Chapter 2 Statistical Learning

```
library(ISLR)
library(ISLR2)
```

```
##
## Attaching package: 'ISLR2'
```

```
## The following objects are masked from 'package:ISLR':
##
##     Auto, Credit
```

```
library(psych)
library(ggplot2)
```

```
##
## Attaching package: 'ggplot2'
```

```
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
```

Exercise 1

a

The performance of a flexible statistical learning method can be better since we might need underlying predictors.

b

An inflexible statistical learning method because the risk of overfitting has already been high, a flexible statistical learning method may exaggerate it

c

A flexible statistical learning method could be better as it provides more flexibility.

d

An inflexible statistical learning method since a flexible model would also fit the noise of the error term.

Exercise 2

a

This is a regression problem and we are interested in inference. Here n = 500 and p = 3.

b

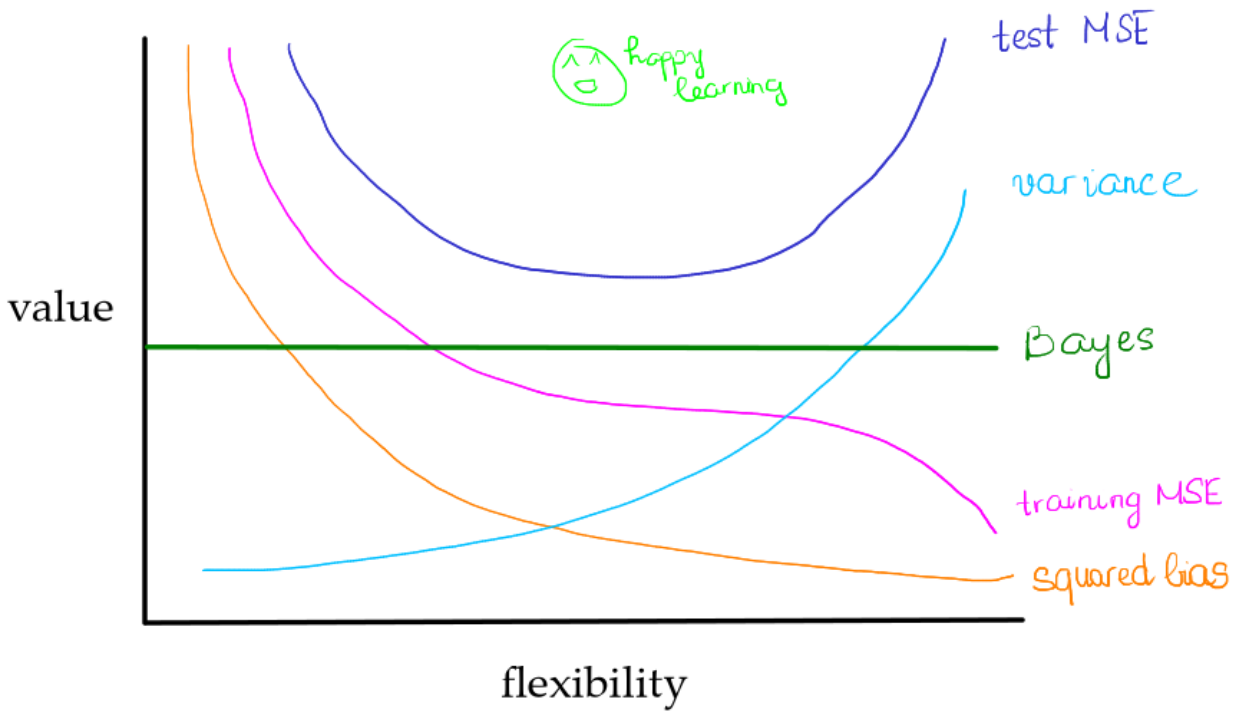
This is a classification problem and we are interested in prediction. Here n = 20 and p = 13.

c

This is a regression problem and we are interested in prediction. Here n = 52 and p = 3.

Exercise 3

a



b

More flexible statistical methods have higher variance because small changes in the training data can result in large changes in

$$\hat{f}$$

. On the other hand, bias refers to the error that is introduced by approximating a real-life problem. Generally, more flexible methods result in less bias.

In terms of training MSE and test MSE, as flexibility increases, the model fits training observations more closely, the training MSE therefore decreases. Test MSE, however, only goes down to a certain point then bounces off since the model poorly generalises on the test set.

And the Bayes curve (the irreducible error) is constant and smaller the the test error.

Exercise 4

a

1/ We want to predict if a data scientist's salary (response) is under or above the average based factors like education level, years of experience or age (predictors) . The goal is prediction.

2/ Traders want to predict whether a particular stock price will increase tomorrow (response) based on prices from previous days (predictors). The goal is prediction.

3/ Researchers want to know which indicators, gender, age at sexual debut or ethnicity (predictors) are highly associated with patients who exposured to HIV (response). The goal is inference.

b

1/ We want to know which factors like education level, years of experience or age (predictors) strongly affect a data scientist's salary (response). The goal is inference.

2/ Traders want to predict a particular stock price tomorrow (response) based on prices from previous days (predictors). The goal is prediction.

3/ Researchers want to predict the probability (response) that a patient has been exposured to HIV from three indicators: gender, age at sexual debut or ethnicity. The goal is prediction.

Exercise 5

A very flexible model might be able to figure out the underlying pattern (when p is small) and be better when the true relationship between the predictors and the response is non-linear. On the other hand, due to its complexity, we may have to sacrifice interpretability and may run the risk of overfitting (A less flexible approach is preferred in general when the goal is inference). Please revise exercise 2 for under which circumstances we prefer a more flexible approach or less flexible approach.

Exercise 6

Non-parametric methods do not make explicit assumptions about the functional form of f. Instead they seek an estimate of f that gets as close to the data points as possible without being too rough or wiggly. Such approaches can have a major advantage over parametric approaches: by avoiding the assumption of a particular functional form for f, they have the potential to accurately fit a wider range of possible shapes for f. Any parametric approach brings with it the possibility that the functional form used to estimate f is very different from the true f, in which case the resulting model will not fit the data well. In contrast, non-parametric approaches completely avoid this danger, since essentially no assumption about the form of f is made. But non-parametric approaches do suffer from a major disadvantage: since they do not reduce the problem of estimating f to a small number of parameters, a very large number of observations (far more than is typically needed for a parametric approach) is required in order to obtain an accurate estimate for f.

Exercise 7

```
data = c(0, 2, 0, 0, -1, 1, 0, 3, 0, 1, 1, 0, 1, 0, 0, 0, 3, 2, 1, 1, 0)
data = matrix(data, nrow = 7, ncol = 3)
data
```

```
##      [,1] [,2] [,3]
## [1,]    0    3    0
## [2,]    2    0    0
## [3,]    0    1    3
## [4,]    0    1    2
## [5,]   -1    0    1
## [6,]    1    1    1
## [7,]    0    0    0
```

a

```
dist(data, method = 'euclidean')
```

```
##          1        2        3        4        5        6
## 2 3.605551
## 3 3.605551 3.741657
## 4 2.828427 3.000000 1.000000
## 5 3.316625 3.162278 2.449490 1.732051
## 6 2.449490 1.732051 2.236068 1.414214 2.236068
## 7 3.000000 2.000000 3.162278 2.236068 1.414214 1.732051
```

b

Because the single nearest neighbour (observation number 5), is green so our prediction is green.

c

Among the three nearest neighbours (observations 5, 6 and 2), two of them are red so our prediction will be red. Or mathematically speaking,

$$P(Y = \text{Red}|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = \text{Red}) = \frac{1}{3}(1+0+1) = \frac{2}{3}$$

$$P(Y = \text{Green}|X = x_0) = \frac{1}{K}\sum_{i\in\mathcal{N}_0} I(y_i = \text{Green}) = \frac{1}{3}(0+1+0) = \frac{1}{3}$$

d

When the relationship is non-linear, we need a more flexible classification method, a larger 1/K which equivalent to a smaller K.

Exercise 8

a

```
head(College)
```

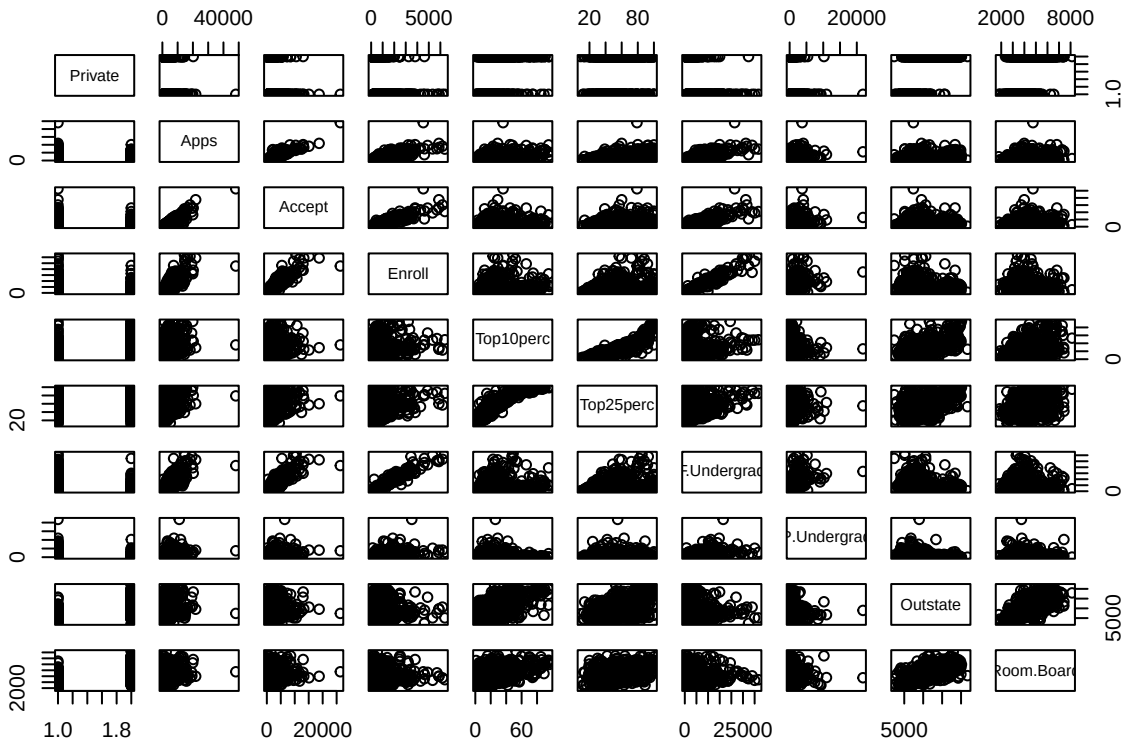
```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                              Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                              Grad.Rate
## Abilene Christian University        60
## Adelphi University                  56
## Adrian College                      54
## Agnes Scott College                 59
## Alaska Pacific University           15
## Albertson College                   55
```

b

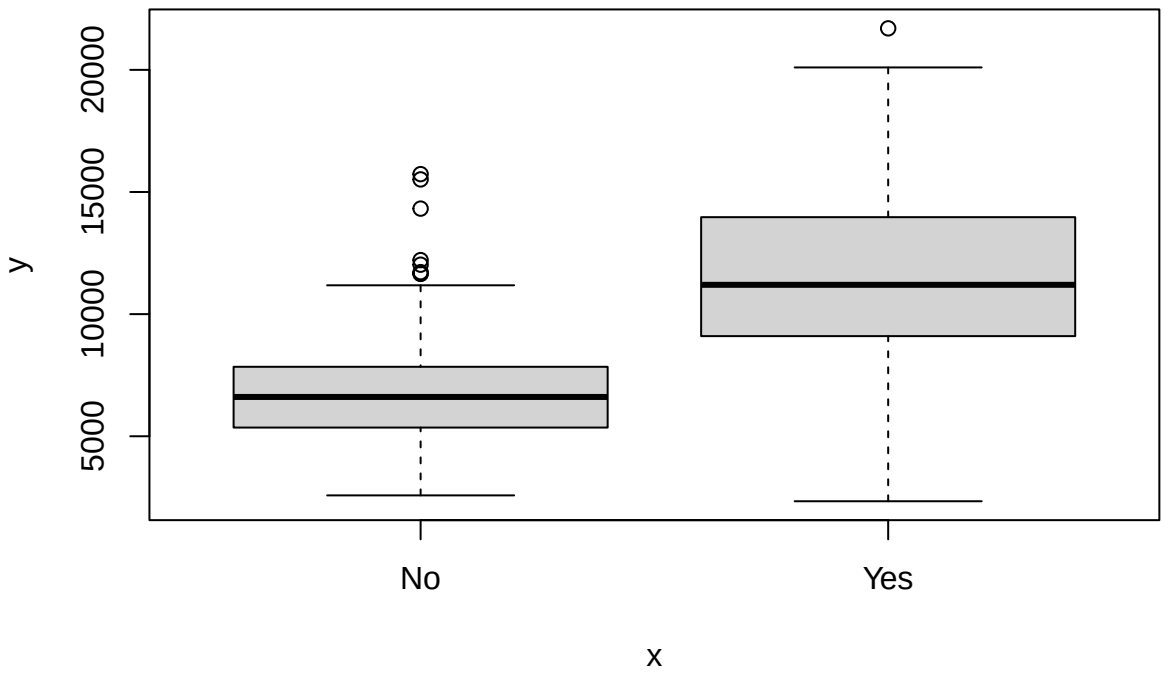
```
View(College)
```

c

```
pairs(College[,1:10])
```




```
plot(College$Private, College$Outstate)
```

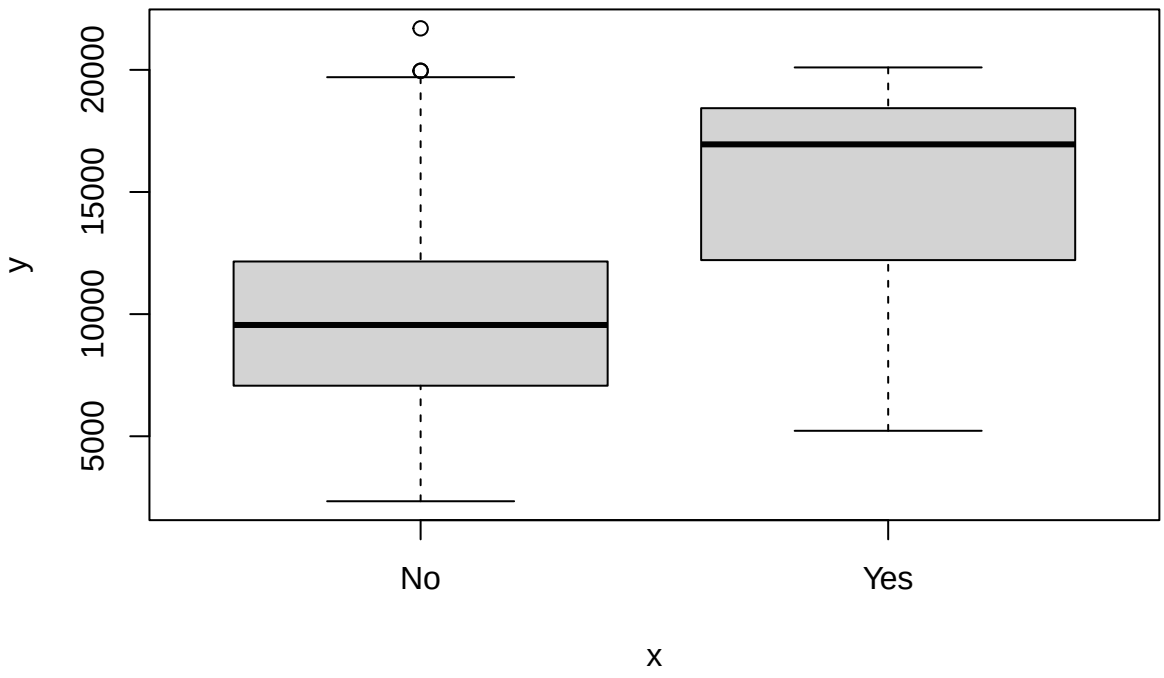



```
Elite = rep('No', nrow(College))
Elite[College$Top10perc > 50] = 'Yes'
Elite = as.factor(Elite)
College = data.frame(College, Elite)
summary(College$Elite)
```

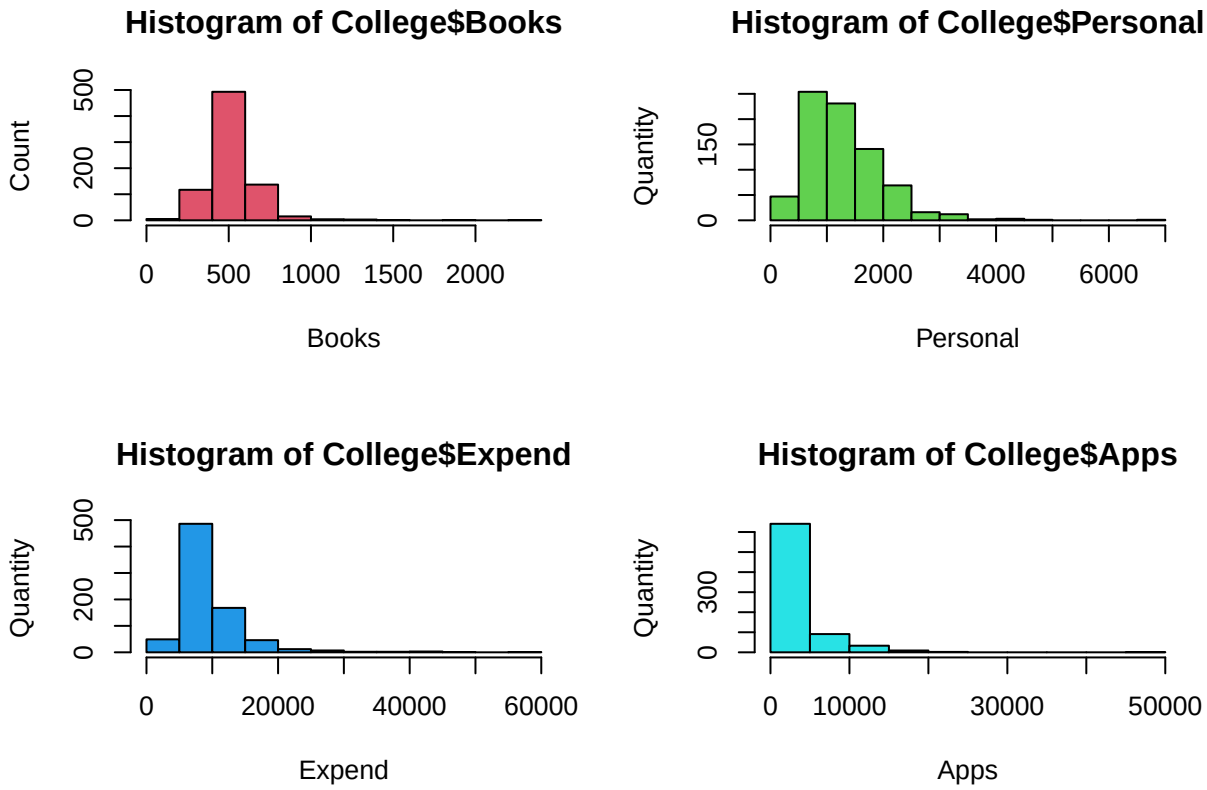
```
##  No Yes
## 699  78
```

```
View(College)

plot( College$Elite, College$Outstate)
```




```
par(mfrow = c(2, 2))
hist(College$Books, col = 2, xlab = 'Books', ylab = 'Count')
hist(College$Personal, col = 3, xlab = 'Personal', ylab = 'Quantity')
hist(College$Expend, col = 4, xlab = 'Expend', ylab = 'Quantity')
hist(College$Apps, col = 5, xlab = 'Apps', ylab = 'Quantity')
```



Exercise 9

a

```
head(Auto)
```

```
##   mpg cylinders displacement horsepower weight acceleration year origin
## 1  18         8          307        130   3504         12.0   70      1
## 2  15         8          350        165   3693         11.5   70      1
## 3  18         8          318        150   3436         11.0   70      1
## 4  16         8          304        150   3433         12.0   70      1
## 5  17         8          302        140   3449         10.5   70      1
## 6  15         8          429        198   4341         10.0   70      1
##                       name
## 1 chevrolet chevelle malibu
## 2         buick skylark 320
## 3        plymouth satellite
## 4             amc rebel sst
## 5               ford torino
## 6          ford galaxie 500
```

```
auto = na.omit(Auto)
summary(auto$cylinders)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   3.000   4.000   4.000   5.472   8.000   8.000
```

```
# cylinders, year, origin and name are qualitative predictors
```

b

```
qualitative_indices = match(c('cylinders', 'year', 'origin', 'name'),
                            colnames(auto))
sapply(auto[, -qualitative_indices], range)
```

```
##       mpg displacement horsepower weight acceleration
## [1,]  9.0           68         46   1613          8.0
## [2,] 46.6          455        230   5140         24.8
```

c

```
sapply(auto[, -qualitative_indices], mean)
```

```
##          mpg displacement   horsepower       weight acceleration
##     23.44592    194.41199    104.46939   2977.58418     15.54133
```

```
sapply(auto[, -qualitative_indices], sd)
```

```
##          mpg displacement   horsepower       weight acceleration
##     7.805007   104.644004    38.491160   849.402560     2.758864
```

d

```
mini_auto = auto[-c(10:85), ]
sapply(mini_auto[, -qualitative_indices], range)
```

```
##       mpg displacement horsepower weight acceleration
## [1,] 11.0           68         46   1649          8.5
## [2,] 46.6          455        230   4997         24.8
```

```
sapply(mini_auto[, -qualitative_indices], mean)
```

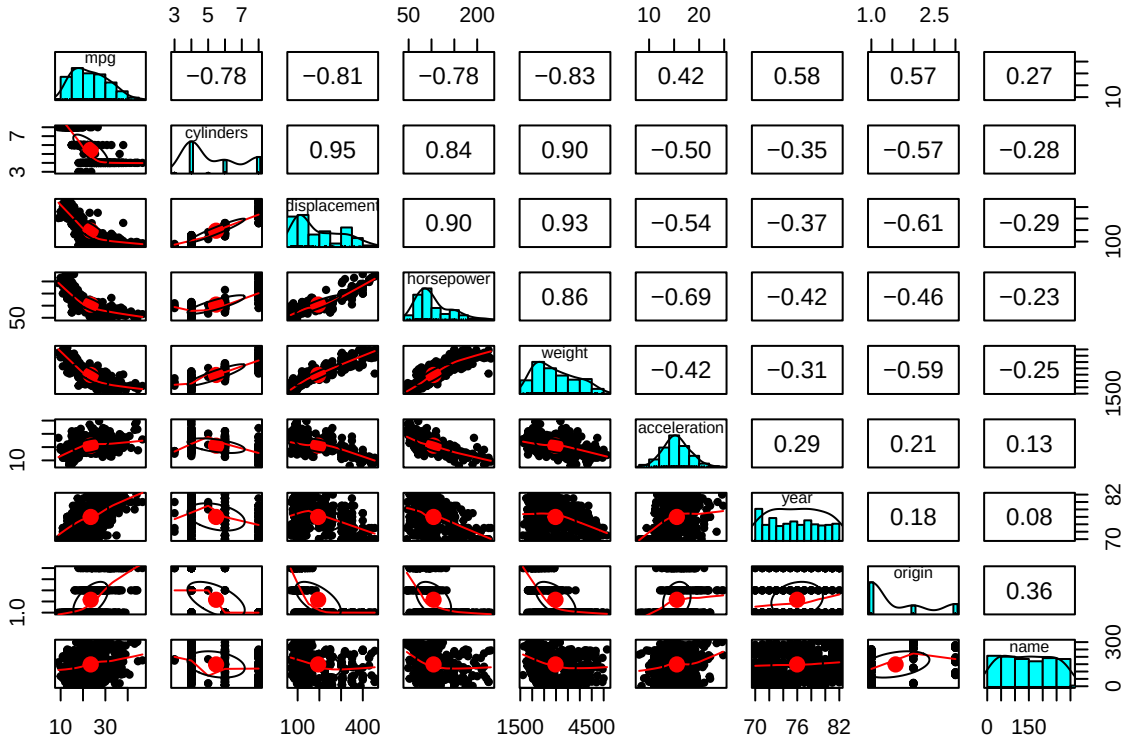
```
##          mpg displacement   horsepower       weight acceleration
##     24.40443    187.24051    100.72152   2935.97152     15.72690
```

```
sapply(mini_auto[, -qualitative_indices], sd)
```

```
##          mpg displacement   horsepower       weight acceleration
##     7.867283    99.678367    35.708853   811.300208     2.693721
```

e

```
pairs.panels(auto)
```



f

Variables like cylinders, displacement, horsepower, weight are linearly related with the target mpg (the correlation coefficients are greater than 0.7). Hence, these predictors might be useful for prediction.

Exercise 10

a

```
head(Boston)
```

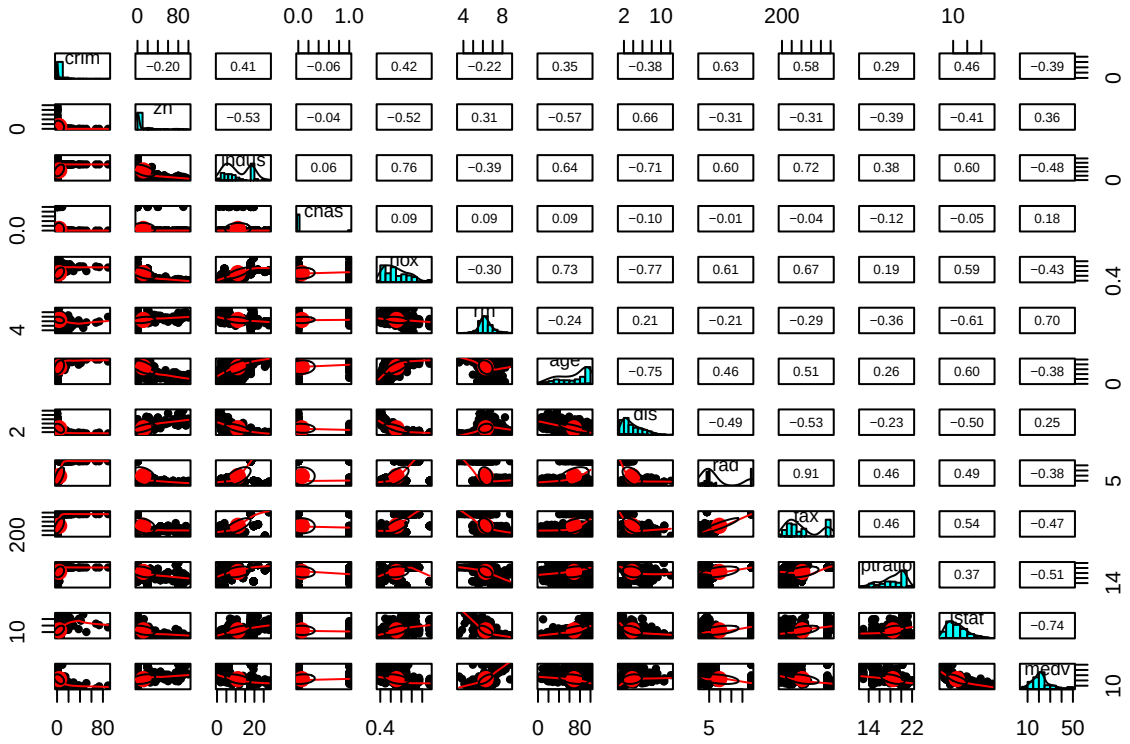
```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
```

```
?Boston
```

```
## starting httpd help server ... done
```

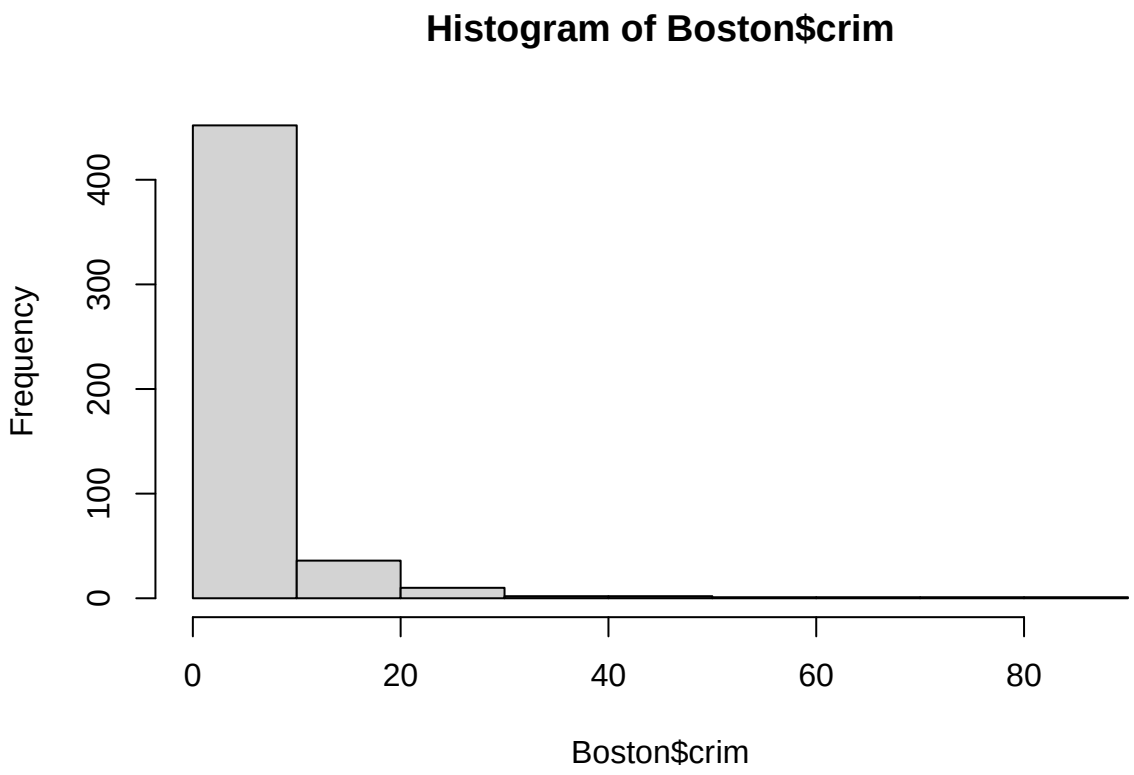
b

```
pairs.panels(Boston)
```



c

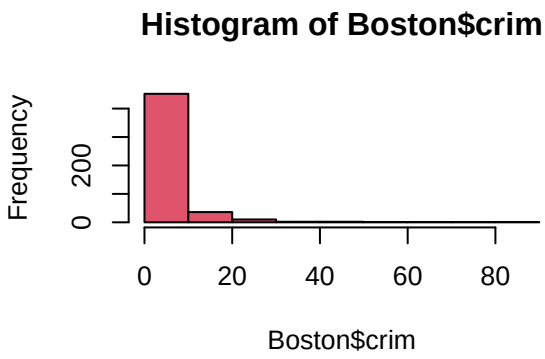
```
hist(Boston$crim)
```

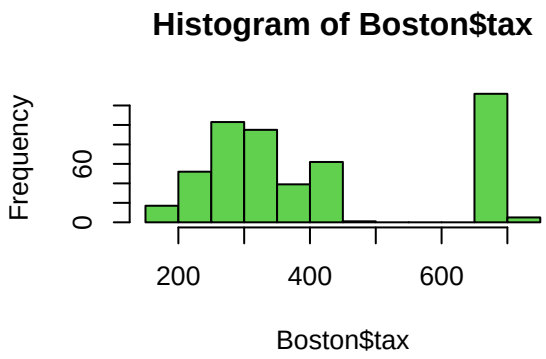


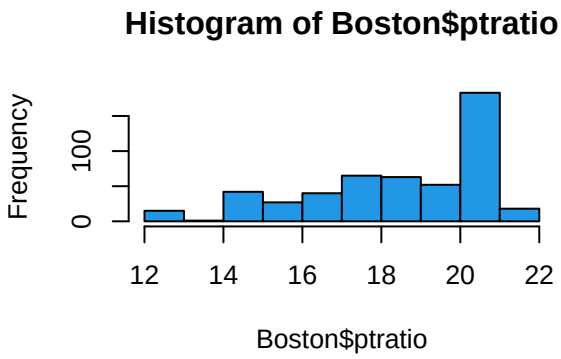
Variables are somewhat linearly related with the target crime rate (the correlation coefficients are around 0.4). Hence, these predictors might be useful for prediction.

d

```
par(mfrow = c(2,2))
hist(Boston$crim, col = 2)
hist(Boston$tax, col = 3)
hist(Boston$ptratio, col = 4)
```







e

```
nrow(Boston[Boston$chas == 1, ])
```

```
## [1] 35
```

f

```
median(Boston$ptratio)
```

```
## [1] 19.05
```

g

```
which.min(Boston$medv)
```

```
## [1] 399
```

```
Boston[399, ]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 30.59    5
```

h

```
nrow(Boston[Boston$rm > 7, ])
```

```
## [1] 64
```

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

```
## [1] 13
```