Kunal Mehta
kunalkam
50352960

Homework 1

Problem 2

a. Grouping Categories of Variables:
   - Crim: On viewing the histogram of the variable's values, it is immediately clear that the bulk of the per capita rates of crime in the dataset lie in the 0-10 range (north of 400 in frequency out of 506 observations), with 10-20 being the only other interval which can be said to have a significant frequency. However, a look at the explicit values of the dataset shows that the values in the interval 0-10 are not well distributed throughout the dataset, but tend to cluster between 0 and 1. Therefore the grouping categories I have chosen are:
     i. Low: 0-0.2
     ii. Moderate: 0.2-0.6
     iii. High: 0.6-1.0
     iv. Very High: 1.0-100.0
   - Zn: Just like crim, a look at the explicit values of the variable provides much more insight into the variable. From the histogram, the 0-10 interval consists of the bulk of the observations, but from the data it is clear that most of these values are 0.0. The lowest non-zero observation in the variable is 12.5, and the highest is a complete 100.0. *It so happens that I keep getting errors that my labels are invalid, that my length 4 should be 1 or 3. I do not understand why this is happening, I thought that some of the values in the variable were not numeric and tried the categorizing after coercing the values to a numeric value but I kept getting the same error. I even fiddled with the melt function of reshape2, admittedly not knowing exactly what it does. Interestingly, if I use the single value numeric as an argument to the function instead of specifying my own ranges, it works. Finally, I decided to NULL this variable since it is not a particularly useful one, as far as I can see.*
     Based on the data, I chose the following categories:
     i. Low: 0-12
     ii. Moderate: 12-25
     iii. High: 25-50
     iv. Very High: 50-100
   - Indus: The proportion of business acres appears to have a normal frequency distribution for the first 3 5-acre intervals, with low frequency (10-20) at both the higher and lower value extremes. The particular interval 18-20 has a very high frequency (north of 150), with even its immediately neighbouring 2-acre intervals having much lower frequencies, indicating that this value is most likely a town suburb planning factor that is implemented successfully in most of the suburbs. Therefore, I have chosen the following categories:
     i. Low: 0-10
     ii. Moderate: 10-18
     iii. High: 18-100
   - Chas: This is a plain boolean variable indicating if the tract of land bounds the Charles River or not. Therefore there are only two categories: True, False.

- Nox: Indicating the nitrogen oxides in parts per 10 million, this variable's histogram shows intervals 0.4-0.45 and 0.5-0.55 having significantly high frequencies, but the other intervals also have significant frequencies, showing a much higher level of variance of this variable as compared to some of the previous ones. Therefore, I think it is prudent to categorize the nox variable into 3 categories with equal break intervals with the labels being Low, Moderate and High. *I tried looking up online what a health value ranges of the nitrogen oxides would be, but the since the values are in parts per 10 million, all of the values in the dataset fall very easily in the very first level of healthy concentration. Therefore, I did not want to arbitrarily decide what high, low or moderate would be for a variable like this so I went with equally spaced out intervals.*
- Rm: This is probably one of the most significant variables in the dataset in terms of its importance. The histogram shows a normal distribution, with the bulk of the observations in the range 5.5-7.0. Since a fractional room in an individual case doesn't make sense, the ranges I have gone for are as follows:
    i. Very Low: 0.0-4.0
    ii. Low: 4.0-5.0
    iii. Moderate: 5.0-6.0
    iv. High: 6.0-7.0
    v. Very High: 7.0-9.0
- Age: The histogram of the variable indicates a left skewed distribution, with the 80-100 interval having a significantly high frequency, particularly the 90-100 range. Therefore I have gone with the categories as follows:
    i. Modern: 0-30
    ii. Balanced: 30-60
    iii. Old: 60-80
    iv. Very Old: 80-100
- Dis: The histogram of the variable indicates a significantly right skewed distribution, with the first three 2-mile intervals having the highest frequencies. The remaining intervals, though, do have considerable frequencies and are varied enough to not be categorized away into one label. Therefore I have chosen the following categories:
    i. Close: 0.0-2.0
    ii. Moderately Close: 2.0-4.0
    iii. Moderately Far: 4.0-6.0
    iv. Far: 6.0-8.0
    v. Very Far: 8.0-13.0
- Rad: This variable has an unusual histogram for its values, which show a normal distribution over the first 4 2-mile intervals, but with a significantly high frequency interval at an extreme right interval. Upon keeping the breaks of shorter intervals, it is seen that there is a high number of observations in the 22-24 mile range. It seems to me that one set of suburbs in particular is geographically situated such that it is very far from the nearest radial highway, since the high distance values are not interspersed among the observations but clusters towards the last 150-170 observations. However, seeing as this variable is part of a dataset where we categorized distance as well, I think it would be

wise to have a similar standard of distance for this variable as well, particularly because it effectively categorizes this variable as well. Therefore the categories are:

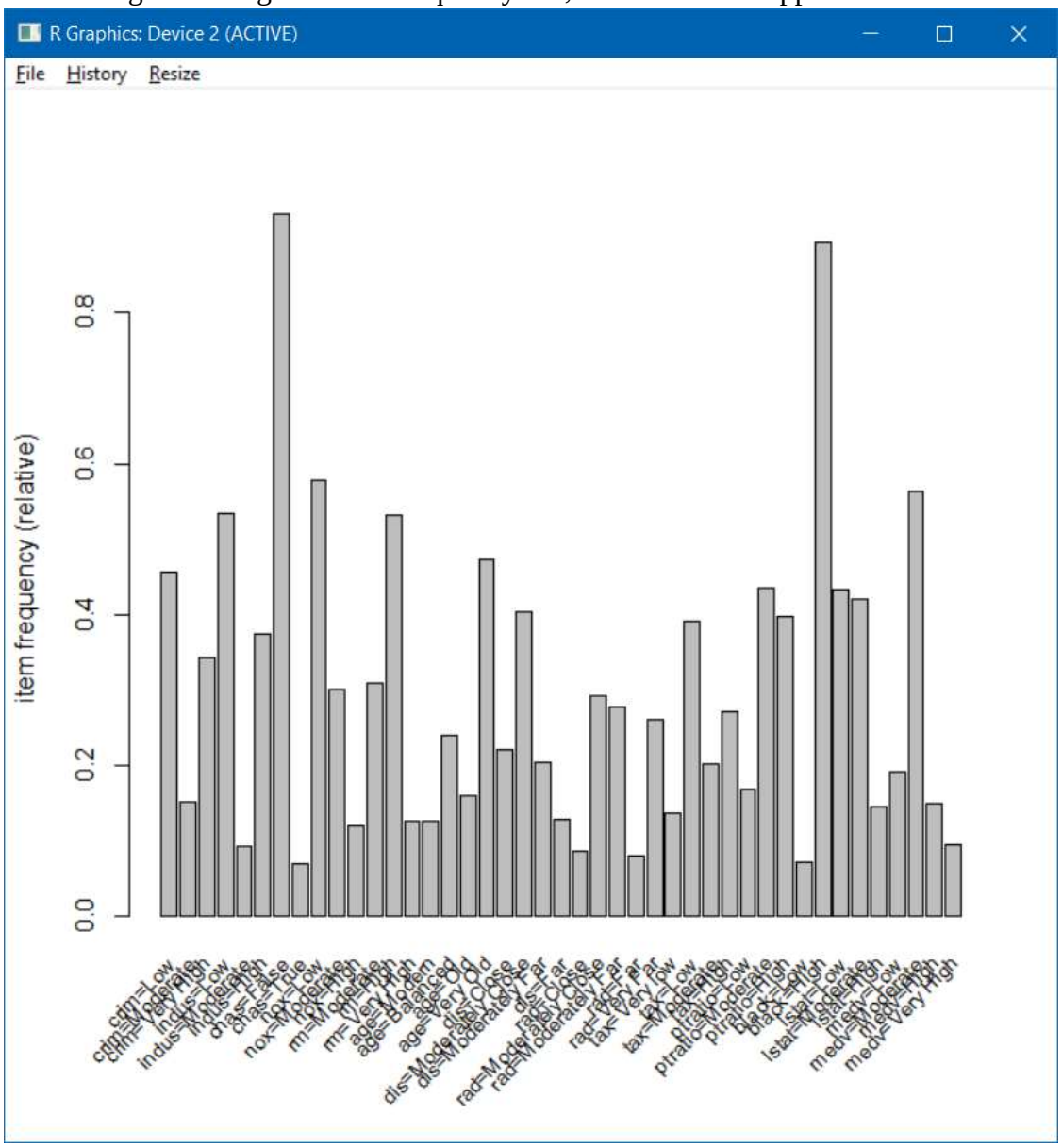
i. Close: 0.0-2.0
ii. Moderately Close: 2.0-4.0
iii. Moderately Far: 4.0-6.0
iv. Far: 6.0-8.0
v. Very Far: 8.0-24.0

- Tax: Quite like the rad variable, the tax variable shows a mostly normal, slightly left skewed distribution over the first seven intervals on the histogram, but with the extreme right 650-700 interval showing the highest frequency of all the intervals. Contrastingly, the 500-550 range has a zero frequency. Based on the data I have chosen the following categories:
  i. Very Low: 0.0-250.0
  ii. Low: 250.0-350.0
  iii. Moderate: 350.0-500.0
  iv. High: 500.0-750.0
- Ptratio: The histogram of ptratio shows a left skewed distribution, with the 20-21 interval having a considerably high frequency. The rest of the intervals have relatively comparable frequencies. Therefore the categories I have chosen are:
  i. Low: 0.0-16.0
  ii. Moderate: 16.0-20.0
  iii. High: 20.0-22.0
- Black: Once again, one interval displaying a very high frequency (350-400), with the rest having almost negligible frequencies. This indicates that there are very few suburbs that have low black populations. The concentration of the high frequency in one narrow interval also suggests the black population is uniformly spread over the Boston suburbs in almost the same proportion. Therefore I have chosen the following categories:
  i. Low: 0.0-150.0
  ii. Moderate: 150.0-300.0
  iii. High: 300.0-400.0
- Lstat: This variable's histogram is right skewed, but not overtly so. The 5-20 interval has the highest frequency. I have chosen the following categories:
  i. Low: 0.0-10.0
  ii. Moderate: 10.0-20.0
  iii. High: 20.0-40.0
- Medv: A left skewed histogram, this variable has the bulk of observations indicating a value in the range 10-35. The categories I have chosen are:
  i. Low: 0.0-15.0
  ii. Moderate: 15.0-25.0
  iii. High: 25.0-35.0
  iv. Very High: 35.0-50.0

Some of the variables have categories that I have chosen based on my subjective view of what the intervals should be. For all the ones above that seem to have little justification as to my choice, this is the reason why. I am pretty certain, though, that the categories I have chosen

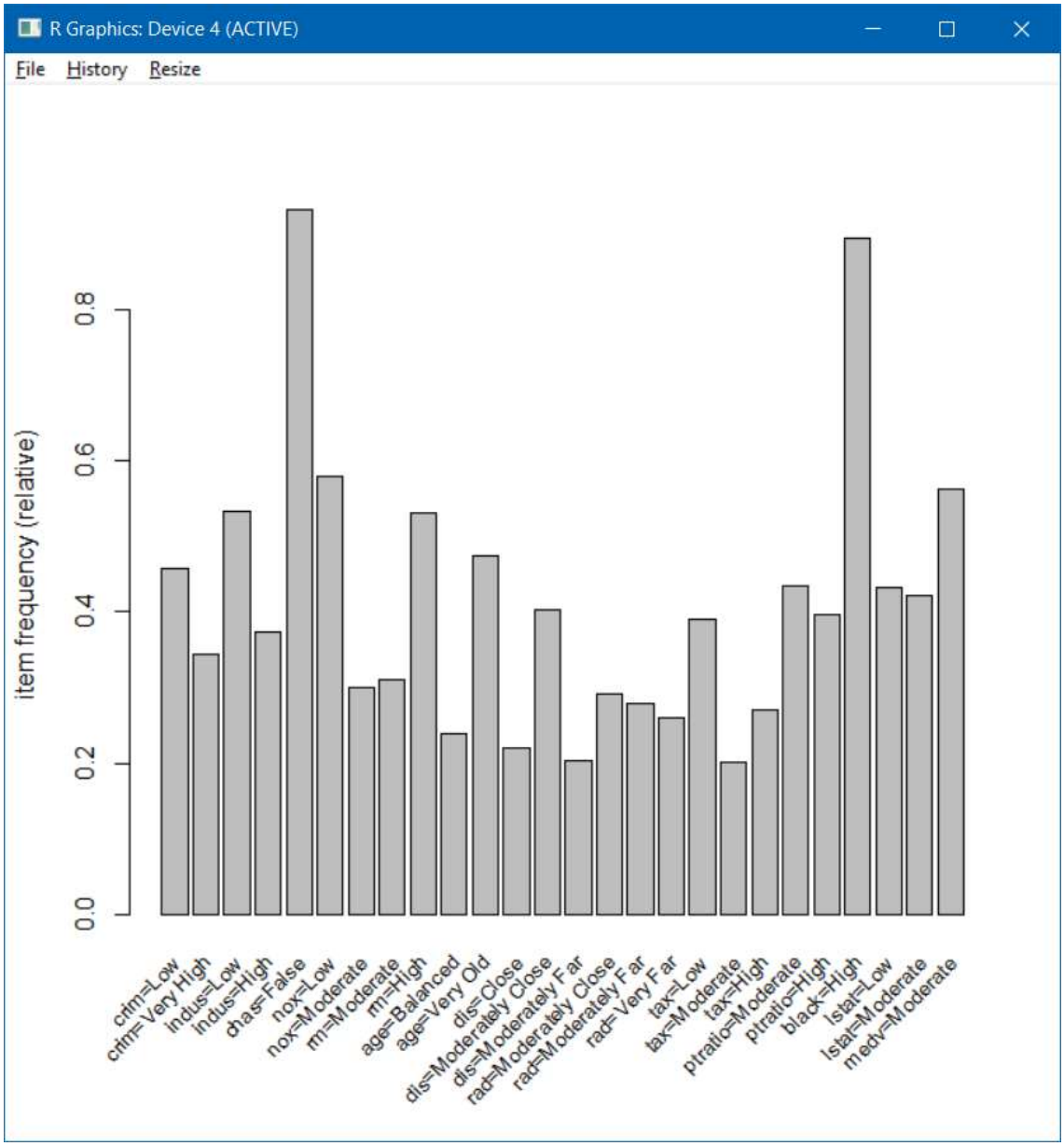
make sense in the context of the variables and will help me get strong association rules that actually indicate causality and real-world parallels unlike the beer-diaper coincidence.

b. Visualizing using itemFrequencyPlot, with support = 0.05:



Visualizing using itemFrequency Plot, with support = 0.2

Kunal Mehta
kunalkam
50352960



I explored around a bit using different values of the support to see how a higher value of the threshold reduced the number of satisfying items. Two of the examples are shown. I wasn't a 100% sure I had to provide just one of the visualizations, but since I did generate many as suggested to in the computational lab video, I figured I'd add another visualization.

On applying the Apriori algorithm, with parameters (support = 0.12, confidence = 0.44):

```
> summary(rules)
set of 8362 rules

rule length distribution (lhs + rhs):sizes
   1    2    3    4    5    6    7    8
   8  295 1402 2543 2392 1304  378   40

   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
  1.000   4.000   4.000   4.512   5.000   8.000

summary of quality measures:
    support         confidence          lift
 Min.   :0.1206   Min.   :0.4400   Min.   :0.7113
 1st Qu.:0.1324   1st Qu.:0.6813   1st Qu.:1.1195
 Median :0.1462   Median :0.8625   Median :1.6427
 Mean   :0.1691   Mean   :0.8233   Mean   :1.8090
 3rd Qu.:0.1838   3rd Qu.:0.9853   3rd Qu.:2.4614
 Max.   :0.9308   Max.   :1.0000   Max.   :3.8333
     count
 Min.   : 61.00
 1st Qu.: 67.00
 Median : 74.00
 Mean   : 85.58
 3rd Qu.: 93.00
 Max.   :471.00

mining info:
 data ntransactions support confidence
   B1           506    0.12       0.44
>
```

I started off with support and confidence parameters as 0.2 and 0.5 respectively. But for the problem part 1.d., I required an association rule indicating a low parent to pupil ratio, and at the parameters that I had chosen, I was not generating a single such rule. After a while of gradually fine-tuning both the values of support, I got these values as the highest 2 decimal point values of support and confidence with which I was able to glean at least one rule as required for 1d. I don't think this is very desirable though, since lowering the support threshold to this level is for sure allowing a lot of needless associations to drive up the computation time and cost.
I think the categorization plays a huge factor in this. Maybe more granularity in my categories could have given a better answer. Shall investigate post submitting the homework.

c. After applying the apriori algorithm, I filtered the rules for those in which the rhs had crim=Low and dis=Close, and sorted them, first by confidence values and second by lift values.
The rule with the highest confidence is
{chas=False, nox=Low, age=Balanced, lstat=Low} -> crim=Low
followed closely by the rule
{chas=False, nox=Low, age=Balanced, lstat=Low, black=High} -> crim=Low
The third rule with almost the same confidence value is

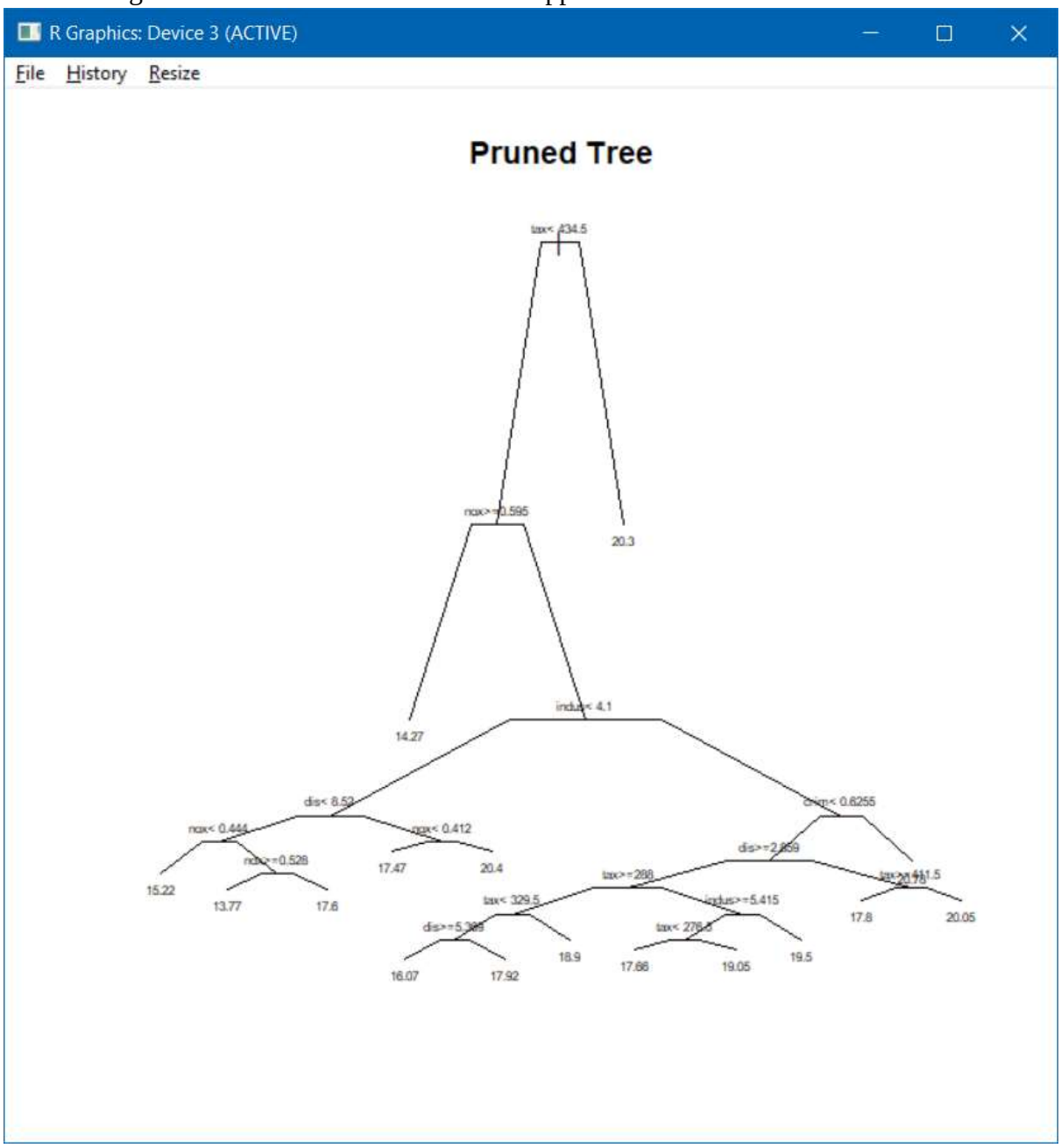
{chas=False, nox = Low, age = Balanced, lstat = Low, indus=Low} -> crim=Low

This shows that the 4 most important factors in locating a low crime neighbourhood are that it must not be adjoining the Charles River, the nitrogen oxides concentration in the air should be low (0.0-0.3), the proportion of owner-occupied units built prior to 1940 being balanced (30%-60%) and the percent of lower status population being low (0%-10%). The second and third rules are insightful too, but one added factor will limit the student's options in choosing a neighbourhood, particularly since these added factors do not increase the confidence value but marginally decrease them in fact.

d. The mined set of association rules has just one single rule that indicates a low pupil to teacher ratio, so naturally it appears as the rule with both the highest confidence as well as the highest lift. The rule is
{rad=Moderately Far} -> ptratio=Low
This indicates that the family must move to a suburb that is moderately far (4.0-6.0) to have a high chance of schooling in an institute with a low pupil-teacher ratio.

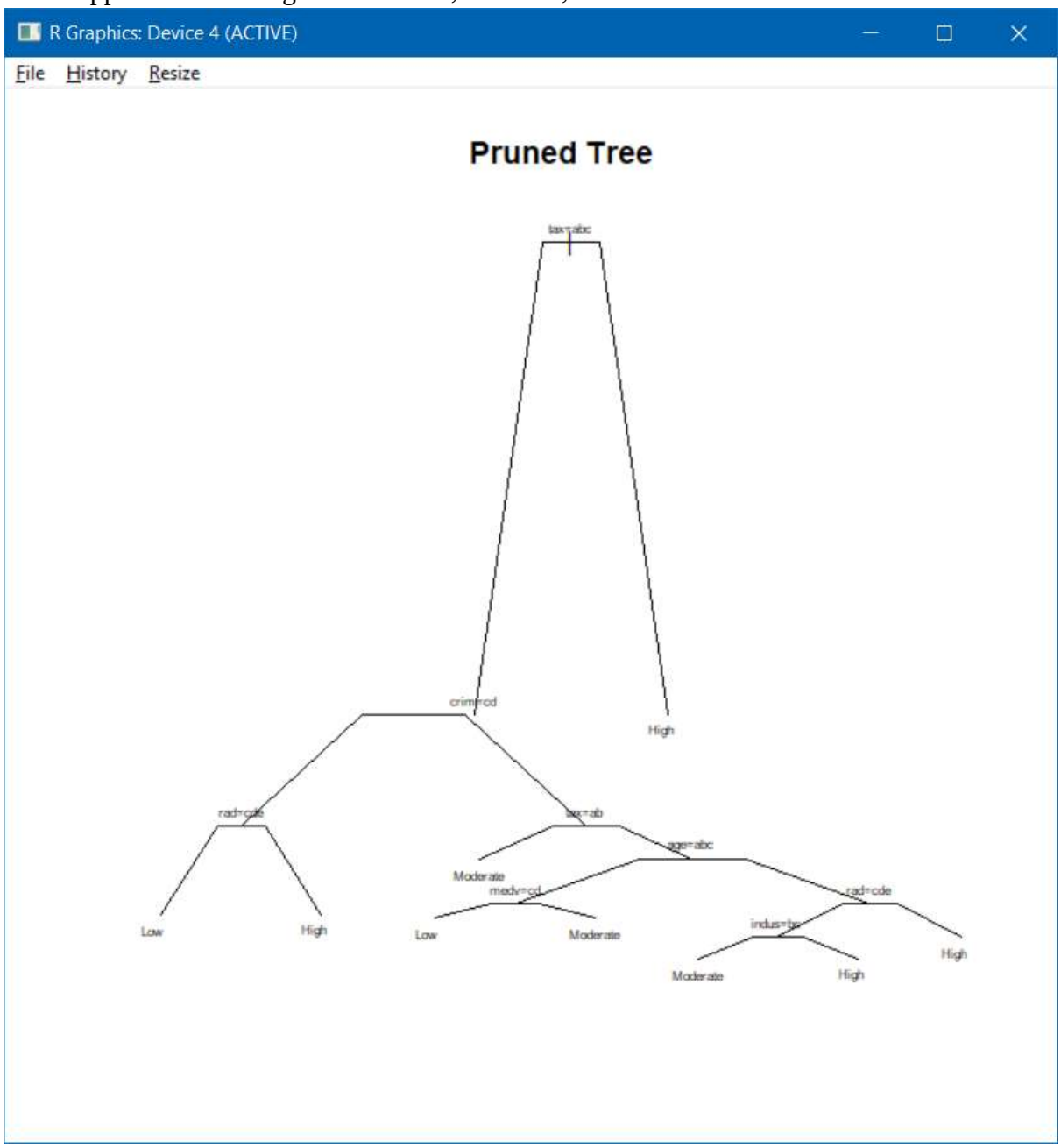
Kunal Mehta
kunalkam
50352960

e. The regression model when applied to Boston dataset



The regression model is in no way comparable to the results I received with the apriori algoritm. The regression model indicates that the biggest factor in having a low pupil teacher ratio is the tax variable, followed by the nox variable. A tax value less than 434.5, followed by a nox value less than 0.595, followed by an indus value less than 4.1, followed by dis less than 8.52, followed yet again by a nox value greater than or equal to 0.526 would give me the lowest pupil-teacher ratio of 13.77.

When applied to the categorized dataset, however,



There is a clear lack of precision due to categorization of the variables, but tax is yet again the main factor. In place of the highly significant nox variable, instead we have the crim variable being the next most influential. If crime is low, then the variable apriori considered the most important comes into play, with a low rad value giving us a low pupil-teacher ratio.

I find it bizarre that what was the most significant variable for the apriori algorithm does not even appear on the pruned tree when the uncategorized Boston dataset is used. Even in the categorized dataset, the rad is the third most influential variable.

Kunal Mehta
kunalkam
50352960

Problem 3

- Loading the dataset was a bit circuitous than the Boston dataset. I had to download the dataset from the UbLearns attachment and change working directory to the folder I had saved it in.
- To create a sample dataset of the same size as the training dataset, the sample function had to be used. To do that I first generated individual lists of the samples with the sample function for each variable, specifying what values it could take.
- The no of records for all of these was the same as that of the marketing dataset, 8993
- Kept the replace argument's value as True since no element must occur twice
- Then I combined all these individual variables into a data frame using the data.frame function. Also attached the labels to the data frame duly.
- Then I made the reference dataset, a replica of the training dataset.
- The next step was to set up the reference data set with randomly permuted values for each variable. This was done with a for loop traversing the reference dataset, which is why the replace argument's value is False
- The next step was to combine the two datasets, which was done using rbind
- The last step was to build a classification tree for the same. This was done using rpart, with the data argument being the combined dataset, the method being class for classification.
- The only issue is that apparently there is no tree, the fit is just a root. On observing the summary, the probabilities are both 0.5, so maybe this is the reason for that.