ISLR
Chapter 2 Ex. 10
Tatiana Romanchishina

a). To begin, load in the Boston data set. The Boston data set is
part of the MASS library in R. How many rows are in this data set?
How many columns? What do the rows and columns represent?

Load the data set:
>library(MASS)
>Boston
>?Boston

Number of rows and columns:
>dim(Boston)
[1] 506 14 --> 506 rows and 14 columns

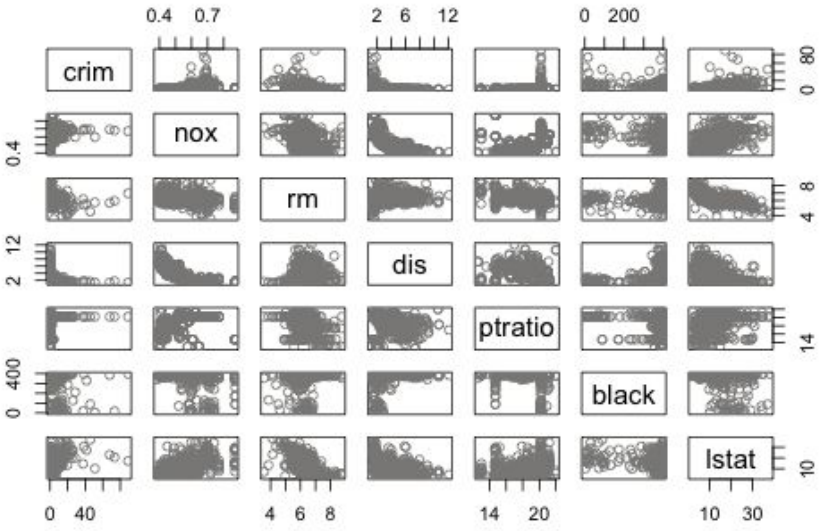
What they represent:
> names(Boston)
"crim" --> crime rate per capita by town
"zn"   --> proportion of residential land zoned for lots over 25,000 sq.ft.
"indus" --> proportion of non-retail business acres per town
"chas" --> Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
"nox"   --> nitrogen oxides concentration (parts per 10 million)
"rm"   --> average number of rooms per dwelling
"age"   --> proportion of owner-occupied units built prior to 1940
"dis"   --> weighted mean of distances to five Boston employment centres
"rad"   --> index of accessibility to radial highways
"tax"   --> full-value property-tax rate per \$10,000
"ptratio" --> pupil-teacher ratio by town
"black" --> 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
"lstat" --> lower status of the population (percent)
"medv" --> median value of owner-occupied homes in \$1000s

(b) Make some pairwise scatterplots of the predictors (columns) in
this data set. Describe your findings.
# Pairwise scatterplots:
attach(Boston)
pairs(Boston)
#All pairs
pairs(Boston, col="snow4")
The previous command creates a very large graph that was hard to read and understand.

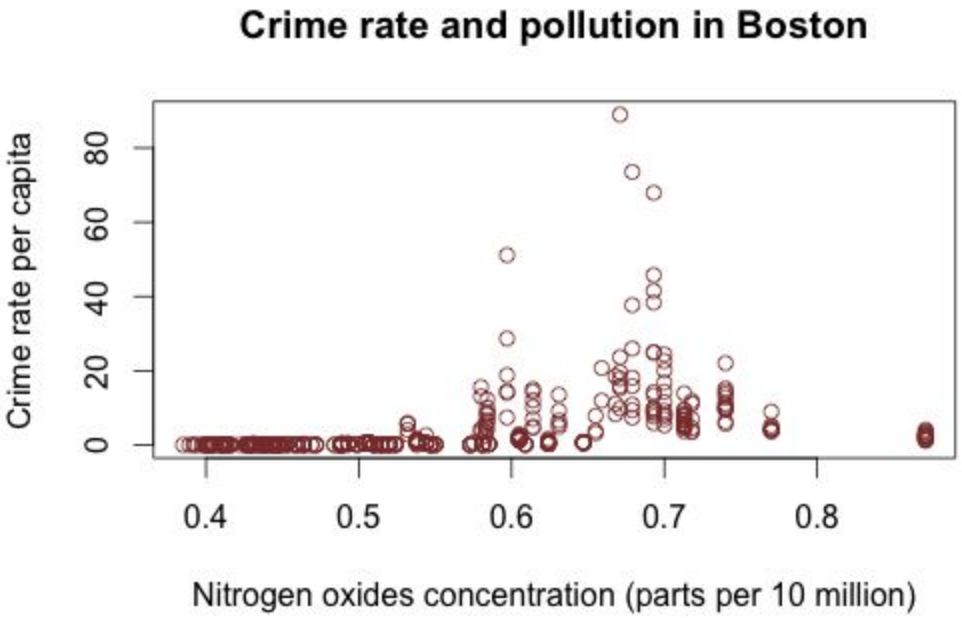
That's why I picked several pairs at a time as demonstrated next, then I considered several pairs individually.

#Some pairs

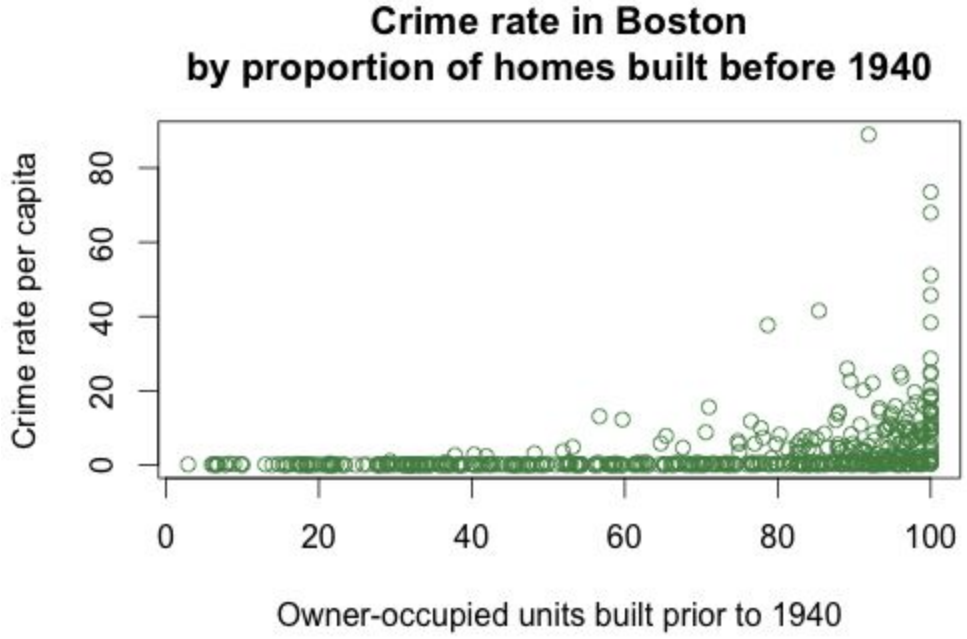
pairs(~ crim + nox + rm + dis + ptratio + black + lstat, Boston, col="snow4")



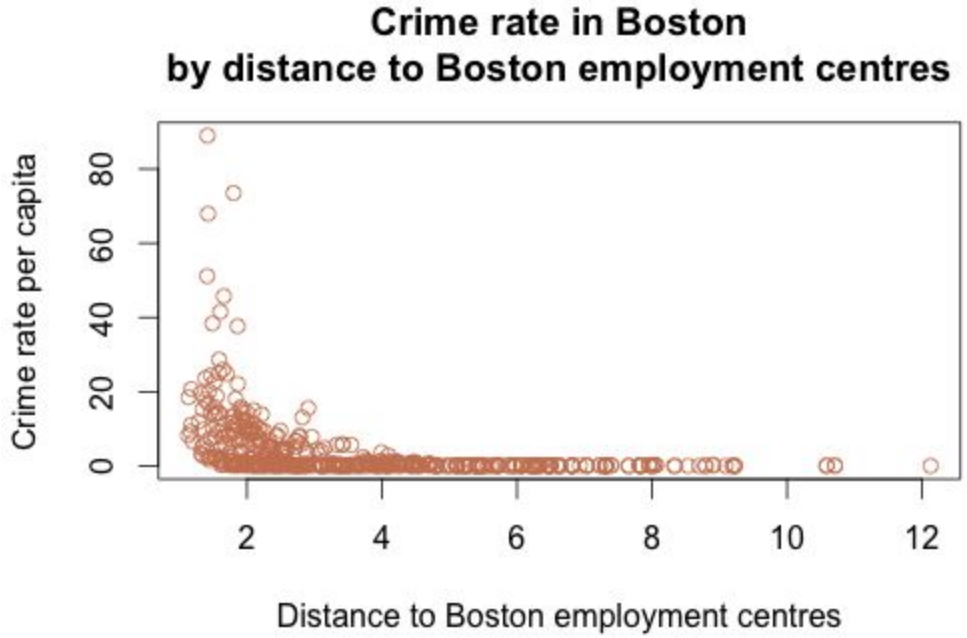
1. > plot(nox, crim, col="indianred4", ylab= "Crime rate per capita", xlab= "Nitrogen oxides concentration (parts per 10 million)", main= "Crime rate and pollution in Boston")



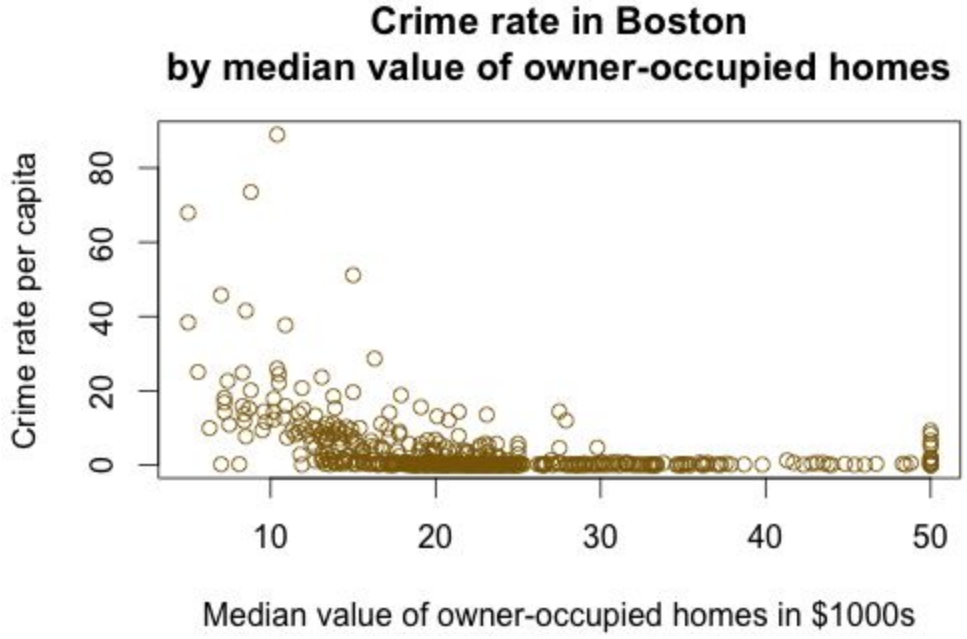
2. >plot(age, crim, col="palegreen4", xlab= "Owner-occupied units built prior to 1940", ylab= "Crime rate per capita", main= "Crime rate in Boston\nby proportion of homes built before 1940")



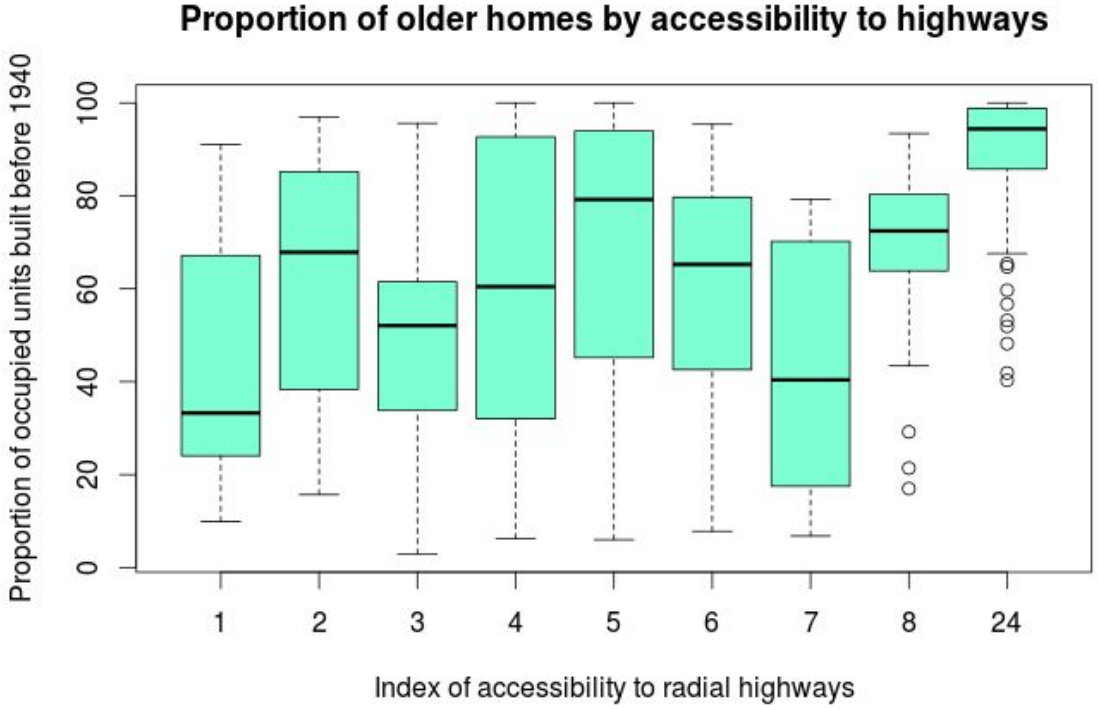
3. > plot(dis, crim, col="lightsalmon3", xlab= "Distance to Boston employment centres", ylab= "Crime rate per capita", main= "Crime rate in Boston\nby distance to Boston employment centres")



4. > plot(medv, crim, col="darkgoldenrod4", xlab= "Median value of owner-occupied homes in $1000s", ylab= "Crime rate per capita", main= "Crime rate in Boston\nby median value of owner-occupied homes")



5. plot(as.factor(rad), age, col="aquamarine")



Observations by chart number:

1. It seems that in the areas of Boston where the concentration of nitrogen oxides is higher than 0.6, the crime rate is higher as well. However, when the concentration gets really high, over 0.8, the crime rate gets close to 0 again.
2. Boxplot 2 shows the correlation between crime rate in Boston and the proportion of owner-occupied houses that were built before 1940. The places with a majority of houses built before 1940 have a higher crime rate.
3. Boxplot 3 shows that the suburbs that are really close to employment centers have a higher crime rate.
4. Boxplot 4 shows that suburbs with a lower median value of owner-occupied houses have a higher crime rate. The crime rate slightly goes up in the suburbs with the highest median value of houses.
5. The proportion of houses built before 1940 is lowest in the suburbs of Boston closest to radial highways. Places that are farthest from radial highways have the highest proportion of houses built before 1940.

(c) Are any of the predictors associated with per capita crime rate?
If so, explain the relationship.

- Boxplot 2 shows the correlation between crime rate in Boston and the proportion of owner-occupied houses that were built before 1940. The places with a majority of houses built before 1940 have a higher crime rate.
- Boxplot 3 shows that the suburbs that are really close to one of the five employment centers have a higher crime rate.
- Boxplot 3 shows that the suburbs that are really close to one of the five employment centers have a higher crime rate.
- Boxplot 4 shows that suburbs with a lower median value of owner-occupied houses have a higher crime rate. The crime rate slightly goes up in the suburbs with the highest median value of houses.

(d) Do any of the suburbs of Boston appear to have particularly high:
Crime rates?

```
> boston_bycrime <- Boston[order(-crim),]
> head(boston_bycrime, n=10)
        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   20.2 396.90 17.21 10.4
419 73.5341  0  18.1  0 0.679 5.957 100.0 1.8026  24 666  20.2  16.45 20.62  8.8
406 67.9208  0  18.1  0 0.693 5.683 100.0 1.4254  24 666  20.2 384.97 22.98  5.0
411 51.1358  0  18.1  0 0.597 5.757 100.0 1.4130  24 666  20.2   2.60 10.11 15.0
415 45.7461  0  18.1  0 0.693 4.519 100.0 1.6582  24 666  20.2  88.27 36.98  7.0
405 41.5292  0  18.1  0 0.693 5.531  85.4 1.6074  24 666   20.2 329.46 27.38  8.5
399 38.3518  0  18.1  0 0.693 5.453 100.0 1.4896  24 666  20.2 396.90 30.59  5.0
428 37.6619  0  18.1  0 0.679 6.202  78.7 1.8629  24 666   20.2  18.82 14.52 10.9
```

414 28.6558 0 18.1 0 0.597 5.155 100.0 1.5894 24 666 20.2 210.97 20.08 16.3
418 25.9406 0 18.1 0 0.679 5.304 89.1 1.6475 24 666 20.2 127.36 26.64 10.4

Tax rates?
> boston_bytax <- Boston[order(-tax),]
> head(boston_bytax, n=10)
       crim   zn indus ch 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
357 8.98296 0 18.10 1 0.770 6.212 97.4 2.1222 24 666   20.2 377.73 17.60 17.8
358 3.84970 0 18.10 1 0.770 6.395 91.0 2.5052 24 666   20.2 391.34 13.27 21.7
359 5.20177 0 18.10 1 0.770 6.127 83.4 2.7227 24 666   20.2 395.43 11.48 22.7
360 4.26131 0 18.10 0 0.770 6.112 81.3 2.5091 24 666   20.2 390.74 12.67 22.6
361 4.54192 0 18.10 0 0.770 6.398 88.0 2.5182 24 666   20.2 374.56 7.79 25.0

Pupil-teacher ratios?
> boston_byptratio <- Boston[order(-ptratio),]
> head(boston_byptratio, n=10)
         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   22.0 382.80 8.05 18.2
356 0.10659 80 1.91 0 0.413 5.936 19.5 10.5857   4 334   22.0 376.04 5.57 20.6
128 0.25915 0 21.89 0 0.624 5.693 96.0 1.7883   4 437   21.2 392.11 17.19 16.2
129 0.32543 0 21.89 0 0.624 6.431 98.8 1.8125   4 437   21.2 396.90 15.39 18.0
130 0.88125 0 21.89 0 0.624 5.637 94.7 1.9799   4 437   21.2 396.90 18.34 14.3
131 0.34006 0 21.89 0 0.624 6.458 98.9 2.1185   4 437   21.2 395.04 12.60 19.2
132 1.19294 0 21.89 0 0.624 6.326 97.7 2.2710   4 437   21.2 396.90 12.26 19.6
133 0.59005 0 21.89 0 0.624 6.372 97.9 2.3274   4 437   21.2 385.76 11.12 23.0
134 0.32982 0 21.89 0 0.624 5.822 95.4 2.4699   4 437   21.2 388.69 15.03 18.4
135 0.97617 0 21.89 0 0.624 5.757 98.4 2.3460   4 437   21.2 262.76 17.31 15.6

Comment on the range of each predictor.
Crime rate range is very wide:
> range(crim)
Min: 0.00632 Max: 88.97620
The tax rate range seems wide, but hard for me to evaluate:
> range(tax)
Min: 187 Max: 711
The pupil-teacher ratio does not seem too wide:
> range(ptratio)
Min: 12.6 Max: 22.0

(e) How many of the suburbs in this data set bound the Charles river?

```
> table(chas)
chas
  0   1
471  35
```

35 suburbs bound the river.

(f) What is the median pupil-teacher ratio among the towns in this data set?

```
> summary(ptratio)
   Min. 1st Qu.  Median          Mean 3rd Qu.  Max.
  12.60   17.40   19.05   18.46   20.20   22.00
```

Pupil-teacher ratio median is 19.05.

(g) Which suburb of Boston has lowest median value of owner-occupied homes?

```
> subset(Boston, medv==min(medv))
        crim    zn  ind    ch  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
```

What are the values of the other predictors for that suburb, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

- Crime rate: Min: 0.00632 Max: 88.97620. #399 is in the lower half of the range, while #406 has a very high crime rate.
- Proportion of residential land zoned for lots over 25,000 sq.ft.: Min: 0 Max: 100. Neither suburb has any residential land zoned for lots over 25000 sq.ft.
- Proportion of non-retail business acres per town: Min: 0.46 Max: 27.74. Both suburbs have the proportion of 18.1.
- Charles River variable (= 1 if tract bounds river; 0 otherwise). Both have 0.
- Nitrogen oxides concentration (parts per 10 million): Min: 0.385 Max: 0.871. Both have a high concentration of NOs.
- Average number of rooms per dwelling: Min: 3.561 Max: 8.780. Both have about 6 rooms per dwelling on average.
- Proportion of owner-occupied units built prior to 1940: Min: 2.9 Max: 100.0. Both suburbs consist of homes built before 1940.
- Weighted mean of distances to five Boston employment centres: Min: 1.1296 Max: 12.1265. Both are almost equally very close to employment centers.
- Index of accessibility to radial highways (ranges 1-8 and 24). Both have the index of 24, which means they are least accessible to radial highways.

- Full-value property-tax rate per \$10,000: Min: 187 Max: 711. Both have a high tax rate.
- Pupil-teacher ratio by town: Min: 12.6 Max: 22.0. Both have the ratio of 20:1, which is very close to the maximum for Boston.
- 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town: Min: 0.32 Max: 396.90. Both have a high proportion of black people.
- Lower status of the population (percent): Min: 1.73 Max: 37.97. Both have values close to the maximum for Boston.
- Median value of owner-occupied homes in $1000s: Min: 5 Max: 50. Both have the minimum value.

(h) In this data set, how many of the suburbs average more than
seven rooms per dwelling?
> mt7rooms <- subset(Boston, rm > 7)
> dim(mt7rooms)
[1] 64 14
64 suburbs have have 7 rooms per dwelling on average.

More than eight rooms per dwelling?
> mt8rooms <- subset(Boston, rm > 8)
> dim(mt8rooms)
[1] 13 14
13 suburbs have 8 rooms per dwelling on average.

Comment on the suburbs that average more than eight rooms per dwelling.
> stargazer(mt8rooms, type="text") ## for slightly better formatting than summary

Statistic	N	Mean	St. Dev.	Min	Max
crim	13	0.719	0.902	0.020	3.474
zn	13	13.615	26.298	0	95
indus	13	7.078	5.393	2.680	19.580
chas	13	0.154	0.376	0	1
nox	13	0.539	0.092	0.416	0.718
rm	13	8.349	0.251	8.034	8.780
age	13	71.538	24.609	8.400	93.900
dis	13	3.430	1.884	1.801	8.907
rad	13	7.462	5.333	2	24
tax	13	325.077	110.971	224	666
ptratio	13	16.362	2.411	13.000	20.200
black	13	385.211	10.529	354.550	396.900
lstat	13	4.310	1.374	2.470	7.440
medv	13	44.200	8.092	21.900	50.000

These suburbs can be characterized by:

- a low crime rate
- maximum proportion of black people for the Boston area
- median value of homes close to the maximum for Boston
- higher NOs concentrations

Other predictors seem to have the ranges close to the whole dataset, thus they are less interesting.