Introduction to statiscal learning

Exercise 1

Name: Jiangxue Han

Computing ID: jh6rg

#1.

(a) The sample size n is extremely large, and the number of predictors p is small.

Better. The flexible method is better with a larger sample size.

(b) The number of predictors p is extremely large, and the number of observations n is small.

Worse. The flexible method is prone to be overfitting. Therefore, it is better to use unflexible method here.

(c) The relationship between the predictors and response is highly non-linear.

Better. The inflexible method will have a higher bias for non-linear relationship than the flexible one.

(d) The variance of the error terms, i.e. σ2 = Var(ε), is extremely high.

Worse. The flexible method would intend to fit the error and will have a higher variance.

#2.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

Regression. Inference. n = 500. p = 3.

The CEO salary is quantitative, therefore this intends to be a regression problem. Since we are going to analyze the relationship between factors and salary, we are more interested in inference.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Classification. Prediction. n = 20. p = 13.

Whether it was a sucess or failure is qualititative, so this is a classification problem. And we are going to predict the success of a new product, therefore we are more interested in prediction.

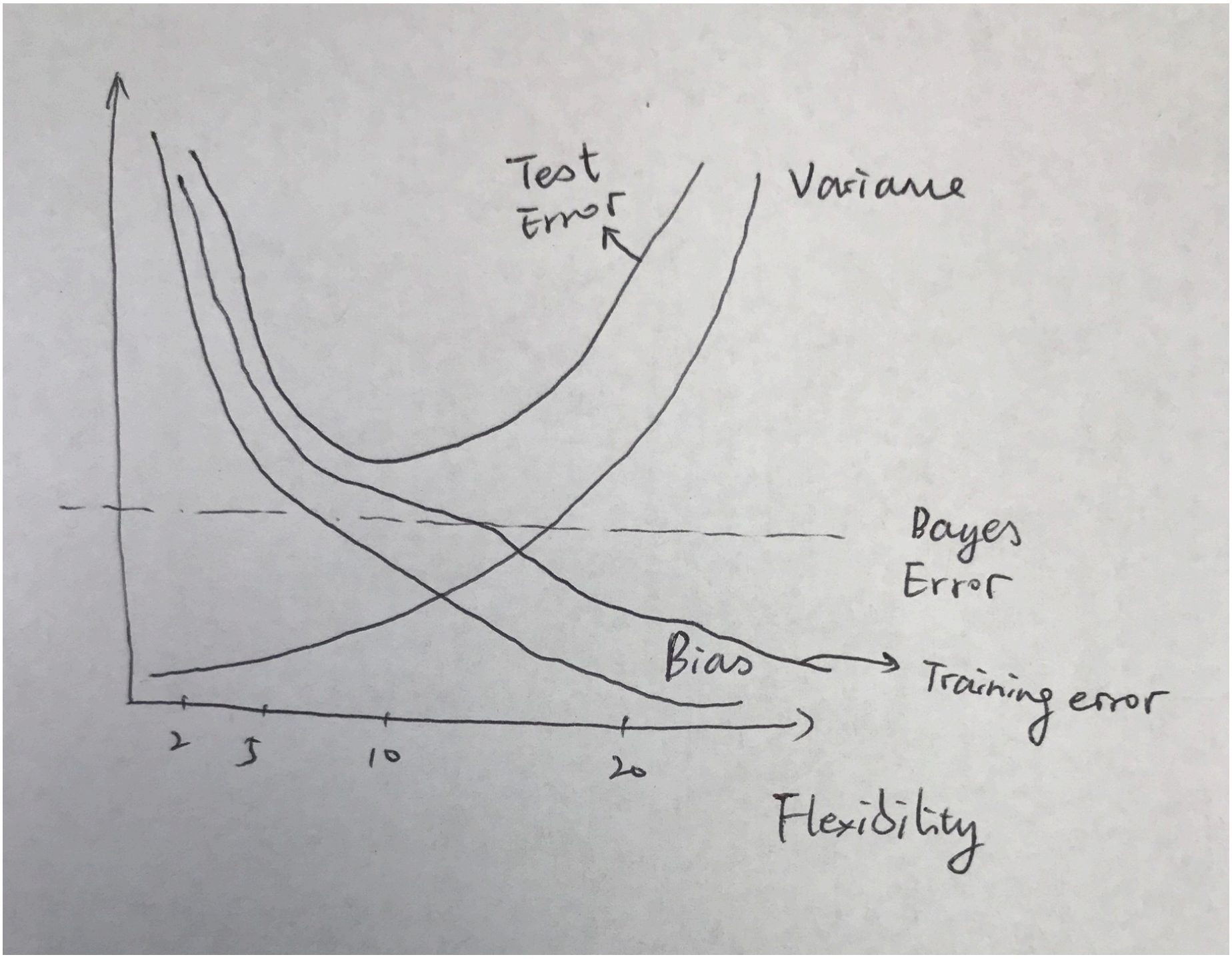
(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Regression. Prediction. n = 52. p = 3.

% change is quantitative, thus this is a regression problem. Since we are going to predict the % change, we are more interested in prediction.

#3.

(a) image:



(b)

i. Bias is decreasing because the method should be better fit with higher flexibility.

ii. Variance is increasing because higher flexibility prones to overfitting, which means the method is less able to adapt to a different dataset.

iii. Training error tends to decrease since the flexible method will fit better, thus reducing the training error.

iv. Test error will decrease at first and start increasing at a certain point. Because at the beginning, with the increasing in flexibility, the model is better fit for the dataset and help test error going down. While much higher flexibility prones to overfitting, which means it cannot predict correctly on the new data and cause increasing test error.

v. Bayes error is irreducible, which equals to the lowest achievable test MSE among all possible methods. Expected test MSE can never lie below bayes error.

#4.

(a)

i. Predicting whether it is going to rain tomorrow or not. Predictors: The history of rainning at the same time period in last 5 years, climate, soil and vegetation. Prediction.

ii. Predicting whether a basketball team is going to win in the next game. Predictors: The history of such team's performance, the history performance of opponent and the location of the game.Prediction.

iii. Determining what factors impact one's blood type. Predictors: Parents and grandparents' blook types, the food that mother consumed during pregnancy and the drug that mother has taken during pregnancy. Inference.

(b)

i. Predicting the school enrollment in the coming year. Predictors: The school enrollment in last 10 years, the number of applicants in last 10 years and the enrollment of school in the same tier. Prediction.

ii. Analyzing the factors that impact the electricity consumption in an area. Predictors: The number of households in that area, the number of factories in that area and the number of infrastructures in that area. Inference.

iii. Predicting the GDP of US in the coming year. Predictors: Wages, corporate profits and interest. Prediction.

(c)

i. Grouping the hotels in New York City to make recommendation.
ii. Grouping the students in a school to improve the teaching method.
iii. Grouping the users based on users' behavior to differentiate advertisement.

#5.

Advantages: A very flexible approach is closer fit to the dataset and has a lower bias.

Disadvantages: A very flexible approach prones to be overfitting with higher variance and is hard to interprete.

A very flexible approach is preferred for prediction and dataset with many variables.While a less flexible approach is preferred for inference and a linear dataset.

#6.

Parametric approach makes an assumption about the function form at first, while non-parametric does not.

Advantages: It simplifies the problem because it is much easier to estimate the parameters then the function form.

Disadvantages: The model will usually not match the true unknown form and prone to be overfitting if we choose flexible approach.

#7.

(a) Obs.1 = 3 Obs.2 = 2 Obs.3 = 3.16 Obs.4 = 2.24 Obs.5 = 1.41 Obs.6 = 1.73

(b) Y is green. Because we predict the Y with only 1 nearest neighbor. Obs.5 is the closest one and its Y is green.

(c) Y is red. Because we predict the Y based on 3 nearest neighbors in this case: Obs.2, Obs.5 and Obs.6. Two of them is red, therefore Y is red.

(d) The best value for K should be small. The boundary with small K would be more flexible and fit better.

8.

(a)

```
library(readr)
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyve
rse 1.2.1 ──
```

```
## ✔ ggplot2 3.0.0     ✔ purrr   0.2.5
## ✔ tibble  1.4.2     ✔ dplyr   0.7.6
## ✔ tidyr   0.8.1     ✔ stringr 1.3.1
## ✔ ggplot2 3.0.0     ✔ forcats 0.3.0
```

```
## ── Conflicts ────────────────────────────────────────── tidyverse_co
nflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
```

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
college <- read.csv('College.csv')
```

(b)

```
rownames(college) <- college[,1]
```

View(college) # Fix does not work

(c)

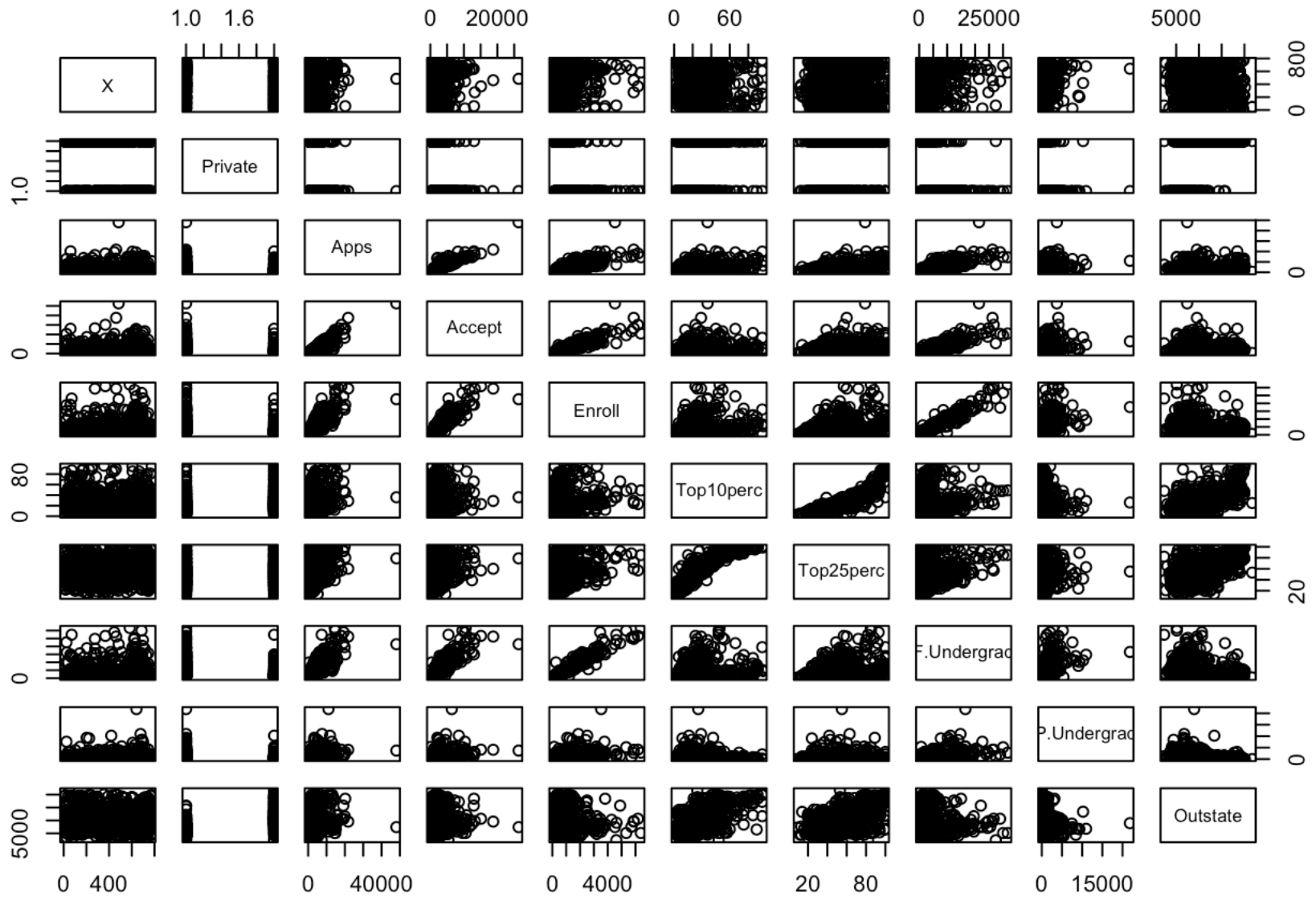
i.

```
summary(college)
```

```
##                              X       Private        Apps
##  Abilene Christian University:  1   No :212   Min.   :   81
##  Adelphi University           :  1   Yes:565   1st Qu.:  776
##  Adrian College               :  1             Median : 1558
##  Agnes Scott College          :  1             Mean   : 3002
##  Alaska Pacific University    :  1             3rd Qu.: 3624
##  Albertson College            :  1             Max.   :48094
##  (Other)                      :771
##      Accept          Enroll       Top10perc       Top25perc
##  Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
##  1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
##  Median : 1110   Median : 434   Median :23.00   Median : 54.0
##  Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
##  3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
##  Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
##
##   F.Undergrad     P.Undergrad         Outstate       Room.Board
##  Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
##  1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
##  Median : 1707   Median :  353.0   Median : 9990   Median :4200
##  Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
##  3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
##  Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
##
##      Books           Personal         PhD            Terminal
##  Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
##  1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
##  Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
##  Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
##  3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
##  Max.   :2340.0   Max.   :6800   Max.   :103.00   Max.   :100.0
##
##    S.F.Ratio      perc.alumni        Expend        Grad.Rate
##  Min.   : 2.50   Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:11.50   1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :13.60   Median :21.00   Median : 8377   Median : 65.00
##  Mean   :14.09   Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :39.80   Max.   :64.00   Max.   :56233   Max.   :118.00
##
```

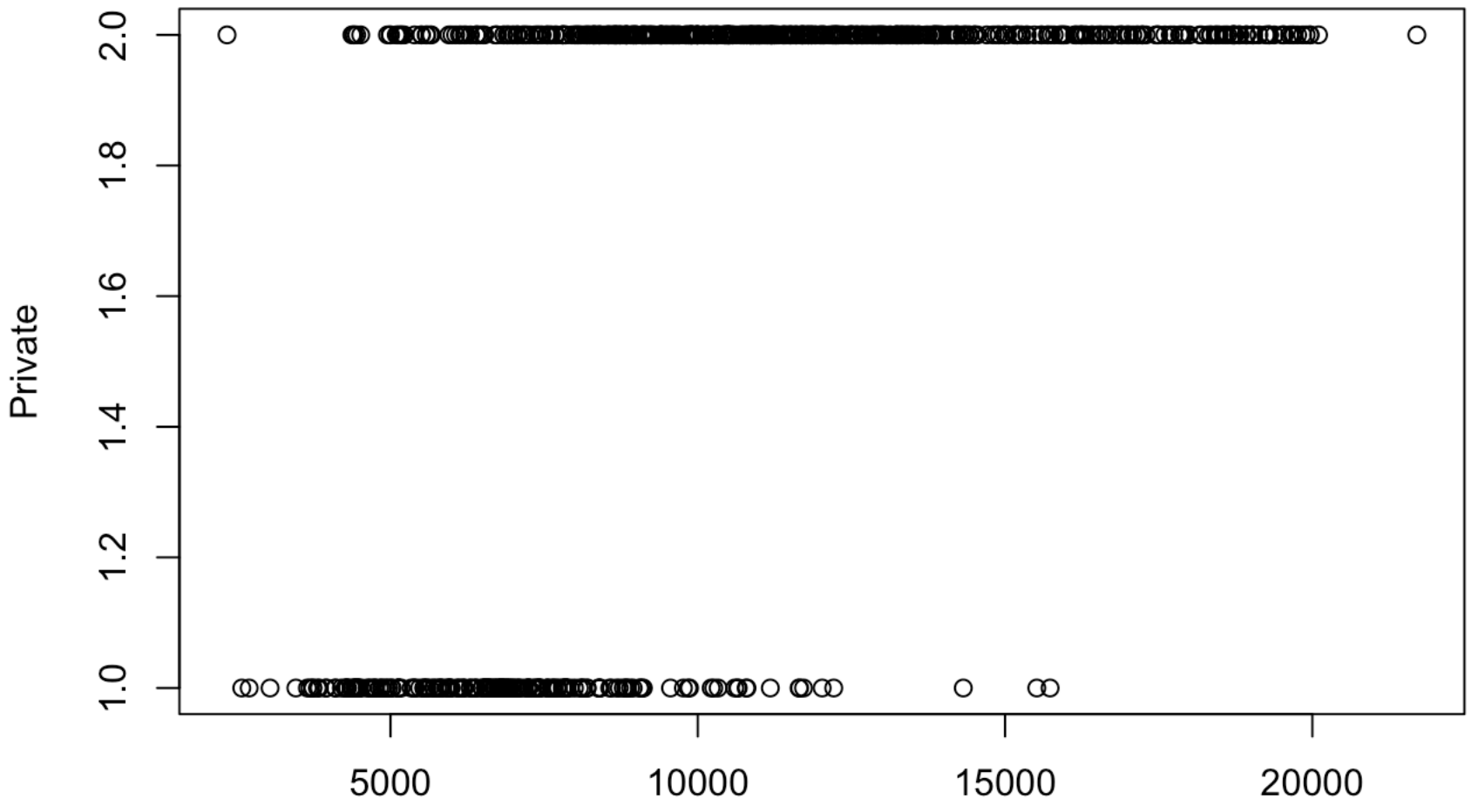
ii.

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



iii.

```
attach(college)
plot(Outstate, Private)
```

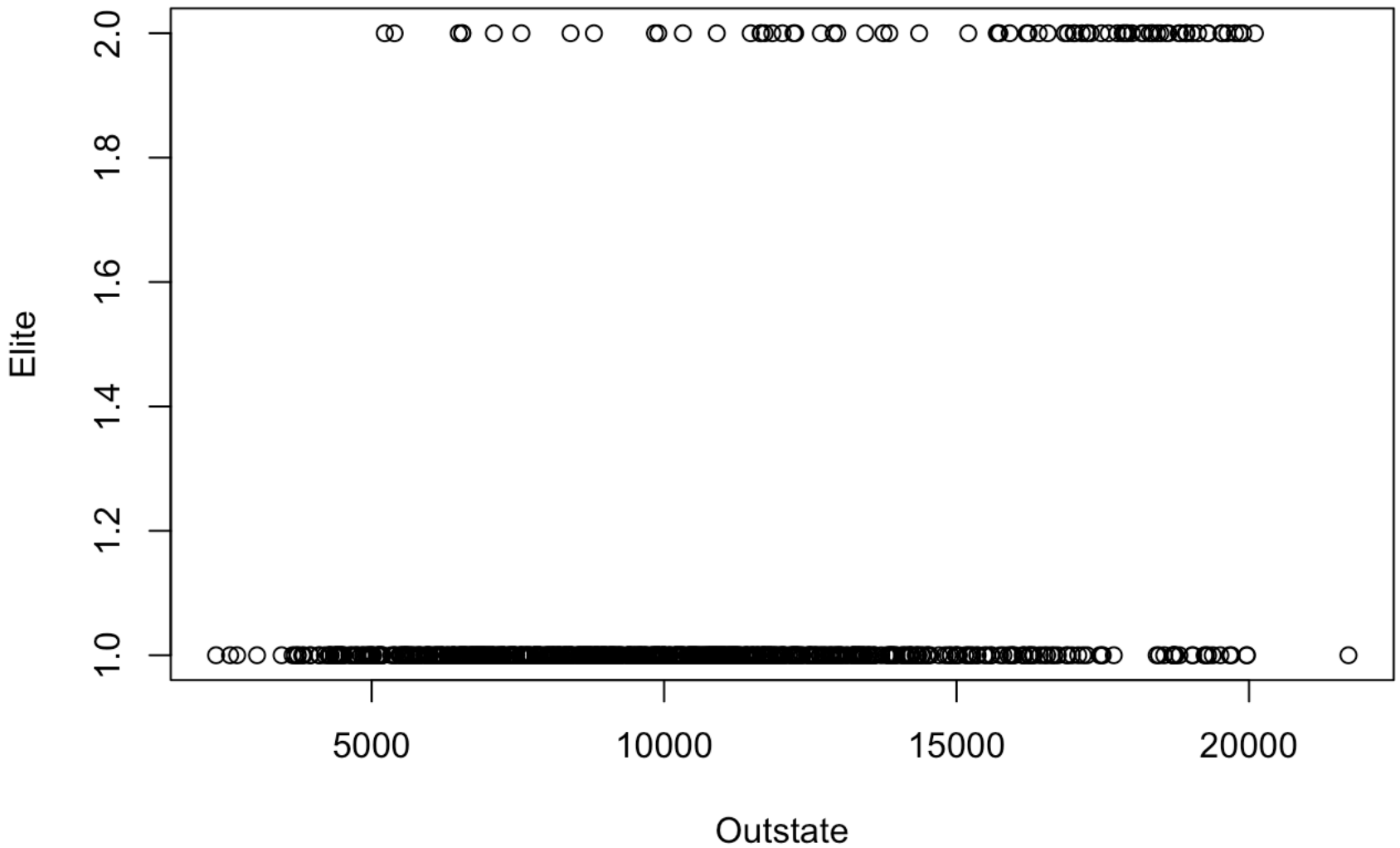


iv.

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

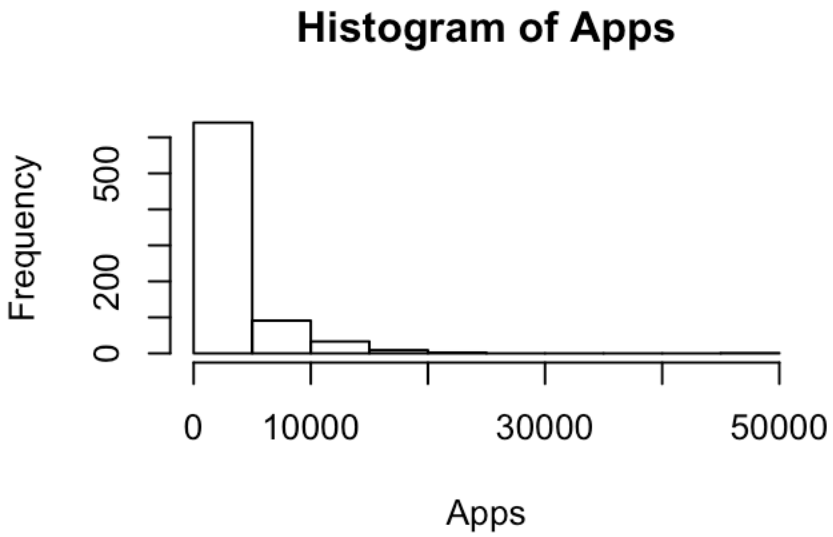
```
##  No Yes
## 699  78
```

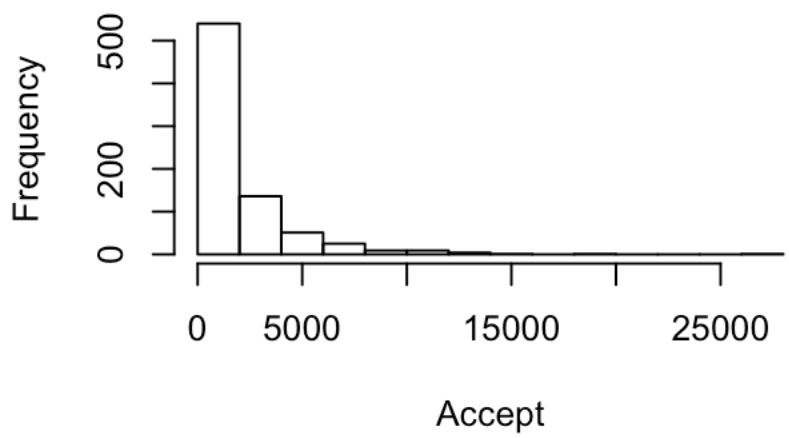
```
plot(Outstate, Elite)
```

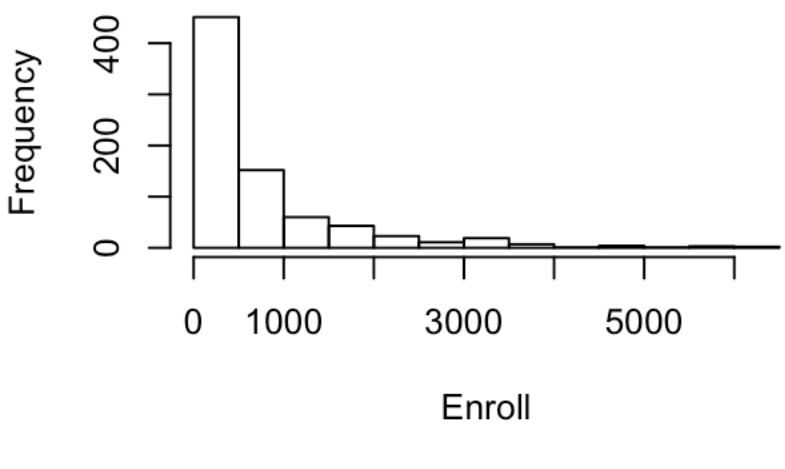


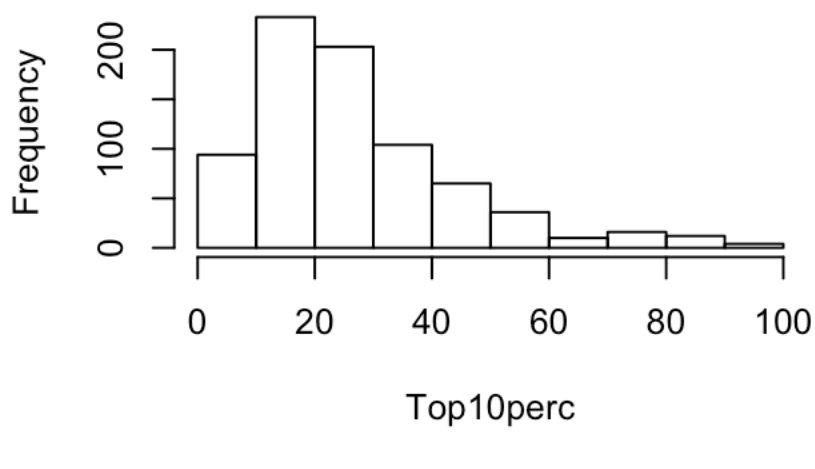
v.

```
par(mfrow=c(2,2))
hist(Apps)
hist(Accept)
hist(Enroll)
hist(Top10perc)
```









vi.

```
by_elite <- group_by(college, Elite)
accpet_means <- summarise(by_elite, acceptRate = mean(Accept/Apps), enrollRate = mean
(Enroll/Accept))
```

Both accpetance rate and enrollment rate of non-elite universities are lower than those of elite universities.

#9.

(a)

```
auto <- read.table('Auto.data.txt',header=T,na.strings="?")
auto <- na.omit(auto)
```

Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year, origin

Qualitative: name # Is origin quantitative or qualitative?

(b)

```
sapply(auto[,1:8], range)       # Applys range function to each column
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,]  9.0         3           68         46   1613          8.0   70
## [2,] 46.6         8          455        230   5140         24.8   82
##      origin
## [1,]      1
## [2,]      3
```

(c)

```
sapply(auto[,1:8], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight
##    23.445918     5.471939   194.411990   104.469388  2977.584184
## acceleration         year       origin
##    15.541327    75.979592     1.576531
```

```
sapply(auto[,1:8], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
##    7.8050075    1.7057832  104.6440039   38.4911599  849.4025600
## acceleration         year       origin
##    2.7588641    3.6837365    0.8055182
```

(d)

```
newAuto <- auto[-(10:85),]
sapply(newAuto[,1:8], range)
```

```
##       mpg cylinders displacement horsepower weight acceleration year
## [1,] 11.0         3           68         46   1649          8.5   70
## [2,] 46.6         8          455        230   4997         24.8   82
##      origin
## [1,]      1
## [2,]      3
```

```
sapply(newAuto[,1:8], mean)
```

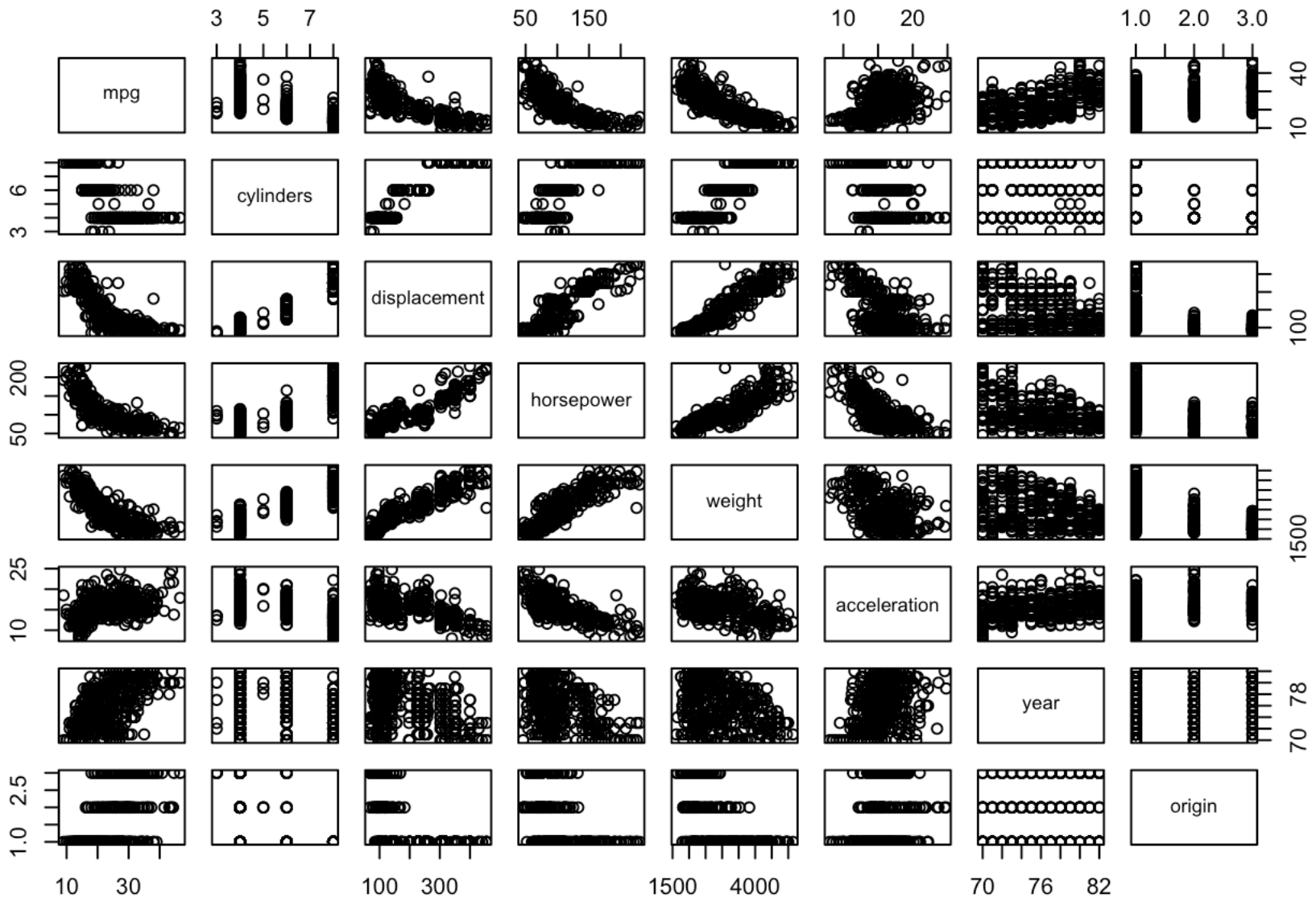
```
##          mpg    cylinders displacement   horsepower       weight
##    24.404430     5.373418   187.240506   100.721519  2935.971519
## acceleration         year       origin
##    15.726899    77.145570     1.601266
```

```
sapply(newAuto[,1:8], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
##     7.867283     1.654179    99.678367    35.708853   811.300208
## acceleration         year       origin
##     2.693721     3.106217     0.819910
```

(e)

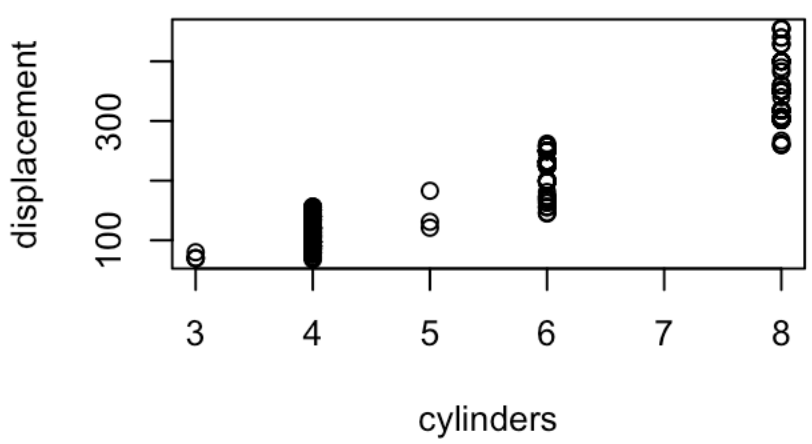
```
pairs(auto[,1:8])
```

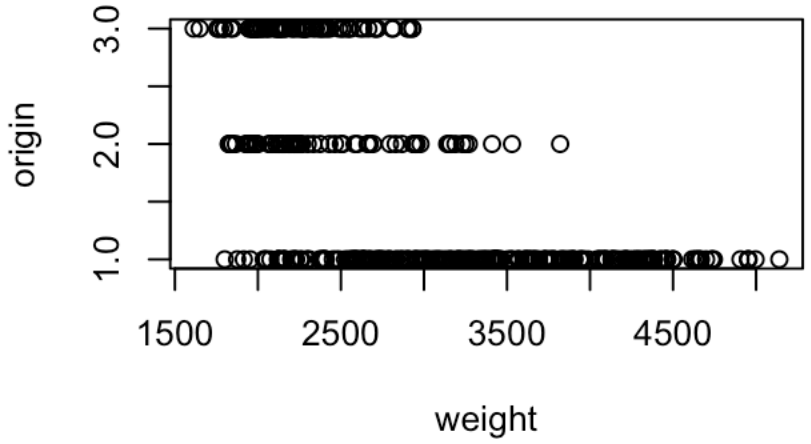


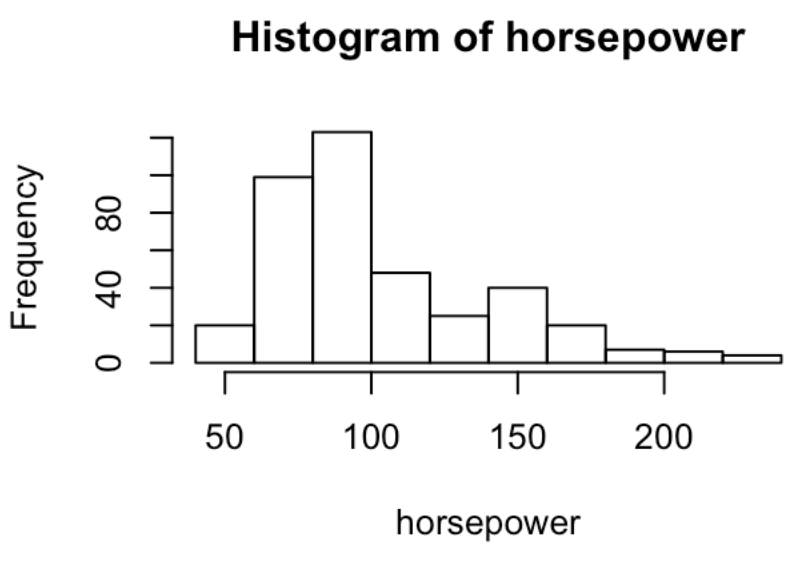

```
attach(auto)
```

```
## The following object is masked from package:ggplot2:
##
##     mpg
```

```
par(mfrow=c(2,2))
plot(cylinders, displacement)
plot(weight, origin)
hist(horsepower)
```







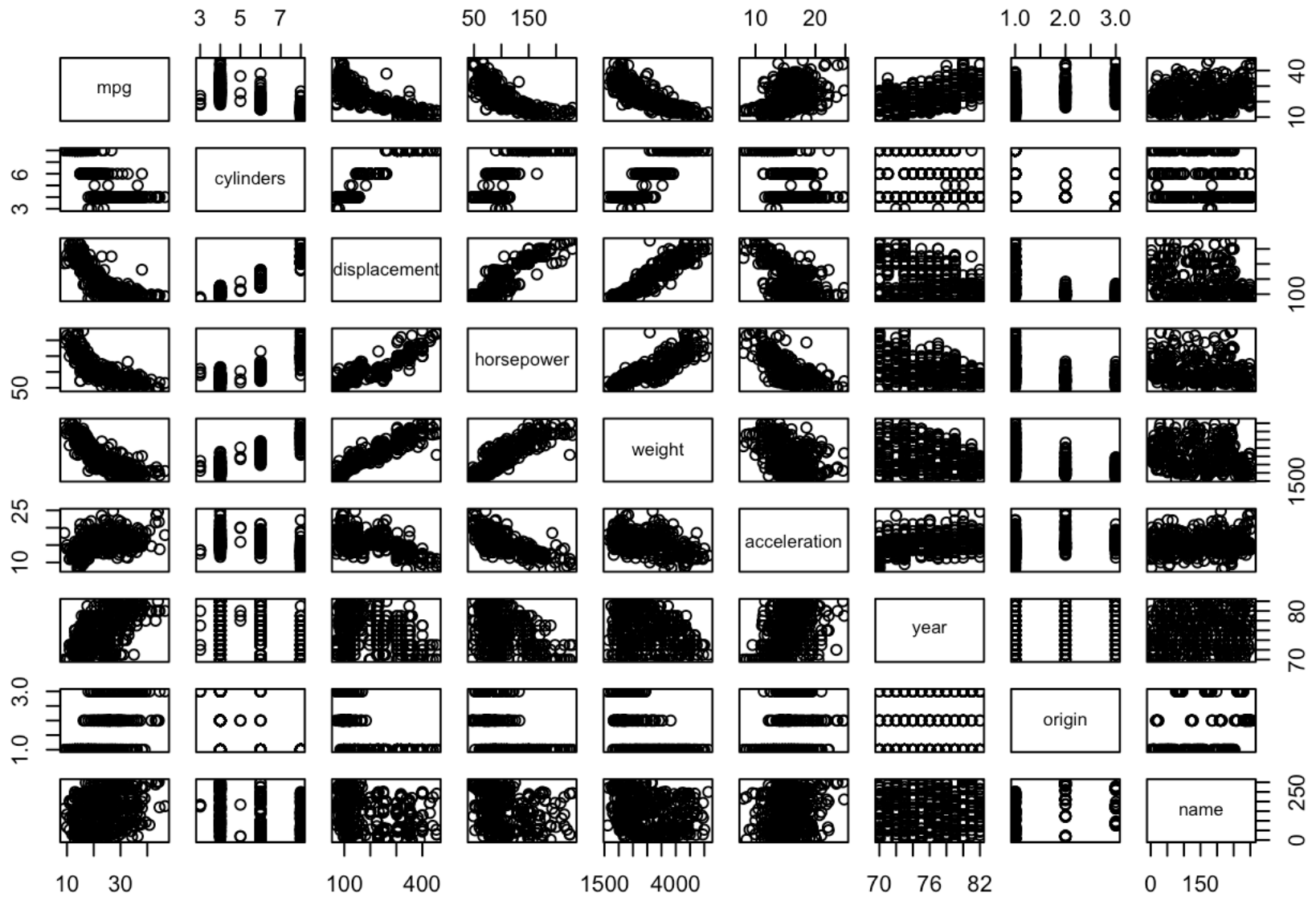
The cylinders and displacement have a poistive relationship.

The range of weight becomes wider with the increasing in origin.

Most of the horsepower is between 70 to 100.

(f)

```
pairs(auto)
```



It

seems like the mpg has correlation with all the variables except acceleration and name. Therefore, I would consider to use the other 6 variables for prediction.

#10.

```
Boston <- Boston
?Boston
```

(a)

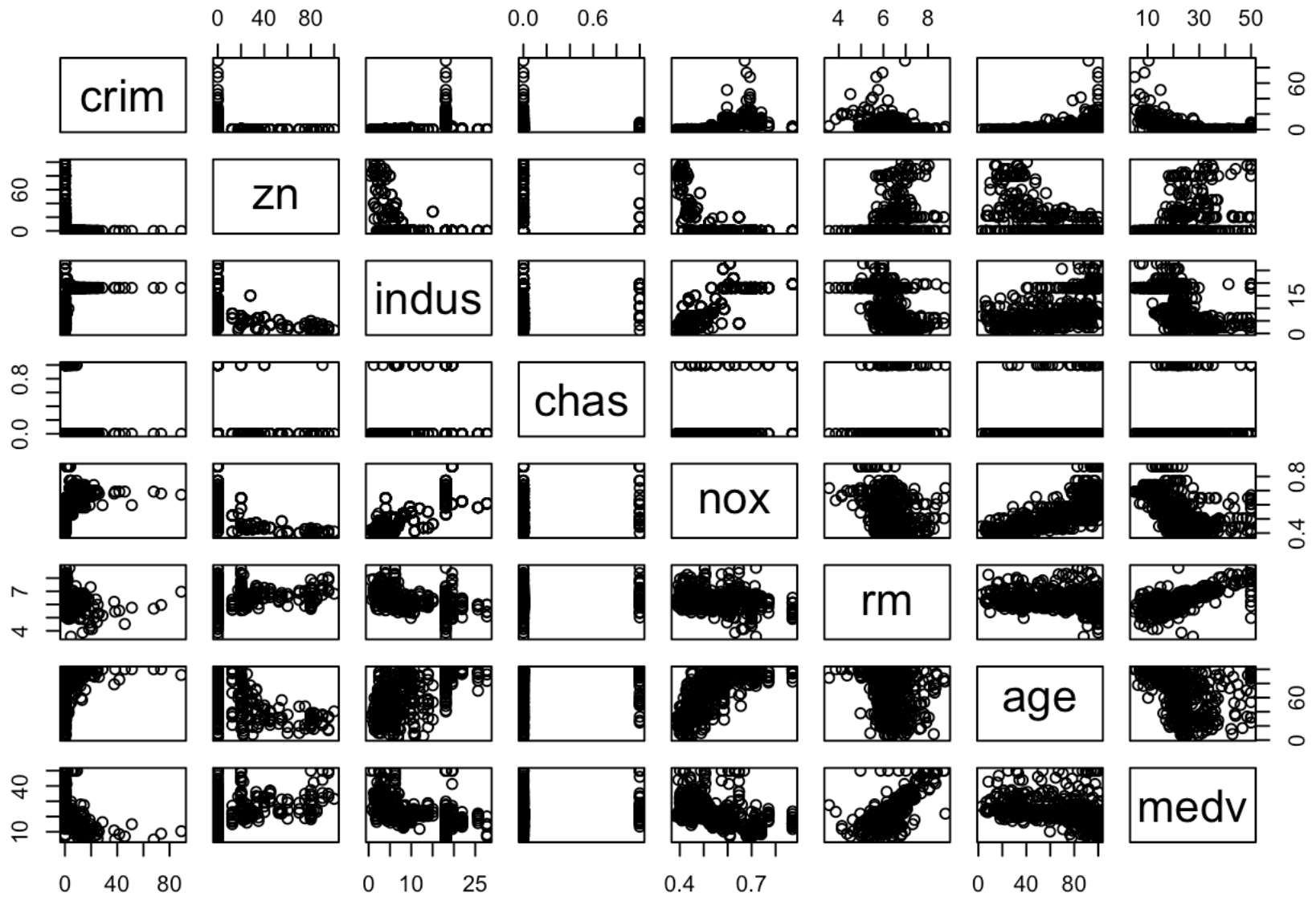
```
dim(Boston)
```

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

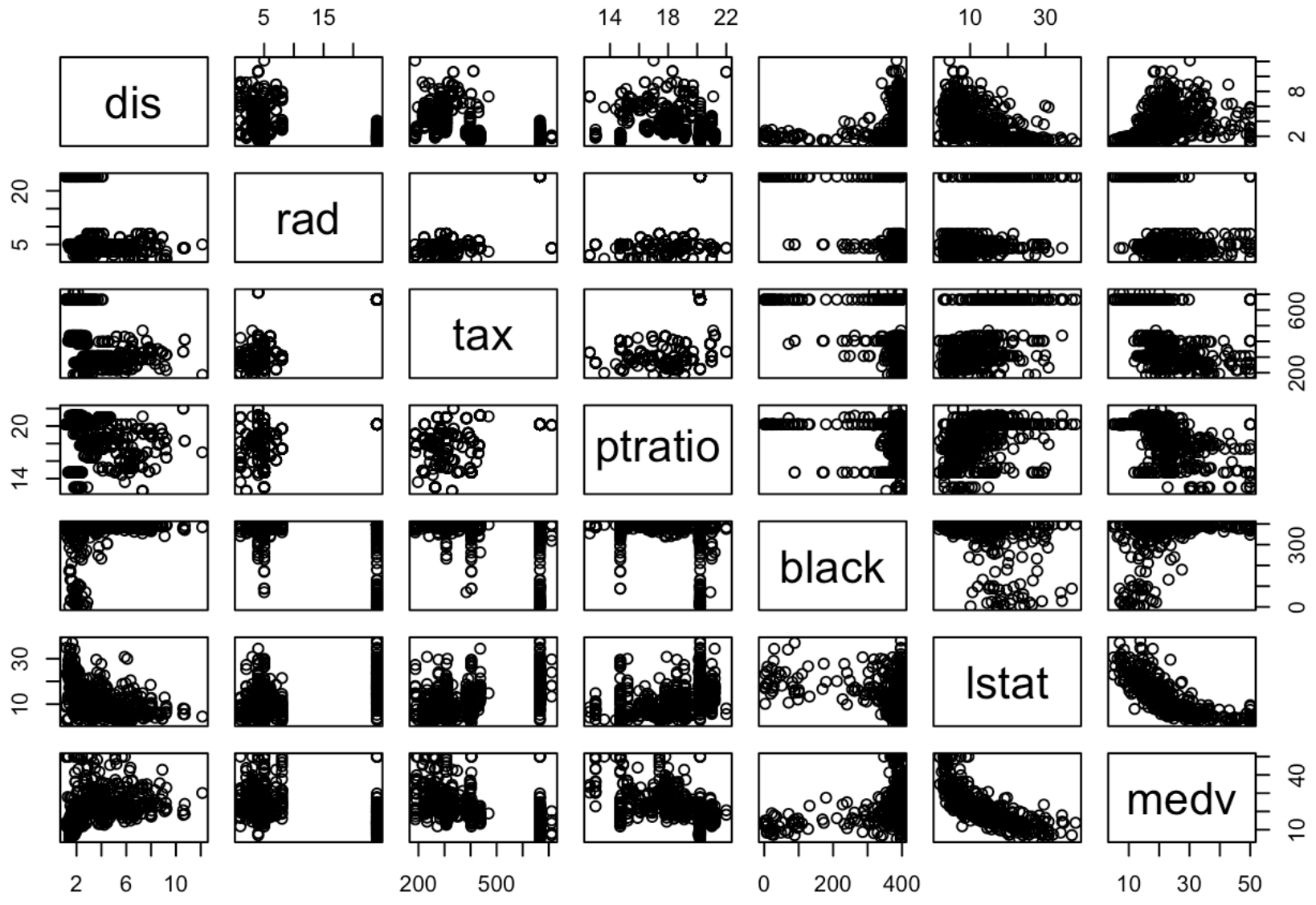
Each row represents an area in the suburbs of Boston and the columns represent the features.

(b)

```
pairs(Boston[,c(1,2,3,4,5,6,7,14)])
```



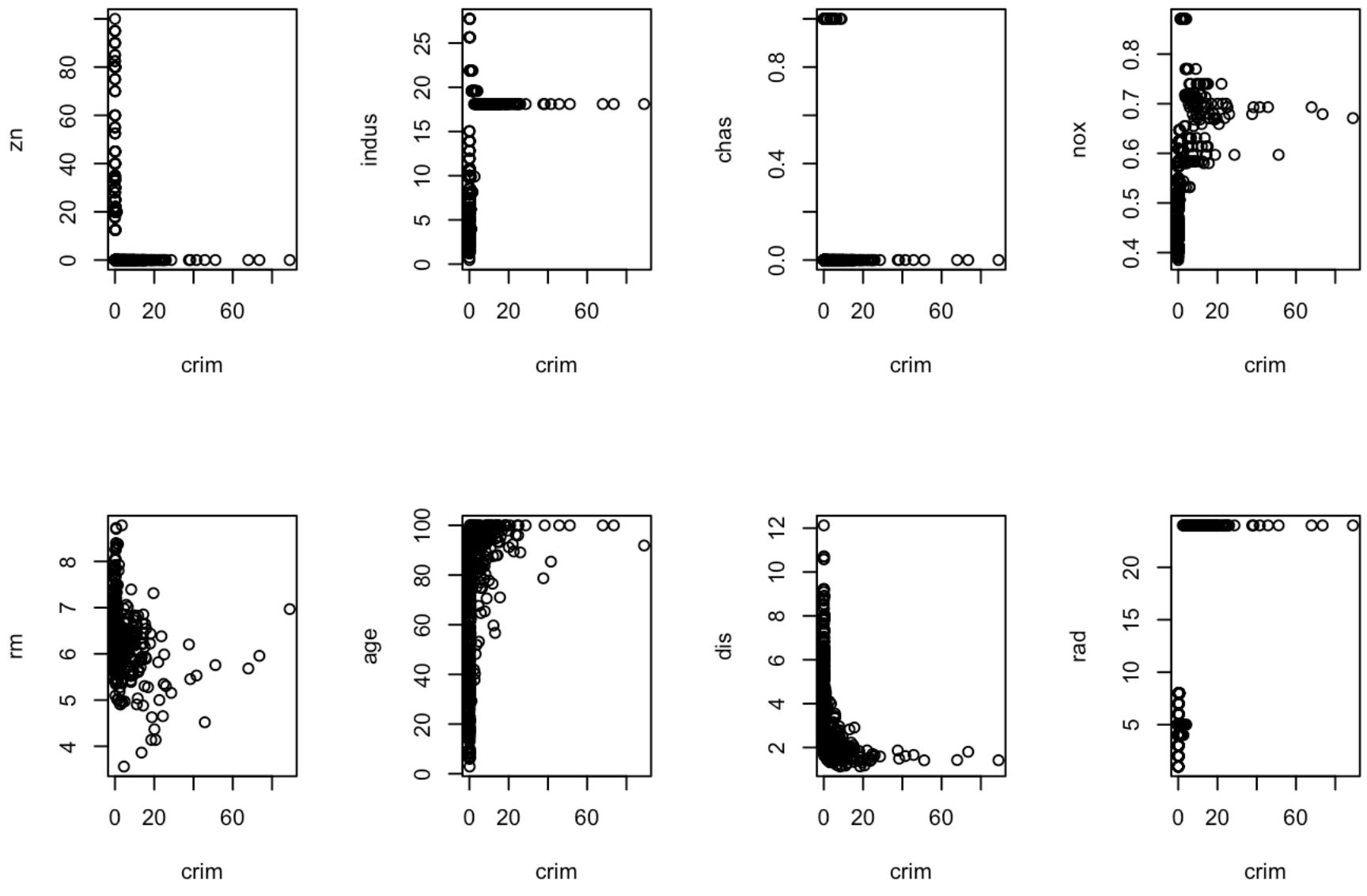

```
pairs(Boston[,8:14])
```



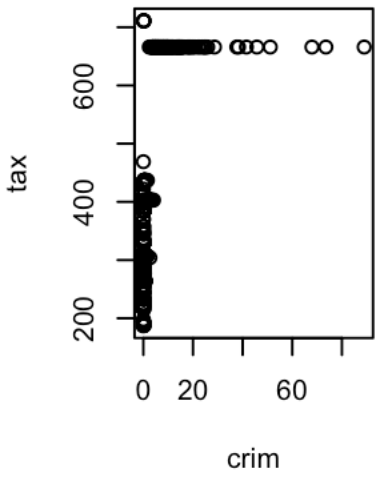
The median value has a strong relationship with per capita crime rate, proprortion of business acres, Charles River, average number of rooms per dwelling, weighted mean of distances to employment centres, accessibility to radial highways, blacks proportion and lower status of the population.

(c)

```
attach(Boston)
par(mfrow=c(2,4))
plot(crim,zn)
plot(crim,indus)
plot(crim,chas)
plot(crim,nox)
plot(crim,rm)
plot(crim,age)
plot(crim,dis)
plot(crim,rad)
```




```
plot(crim,tax)
```



Per capita crime rate tends to be higher in the area with few lots over 25,000 sq.ft or close to five employement centres.

Per capita crime tends to be higher when proportion of non-retail business acres per town is around 13 or nitrogen oxides concentration is around 0.6.

Per capita crime tends to be higher when tract bounds river.

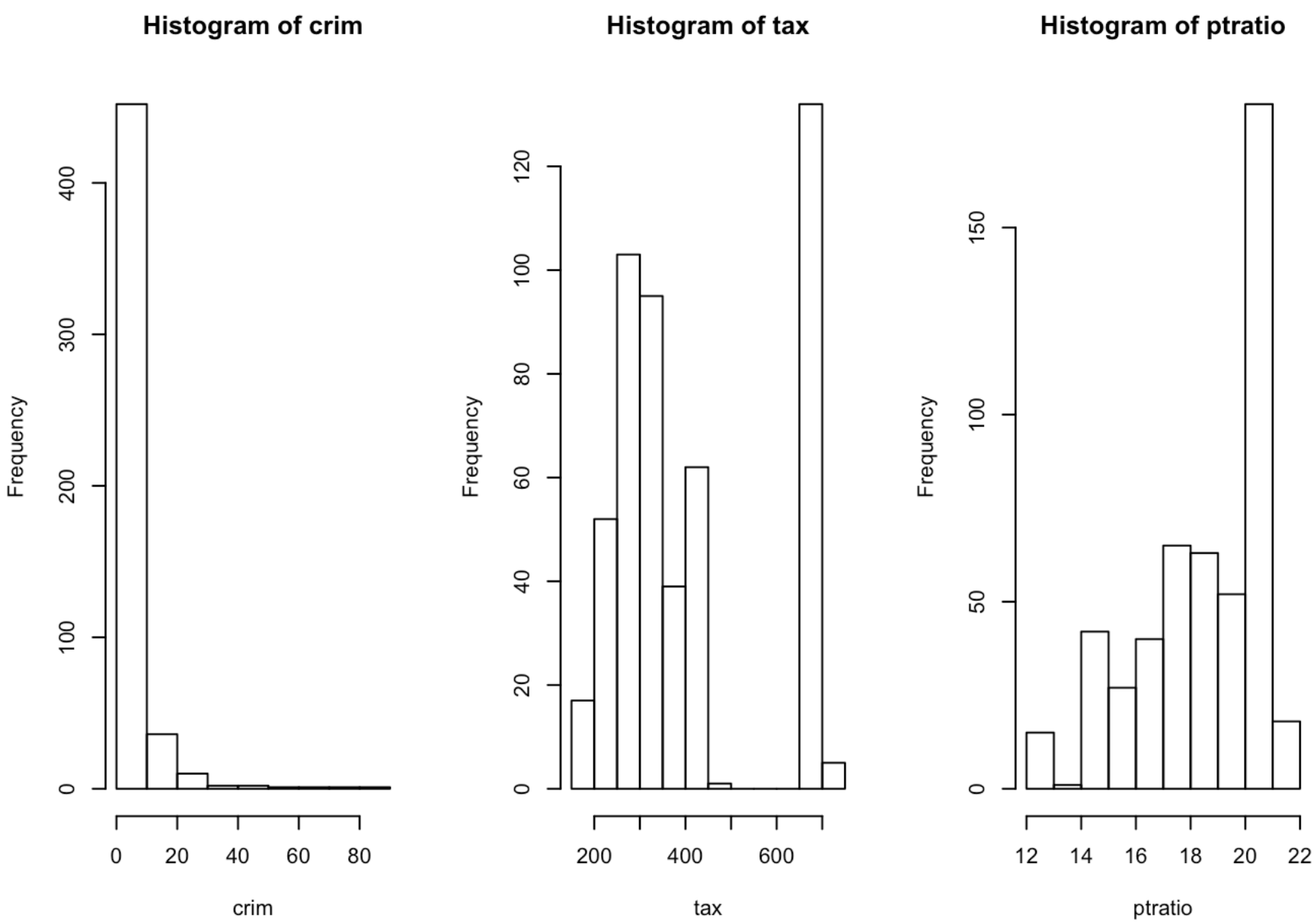
Per capita crime has a positive relationship with proportion old units, accessibility to radial highways and tax rate.

(d)

```
sapply(Boston, range)
```

```
##          crim  zn indus chas   nox    rm   age     dis rad tax ptratio
## [1,]  0.00632   0  0.46    0 0.385 3.561   2.9  1.1296   1 187    12.6
## [2,] 88.97620 100 27.74    1 0.871 8.780 100.0 12.1265  24 711    22.0
##       black lstat medv
## [1,]   0.32  1.73    5
## [2,] 396.90 37.97   50
```

```
par(mfrow=c(1,3))
hist(crim)
hist(tax)
hist(ptratio)
```



crime rate is lower (below 1) in most of the area, while some area even have crime rate over 40.

tax rate has two divisions: One is between 200 to 500, and another one is around 680.

Some areas have extreme high pupil-teacher ratios over 20.

(e)

```
sum(Boston$chas == 1)
```

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

35

(f)

```
median(ptratio)
```

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

19.05

(g)

```
Boston[Boston$medv == min(medv),]
```

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

```
summary(Boston)
```

```
##       crim                zn               indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08204   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox                rm              age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax             ptratio          black       
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   :  0.32  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.:375.38  
##  Median : 5.000   Median :330.0   Median :19.05   Median :391.44  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :356.67  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:396.23  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :396.90  
##      lstat            medv      
##  Min.   : 1.73   Min.   : 5.00  
##  1st Qu.: 6.95   1st Qu.:17.02  
##  Median :11.36   Median :21.20  
##  Mean   :12.65   Mean   :22.53  
##  3rd Qu.:16.95   3rd Qu.:25.00  
##  Max.   :37.97   Max.   :50.00
```

(h)

```
sum(rm>7)
```

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

```
sum(rm>8)
```

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

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

```
##       crim                zn             indus            chas        
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000  
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000  
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000  
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538  
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000  
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000  
##       nox               rm             age              dis        
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801  
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288  
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894  
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430  
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652  
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907  
##       rad              tax           ptratio          black       
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :354.6  
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:384.5  
##  Median : 7.000   Median :307.0   Median :17.40   Median :386.9  
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :385.2  
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:389.7  
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :396.9  
##      lstat            medv      
##  Min.   :2.47   Min.   :21.9  
##  1st Qu.:3.32   1st Qu.:41.7  
##  Median :4.14   Median :48.3  
##  Mean   :4.31   Mean   :44.2  
##  3rd Qu.:5.12   3rd Qu.:50.0  
##  Max.   :7.44   Max.   :50.0
```

The crime rate tends to be lower. Households are older. The proportion of non-retail business is lower. Tax rate is lower. Households are away from the highways. Lower status of the population is lower. Median value is higher.