Exercise 16, p. 195: logistic regression only.

```
library(ISLR)
Boston = read.csv("/Volumes/work/MTH522/data/Boston.csv")
head(Boston)
```

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

Adding the crim01 variable into the dataset to identify if the crime rate is above or below the median(crim).

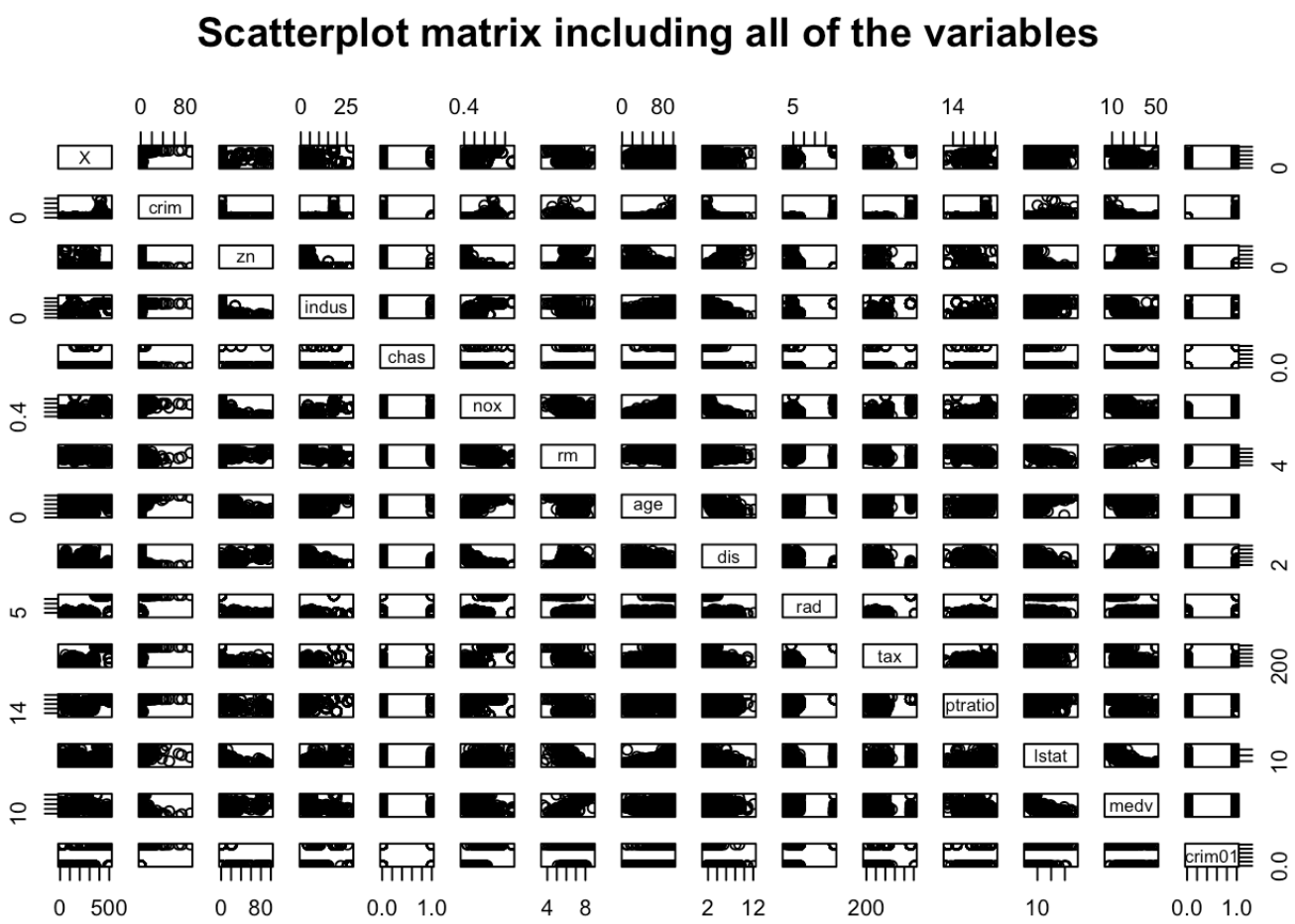
```
Boston1 <- Boston
Boston1$crim01 <- NA
crim_median = median(Boston1$crim)
print(crim_median)
```

```
## [1] 0.25651
```

```
for(i in 1:dim(Boston1)[1]){
  if (Boston1$crim[i] > crim_median){
    Boston1$crim01[i] = 1
  }else{
    Boston1$crim01[i] = 0
  }
}
head(Boston1)
```

```
##   X    crim zn indus chas   nox    rm  age    dis rad tax ptratio lstat medv
## 1 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3  4.98 24.0
## 2 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8  9.14 21.6
## 3 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8  4.03 34.7
## 4 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7  2.94 33.4
## 5 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7  5.33 36.2
## 6 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7  5.21 28.7
##   crim01
## 1      0
## 2      0
## 3      0
## 4      0
## 5      0
## 6      0
```

```
pairs(Boston1[,1:15], main="Scatterplot matrix including all of the variables")
```



```
autocorr = cor(Boston1$crim01,Boston1)
autocorr
```

```
##              X      crim        zn     indus       chas       nox         rm
## [1,] 0.3694304 0.4093955 -0.436151 0.6032602 0.07009677 0.7232348 -0.1563718
##            age        dis       rad       tax   ptratio     lstat       medv
## [1,] 0.6139399 -0.6163416 0.6197862 0.6087413 0.2535684 0.4532627 -0.2630167
##      crim01
## [1,]      1
```

Observations:

1. From the above table we can observe that, `indus` , `nox` , `age` , `rad` , `tax` are the variables that are statistically significant to `crim01`

```
require(caTools)
```

```
## Loading required package: caTools
```

```
set.seed(123)
Boston_split = sample.split(Boston1$crim01, SplitRatio = 0.80)
Boston_train = subset(Boston1, Boston_split==TRUE)
Boston_test = subset(Boston1, Boston_split==FALSE)
```

Observations:

1. logistic regression for the variables which are statistically significant to `crim01` .
i.e., `indus` , `nox` , `age` , `rad` and `tax`

```
Boston1_glm = glm(crim01 ~ indus+nox+age+rad+tax, data=Boston_train, family=binomial)
Boston1_prob = predict(Boston1_glm,Boston_test,type="response")
Boston1_pred = rep(0,length(Boston_test$crim01))
Boston1_pred[Boston1_prob >0.5] = 1

table(Boston1_pred,Boston_test$crim01)
```

```
##
## Boston1_pred  0  1
##            0 48  9
##            1  3 42
```

```
mean(Boston1_pred != Boston_test$crim01)
```

```
## [1] 0.1176471
```

Observations:

1. The logistic model has a 11.76% error rate. It means it has a correct prediction of 88.24%