Lab 1: Exploratory Data Analysis

This Bitter Earth

YILIN LI

Exercise 1

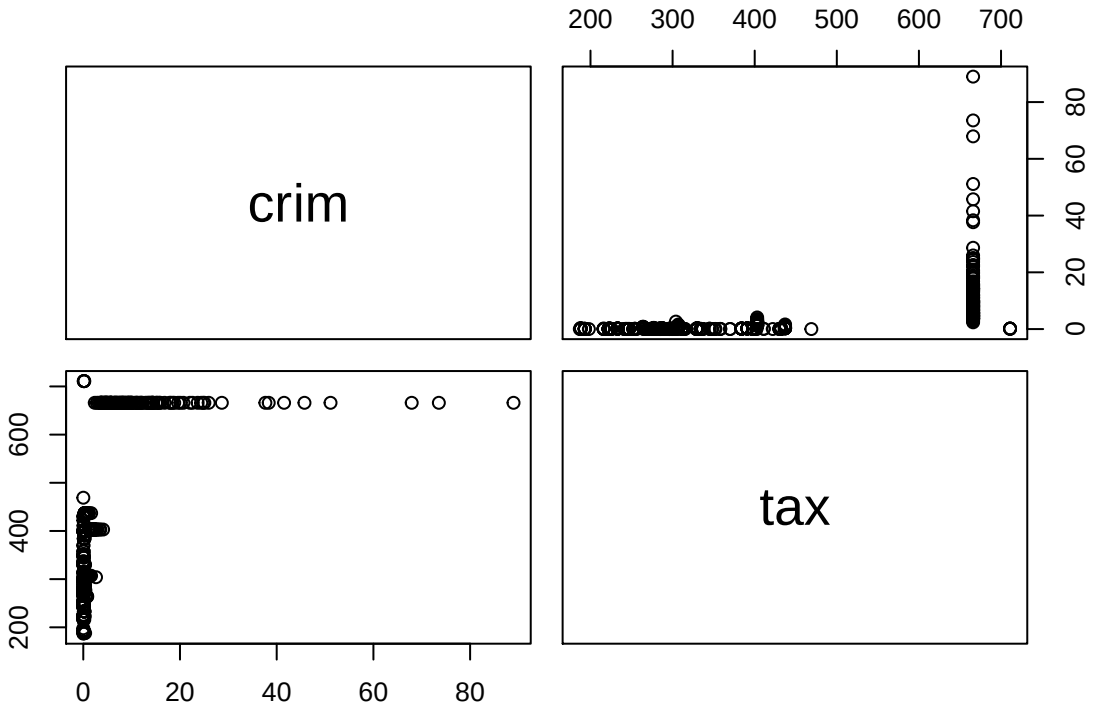
```
library(MASS)
dim(Boston)
```

```
## [1] 506  14
```

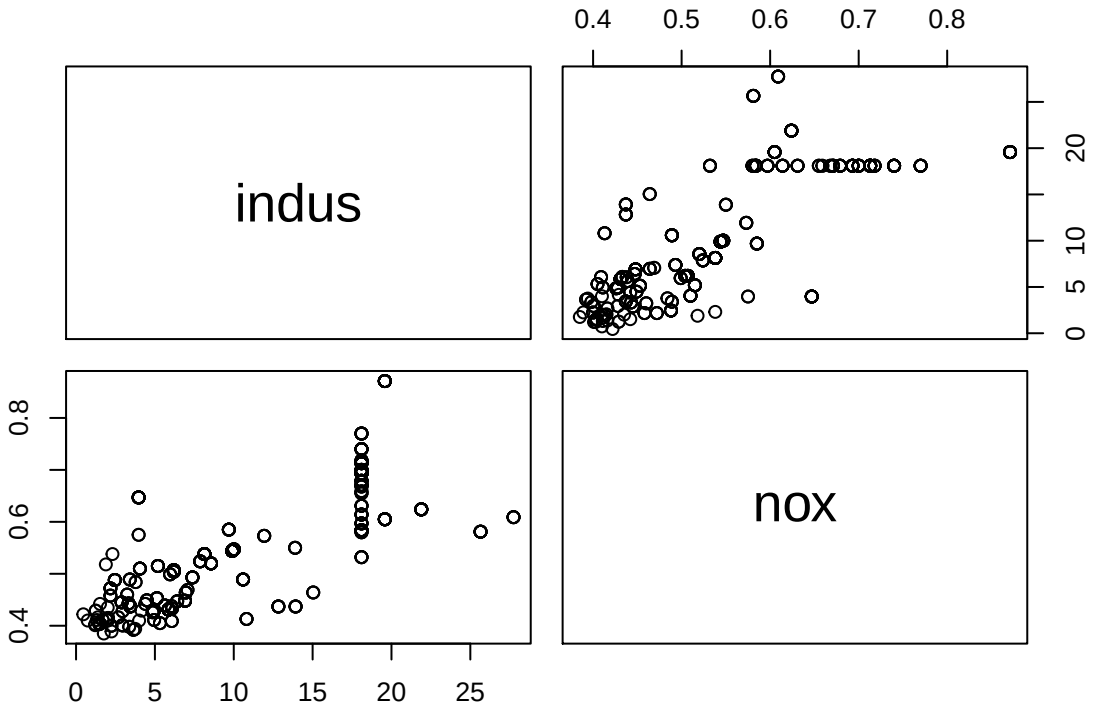
The Boston Dataframe has 506 rows and 14 columns. Each row represents 1 observation, which is Boston housing value and each column represents 1 feature recorded.

Exercise 2

```
pairs(Boston[,c(1,10)])
```



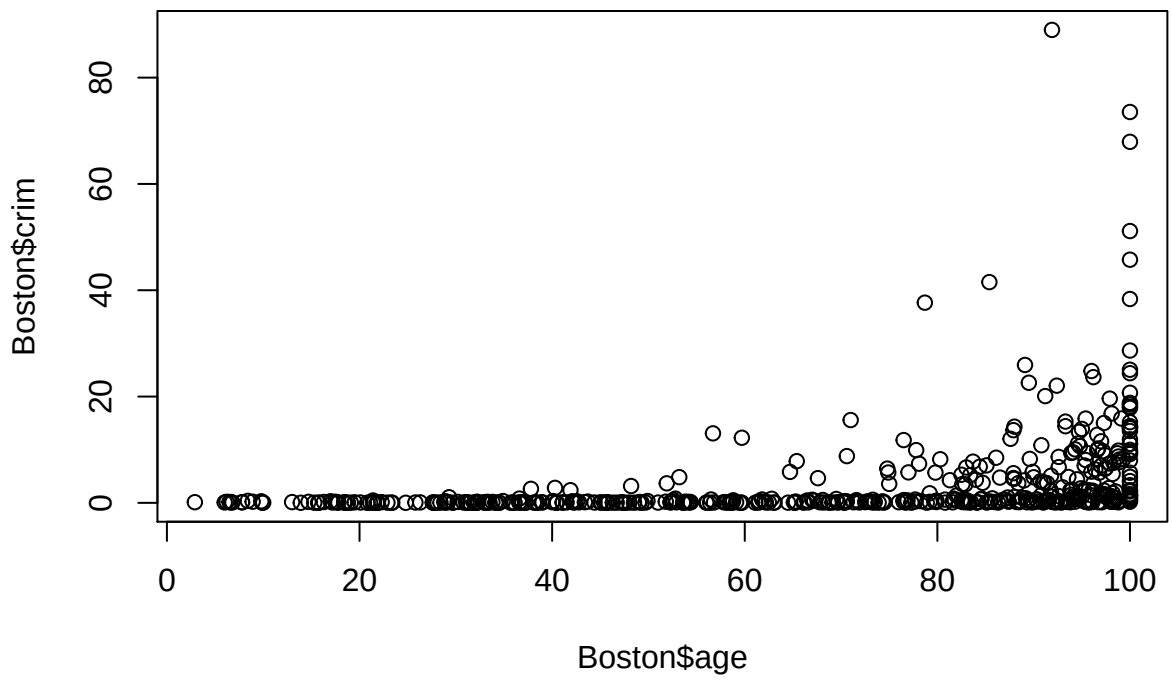
```
pairs(Boston[,c(3,5)])
```



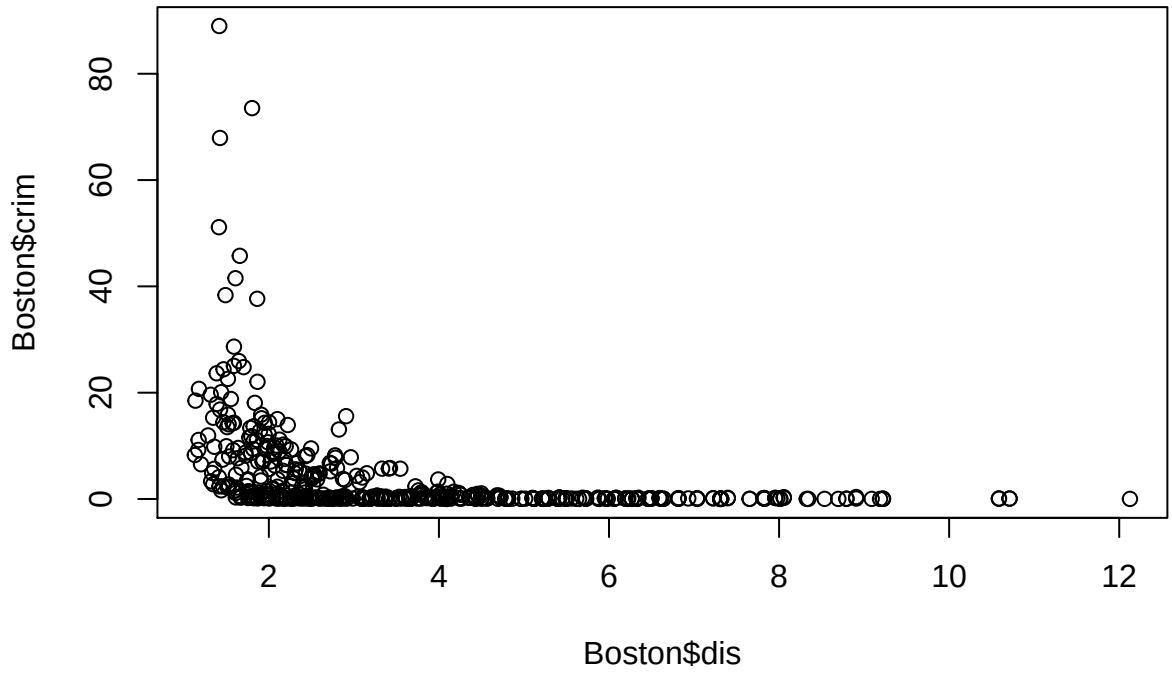
- crime rate is correlated to the level of tax
- the higher industry density, the higher nitrogen oxides concentration

Exercise 3

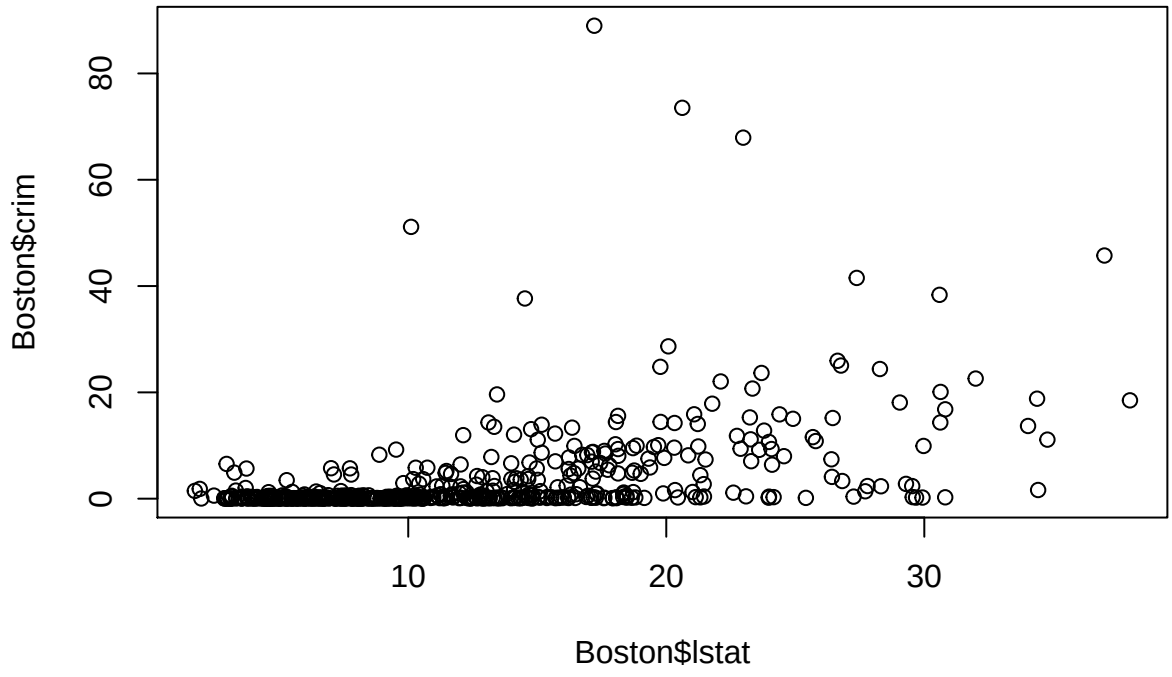
```
plot(Boston$age, Boston$crim)
```



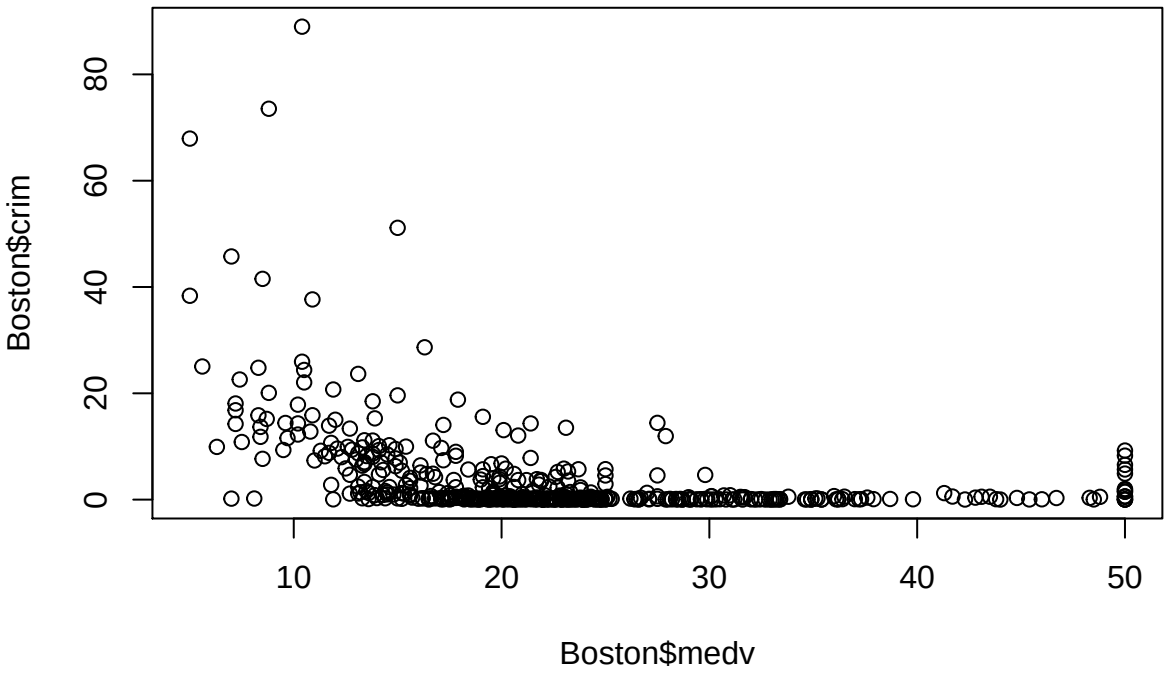
```r
plot(Boston$dis, Boston$crim)
```




```r
plot(Boston$lstat, Boston$crim)
```




```r
plot(Boston$medv, Boston$crim)
```



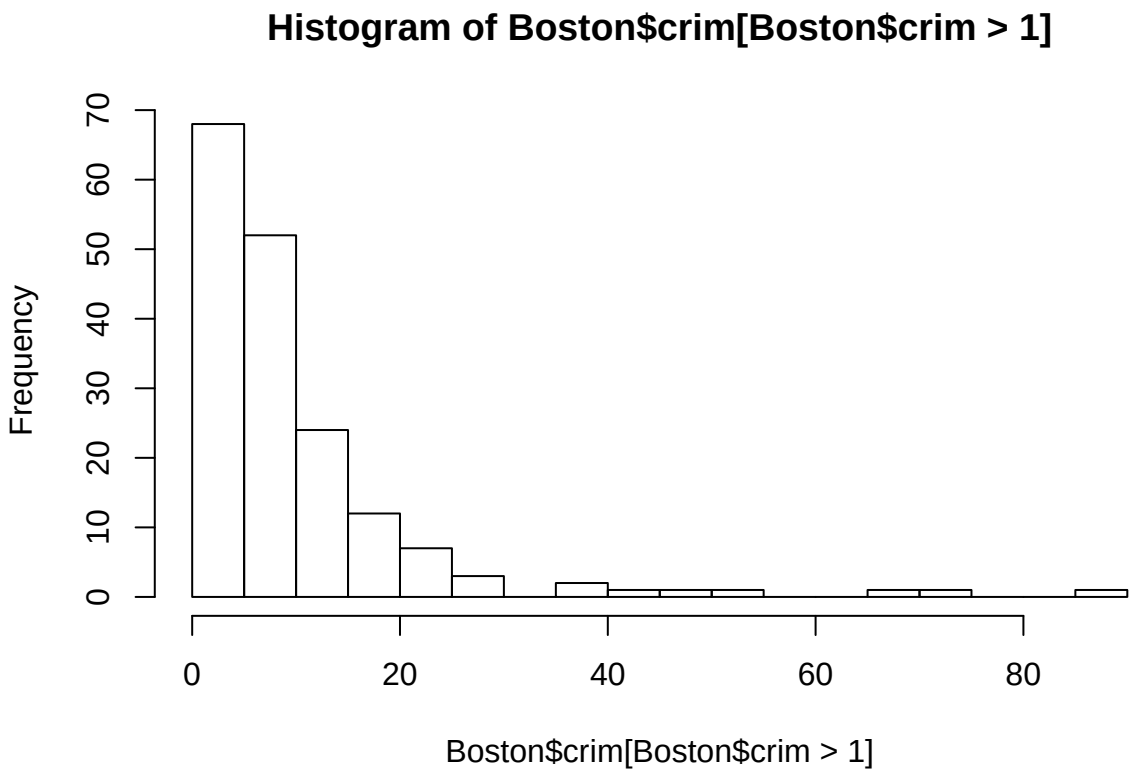
- more built prior to 1940, the higher the criminal rate.
- closer to the employment centres, the higher the criminal rate.
- more populations are in lower status, the higher the criminal rate.
- the less the owner-occupied homes in $1000, the higher the criminal rate.

Exercise 4

```
length(Boston$crim[Boston$crim>20])
```

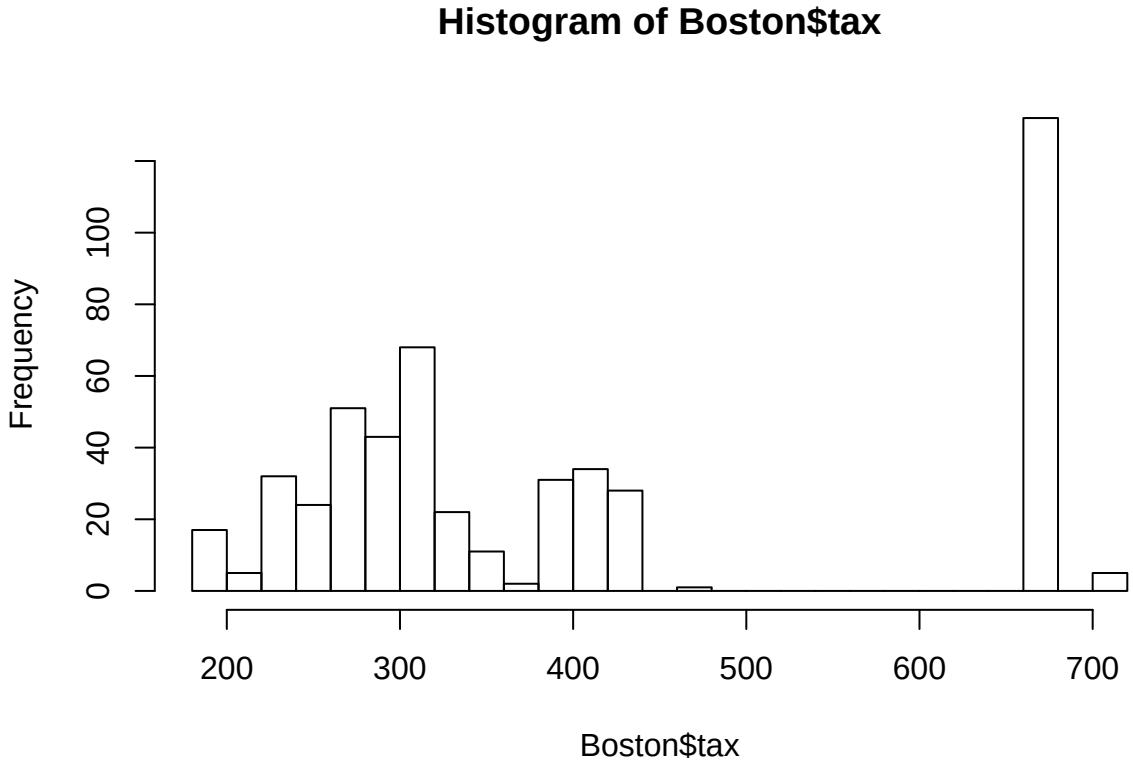
```
## [1] 18
```

```
hist(Boston$crim[Boston$crim>1],breaks = 25)
```



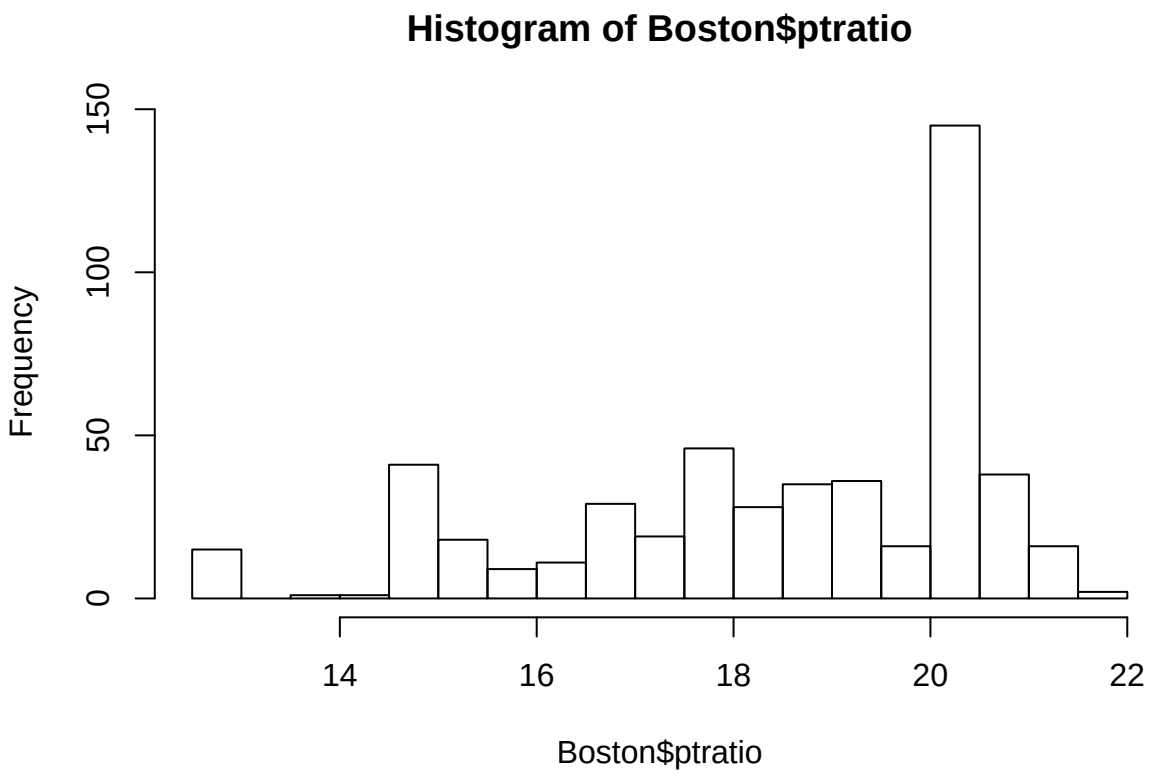
Most suburbs have the criminal rate lower than 20%, but there are 18 suburbs have the criminal rate higher than 20% even reaching 80%.

```
hist(Boston$tax,breaks = 25)
```



There's a large divide between suburbs taxing from 200 to 400 and a huge peak taxing at 660-680.

```r
hist(Boston$ptratio,breaks = 25)
```



The pupil-teacher ratio skew towards higher ratio and has a peak at 20-21.

Exercise 5

```r
dim(subset(Boston, chas==1))[1]
```

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

35 suburbs bound the Charles River.

Exercise 6

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

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

The median pupil-teacher ratio is 19.05.

Exercise 7

The input is a 506×14 matrix $X \in R^{506 \times 14}$, where each row is a suburb data. The response is a column vector $y \in R^{506}$ represents the predicted average value of the home for each suburb data.