W4240 Data Mining Section 2

Rui Fan

############ 2.8 ############

##(a)

```
college<-read.csv("http://www-bcf.usc.edu/~gareth/ISL/College.csv",
header=TRUE)
```

##(b)

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

##(c)

#i.

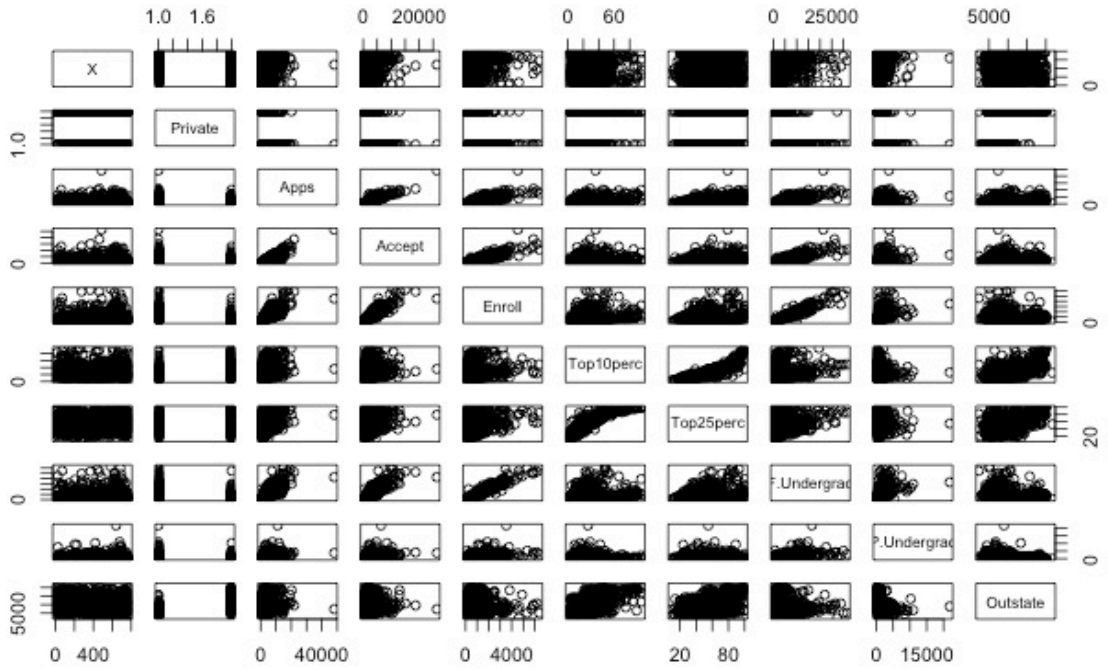
```
summary(college)
```

```
                          X       Private        Apps           Accept          Enroll       Top10perc      Top25perc
 Abilene Christian University:  1   No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
 Adelphi University          :  1   Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
 Adrian College              :  1             Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
 Agnes Scott College         :  1             Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
 Alaska Pacific University   :  1             3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
 Albertson College           :  1             Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
 (Other)                     :771
  F.Undergrad     P.Undergrad         Outstate       Room.Board       Books           Personal         PhD            Terminal
 Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0
 1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0
 Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0   Median :1200   Median : 75.00   Median : 82.0
 Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7
 3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0
 Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   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.

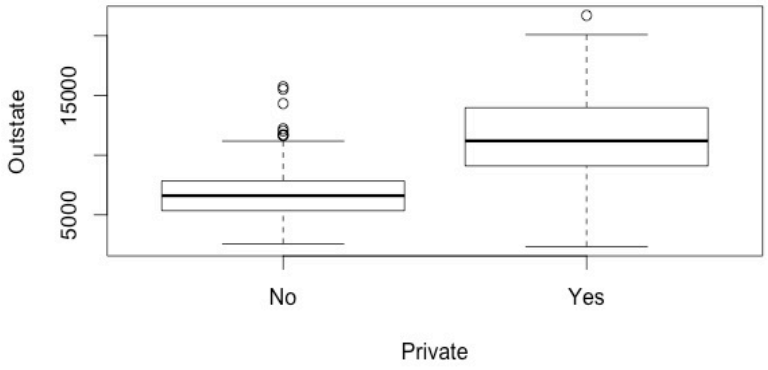
```
pairs(college[,1:10])
```



#iii.

attach(college)

plot(Outstate~Private)



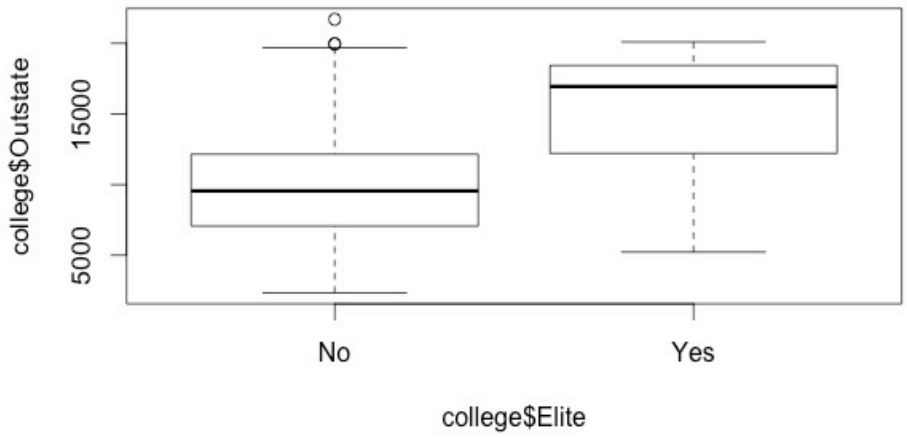
#iv.

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

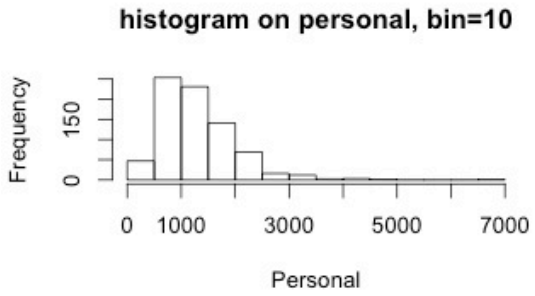
Elite[college$Top10perc >50]="Yes"

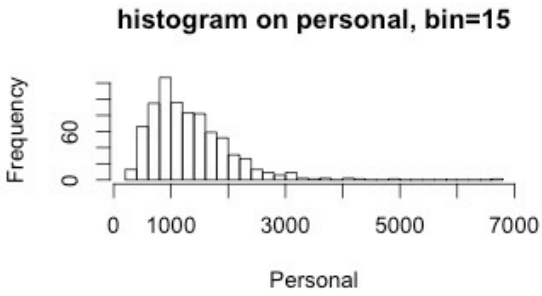
Elite=as.factor(Elite)

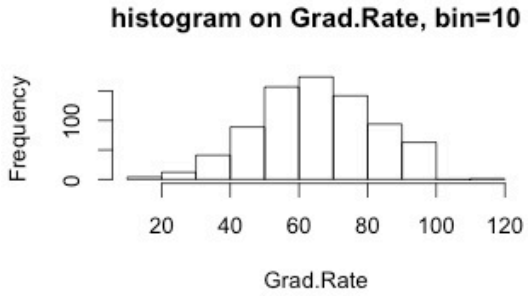
```
college=data.frame(college ,Elite)
summary(Elite)
   No Yes
  699  78
plot(college$Outstate~college$Elite)
```

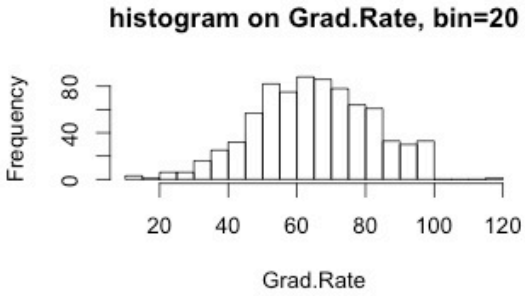



```
# v.
names(college)
par(mfrow=c(2,2))
hist(Personal, main="histogram on personal, bin=10",breaks=10)
hist(Personal, main="histogram on personal, bin=15",breaks=25)
hist(Grad.Rate, main="histogram on Grad.Rate, bin=10",breaks=10)
hist(Grad.Rate, main="histogram on Grad.Rate, bin=20",breaks=20)
```









#vi.

dim(college)

[1] 777 19

class(college$)

There are 777 subjects and 19 variables, of which 2(x and private) are categorical variable and the rest are numerical.

Also from the scatterplot we can see that some of the variables are highly positively related.

############ 2.9 #############

a.

Auto<-read.table(file.choose(), header=T, na.strings="?")

Auto=na.omit(Auto)

names(Auto)

summary(Auto)

attach(Auto)

```
class(mpg)
[1] "numeric"
class(cylinders)
[1] "integer"
class(displacement)
[1] "numeric"
class(horsepower)
[1] "numeric"
class(weight)
[1] "numeric"
> class(acceleration)
[1] "numeric"
class(year)
[1] "integer"
class(origin)
[1] "integer"
class(name)
[1] "factor"
```

So except for variable “name”, the rest of these eight variable are all numeric variable and “name” is a categorical variable.

```
## b.
allrange<-matrix(NA,8,2)
for(i in 1:8) {
   allrange[i,1]<-range(Auto[,i])[1]
   allrange[i,2]<-range(Auto[,i])[2]
```

Set the lower bound and the upper bound for range of each variable. And the first column of following matrix is the lower bound of each variable, the second column of following matrix is upper bound of each variable.

```
}
```

allrange

```
      [,1]   [,2]
[1,]    9   46.6
[2,]    3    8.0
[3,]   68  455.0
[4,]   46  230.0
[5,] 1613 5140.0
[6,]    8   24.8
[7,]   70   82.0
[8,]    1    3.0
```

```
## c.
meanandsd<-matrix(NA,8,2)
for(i in 1:8) {
  meanandsd[i,1]<-mean(Auto[,i])
  meanandsd[i,2]<-sd(Auto[,i])
```

Set the mean and standard deviation for range of each variable. And the first of the column of following matrix is the mean of each variable; the second column of the following matrix is standard deviation of each variable.

```
}
meanandsd
```

```
              [,1]          [,2]
[1,]     23.445918     7.8050075
[2,]      5.471939     1.7057832
[3,]    194.411990 104.6440039
[4,]    104.469388    38.4911599
[5,] 2977.584184 849.4025600
[6,]     15.541327     2.7588641
[7,]     75.979592     3.6837365
[8,]      1.576531     0.8055182
```

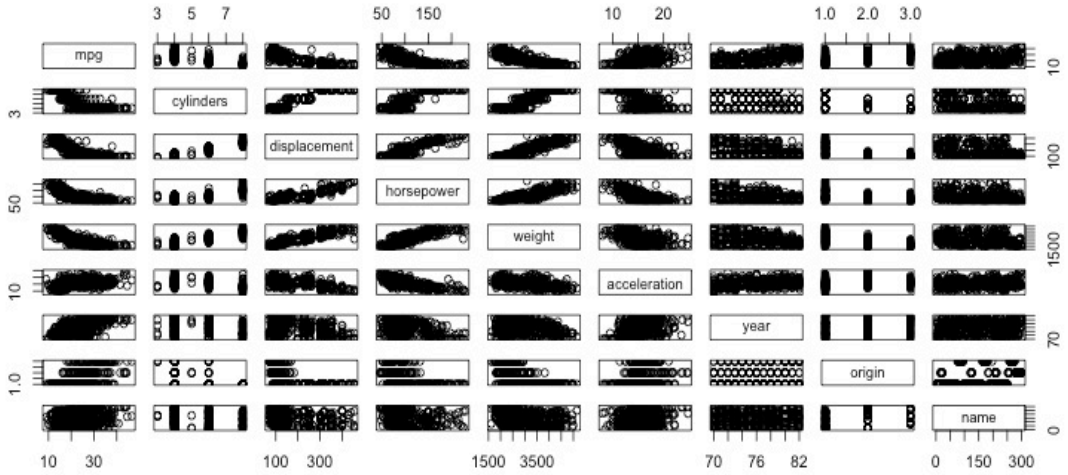
```
## d.
Autoremoved<-Auto[-(10:85),]
Autoremoved
summarywork<-matrix(NA,8,4)
for(i in 1:8) {
  summarywork[i,1]<-mean(Autoremoved[,i])
  summarywork[i,2]<-sd(Autoremoved[,i])
  summarywork[i,3]<-range(Autoremoved[,i])[1]
  summarywork[i,4]<-range(Autoremoved[,i])[2]
```

Set the mean, SD, lower bound and the upper bound for range of each variable. And the first column of following matrix is the mean of each variable, the second column is standard deviation, third column is the lower bound and the forth column is upper bound of each variable.

```
}
summarywork
```

	[,1]	[,2]	[,3]	[,4]
[1,]	24.404430	7.867283	11.0	46.6
[2,]	5.373418	1.654179	3.0	8.0
[3,]	187.240506	99.678367	68.0	455.0
[4,]	100.721519	35.708853	46.0	230.0
[5,]	2935.971519	811.300208	1649.0	4997.0
[6,]	15.726899	2.693721	8.5	24.8
[7,]	77.145570	3.106217	70.0	82.0
[8,]	1.601266	0.819910	1.0	3.0

```
## e.
pairs(Auto)
```



#From the scatterplot above, we can see that there are negative relationships between the mpg and displacement, weight, year as well as origin.

f.

```
fit<-lm(mpg~cylinders+displacement+horsepower+weight+acceleration+year+origin)
summary(fit)
```

Call:

```
lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
    acceleration + year + origin)
```

Residuals:

Min	1Q	Median	3Q	Max
-9.6524	-2.0723	-0.0504	1.8022	13.0248

Coefficients:

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-2.182e+01	4.266e+00	-5.115	4.94e-07 ***
cylinders	-2.413e-01	3.382e-01	-0.714	0.4760
displacement	1.558e-02	7.265e-03	2.144	0.0326 *

horsepower 1.129e-02 6.991e-03 1.615 0.1071

weight -6.837e-03 5.797e-04 -11.795 < 2e-16 ***

acceleration 1.520e-01 7.738e-02 1.964 0.0503 .

year 7.745e-01 4.930e-02 15.709 < 2e-16 ***

origin 1.359e+00 2.685e-01 5.062 6.41e-07 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.326 on 389 degrees of freedom

Multiple R-squared: 0.8226, Adjusted R-squared: 0.8194

F-statistic: 257.7 on 7 and 389 DF, p-value: < 2.2e-16

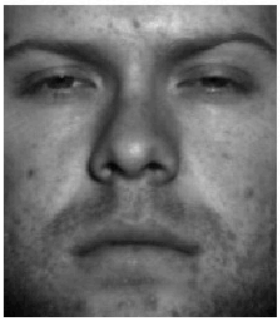
The result confirms what I expected.

########## 3. ##############

a.

```
> face_01
Pixmap image
  Type          : pixmapGrey
  Size          : 192x168
  Resolution    : 1x1
  Bounding box  : 0 0 168 192
```

hw01_01a: the first face



> class(x)

[1] "numeric"

So x is numerical and the size of the original image in pixels is 192*168.

b.

```
> min(faces_matrix)
[1] 0.007843137
> max(faces_matrix)
[1] 1
```

##0.007843137 corresponds to almost transparent, and 1 corresponds to black.

c.

```
length(my.dir.list.1)
 [1] 38
length(my.dir.list.2)
 [1] 2547
some(my.dir.list.1)
 [1] "yaleB03" "yaleB05" "yaleB12" "yaleB19" "yaleB26" "yaleB31" "yaleB33"
 [8] "yaleB35" "yaleB37" "yaleB38"
some(my.dir.list.2)
 [1] "yaleB04/yaleB04_P00A+085E-20.pgm" "yaleB06/yaleB06_P00A+060E-20.pgm"
 [3] "yaleB06/yaleB06_P00A-070E+45.pgm" "yaleB09/yaleB09_P00A+050E-40.pgm"
 [5] "yaleB12/yaleB12_P00A-025E+00.pgm" "yaleB13/yaleB13_P00A+005E-10.pgm"
 [7] "yaleB16/WS_FTP.LOG"               "yaleB26/yaleB26_P00A-015E+20.pgm"
 [9] "yaleB28/yaleB28_P00A+005E+10.pgm" "yaleB39/yaleB39_P00A+050E+00.pgm"
```

d.