# Assignment1.R

*afuyo*

*Sun Apr 02 07:10:36 2017*

```
#PREDICT 422
#Programming Assignment 1
#

###################################
#Excercise 8 (ISLR Section 2.4)

#8(a)
#Use the read.csv() function to read the data into R.
eval=FALSE
college=read.csv("C:\\playground\\Predict422\\R\\week1\\college.csv")
summary(college)
```

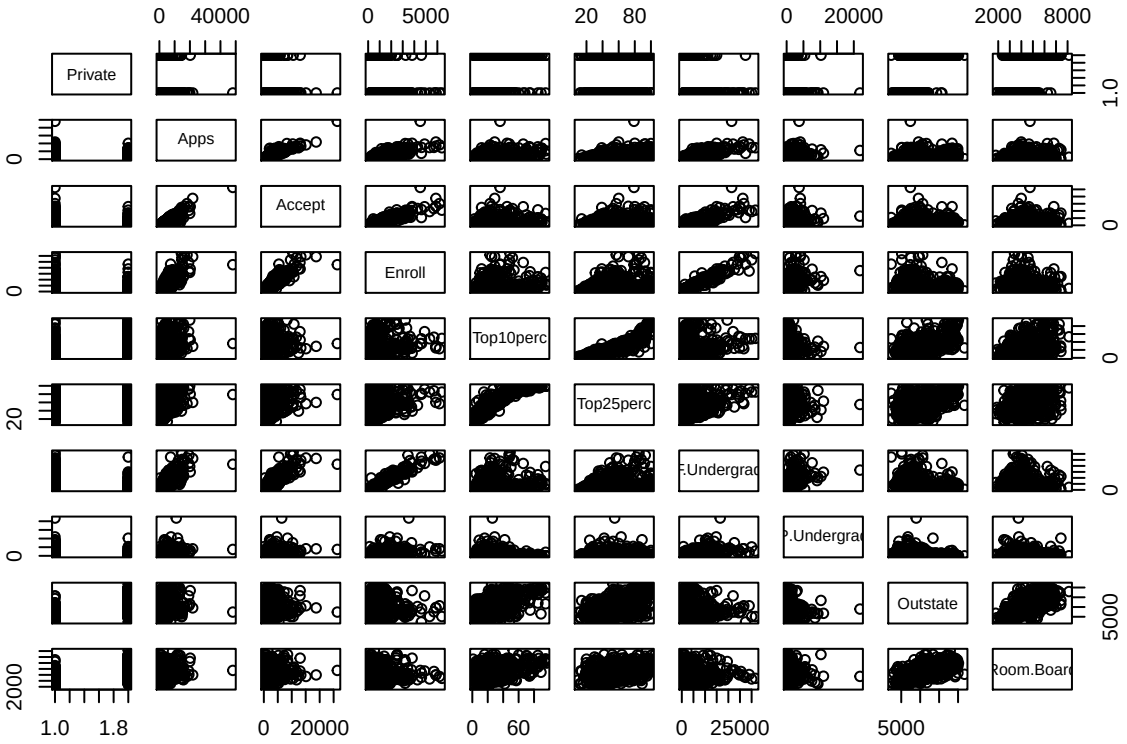
```
##                              X       Private        Apps
##  Abilene Christian University:  1   No :212   Min.   :   81
##  Adelphi University           :  1   Yes:565   1st Qu.:  776
##  Adrian College               :  1             Median : 1558
##  Agnes Scott College          :  1             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
##
```

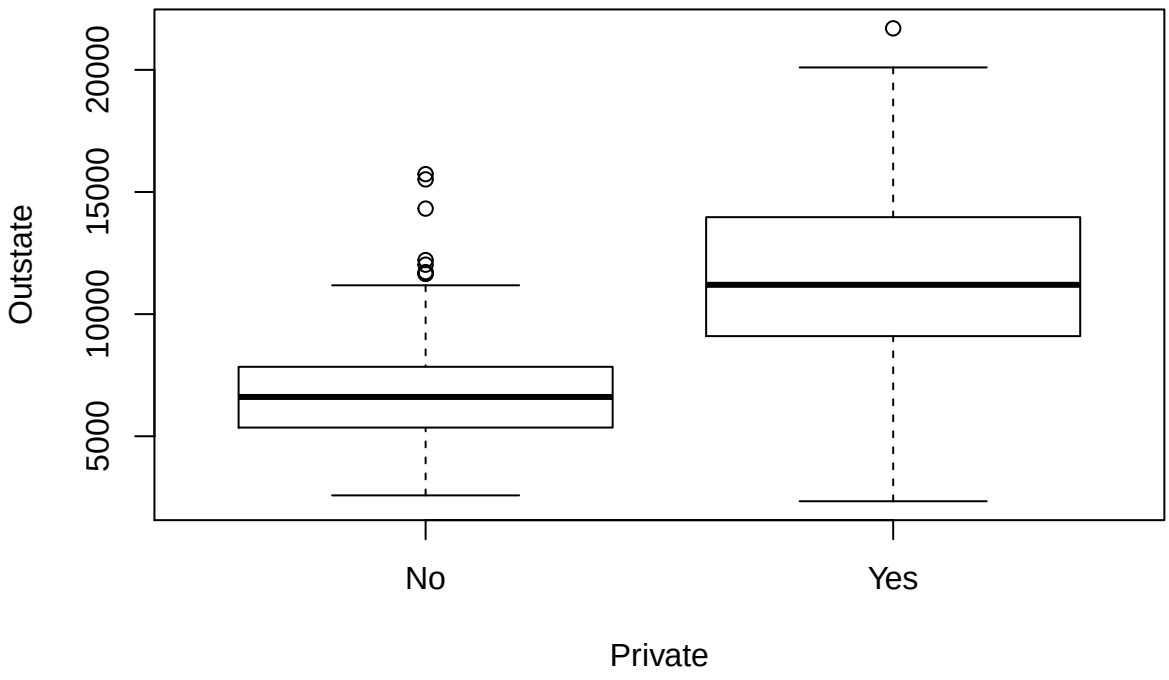
```
#8(b)
fix(college)
rownames(college) = college[,1]
fix(college)
college = college[,-1]
fix(college)
#8(c)
# 8 (c) i.
summary(college)
```

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

```r
# 8 (c) ii.
pairs(college[,1:10])
```



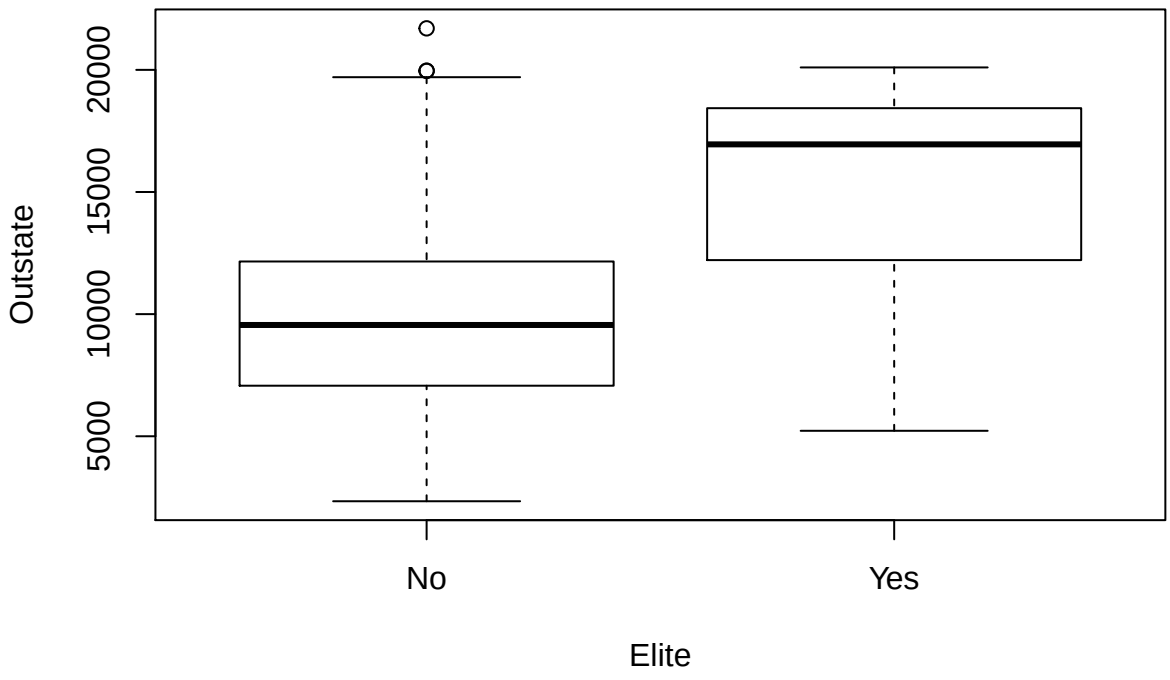

```r
# 8 (c) iii.
plot(college$Private,college$Outstate,xlab="Private",ylab="Outstate")
```




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

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

```
plot(college$Elite,college$Outstate,xlab="Elite",ylab="Outstate")
```




```
# 8 (c) v.
par(mfrow=c(2,2))
hist(college$Apps,breaks=50,col="gray",xlab="Apps",main="")
hist(college$Accept,breaks=20,col="gray",xlab="Accept",main="")
hist(college$Enroll,breaks=15,col="gray",xlab="Enroll",main="")
# 8 (c) vi.
# additional EDA

########################################
# Exercise 9 (ISLR Section 2.4)
inPath = file.path("C:","playground","Predict422",
                   "R","week1")
# Load data and remove missing values per the lab (Section 2.3.4)
Auto = read.table(file.path(inPath,"Auto.data"),header=TRUE,na.strings="?")
dim(Auto)
```

```
## [1] 397   9
```

```
Auto = na.omit(Auto)
dim(Auto)
```

```
## [1] 392   9
```

```r
# 9 (a)
str(Auto)
```

```
## 'data.frame':    392 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : int  8 8 8 8 8 8 8 8 8 8 ...
##  $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
##  $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
##  $ weight      : num  3504 3693 3436 3433 3449 ...
##  $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
##  $ year        : int  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223 241 1
##  - attr(*, "na.action")=Class 'omit'  Named int [1:5] 33 127 331 337 355
##   .. ..- attr(*, "names")= chr [1:5] "33" "127" "331" "337" ...
```

```r
summary(Auto)
```

```
##       mpg          cylinders      displacement     horsepower
##  Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   : 46.0
##  1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.: 75.0
##  Median :22.75   Median :4.000   Median :151.0   Median : 93.5
##  Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :104.5
##  3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:126.0
##  Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :230.0
##
##      weight      acceleration        year           origin
##  Min.   :1613   Min.   : 8.00   Min.   :70.00   Min.   :1.000
##  1st Qu.:2225   1st Qu.:13.78   1st Qu.:73.00   1st Qu.:1.000
##  Median :2804   Median :15.50   Median :76.00   Median :1.000
##  Mean   :2978   Mean   :15.54   Mean   :75.98   Mean   :1.577
##  3rd Qu.:3615   3rd Qu.:17.02   3rd Qu.:79.00   3rd Qu.:2.000
##  Max.   :5140   Max.   :24.80   Max.   :82.00   Max.   :3.000
##
##                  name
##  amc matador       :  5
##  ford pinto        :  5
##  toyota corolla    :  5
##  amc gremlin       :  4
##  amc hornet        :  4
##  chevrolet chevette:  4
##  (Other)           :365
```

```r
# 9 (b)
#sapply(Auto[,1:7],range)
for (ii in 1:7)
{
  rng = range(Auto[,ii])
  print(paste(names(Auto)[ii],": ",rng[1]," to ",rng[2],sep=""))
}
```

```
## [1] "mpg: 9 to 46.6"
```

```
## [1] "cylinders: 3 to 8"
## [1] "displacement: 68 to 455"
## [1] "horsepower: 46 to 230"
## [1] "weight: 1613 to 5140"
## [1] "acceleration: 8 to 24.8"
## [1] "year: 70 to 82"
```

```
# 9 (c)
#apply(Auto[,1:7],2,mean)
#apply(Auto[,1:7],2,sd)
sapply(Auto[,1:7],mean)
```

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

```
sapply(Auto[,1:7],sd)
```

```
##          mpg    cylinders displacement   horsepower       weight
##     7.805007     1.705783   104.644004    38.491160   849.402560
## acceleration         year
##     2.758864     3.683737
```

```
# 9 (d) Now remove the 10th through 85th observations.
AutoSubset = Auto[-(10:85),]
sapply(AutoSubset[,1:7],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
```

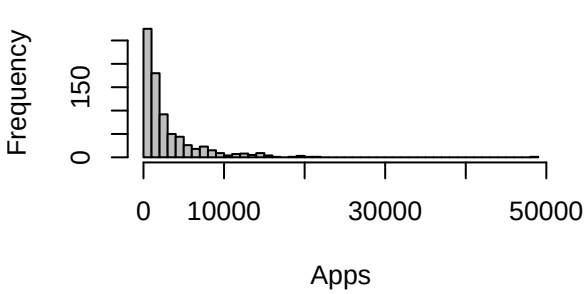
```
sapply(AutoSubset[,1:7],mean)
```

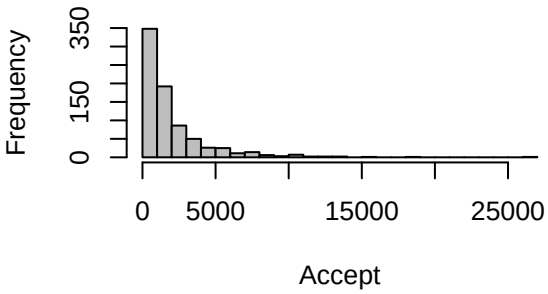
```
##          mpg    cylinders displacement   horsepower       weight
##    24.404430     5.373418   187.240506   100.721519  2935.971519
## acceleration         year
##    15.726899    77.145570
```

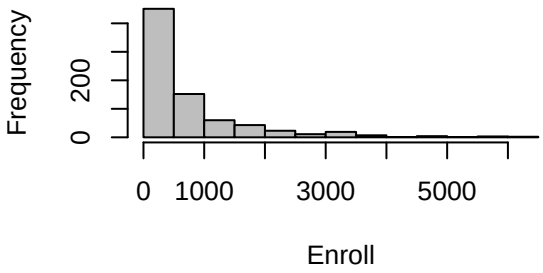
```
sapply(AutoSubset[,1:7],sd)
```

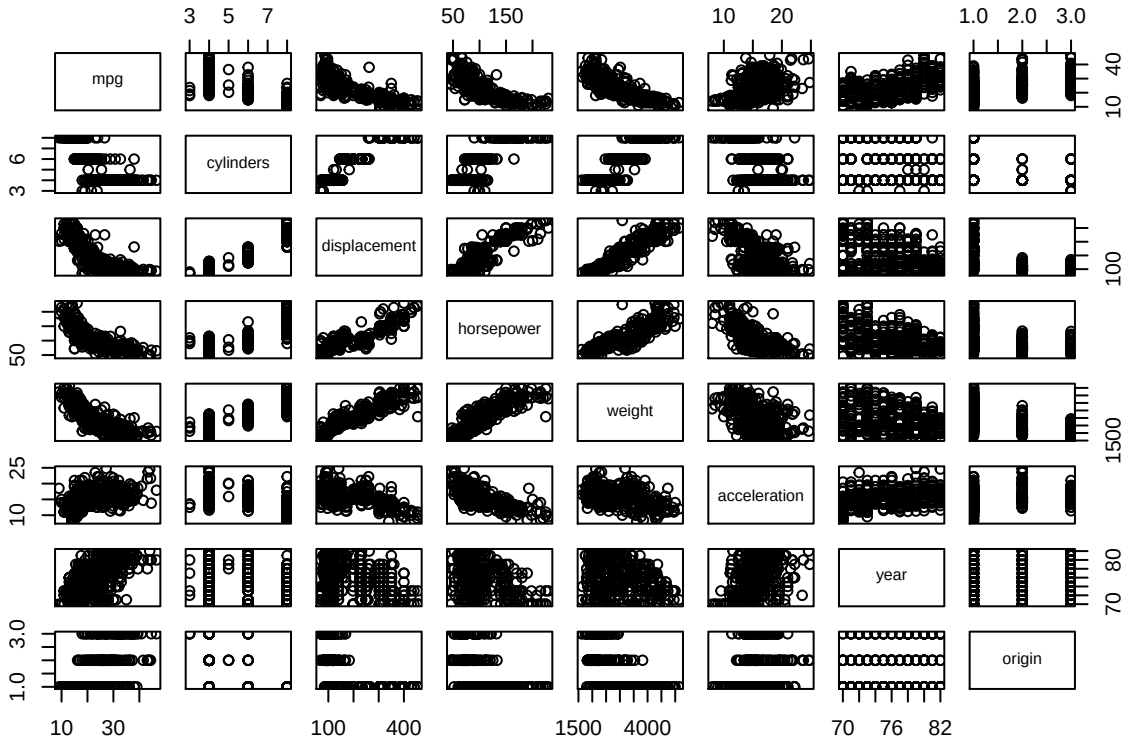
```
##          mpg    cylinders displacement   horsepower       weight
##     7.867283     1.654179    99.678367    35.708853   811.300208
## acceleration         year
##     2.693721     3.106217
```

```
# 9 (e) Using the full data set, investigate the predictors graphically,
# scatterplots or other tools of your choice. Create some plots
#highlighting the relationships among the predictors. Comment
#zon your findings.
pairs(Auto[,1:8])
```

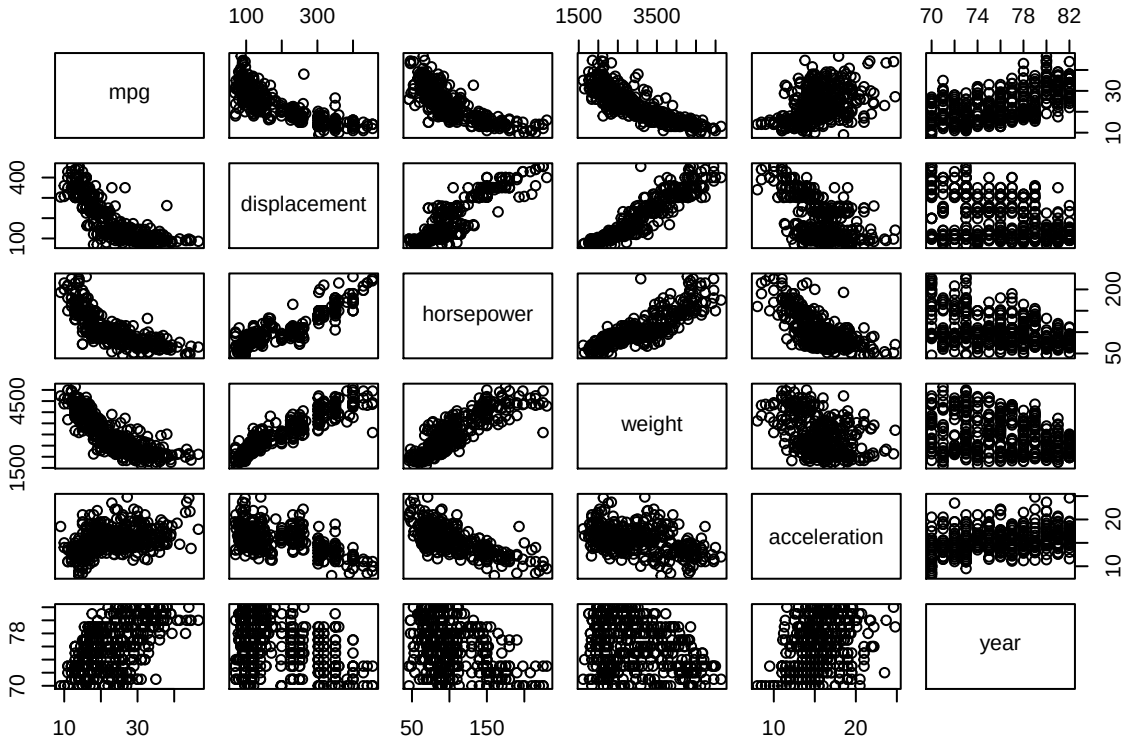

Frequency
0
150
0
10000
30000
50000
Apps


Frequency
0
150
350
0
5000
15000
25000
Accept

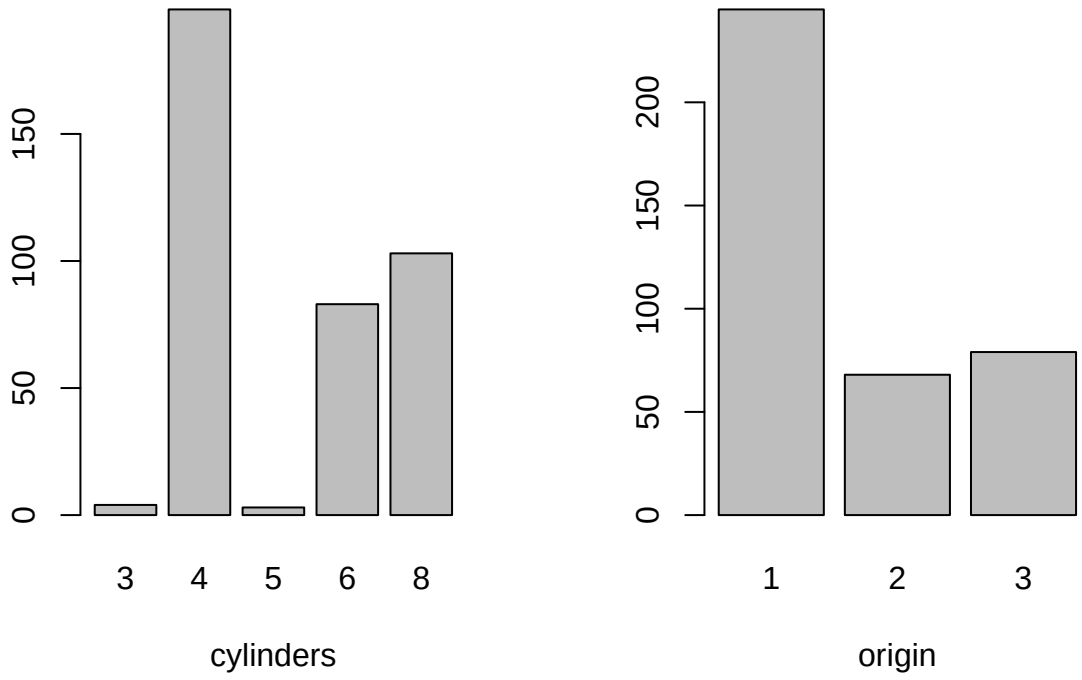

Frequency
0
200
0
1000
3000
5000
Enroll




```
Auto$cylinders = as.factor(Auto$cylinders)
Auto$origin = as.factor(Auto$origin)
pairs(Auto[,c(1,3:7)])
```

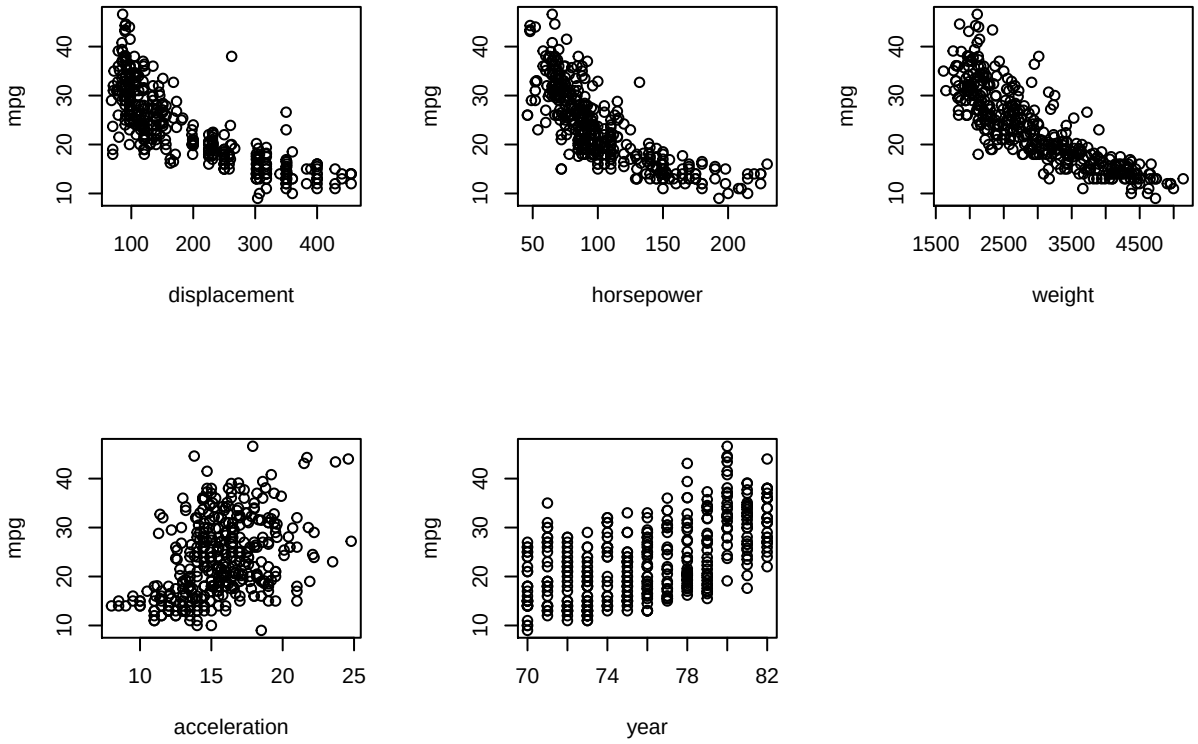



```
par(mfrow=c(1,2))
barplot(table(Auto$cylinders),xlab="cylinders")
barplot(table(Auto$origin),xlab="origin")
```

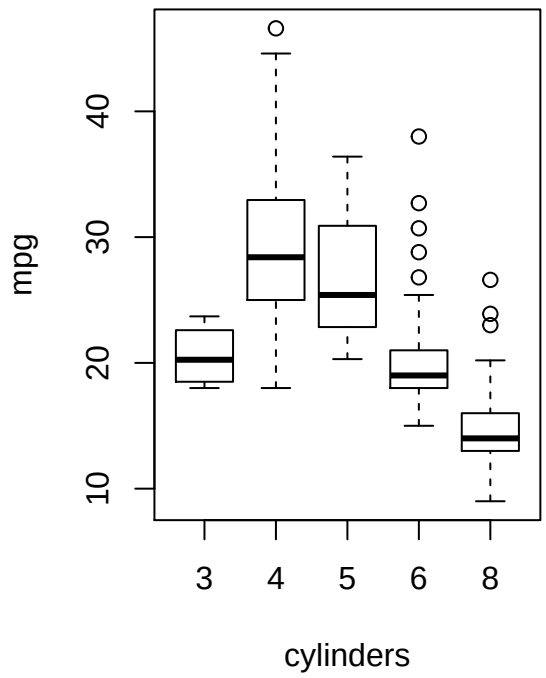


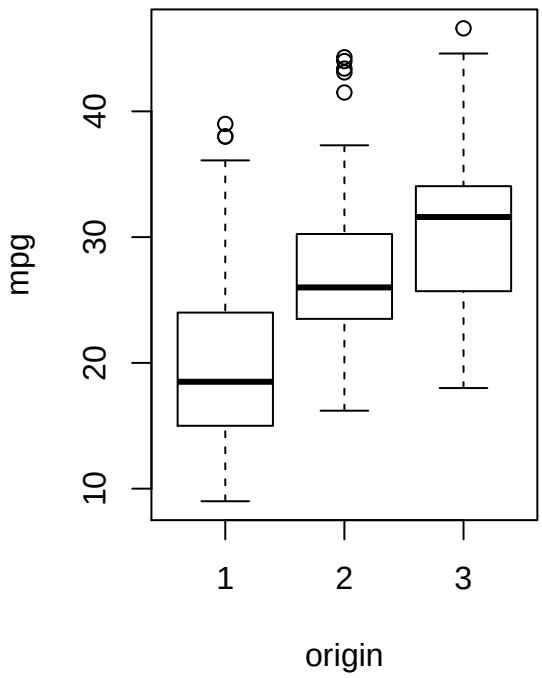

```
# 9 (f)Suppose that we wish to predict gas mileage (mpg) on the basis
#of the other variables. Do your plots suggest that any of the
#other variables might be useful in predicting mpg? Justify your
#answer.
par(mfrow=c(2,3))
plot(Auto$displacement,Auto$mpg,xlab="displacement",ylab="mpg")
# more horsepower correlates with lower mpg
plot(Auto$horsepower,Auto$mpg,xlab="horsepower",ylab="mpg")
#heavier weight correlates with lower mpg
plot(Auto$weight,Auto$mpg,xlab="weight",ylab="mpg")
plot(Auto$acceleration,Auto$mpg,xlab="acceleration",ylab="mpg")
plot(Auto$year,Auto$mpg,xlab="year",ylab="mpg")
#mpg increases over time

par(mfrow=c(1,2))
```




```
plot(Auto$cylinders,Auto$mpg,xlab="cylinders",ylab="mpg")
plot(Auto$origin,Auto$mpg,xlab="origin",ylab="mpg")
```


mpg
10
20
30
40
3
4
5
6
8
cylinders


mpg
10
20
30
40
1
2
3
origin