Patrick Rogan
UNI: psr2125

STAT4702 Homework 7

2.4.1

a. Given an extremely large sample size n and a small number of predictors p, a more flexible method would generally be better as we have sufficient data to make an accurate estimate of f.

b. In the opposite case of a, a more inflexible method would generally be better as a more flexible model has the risk of overfitting in this situation.

c. Given a highly nonlinear relationship between predictors and response, a more flexible model is desirable as an inflexible model may not be able to accurately reflect the behaviour of f.

d. If $\sigma^2 = Var(\epsilon)$ a less flexible method would be better as a more flexible method would tend to follow the "noise" instead of the underlying behaviour of f.

2.4.2

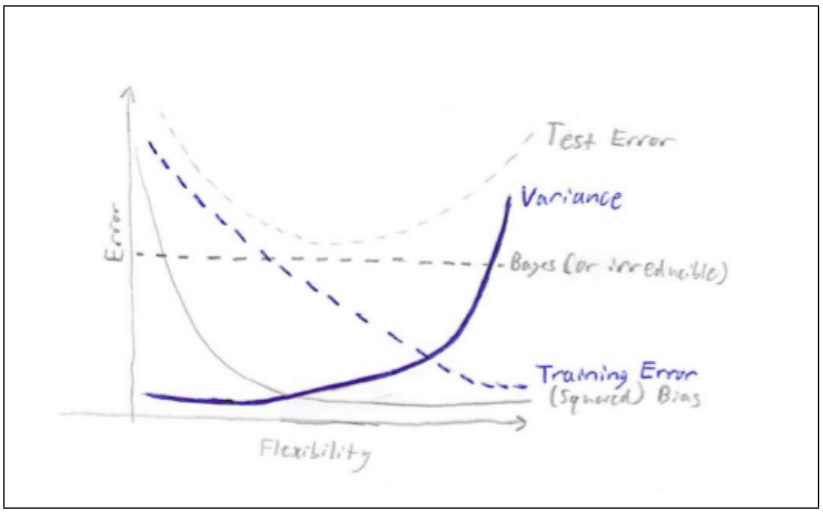
a. As CEO salary is a quantitative response, this is a regression problem and since we are interested in the factors that determine salary this is an inference problem.
$n = 500$
$p = 3$

b. As success and failure are qualitative responses, this is a classification problem; since we are trying to determine an outcome this is a prediction problem.
$n = 20$
$p = 14$

c. As we are interested in predicting % change, a quantitative prediction, we have a regression-prediction problem.
$n = 52$
$p = 4$

2.4.3

a.



b.

- Bayes (or irreducible) error is constant as this is the minimum achievable error.
- Training error goes down constantly as the more flexible our model is the closer it will fit the data.
- Test error will decrease up to a point and then begin to increase in situations where the model has become too flexible and overfits the data.
- Variance will tend to increase with the complexity of the model.

- (Squared) Bias will typically decrease as the flexibility increases as more features of the data will be reflected in the model.

2.4.4

a.

i. Determining whether someone should receive a loan. The response would be yes/no and predictors would include income, credit score and criminal history. This is a prediction problem as the goal is to determine if the person will pay back the loan.

ii. Predicting whether an employee will leave a job. The response would be likely/unlikely and the predictors would be salary, number of sick days taken and performance review ratings. This is a prediction problem.

iii. Determining what factors contribute to someone completing a college education. The response would be yes/no for completing a four year degree and the predictors would be SAT scores, parent's household income and high school GPA. This is a classification problem.

b.

i. Predicting future gasoline price, the response would be the price/gallon and predictors would include number of active oil wells, global demand for crude oil and refinery output (the latter two given in number of barrels/time period).

ii. Identifying which factors are important in determining how favourably someone reviews a movie. The response would a rating of the film from 1-10 and the predictors would be which actors were involved, movie run time, and genre. Since we are interested in the contributing factors this is an inference problem.

iii. Predicting ambulance response time. The response would be time in minutes, predictors would be time of day, weather conditions and ambulance driver seniority. This is also a prediction problem.

c.

i. Grouping adults by age, body mass index, smoker/nonsmoker and consumer of alcoholic beverages/nonconsumer. This could be used to group individuals to identify cancer risk.

ii. Grouping items purchased by customers on an online shopping website to provide suggestions for other customers.

iii. Grouping individuals based on number of bank accounts, frequency of withdrawals/deposits, income and past criminal history to detect potential tax fraud.

2.4.5

Very flexible approaches to regression/classification produce models that follow the data closely. This has the advantage of potentially more closely following the underlying behaviour of f. A more flexible approach would be more appropriate when attempting to fit a model that had erratic behaviour or when there are a large number of observations. A less flexible approach would be preferred when there are few observations and the underlying behaviour of f is well behaved (say highly linear).

2.4.6

Parametric statistical reduces the problem of attributing the behaviour of f to a select number of parameters. Non-parametric approaches rely solely on the behaviour of the observations and assume nothing of f. Parametric regression/classification will provide a model that is easier to interpret and will produce a model that is better for prediction. The disadvantages of parametric approaches come from the need to assume something about the underlying form of f. If the parametric model is incorrect, conclusions drawn could also be incorrect.

2.4.7

a.

Obs.	X_1	X_2	X_3	Y	d
1	0	3	0	Red	3
2	2	0	0	Red	2
3	0	1	3	Red	$\sqrt{10}$
4	0	1	2	Green	$\sqrt{5}$
5	-1	0	1	Green	$\sqrt{2}$
6	1	1	1	Red	$\sqrt{3}$

b. For $K = 1$ we classify the point as green as Observation #5 is the closest point to 0,0,0.

c. For $K = 3$ we classify the point as red as Observations #2, #5 and #6 are closest to 0,0,0.

d. For a highly nonlinear Bayes decision boundary, we would expect the best value of K would be small allowing for a more flexible boundary creation by KNN.

2.4.8

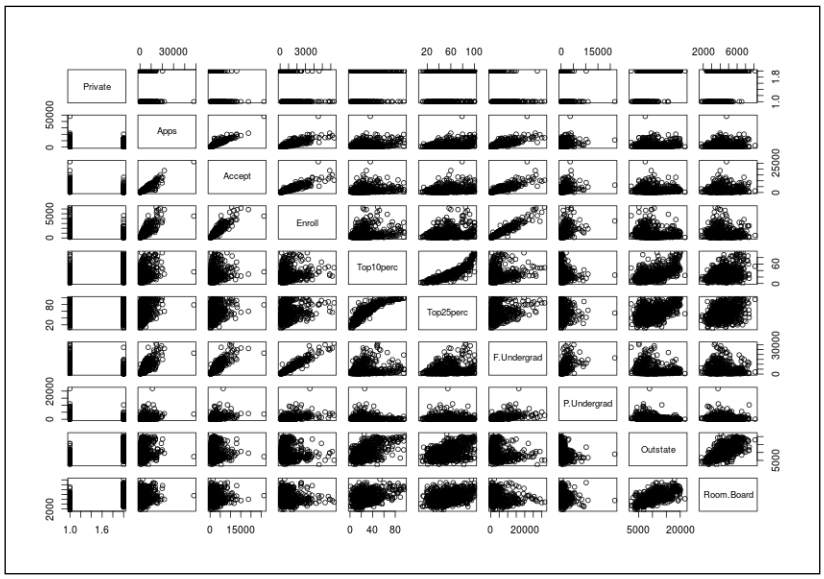
a. See Supporting Code Section

b. See Supporting Code Section

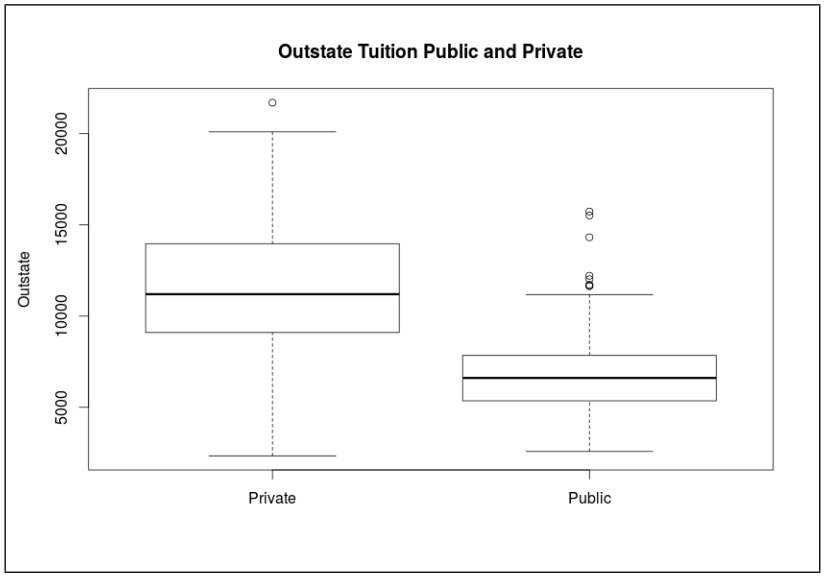
c. See Supporting Code Section

i. See Supporting Code Section

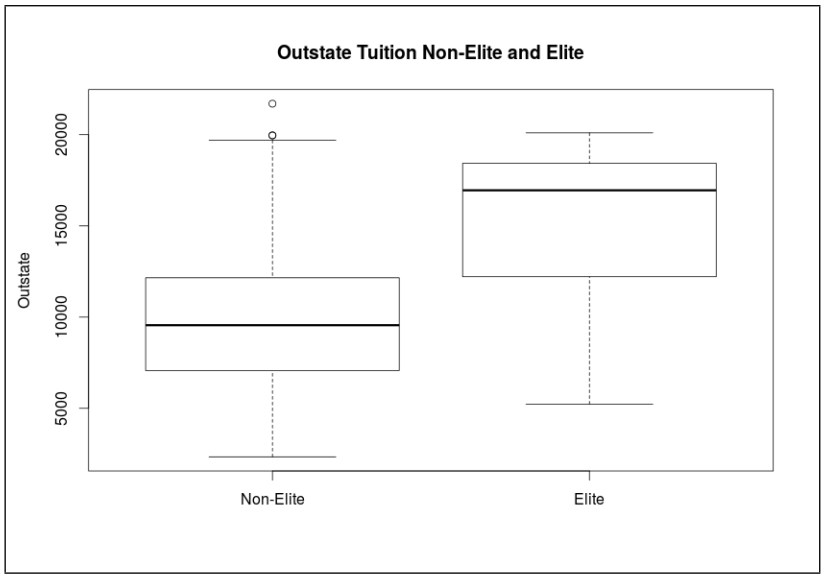
ii.



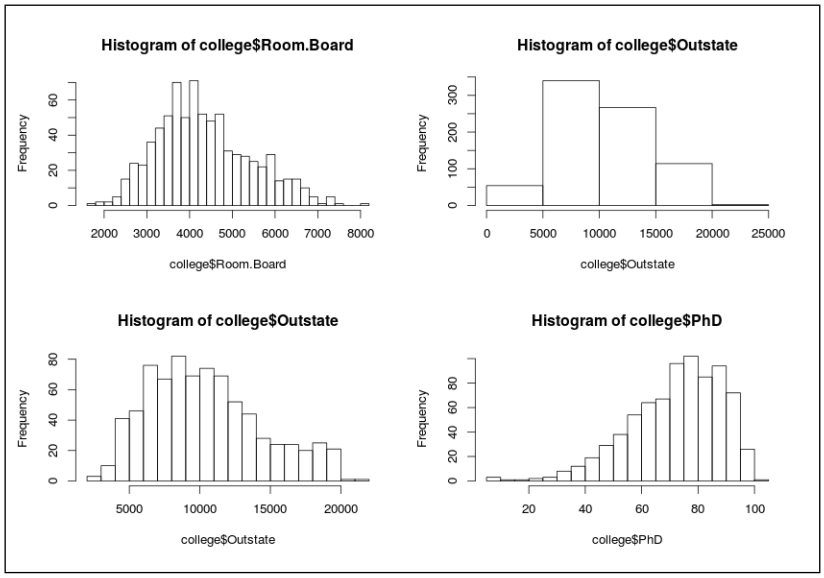
iii.



iv.



v.



vi. The graduation rate at elite institutions is higher than that of non-elite institutions, similarly the graduation rate is higher at private institutions than at non-private. Cost of tuition, room and board, expenses and even the cost of books appears to be higher at elite institutions.

2.4.9

a.

Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year
Qualitative: name, origin (numeric values, but a categorical variable)

b., c.

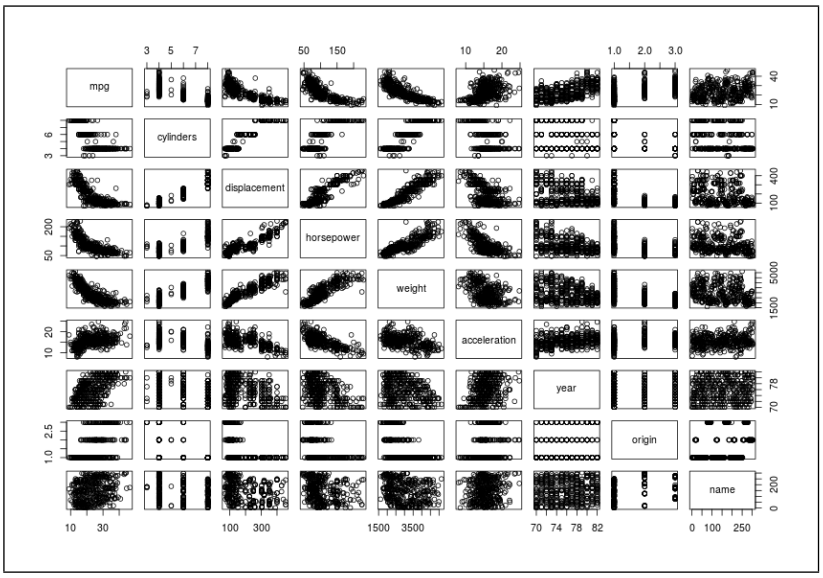
	Range	μ	σ
mpg	37.6	23.45	7.81
cylinders	5	5.5	1.7
displacement	387	194.4	104.6
horsepower	184	104.5	38.5
weight	3537	2977.6	849.4
acceleration	16.8	15.54	2.76
year	12	76.0	3.7

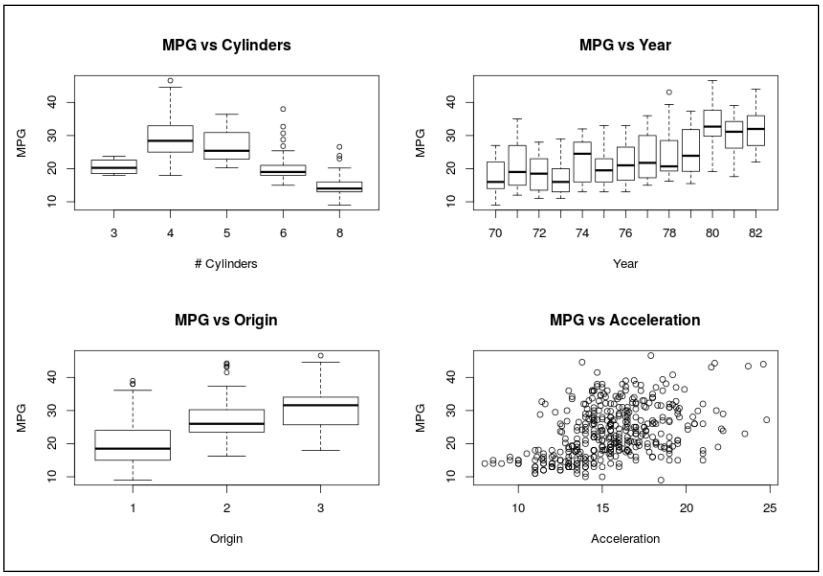
d.

	Range	μ	σ
mpg	35.6	24.40	7.87
cylinders	5	5.4	1.7
displacement	387	187.2	99.7
horsepower	184	100.7	35.7
weight	3348	2936.0	811.3
acceleration	16.3	15.73	2.69
year	12	77.1	3.1

e.

There appear to be many correlated predictors (both positively and negatively correlations. For example, displacement, weight and horsepower all appear positively correlated while these three variables appear negatively correlated with acceleration. MPG also has several correlations (see f.).





f.

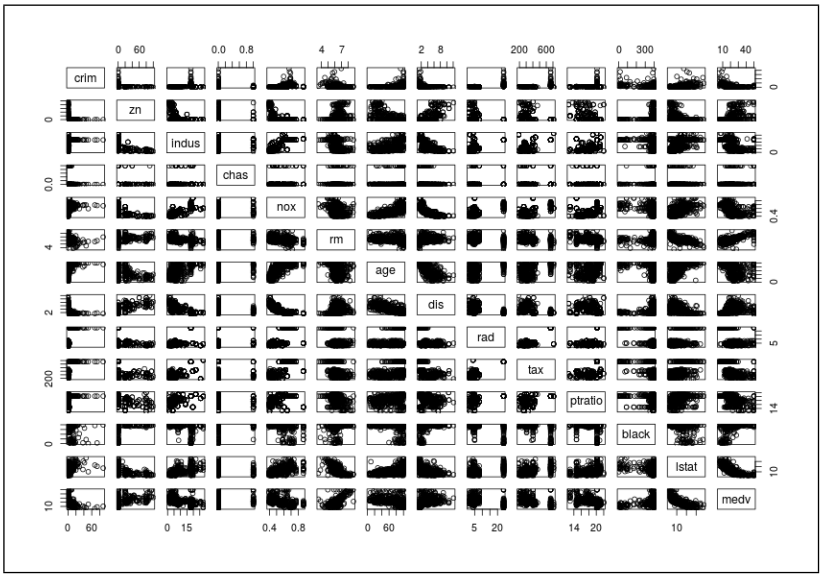
With the exception of predictors "name" and possibly "acceleration", most of the remaining variables appear to have at least a weak correlation with mpg, suggesting they might be useful for prediction.

2.4.10

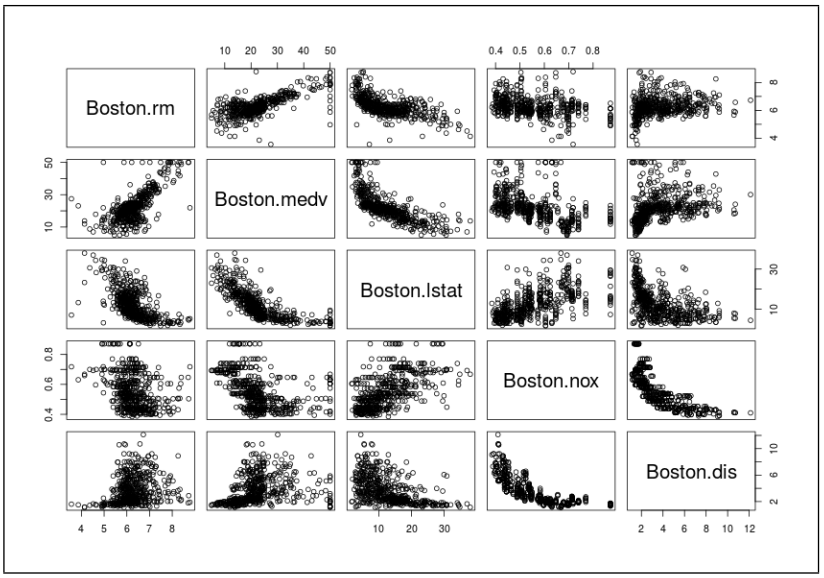
a.

The dataset has 506 rows and 14 columns. The data describes housing values in the regions around Boston; columns represent characteristics of the regions and rows correspond to individual entries.

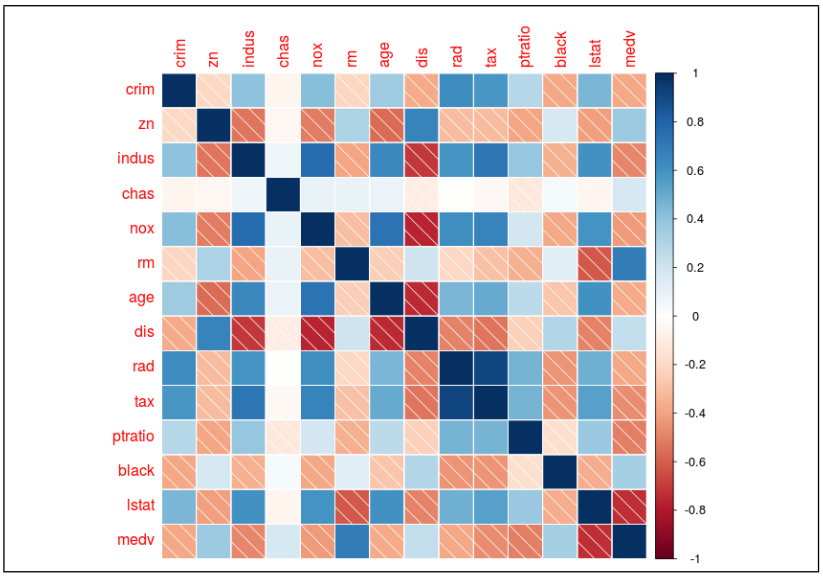
b.



Some of the more pronounced correlations in the data are between rm, medv, lstat, nox and dis, these are separately plotted below:

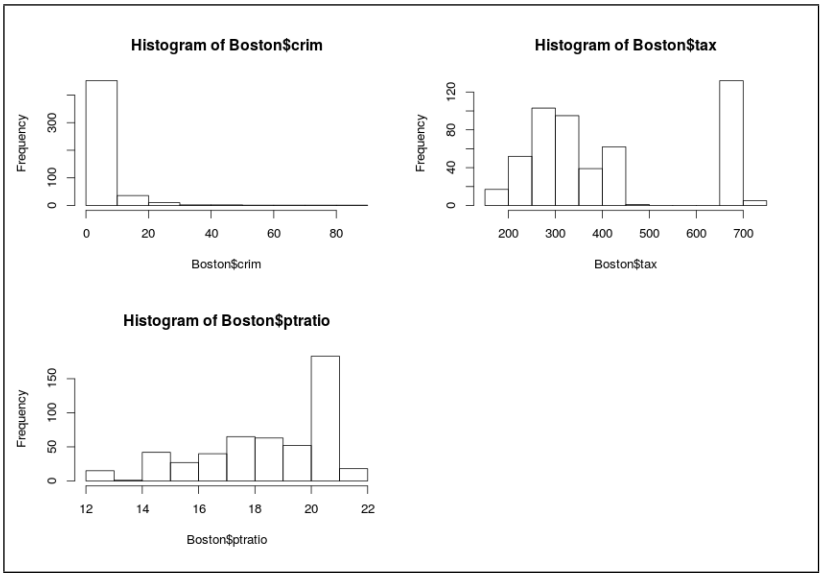


c. The following variables are associated with crime per capita: indus, nox, age, rad, tax, pratio and lstat. These predictors appear to indicate that older, industrial areas with lower population status tend to have more crime.



d.

	Range	Min	Max	Median
crime rate	88.97	0.00632	88.9762	0.25651
tax rate	524	187	711	330
pupil teacher ratios	9.4	12.6	22	19.05



Looking at the 97th percentile (which we'll call particularly high) of each variable:

```
       crim zn indus chas   nox    rm  age    dis rad tax ptratio black lstat medv
381 88.9762  0  18.1    0 0.671 6.968 91.9 1.4165  24 666    20.2 396.9 17.21 10.4
```

```
       crim zn indus chas   nox    rm  age    dis rad tax ptratio  black lstat medv
489 0.15086  0 27.74    0 0.609 5.454 92.7 1.8209   4 711    20.1 395.09 18.06 15.2
490 0.18337  0 27.74    0 0.609 5.414 98.3 1.7554   4 711    20.1 344.05 23.97  7.0
491 0.20746  0 27.74    0 0.609 5.093 98.0 1.8226   4 711    20.1 318.43 29.68  8.1
492 0.10574  0 27.74    0 0.609 5.983 98.8 1.8681   4 711    20.1 390.11 18.07 13.6
493 0.11132  0 27.74    0 0.609 5.983 83.5 2.1099   4 711    20.1 396.90 13.35 20.1
```

```
       crim zn indus chas   nox    rm  age     dis rad tax ptratio  black lstat medv
355 0.04301 80  1.91    0 0.413 5.663 21.9 10.5857   4 334      22 382.80  8.05 18.2
356 0.10659 80  1.91    0 0.413 5.936 19.5 10.5857   4 334      22 376.04  5.57 20.6
```

- Crime rate has an exceptionally large range, looking into the data there appears to be a large number of very low crime rate regions, several higher crime rate regions with the maximum appearing to be an outlier.
- Tax rate also has a large range with most rates either being in a low/medium range < 500 or in a high range > 600.
- The range of student to teacher ratio also has a fairly large range with the least taxed teacher having nearly 10 students less than the most.

e. 35

f. 19.05

g. Both 399 and 406 have the lowest median value of owner-occupied homes.

```
       crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98    5
```

	Range	Min	Max	Median
crim	88.96988	0.00632	88.9762	0.25651
zn	100	0	100	0
indus	27.28	0.46	27.74	9.69
chas	1	0	1	0
nox	0.486	0.385	0.871	0.538
rm	5.219	3.561	8.78	6.2085
age	97.1	2.9	100	77.5
dis	10.9969	1.1296	12.1265	3.20745
rad	23	1	24	5
tax	524	187	711	330
ptratio	9.4	12.6	22	19.05
black	396.58	0.32	396.9	391.44
lstat	36.24	1.73	37.97	11.36
medv	45	5	50	21.2

These two suburbs have higher crime rates, are more industrial, have older buildings, have greater access to radial highways and contains a higher percentage of lower status population than most other regions.

h. > 7 Rooms/dwelling: 64
> 7 Rooms/dwelling: 13

Looking at the the suburbs with greater than eight rooms per dwelling on average, there don't appear to be any factors that are substantially different from the population of suburbs as a whole.

```
       crim zn indus chas    nox    rm  age    dis rad tax ptratio  black lstat medv
98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0 396.90  4.21 38.7
164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7 388.45  3.32 50.0
205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7 390.55  2.88 50.0
225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4 385.05  4.14 44.8
226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4 382.00  4.63 50.0
227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4 387.38  3.13 37.6
233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4 385.91  2.47 41.7
234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4 378.95  3.95 48.3
254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1 396.90  3.54 42.8
258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0 389.70  5.12 50.0
263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0 386.86  5.91 48.8
268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0 384.54  7.44 50.0
365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2 354.55  5.29 21.9
```

```
#########################################################################################################
#################################### Supporting Code ####################################################
#########################################################################################################
### 2.4.8 a
college = read.csv('/home/pat/Desktop/STAT4702/ISL_Datasets/College.csv')

### 2.4.8 b
fix(college)
rownames(college) = college[,1]
fix(college)
college = college[,-1]
fix(college)

### 2.4.8 c
summary(college)

### 2.4.8 c ii
pairs(college[,1:10])

### 2.4.8 c iii
boxplot(college[college$Private == "Yes",9],
        college[college$Private =="No",9],
        names = c("Private","Public"),ylab="Outstate",
        main="Outstate Tuition Public and Private")

### 2.4.8 c iv
Elite = rep("No", nrow(college))
Elite[college$Top10perc >50] = " Yes "
Elite = as.factor(Elite)
college = data.frame(college, Elite)
boxplot(college$Outstate ~ college$Elite,
        names = c("Non-Elite","Elite"),
        ylab="Outstate",
        main="Outstate Tuition Non-Elite and Elite")

### 2.4.8 c v
par(mfrow = c(2,2))
hist(college$Room.Board, breaks = 30)
hist(college$Outstate, breaks = 5)
hist(college$Outstate, breaks = 18)
hist(college$PhD, breaks = 20)

### 2.4.9
Auto = read.csv("/home/pat/Desktop/STAT4702/ISL_Datasets/Auto.csv",
                header = T, na.strings = "?")
Auto = na.omit(Auto)

### 2.4.9 b/c
for(i in 1:(length(Auto[1,])-2)){
  print(c(diff(range(Auto[,i])),"&", mean(Auto[,i]),"&",sqrt(var(Auto[,i]))))
  }

### 2.4.9 d
Auto2 = Auto[-(10:85),]
for(i in 1:(length(Auto2[1,])-2)){
  print(c(diff(range(Auto2[,i])),"&", mean(Auto2[,i]),"&",sqrt(var(Auto2[,i]))))
  }

pairs(Auto)
```

```
### 2.4.9 e
par(mfrow = c(2,2))
boxplot(Auto$mpg ~ Auto$cylinders,
        ylab="MPG", xlab = "# Cylinders",
        main="MPG vs Cylinders")

boxplot(Auto$mpg ~ Auto$year,
        ylab="MPG", xlab = "Year",
        main="MPG vs Year")

boxplot(Auto$mpg ~ Auto$origin,
        ylab="MPG", xlab = "Origin",
        main="MPG vs Origin")

plot(Auto$acceleration, Auto$mpg,
     main = "MPG vs Acceleration",
     ylab = "MPG", xlab = "Acceleration")

### 2.4.10 a
library(MASS)
Boston
dim(Boston)
?Boston

### 2.4.10 c
library(corrplot)
corrplot(cor(Boston),method = "shade")

### 2.4.10 d
Boston[which(Boston$crim > 0.97*max(Boston$crim)),]
Boston[which(Boston$tax > 0.97*max(Boston$tax)),]
Boston[which(Boston$ptratio > 0.97*max(Boston$ptratio)),]

### 2.4.10 e
sum(Boston$chas)

### 2.4.10 f
median(Boston$ptratio)

### 2.4.10 g
Boston[which(Boston$medv == min(Boston$medv)),]

### 2.4.10 h
summary(Boston$rm>7)
summary(Boston$rm>8)

Boston[which(Boston$rm > 8),]
```