HW1 Solution Outline CS6220-Data Mining

1.JWHT problem 8, p. 54

(a)

```
setwd("/Users/ovitek/Dropbox/Olga/Teaching/CS6220/Fall15/Homeworks/Hw1")
college=read.csv("College.csv")
head(college)
```

```
##                              X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3              Adrian College     Yes 1428   1097    336        22
## 4          Agnes Scott College     Yes  417    349    137        60
## 5    Alaska Pacific University     Yes  193    146     55        16
## 6            Albertson College     Yes  587    479    158        38
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
## 4        89         510          63    12960       5450   450      875  92
## 5        44         249         869     7560       4120   800     1500  76
## 6        62         678          41    13500       3335   500      675  67
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
## 4       97       7.7          37  19016        59
## 5       72      11.9           2  10922        15
## 6       73       9.4          11   9727        55
```

(b)

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

```
##                              Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
## Agnes Scott College              Yes  417    349    137        60
## Alaska Pacific University        Yes  193    146     55        16
## Albertson College                Yes  587    479    158        38
##                              Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University        52        2885         537     7440
## Adelphi University                  29        2683        1227    12280
## Adrian College                      50        1036          99    11250
## Agnes Scott College                 89         510          63    12960
## Alaska Pacific University           44         249         869     7560
```

```
## Albertson College                   62         678          41    13500
##                              Room.Board Books Personal PhD Terminal
## Abilene Christian University       3300   450     2200  70       78
## Adelphi University                 6450   750     1500  29       30
## Adrian College                     3750   400     1165  53       66
## Agnes Scott College                5450   450      875  92       97
## Alaska Pacific University          4120   800     1500  76       72
## Albertson College                  3335   500      675  67       73
##                              S.F.Ratio perc.alumni Expend Grad.Rate
## Abilene Christian University      18.1          12   7041        60
## Adelphi University                12.2          16  10527        56
## Adrian College                    12.9          30   8735        54
## Agnes Scott College                7.7          37  19016        59
## Alaska Pacific University         11.9           2  10922        15
## Albertson College                  9.4          11   9727        55
```

(c)

i.

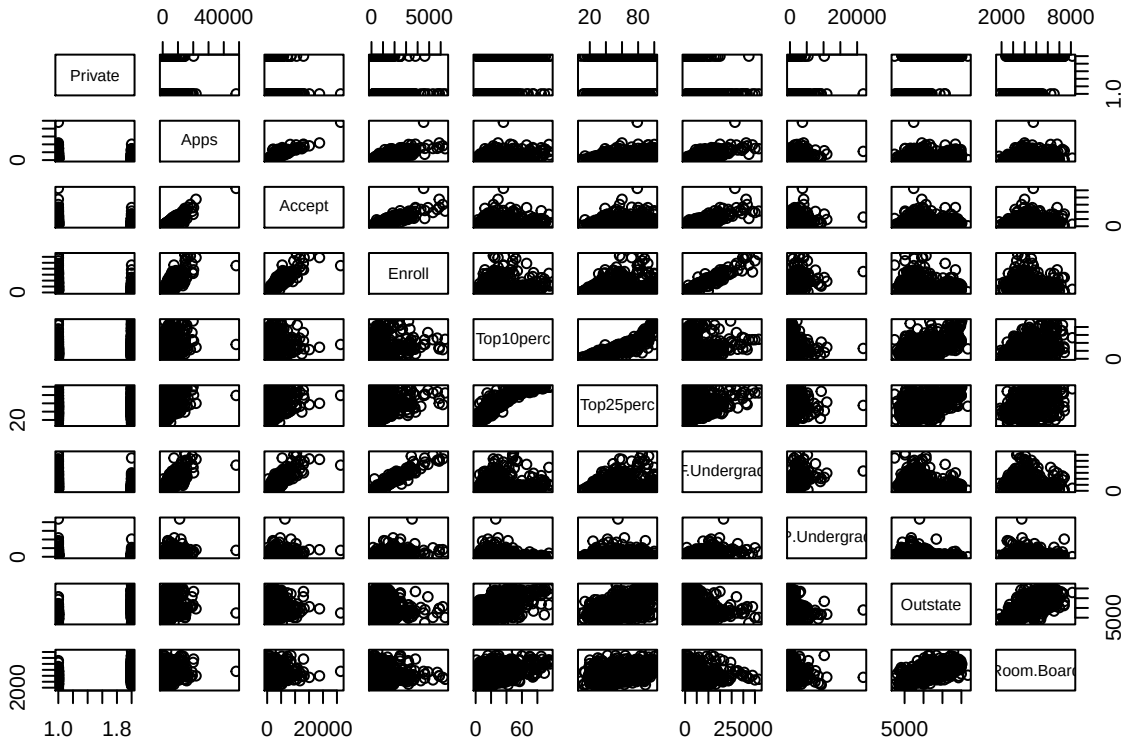
```
summary(college)
```

```
##  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
```

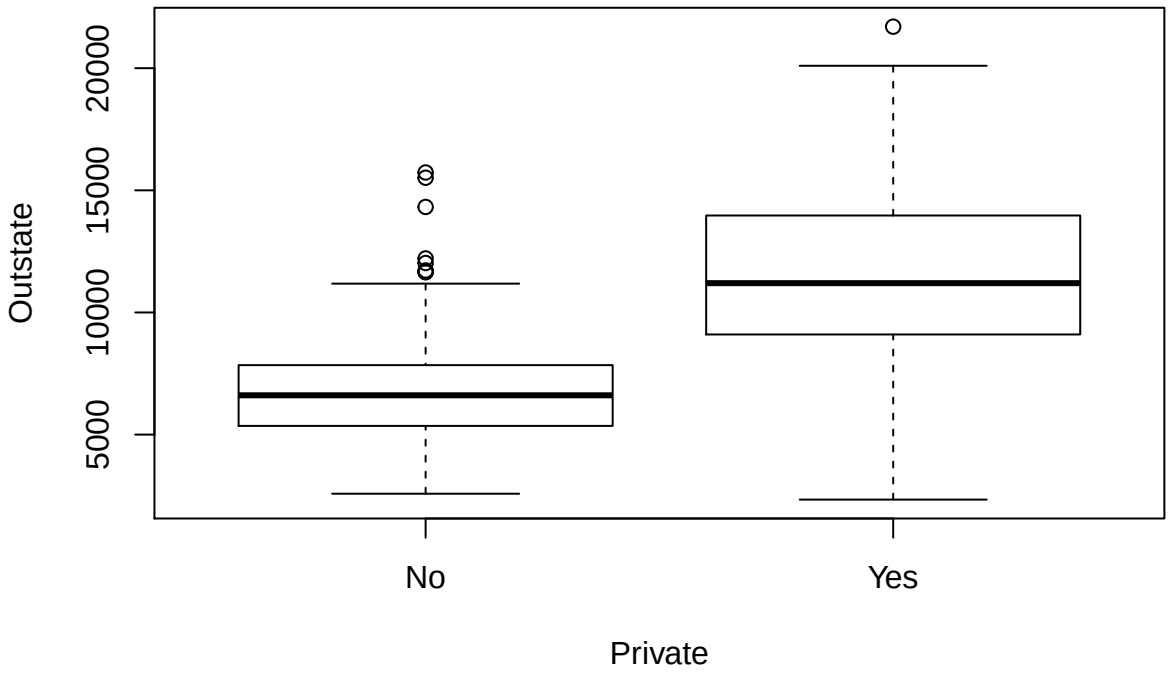
ii.

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



iii.

```
plot(Outstate~Private, data=college)
```




```
#or you can use: plot(college$Outstate~college$Private)
# You can see the outstate.tuition is higher for Private school compared to Public school
```

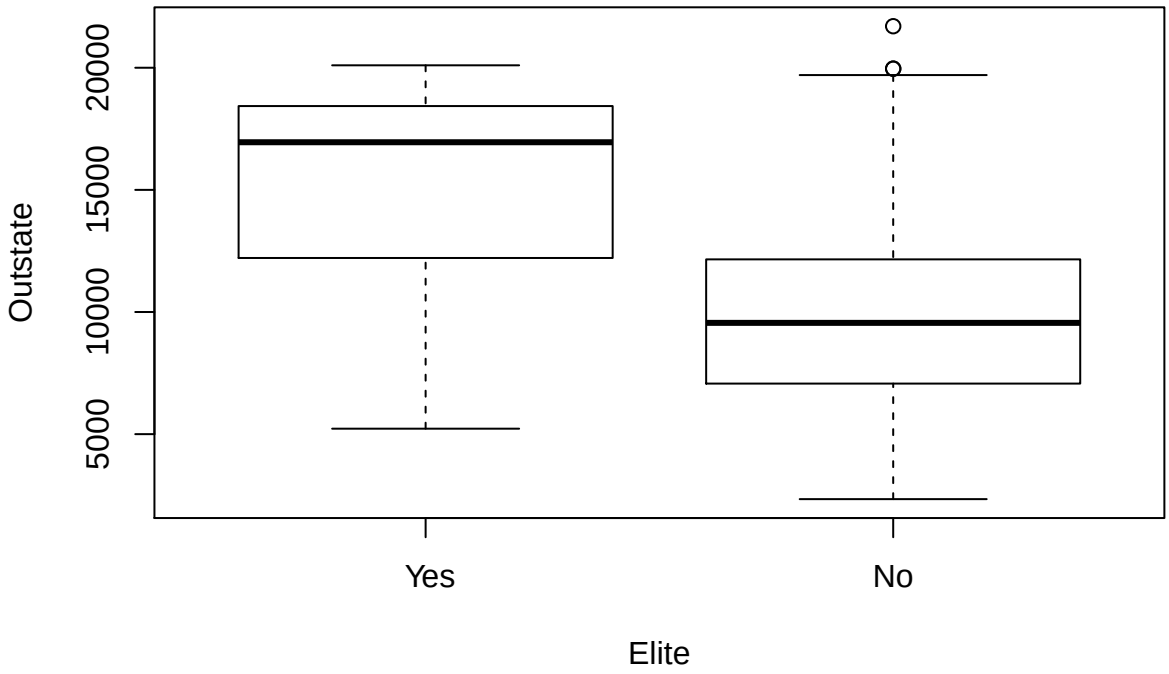
iv.

```
Elite =rep ("No",nrow(college ))
Elite [college$Top10perc >50]=" Yes"
Elite =as.factor (Elite)
college =data.frame(college ,Elite)
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       Elite
##  Min.   : 10.00    Yes: 78
##  1st Qu.: 53.00   No  :699
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

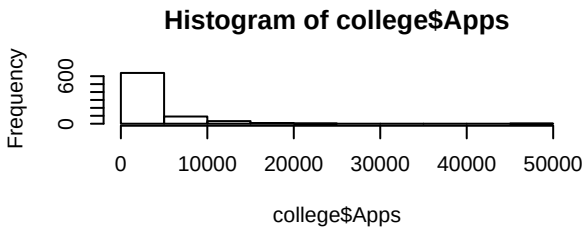
```
plot(Outstate~Elite, data=college)
```

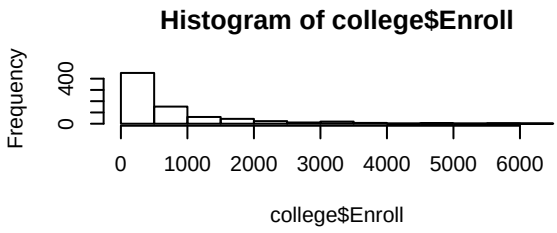


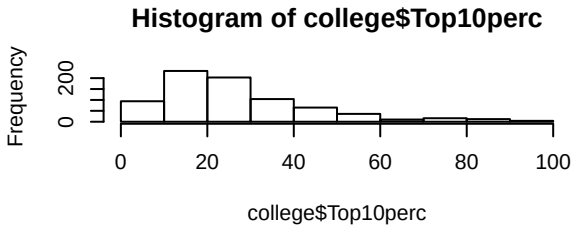

```
# You can see the outstate.tuition is higher for Elite school compared to public school
```

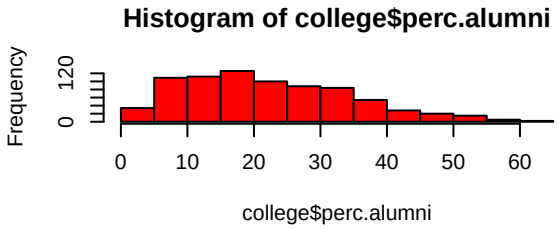
v.

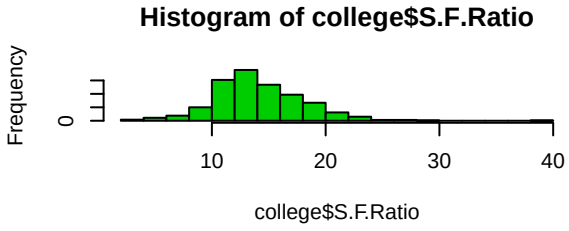
```
par(mfrow=c(3,2))
hist(college$Apps)
hist(college$Enroll)
hist(college$Top10perc)
hist(college$perc.alumni, col=2)
hist(college$S.F.Ratio, col=3, breaks=15)
hist(college$Expend, col=4, breaks=50)
```

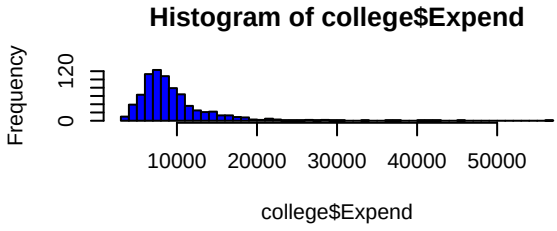








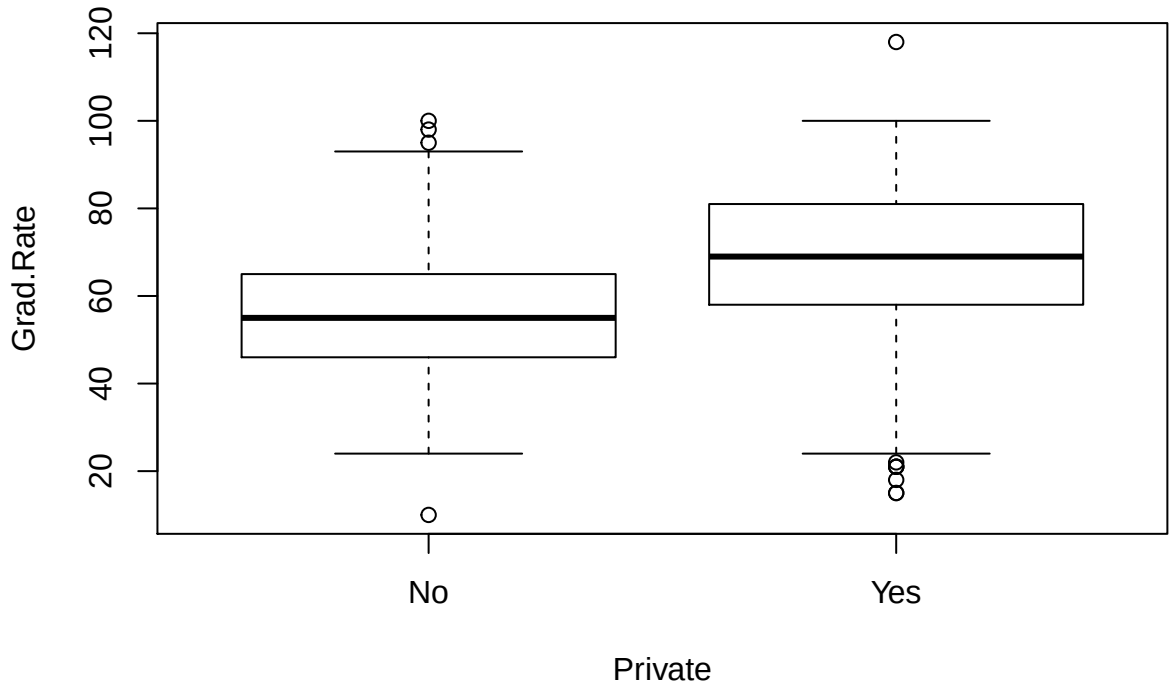




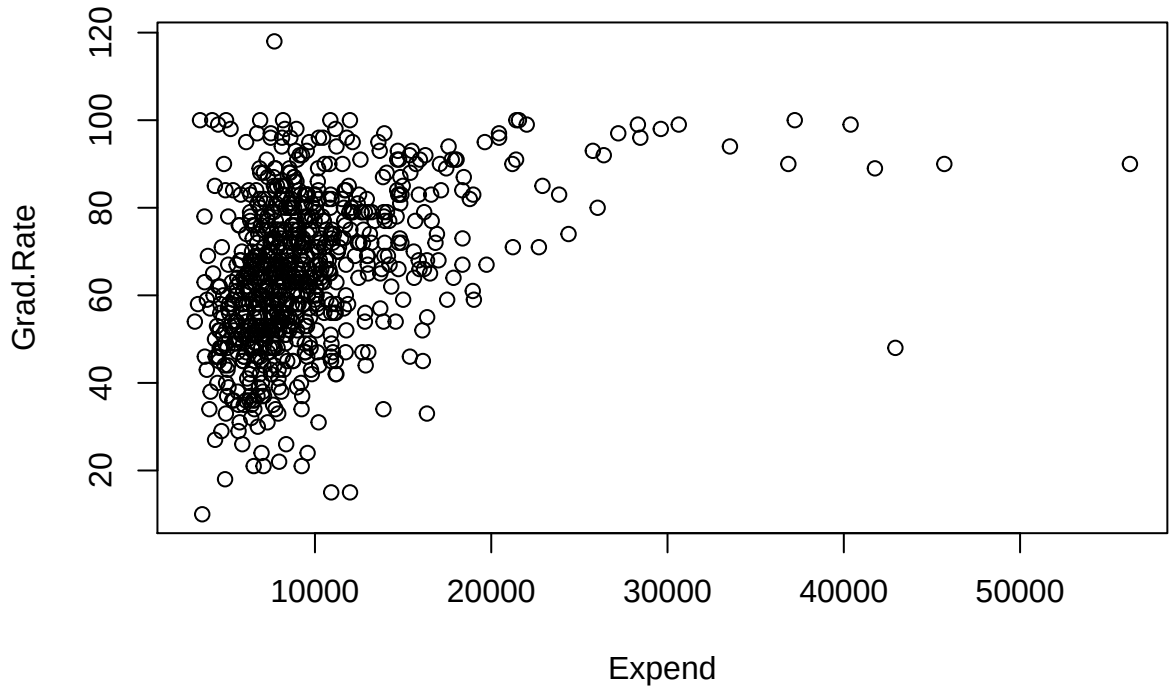

```
# Many variables have skewed distributions
```

vi.

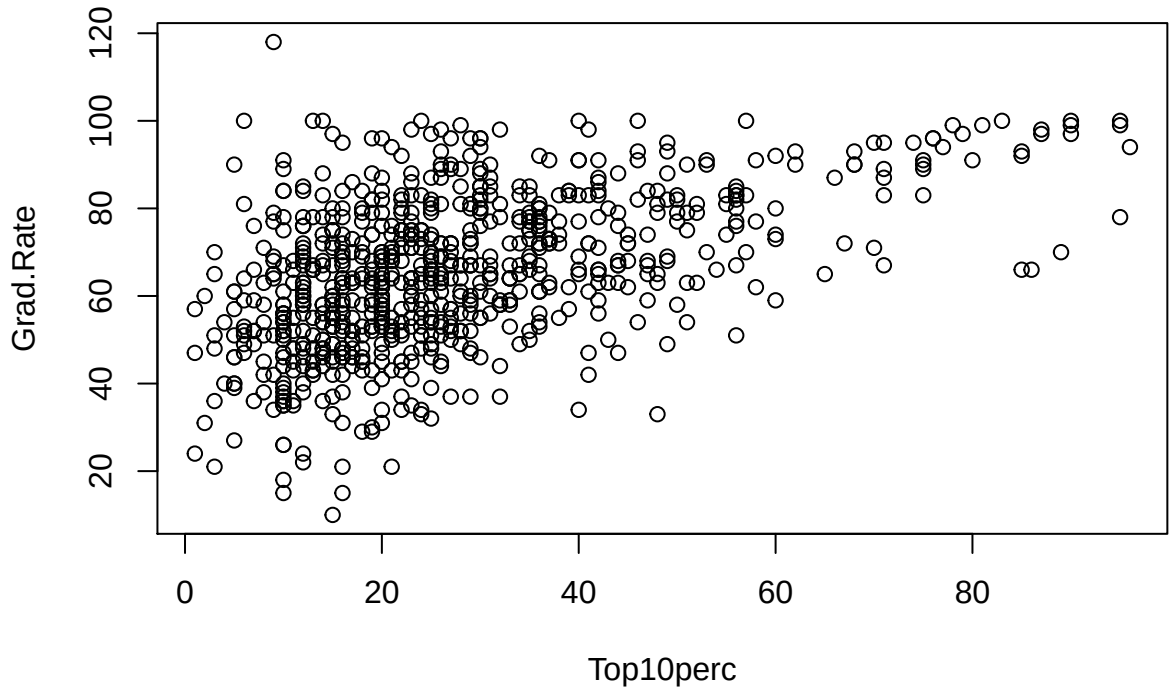
```
par(mfrow=c(1,1))
# Private schools seem to have higher graduation rate.
plot(Grad.Rate~Private, data=college)
```



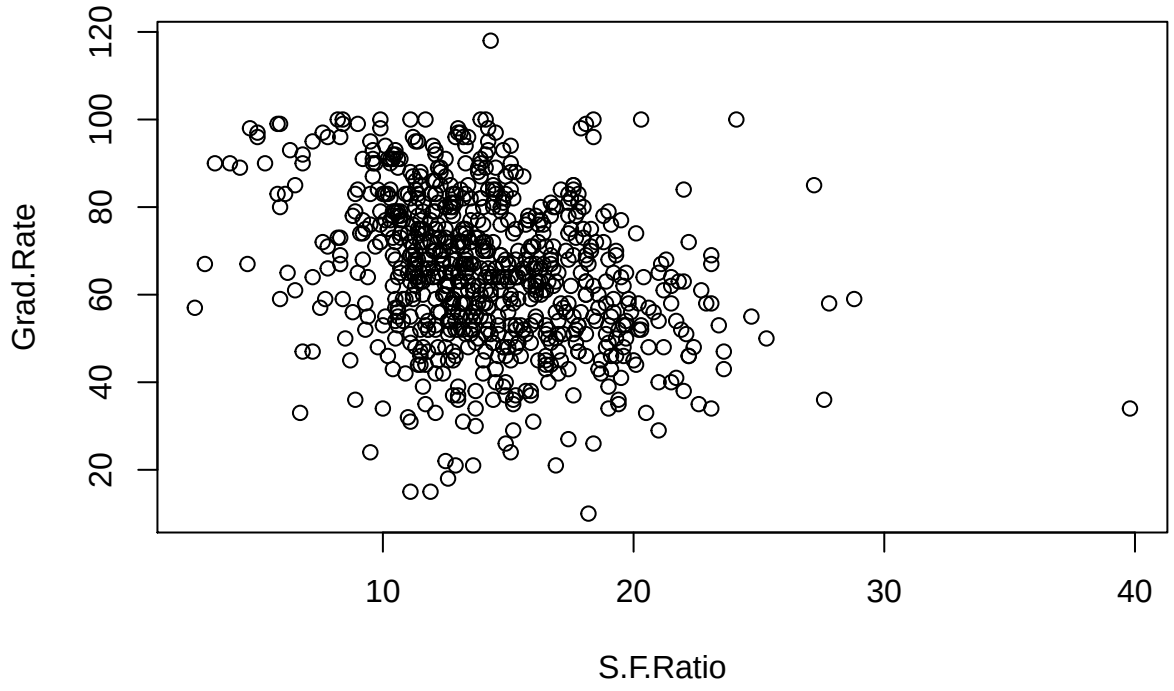

```
# High Expend is associated with high graduation rate however the graduation rate appears to ne non-lin
plot(Grad.Rate~Expend, data=college)
```



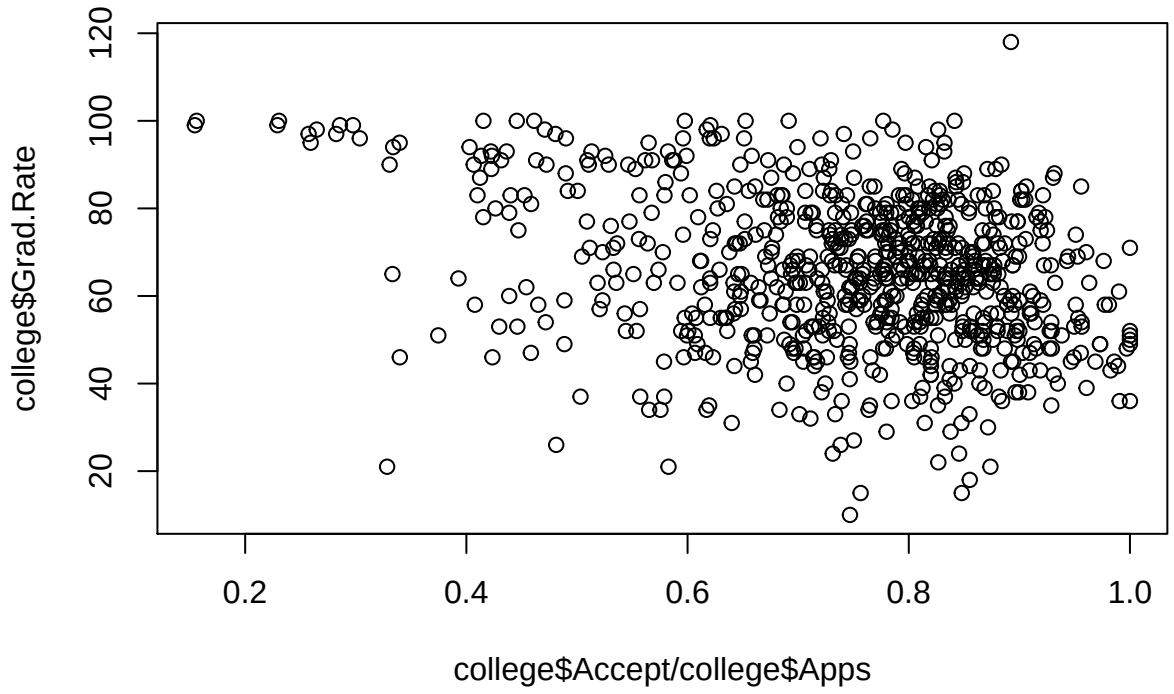

```
# Colleges with the most students from top 10% seem to have higher Grad.rate
plot(Grad.Rate~Top10perc, data=college)
```




```
# Colleges with low Student to Faculty ratio have better graduation rate
plot(Grad.Rate~S.F.Ratio, data=college)
```




```
# Colleges with low acceptance rate tend to have high Grad.Rate.
plot(college$Accept / college$Apps, college$Grad.Rate)
```



2. JWHT problem 9, p. 56

```
setwd("/Users/ovitek/Dropbox/Olga/Teaching/CS6220/Fall15/Homeworks/Hw1")
Auto = read.csv("Auto.csv", header=T, na.strings="?")
Auto = na.omit(Auto)
dim(Auto)
```

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

```
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
```

(a) quantitative: mpg, displacement, horsepower, weight, acceleration, year

qualitative: name, origin,cylinders

(b)

```
quant = c(1, 3:7)
sapply(Auto[quant], range)
```

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

(c)

```
sapply(Auto[quant], mean)
```

```
##          mpg displacement   horsepower       weight acceleration
##     23.44592    194.41199    104.46939   2977.58418     15.54133
##         year
##     75.97959
```

```
sapply(Auto[quant], sd)
```

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

(d)

```
Auto2 = Auto[-(10:85),]
sapply(Auto2[quant], range)
```

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

```
sapply(Auto2[quant], mean)
```

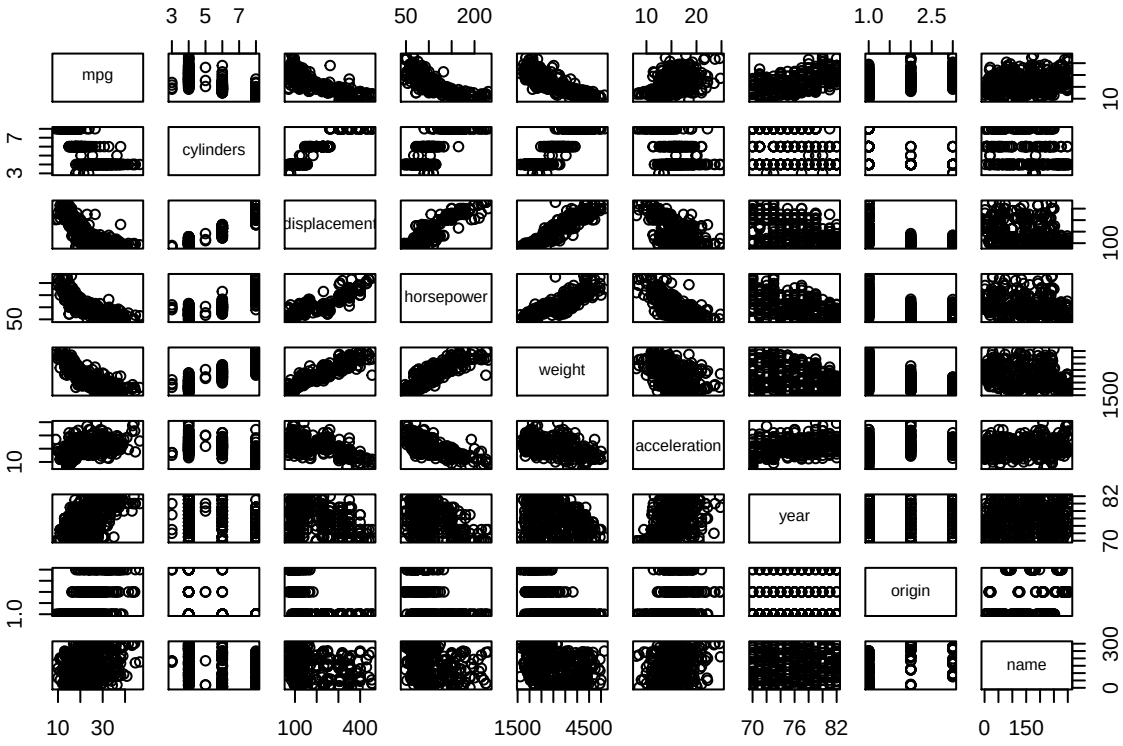
```
##          mpg displacement   horsepower       weight acceleration
##     24.40443    187.24051    100.72152   2935.97152     15.72690
##         year
##     77.14557
```

```
sapply(Auto2[quant], sd)
```

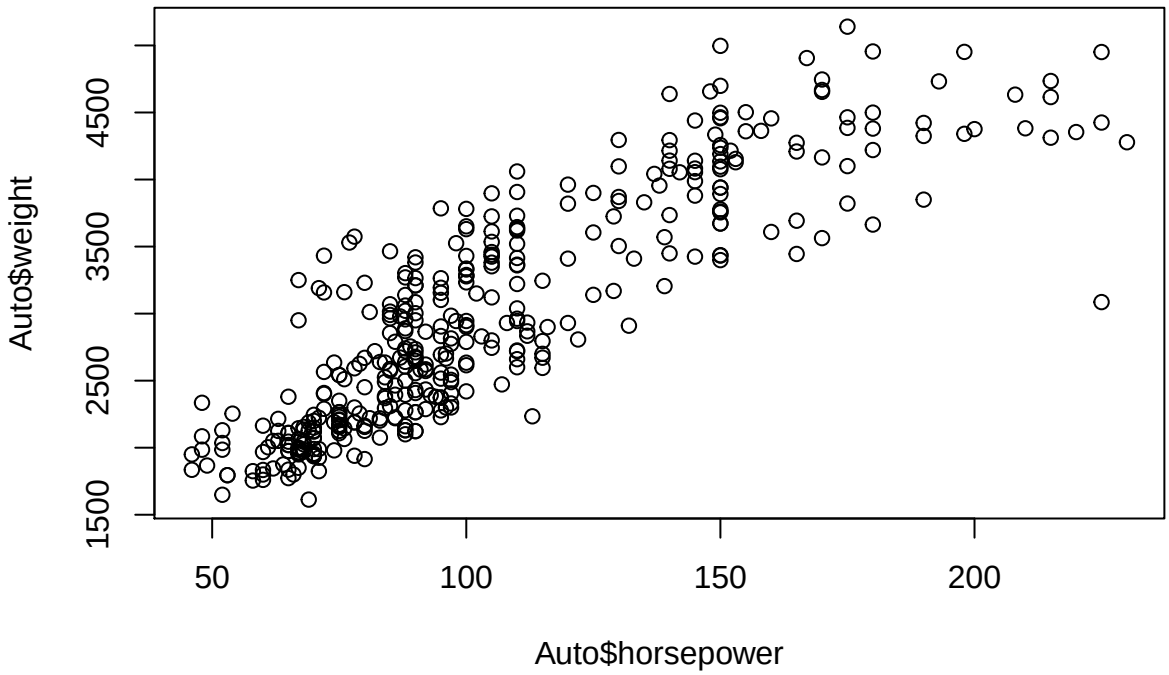
```
##          mpg displacement   horsepower       weight acceleration
##     7.867283    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

(e)

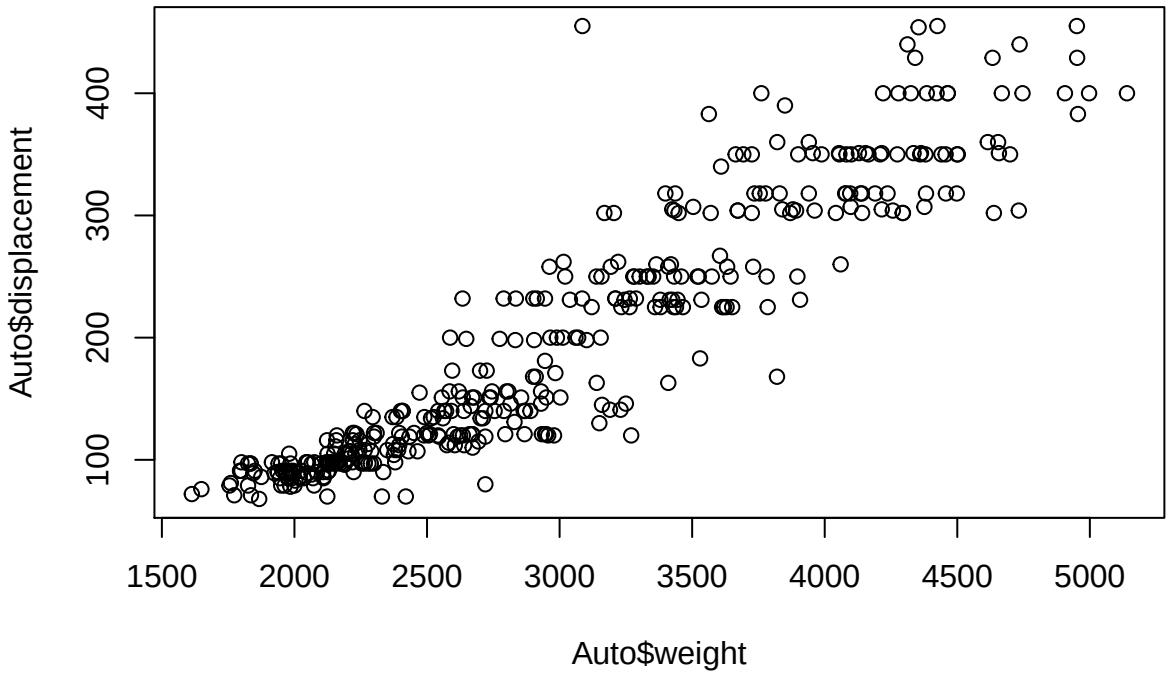
```
#pairs show us some possible correlations between predictors
pairs(Auto)
```



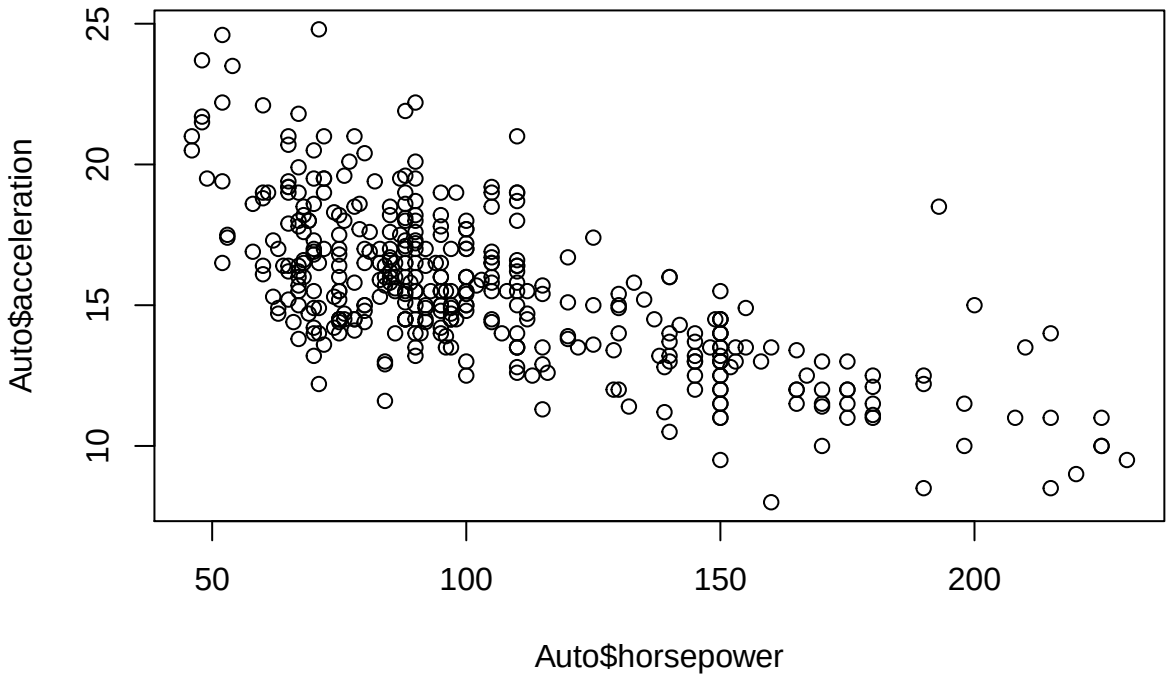

```
#weight and horesepower have positive correlation
plot(Auto$weight ~ Auto$horsepower)
```




```
#positive correlation between displacement and weight
plot(Auto$displacement ~ Auto$weight)
```

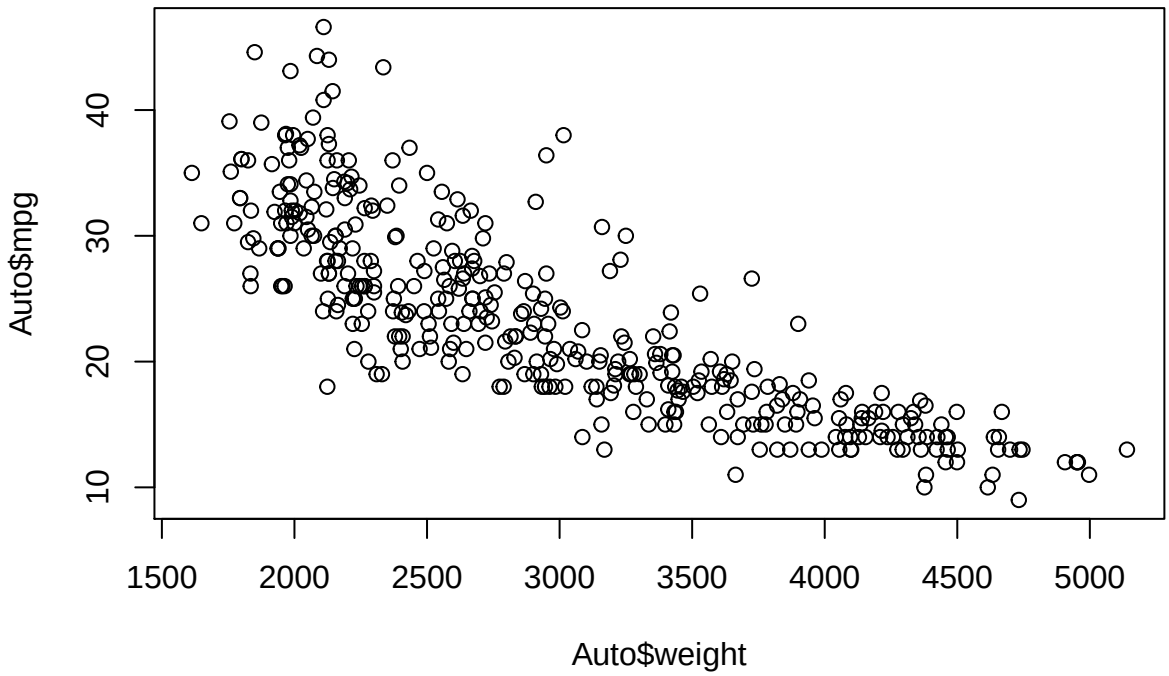



```
#negative correlation between acceleration and horsepower
plot(Auto$acceleration ~ Auto$horsepower)
```

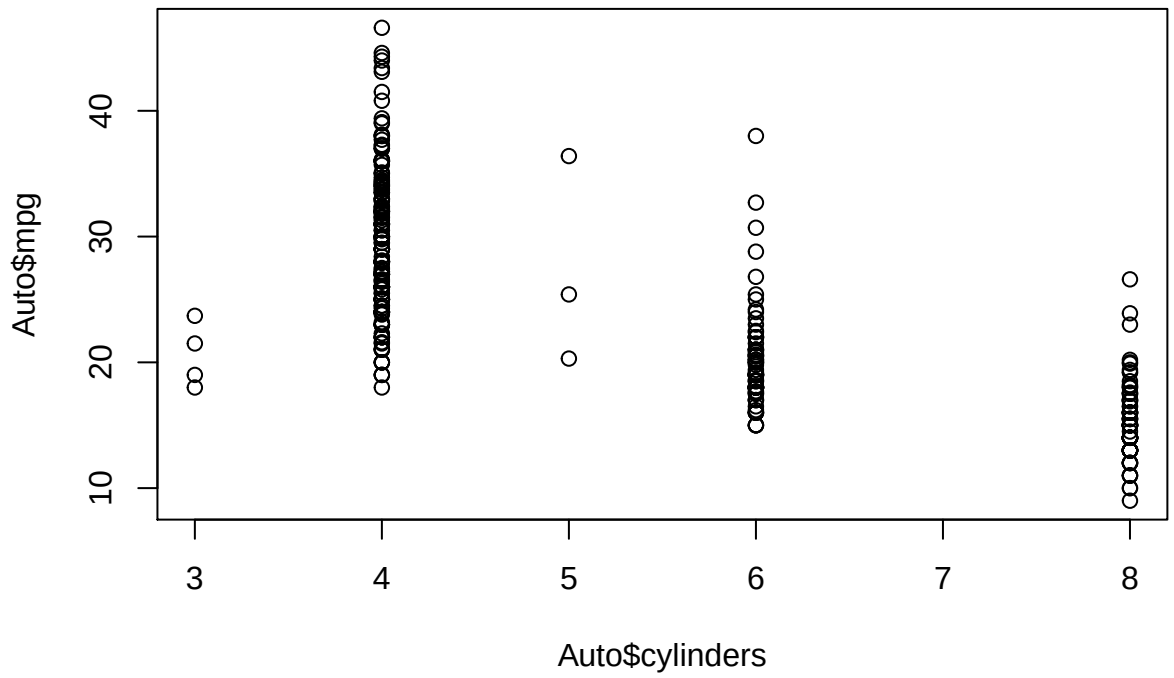


(f)

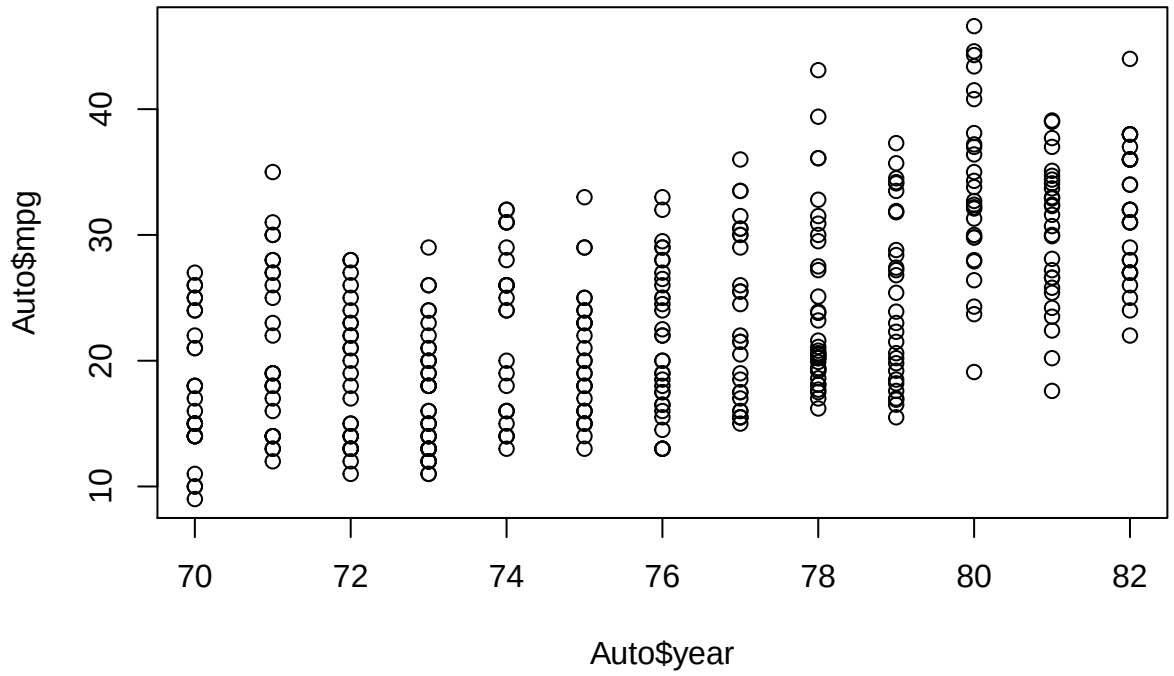
```
# pairs plot shows that Several parameters seem to have correlation with mpg. I will highlight some of
#Since weight, horsepower, displacement and acceleration seem to have correlation with eachother -as sh

# Heavier weight correlates with lower mpg.
plot(Auto$mpg ~ Auto$weight)
```



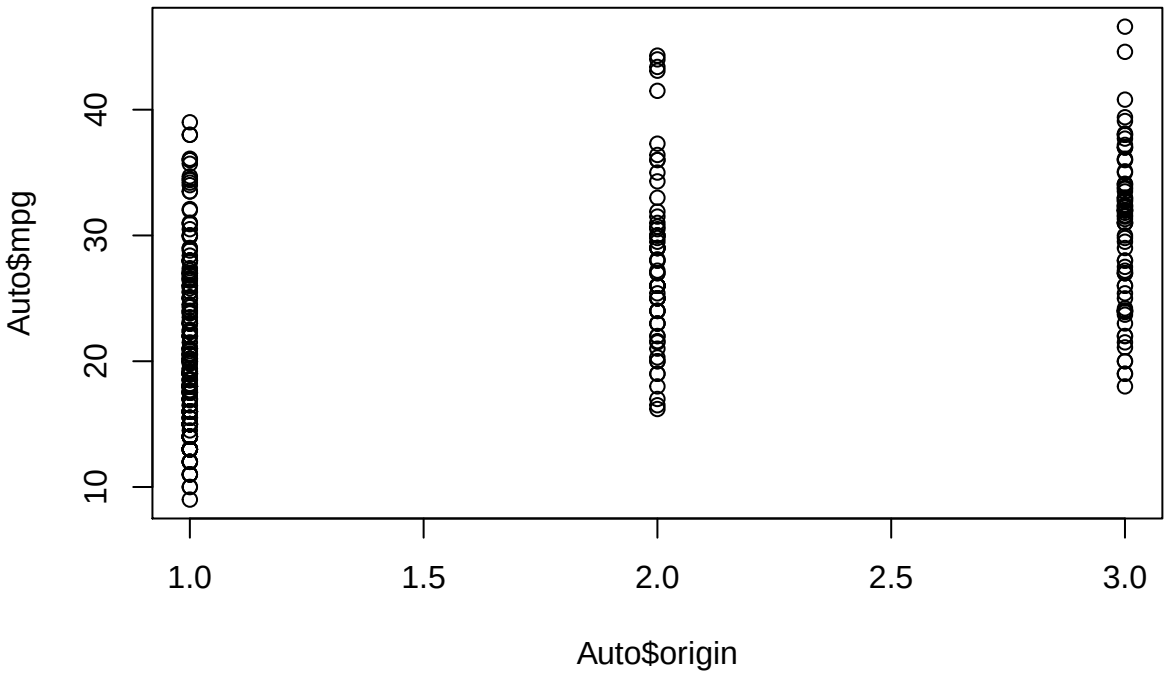
```
# More cylinders correlates with lower mpg.
plot(Auto$mpg ~ Auto$cylinders)
```




```
# Recent cars has higher mpg.
plot(Auto$mpg ~ Auto$year)
```




```
# Cars with origin 3 are the most efficient.
plot(Auto$mpg ~ Auto$origin)
```


Auto$mpg
10
20
30
40
1.0
1.5
2.0
2.5
3.0
Auto$origin