**HW 1 (Chapter 2)**

**Luyao Zhang (NetID: lzhang94)**

**Ex1:**

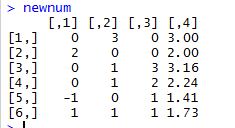
1. Better: a more flexible approach will fit the data more accurately than an inflexible approach, and with the large sample size, we don’t need to worry about overfitting
2. Worse: with the small number of observations, using a flexible approach will have the problem of overfitting, compared with an inflexible approach
3. Better: using a flexible approach will fit the data better than an inflexible approach, because it captures better the curvature
4. Worse: a flexible approach follows the noise, or the error term too closely, and therefore will increase the variance

**Ex4:**

1. **Examples:**
2. Illness diagnosis: response: ill, not ill; predictors: heart rate, blood pressure, weight, gender, medical history; goal: inference
3. Product quality check: response: good, fail; predictors: size, weight, working life span; goal: inference
4. Stock price: response: price up, price down; predictors: stock price of a period of time from the recent past (e.g., for the past month, season…); goal: prediction
5. **Examples:**
6. Students’ grades: response: final exam grades; predictors: past GPA, hours spent studying, gender, district, experience of teachers; goal: prediction
7. Manager salary: response: salary of managers; predictors: experience, industry, age, education background; goal: inference
8. Organizational turnover: response: turnover rate; predictors: age, department, pay, leadership style, perceived relationship with coworkers; goal: prediction
9. **Examples:**
10. Amazon product recommendations: recommend products to customers according to customers who have purchased or browsed the same or similar products
11. TV show audience survey: clustering of ages for a TV show to figure out among which group of people the show is the most and the least popular
12. Deciding grocery store targeted customers: clustering of socioeconomic status (SES) of consumers to see which clusters of consumers shop at which grocery store, and the result can be used to design promoting activities

**Ex7:**

1. The Euclidean distance are shown below in the 4th column of the table:



1. Our prediction with K=1 would be “Green”, because the 5th observation has the smallest Euclidean distance of all, and the color of is “Green”.
2. Our prediction with K=3 would be “Red”, because among the 3 smallest Euclidean distances (observation No.5, 6, and 2), two of them are “Red”, while one is “Green”. Therefore, it’s more likely to be “Red” than “Green”.
3. Small. A small K allows for a flexible model, which is suitable for a highly non-linear relationship, while a large K take more observations into consideration, and therefore is less flexible, which is more suitable for modeling a linear relationship.

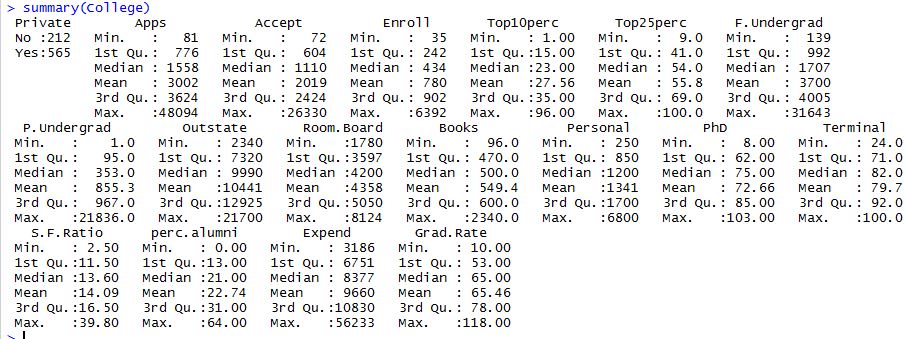
**Ex8:**

1. and (b) are shown below:

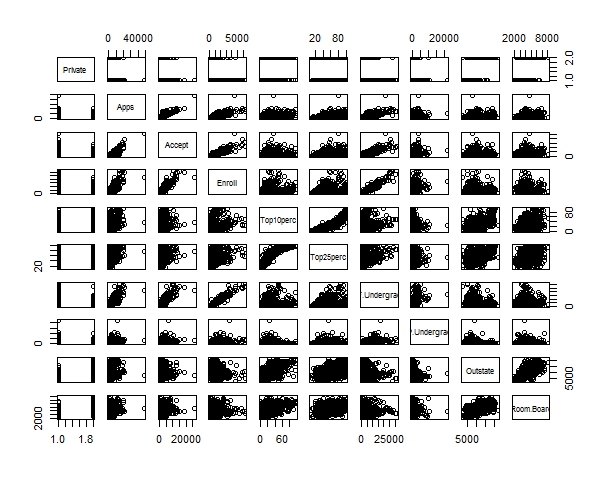


**(c)**

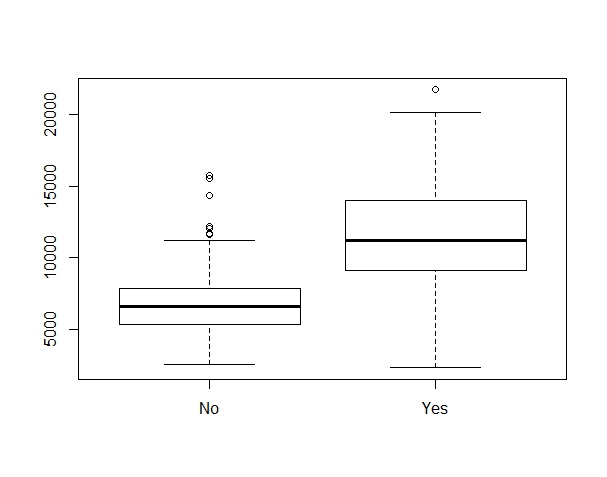
(i) The summary is shown below:



(ii) The plots are as below:



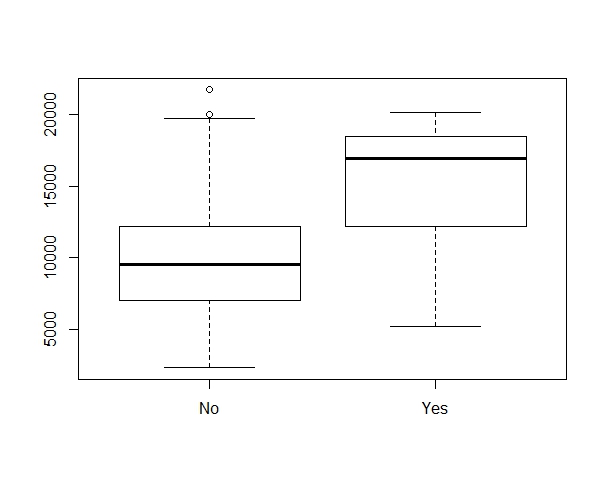
(iii) The plot is as below:



(iv) As the results shown below, there are 78 elite universities.

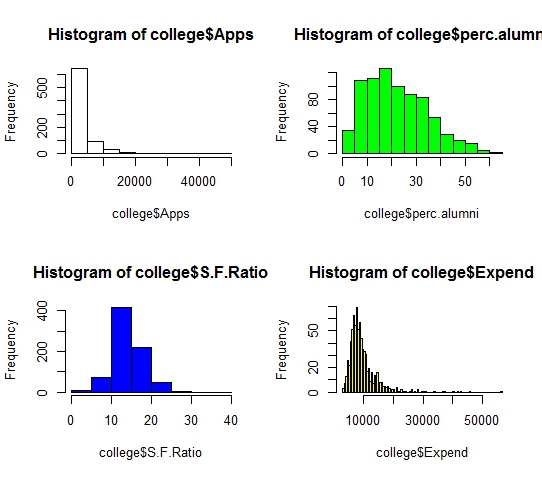


The plot is as below:



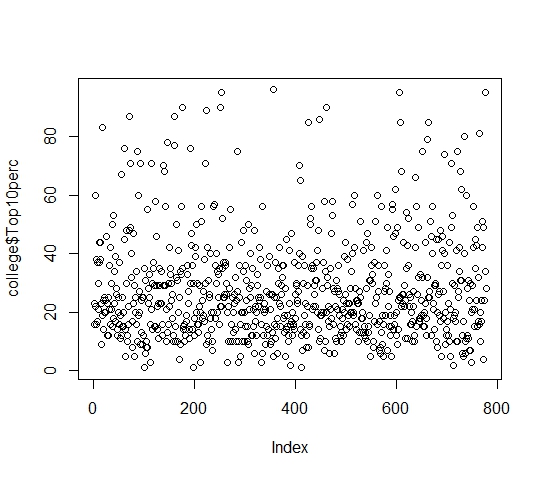
(V)

The 2\*2 histograms of some quantitative variables are as below:

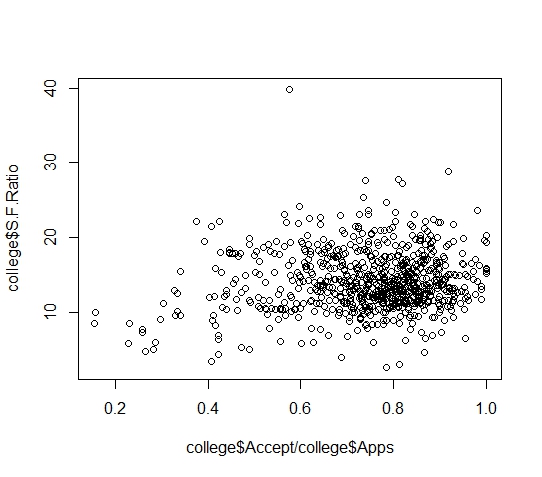


(vi)

Colleges with the most students from top 10% of their high school classes do not necessarily have the highest graduation rate

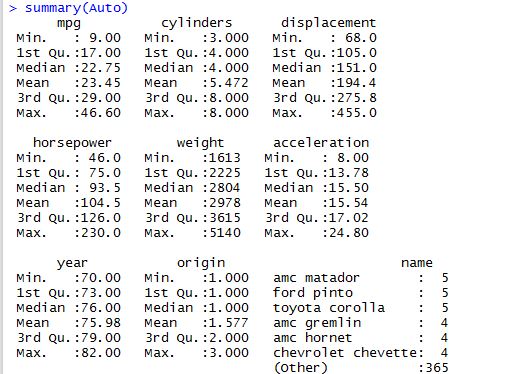


Colleges where the acceptance rates are low are more likely to have lower pupil-teacher ratios, compared with colleges where the acceptance rates are high.

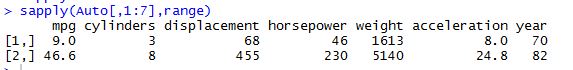


**Ex9:**

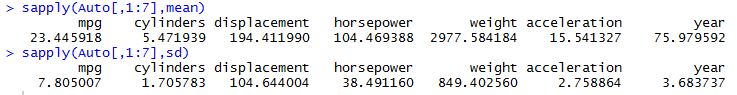
1. na.omit was used to deal with missing values. And according to the summary of the data below, quantitative variables are: mpg, cylinders, displacement, horsepower, weight, acceleration, and year; qualitative variables are: name, origin



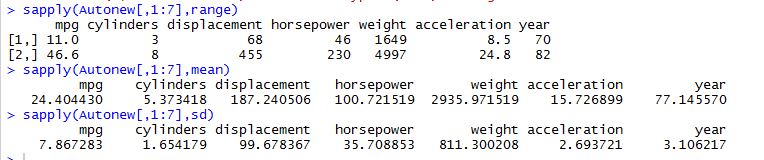
1. Range of the quantitative variables are as below:



1. Mean and s.d. of the quantitative variables are as below:

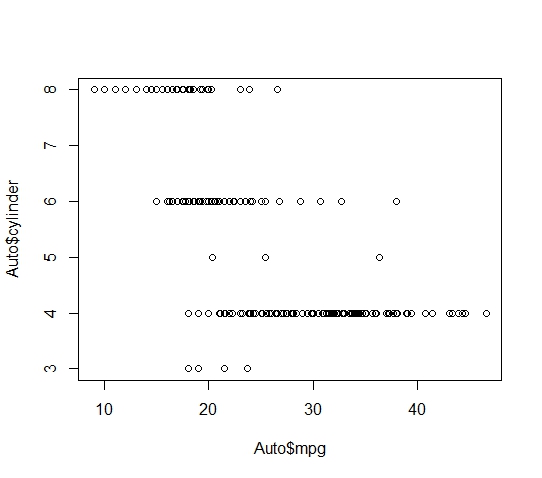


1. The range, mean, and s.d. of the subset of data are as below:

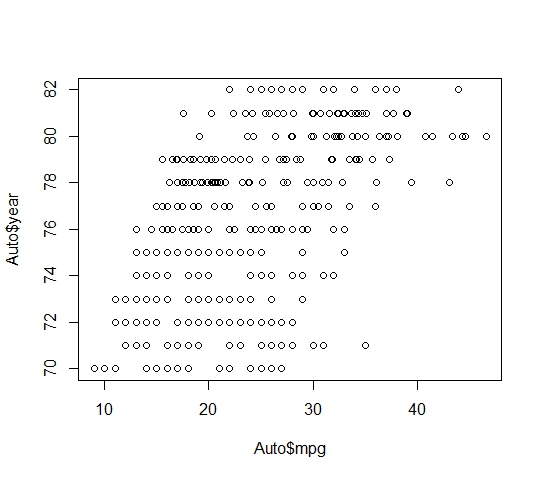


1. Plots and interpretations are as below:

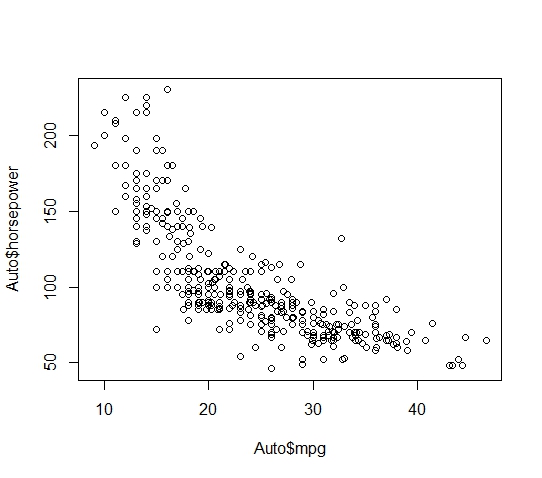
The plot below indicates that more cylinders a car has, the less mpg it has.



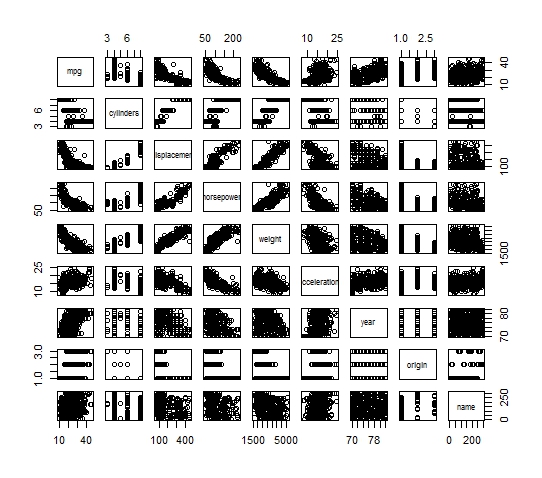
The plot shows that as time goes by, mpg of cars tend to increase.



According to the plot below, the more horsepower a car has, the less efficient the car is.



1. According to the plots by pairs below, it looks like every independent variable is somewhat correlated with mpg. However, the variable “name” has too few observations under each name, which means if we use “name” as a predictor, we are probably going to have a problem of overfitting, and generalizability will be suspicious, too.



**Ex10:**

1. According to R, the dataset “Boston” has 506 rows and 14 columns. The 14 columns represent 14 features of the neighborhood, and the rows stand for housing values in Boston suburbs.
2. According to the plots below, we can find that:

The variable “crim” may correlate with: age, tax, rad, dis, and ptratio

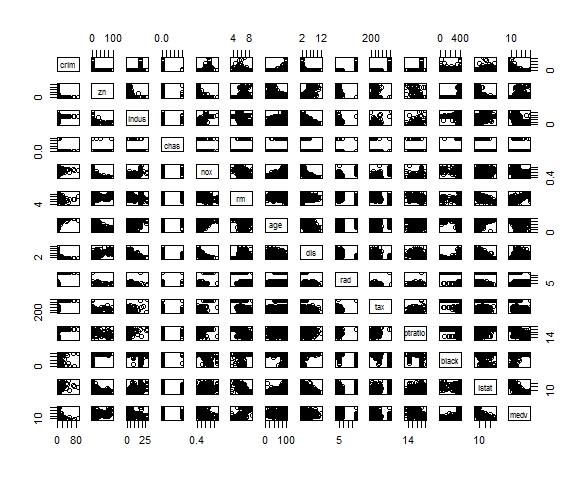
The variable “zn” may correlate with: age, Indus, nox, age, and lstat

The variable “indus” may correlate with: age, nox, and dis

The variable “nox” may correlate with: age, Indus, and dis

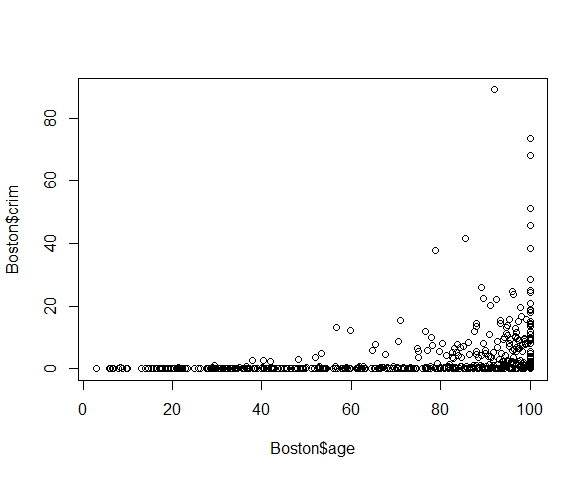
The variable “dis” may correlate with: lstat, and zn

The variable “lstat” may correlate with: medv, zn, and dis

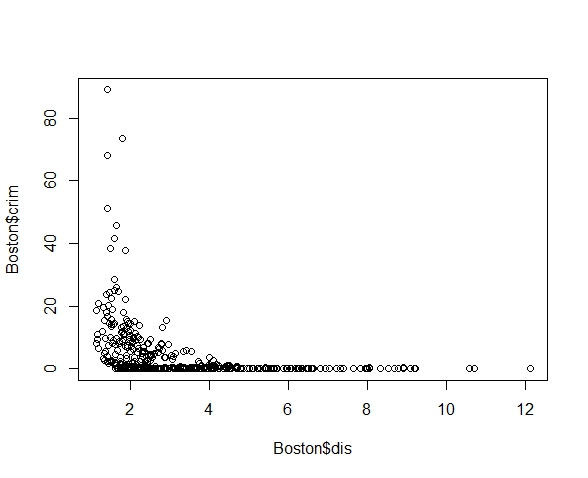


1. The plots and interpretations are as below:

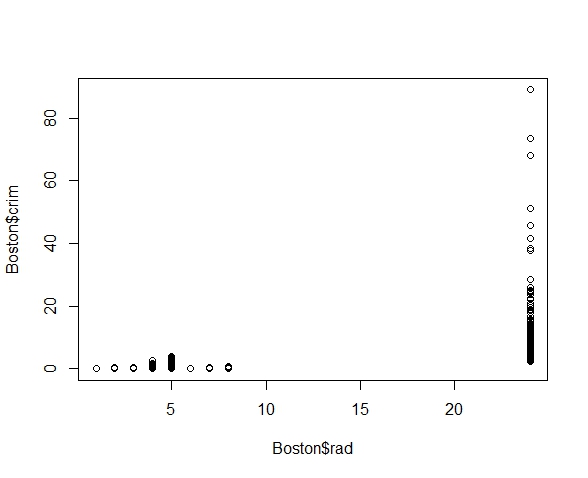
The neighborhood with more old homes tends to have higher crime rate.



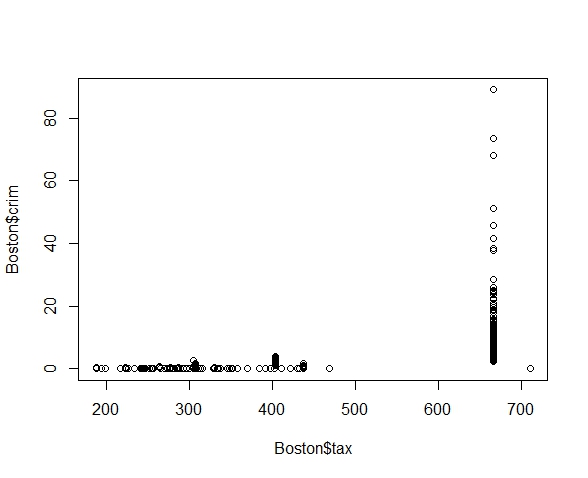
Being closer to work area is associated with higher crime rate.



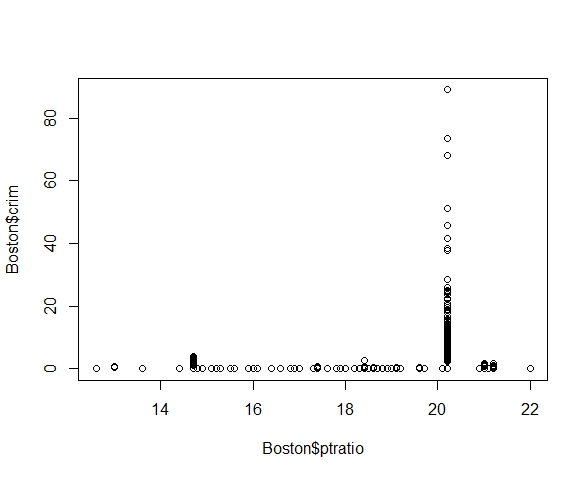
Higher index of accessibility to radial highways is associated with higher crime rate.



Higher tax rate is associated with higher crime rate.



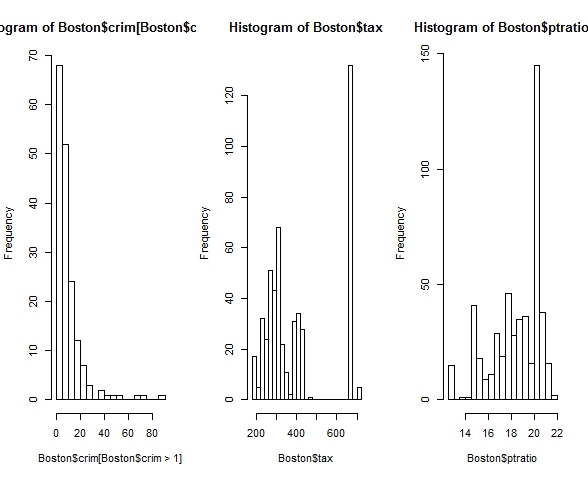
Higher student-teacher ratio is associated with higher crime rate.



1. As the three plots have shown below, most towns have crime rates that are not particularly high, but the histogram does have a pretty long tail, which indicates that there are still some towns that have very high crime rates. There are 18 towns that have crime rate higher than 20.

In terms of tax rates, a large gap can be observed between towns where tax rates are relatively low, and where tax rates are higher than 660.

In terms of student-teacher ratio, it can be observed from the histogram that it’s skewed towards the high ratio end, but there doesn’t seem to be a particularly high ratio (range: 14-22)



1. 35 of the suburbs in this data set bound the Charles River.

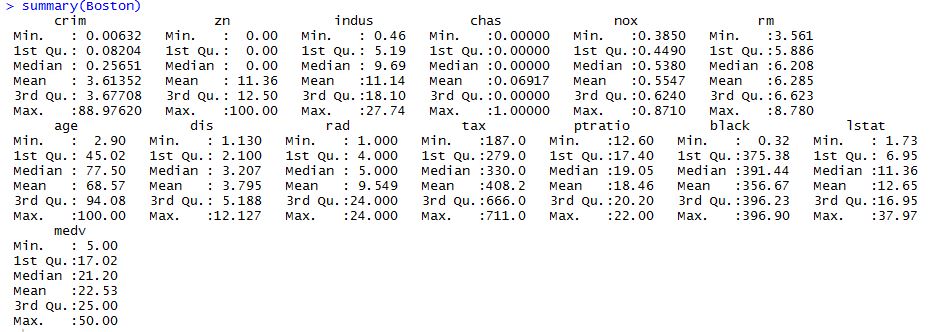
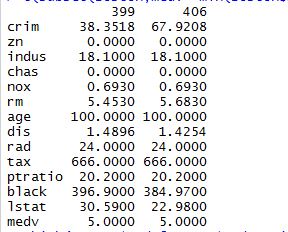


1. The median is 19.05.

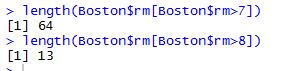


1. Observations 399 and 406 have the lowest median value of owner-occupied homes. The values of the other predictors for these suburbs are shown below. Compared to the overall ranges for those predictors, these two suburbs have:
   1. crim – larger than the 3rd quartile
   2. zn – is the minimum
   3. indus – is the 3rd quartile
   4. not bounded by Charles river
   5. nox – larger than the 3rd quartile
   6. rm – even smaller than the 1st quartile
   7. age – is the maximum
   8. dist – even smaller than the 1st quartile
   9. rad -- is the maximum
   10. tax – is the 3rd quartile
   11. ptratio – is the 3rd quartile
   12. black at max for observation 399 and larger than the 1st quartile for observation 406
   13. lstat – larger than the 3rd quartile
   14. medv – is the minimum

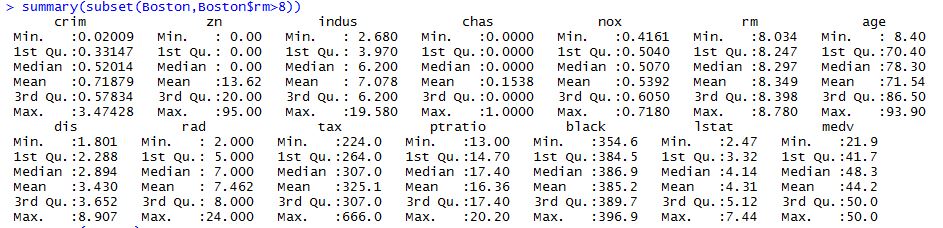
It looks to me that these two suburbs are neither the best nor the worst to live in; the two suburbs have their own pros and cons. The two suburbs are high in some of the features positively relating to crime rate, but are also low in some of the features positively relating to crime rate.



1. 64 suburbs average more than 7 rooms, and 13 suburbs average more than 8 rooms



Below is the summary of the subset with more than 8 rooms per dwelling:



By comparing the summary above with the summary of the whole dataset in (g), we can find that suburbs average more than 8 rooms per dwelling tend to have lower crime rate and lstat.