Statistical learning assignment 1 - chapter 2

孫浩哲　　M072040002

September 20, 2018

8.

a.
```
setwd("C:/Users/asus/Desktop")
college=read.csv("college.csv")
```

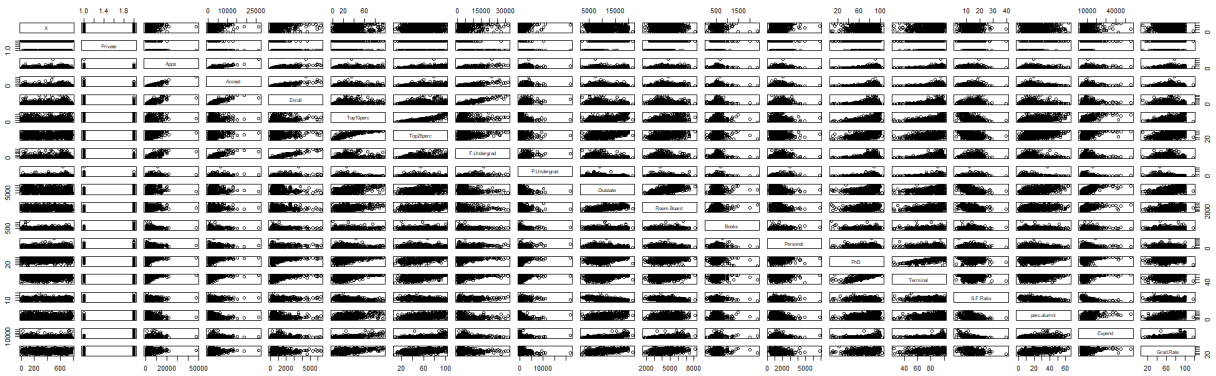
b.
```
rownames(college)=college[,1]
fix(college)
college=college[,-1]
fix(college)
```

c.

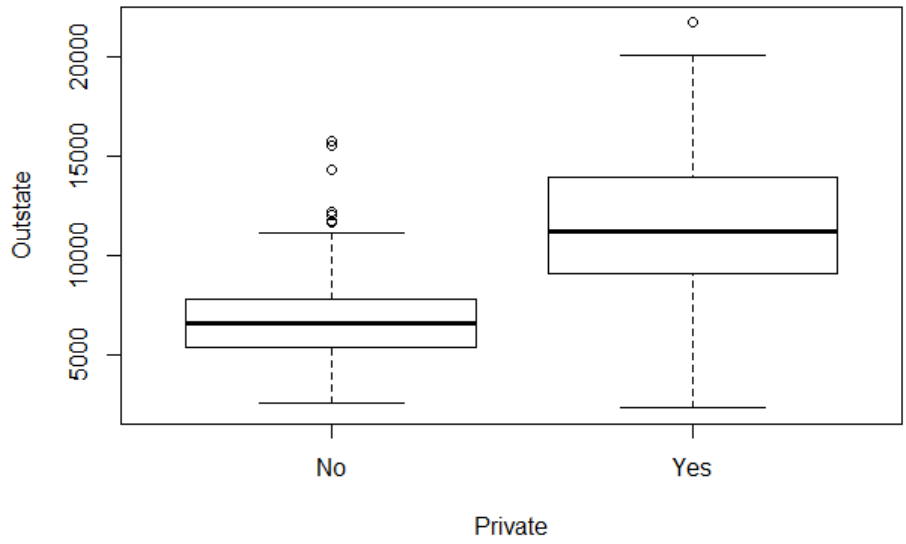
i.

```
summary(college)
```

ii.



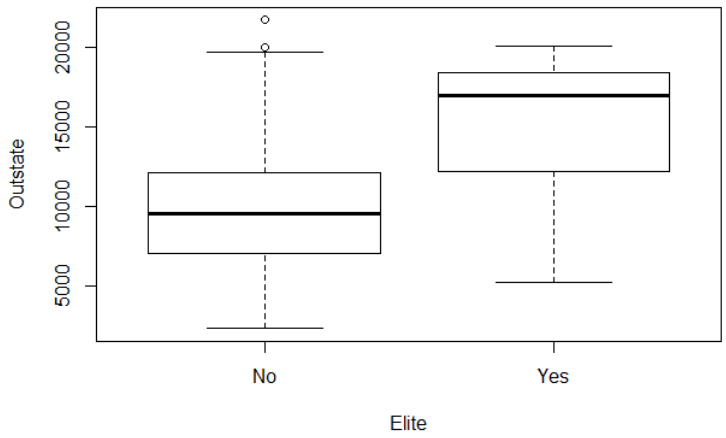
iii.



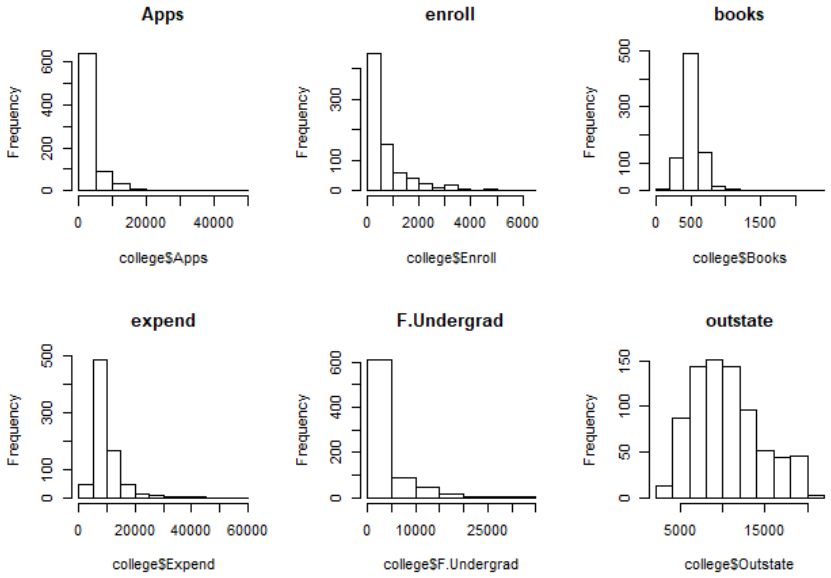
iv.

```
Elite=rep("No",nrow(college ))
```

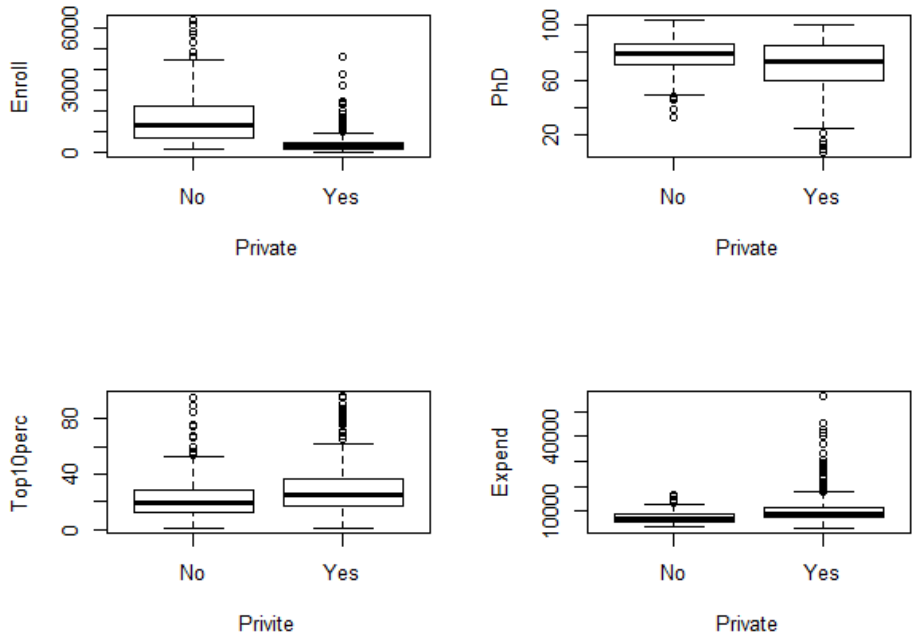
```
Elite[college$Top10perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college,Elite)
summary(college)
```



V.



vi.



As shown in figure, we compare the schools which are private to public ones.We want to know which class of variance is larger, so we check four features of the dataset through coding.

Class	Enroll	PhD	Top 10%	Expend
Private	209332.9	301.05	318.67	32291667
Public	1591614.4	151.72	261.81	7265945

According to the result, the variance of private schools is larger than public ones except for Enroll.

9.

(a)

Quantitative : mpg, displacment, horsepower, weight, acceleration

Qualitative : cylinders, year, origin, name

(b)

	mpg	displacement	horsepower	weight	acceleration
range	$9.0 \sim 46.60$	$68.0 \sim 455.0$	$46.0 \sim 230.0$	$1613 \sim 5140$	$8.00 \sim 24.80$

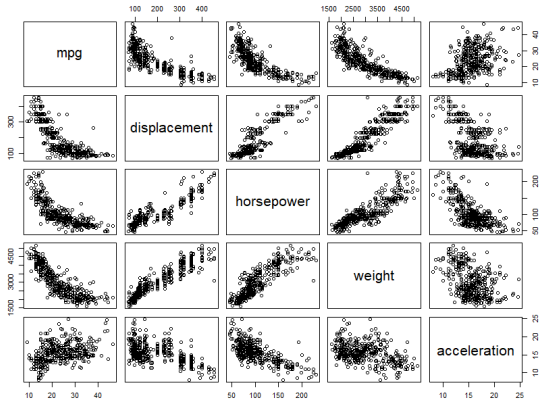
(c)

statistics	mpg	displacement	horsepower	weight	acceleration
mean	23.45	194.41	104.47	2977.58	15.54
std	7.81	104.64	38.49	849.40	2.76

(d)

statistics	mpg	displacement	horsepower	weight	acceleration
range	$11.00 \sim 46.60$	$68.0 \sim 455.0$	$46 \sim 230$	$1649 \sim 4997$	$8.50 \sim 24.80$
mean	24.37	5.38	187.75	100.96	2939.64
std	7.88	1.66	99.94	35.90	812.65

(e)



As shown in figure, we can see that mpg is negative correlated with many of predictors. However, the acceleration is not highly correlated with mpg.

(f)

```
> cor(auto$mpg,auto$weight)
[1] -0.8322442
> cor(auto$mpg,auto$horsepower)
[1] -0.7784268
> cor(auto$mpg,auto$displacement)
[1] -0.8051269
```

We can use weight to predict mpg, because the correlation between weight and mpg is -0.83, that means they are highly correlated.

10.

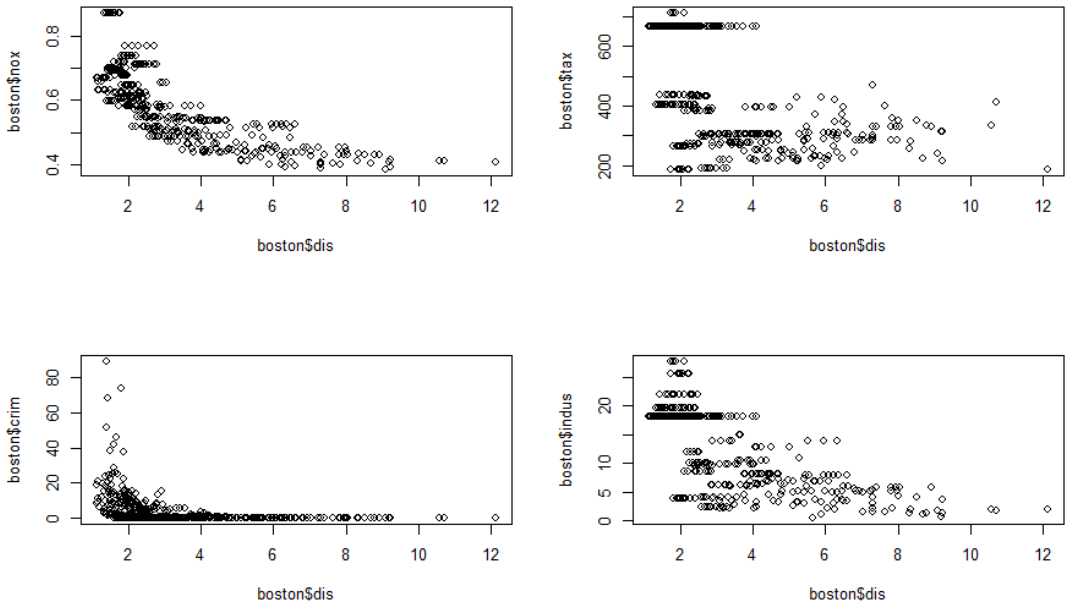
(a)

506 rows, 14 columns.

The rows represent how many suburbs observed in this dataset.

The columns represent the features of the dataset.

(b)



We want to know the predictors associated with distance, as shown in figure, only Nitrogen Oxides Concentration(nox) is negative correlated with distance.

(c)

The index of accessibility to radial highways(rad) is associated with per capita crime rate. Because it has the highest correlation coefficient between crim than any other predictor.

```
for(i in 2:14)
{
 x[i-1]=cor(Boston$crim,Boston[,i])
}
> max(x)
[1] 0.6255051
> cor(Boston$crim,Boston$rad)
[1] 0.6255051
```

(d) `summary(Boston)`

Through coding, we can find that there are some suburbs appear to have particularly high crime rates and tax rates.

(e) There are 35 suburbs in this data set bound the Charles river.

```
> summary(as.factor(Boston$chas))
  0   1
471  35
```

(f) 19.05

```
> median(Boston$ptratio)
[1] 19.05
```

(g) `summary(Boston)`
`min(Boston$medv)`

Through coding, we discover that the suburb has lowest median value of owner occupied is high in most of features.

(h) 64 suburbs average more than seven rooms per dwelling.

13 suburbs average more than eight rooms per dwelling.

```
> bo=boston[which(boston$rm>7),]
> nrow(bo)
[1] 64
> bo=boston[which(boston$rm>8),]
> nrow(bo)
[1] 13
```

The suburbs which average more than eight rooms per dwelling are high in black and low in dis, lstat and crime.