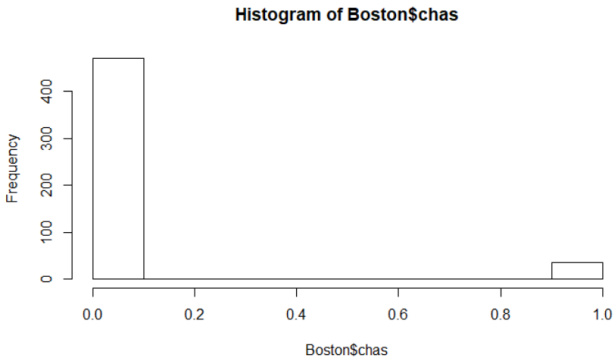
Since the distribution for this feature is uniform spliting the values into two equal half would be better.The values from 0 to 0.6 is considered as ‘low ’ and remaining values is considered as ‘high’.

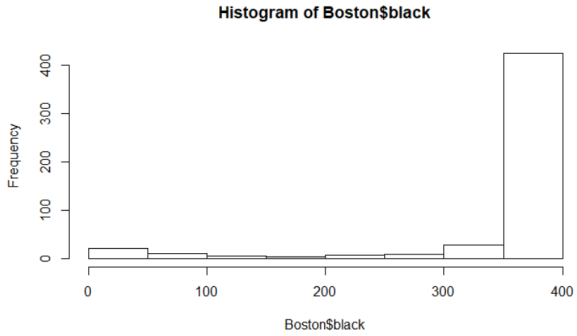


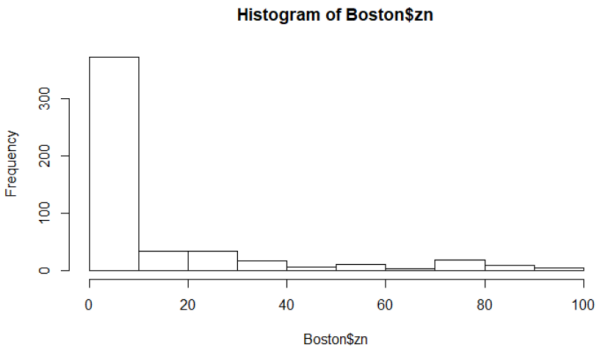
When we see the average number of rooms per dwelling the distribution is normally distributed so the I’m splitting the values into two equal halves (i.e. from 0 to 6 as ‘low’ and remaining set of values as ‘high’)



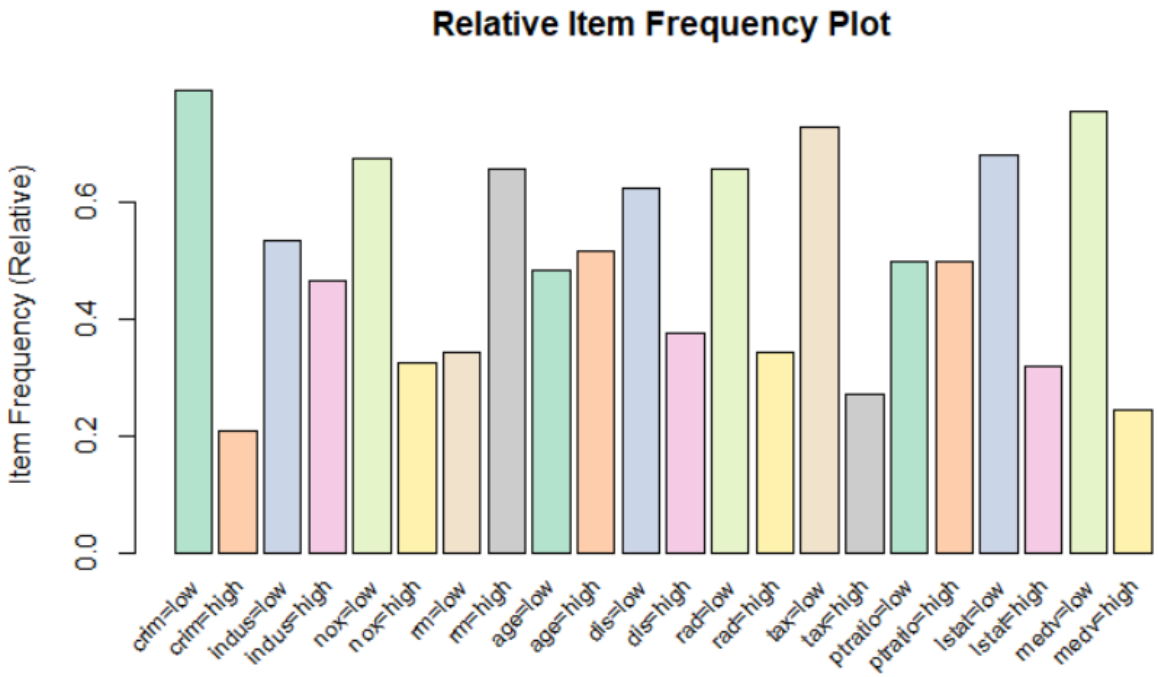
Since more number of people is from 80 to 100 so I'm considering those as values as 'high' for equal distribution and remaining set of values as 'low'







We can see that variables are heavily skewed or discontinuous towards one side. By building the building the multiple regression we find that these variables are less correlated towards our output dependent variable so I removed those variables.



Above is the plot for number of items repeated. Based upon the item frequency plot we can see that some items are more prevalent compared to others items.

From the graph above crim = 'low' and medv = 'high' have high frequency compared to other variables.

For the apriori algorithm I had taken the parameters like support as 0.03 and confidence as 0.8.

Taking support = 0.03 and confidence = 0.8 there were around 98547 rules

C) The student is interested in a low crime area with the proximity to the closest distance, The rules are

```
    lhs                                 rhs        support    confidence lift  count
[1] {age=low,dis=low,medv=high}      => {crim=low} 0.04743083 1          1.265 24
[2] {dis=low,ptratio=low,medv=high}  => {crim=low} 0.10079051 1          1.265 51
[3] {indus=low,dis=low,medv=high}    => {crim=low} 0.08893281 1          1.265 45
[4] {dis=low,rad=low,medv=high}      => {crim=low} 0.07312253 1          1.265 37
[5] {dis=low,tax=low,medv=high}      => {crim=low} 0.10474308 1          1.265 53
[6] {dis=low,ptratio=low,lstat=high} => {crim=low} 0.04743083 1          1.265 24
```

Inference: One interesting observation for the crime to be low opt for the place that is very far from the Boston employee Centre's because out of 6 top rules all the rules contains dis='low' so it is recommended.

Inference: one more interesting observation found was the median value of owner occupied home is very high (i.e. they can get the home for around ($20K – $40K) with an average number of rooms around is 5 to 7

D) There were around 2653 rules for the low pupil-teacher ratio to be low sorted based upon the confidence.

```
> inspect(head(sort(rulesLowPTratio, by = "confidence"),n=3))
    lhs                                rhs           support    confidence lift count
[1] {indus=low,nox=high,medv=high} => {ptratio=low} 0.01778656 1          2     9
[2] {nox=high,rad=low,medv=high}   => {ptratio=low} 0.02964427 1          2    15
[3] {nox=high,tax=low,medv=high}   => {ptratio=low} 0.02964427 1          2    15
```

Inference: From the set of rules If opt for the place where the Nitrogen oxide concentration is high then there will be low student to teacher ratio but that is not recommended for the family since there is high nitrogen concentrate.

```
    lhs                                      rhs           support  confidence lift     count
[1] {rm=high,age=low,tax=low}             => {ptratio=low} 0.270751 0.8058824  1.611765 137
[2] {nox=low,rm=high,age=low,tax=low}     => {ptratio=low} 0.270751 0.8058824  1.611765 137
[3] {rm=high,age=low,tax=low,lstat=low}   => {ptratio=low} 0.270751 0.8058824  1.611765 137
```

So based upon the lift it is better to opt for the place where public tax is low and the people age is less than 40 and property tax is 'low'.

E)

```
             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
```

The above is the summary of the multiple regression applied over the data frame.

Interesting thing we found was linear regression also follows the same set of rules which we got in apriori algorithm. The trend is Nitrogen oxide concentration should be high and Median value of owner occupied should also be high for the pupil-teacher ratio to low since in the summary of the regression those two features are negatively correlated and those variables highly dependent upon our output variable in this case our output variable is pupil to teacher ratio.

In the current case since the variables are categorical the apriori algorithm is preferred. In general, the regression is preferred when the features are non-categorical since lm model performs better on continuous variables. Apriori algorithm is usually preferred when we want to know the interaction between the variables and also if we want to know the range of value we will be using this algorithm but the algorithm is computationally very expensive for very large data.

Question – 4

The problem is to predict the set of rules for the given condition by using Generalized Association Rules and CART method in the Marketing data set.

When we are checking for any NA values in the data we came to know there were some NA values in some of the column in the data frame

```
> apply(marketingData, 2, function(x) any(is.na(x)))
      Income          Sex      Marital          Age          Edu   Occupation         Lived
       FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE
 Dual_Income    Household Householdu18       Status    Home_Type       Ethnic     Language
       FALSE        FALSE        FALSE        FALSE        FALSE        FALSE        FALSE
      Target
       FALSE
```

Since Lived column almost have the 10% values as NA's, replace them with random sampling form the allowed list of values 1,2,3,4,5. Since all the features are categorical type, used the mode of the respective columns to replace the NA's

I created separate target variable and kept the default value as 1 for the given data sample

When we are doing EDA given data sample follows the uniform distribution so are assuming reference sample will also follow the uniform distribution so I randomly sampled the data for each individual column and created the sample target samples with default values as 0.

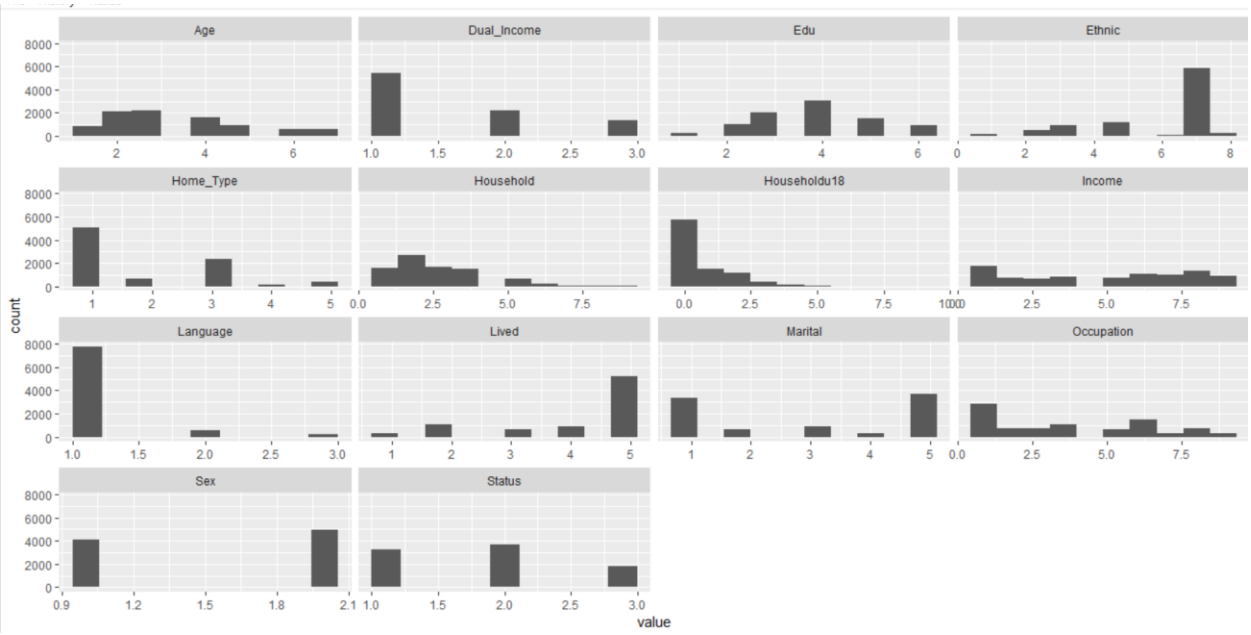
And merged both of the given data with target variable as 1 with the sampled generated data with target variable as 0.

For the result to be accurate and for the resultant values to be integer I had converted numeric values into factor variables.

Then build the CART model using r-part library. The output variable is the 'Target Value' because that separates the actual value with the generated value. From the summary of the we will be getting the prediction accuracy for individual leaf node. Out of all those leaf nodes 15 has the highest probability of class 1

```
node number: 15
  root
  Householdu18=0,1,2
  Language=1
  Household=1,2,3,4,5
```

From the observed rule we can say that household can have from 1 to 5 persons and the number of people with age lesser than 18 can be 0,1, or 2 and can speak one language.



The inference is because of the reason that when we look into the histogram we can see that for household ,household 18 and language the data is skewed towards one particular value so that probability for 1 is maximum for those features.