HW 1

Austin Pesina

2021-02-04

#2

a) This is a regression problem because all the variables have numerical values. This is also an inference problem because we are looking at the relationship between the variables and the CEO's salary. n = top 500 U.S. firms; p = 3
b) classification; prediction; n = 20; p = 3
c) regression; prediction; n = 52 p=3

#5

A flexible approach can give a better fit for non-linear models and reduces the bias. On the other hand, it requires a greater number of parameters, it overfits the model, and there is an increase in variance.

A more flexible approach is preferred when we want a prediction while a less flexible approach is preferred when we want an inference and interpretability.

#6

A parametric approach reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f.

A non-parametric approach does not assume a particular form of f and requires a large sample to accurately estimate f.

The advantages of a parametric approach over a non-parametric one is that it simplifies the modeling of f down to fewer parameters so not as many observations are needed. The disadvantages are that there is a potentially inaccurate estimate of f if the form of f is wrong; it could also overfit the observations if a more flexible model is used.

#8

```
library(ISLR)
data(College)
college <- read.csv("~/Data_Mining/Data/College.csv", header=T)

head(college[, 1:5])
```

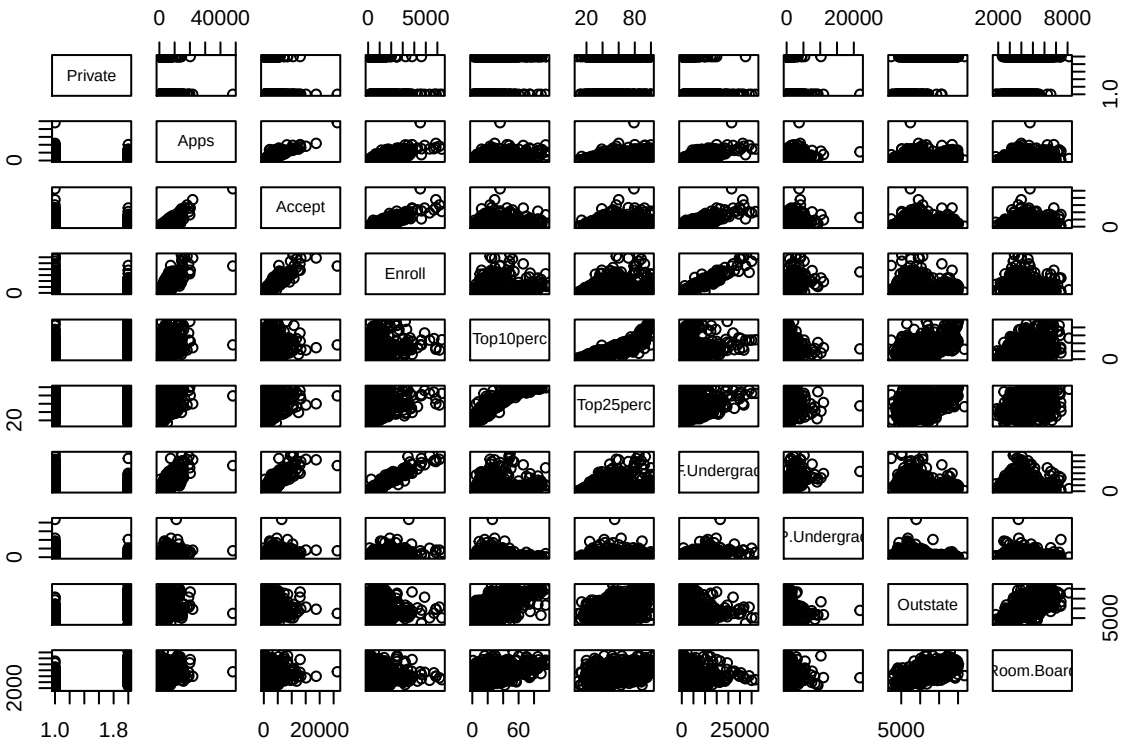
```
##                              X Private Apps Accept Enroll
## 1 Abilene Christian University     Yes 1660   1232    721
```

```
## 2           Adelphi University     Yes 2186   1924    512
## 3              Adrian College     Yes 1428   1097    336
## 4          Agnes Scott College     Yes  417    349    137
## 5    Alaska Pacific University     Yes  193    146     55
## 6            Albertson College     Yes  587    479    158
```

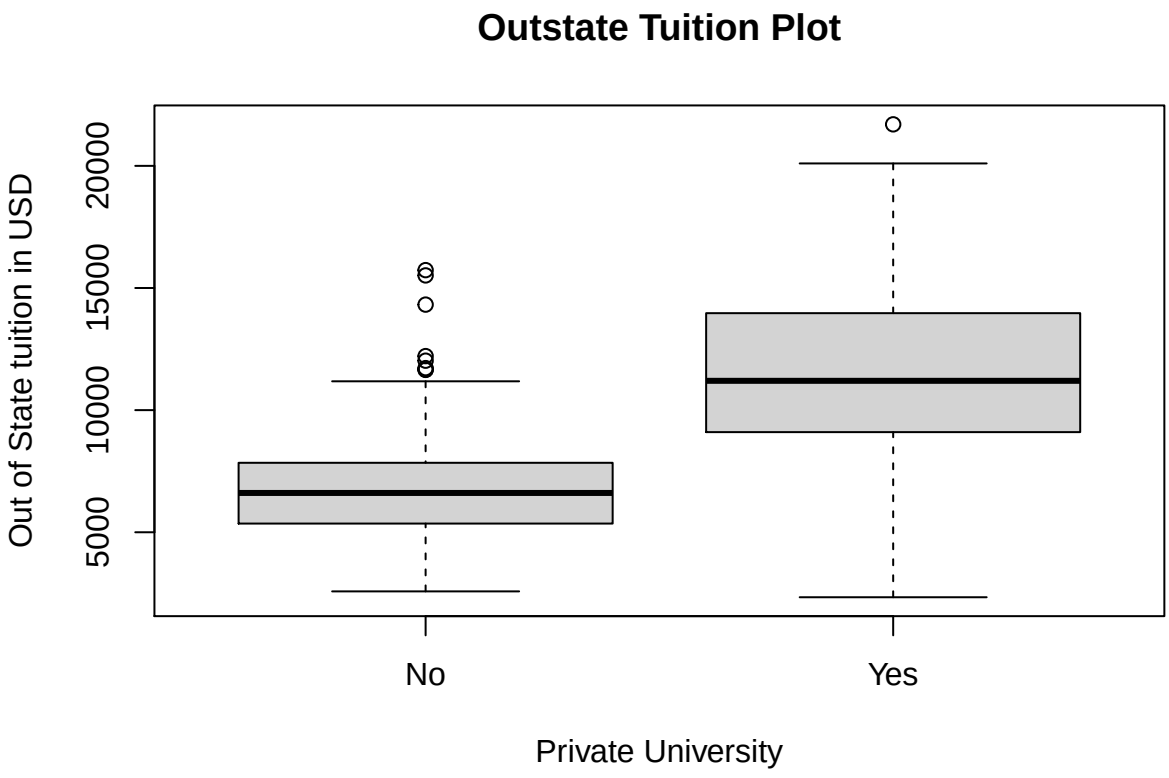
```
rownames <- college[,1]
fix(college)
college <- college[,-1]
fix(college)

college$Private<-as.factor(college$Private)

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
```

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



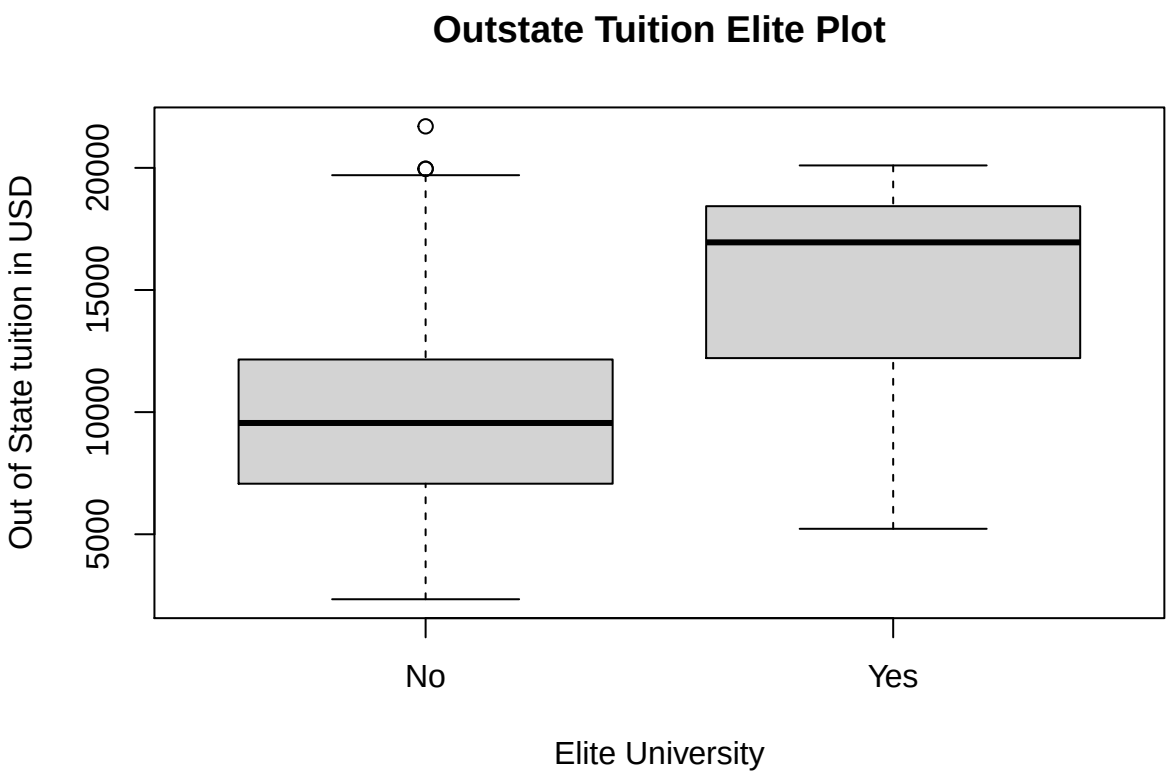

```
plot(college$Private, college$Outstate, xlab = "Private University",
     ylab ="Out of State tuition in USD", main = "Outstate Tuition Plot")
```



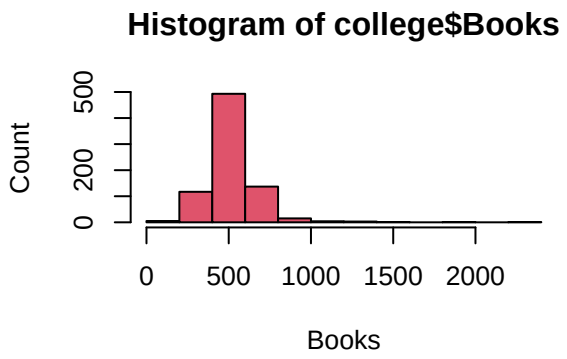

```
Elite <-rep("No", nrow(college))
Elite [college$Top10perc >50] = "Yes"
Elite <-as.factor(Elite)
college <- data.frame(college, Elite)
fix(college)
summary(college$Elite)
```

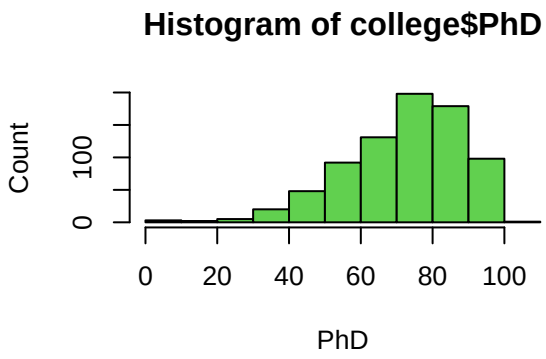
```
##  No Yes
## 699  78
```

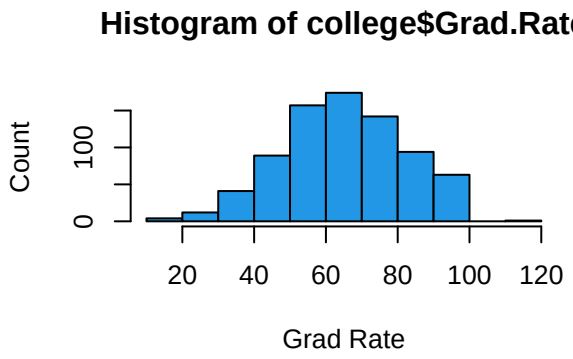
```
plot(college$Elite, college$Outstate, xlab = "Elite University",
     ylab ="Out of State tuition in USD", main = "Outstate Tuition Elite Plot")
```

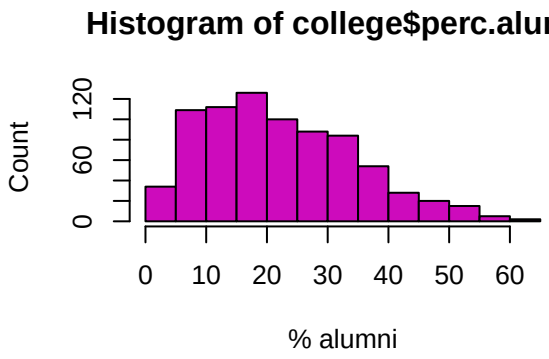



```
par(mfrow = c(2,2))
hist(college$Books, col = 2, xlab = "Books", ylab = "Count")
hist(college$PhD, col = 3, xlab = "PhD", ylab = "Count")
hist(college$Grad.Rate, col = 4, xlab = "Grad Rate", ylab = "Count")
hist(college$perc.alumni, col = 6, xlab = "% alumni", ylab = "Count")
```










```
summary(college$PhD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

#9

```
library(ISLR)
data(Auto)
auto <-read.csv("~/Data_Mining/Data/Auto.csv", header=T, na.strings="?")
fix(auto)

str(auto)
```

```
## 'data.frame':    397 obs. of  9 variables:
##  $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
##  $ cylinders   : num  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        : num  70 70 70 70 70 70 70 70 70 70 ...
##  $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
##  $ name        : chr  "chevrolet chevelle malibu" "buick skylark 320" "plymouth satellite" "amc rebel
```

```
quant <- sapply(Auto, is.numeric)
quant
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##         TRUE         TRUE         TRUE         TRUE         TRUE         TRUE
##         year       origin         name
##         TRUE         TRUE        FALSE
```

```
sapply(auto[quant], range)
```

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

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

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.515869     5.458438   193.532746           NA  2970.261965    15.555668
##         year       origin
##    75.994962     1.574307
```

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

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    7.8258039    1.7015770  104.3795833           NA  847.9041195    2.7499953
##         year       origin
##    3.6900049    0.8025495
```

```
sub <-auto[-c(10:85), -c(4,9)]
sapply(sub, range)
```

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

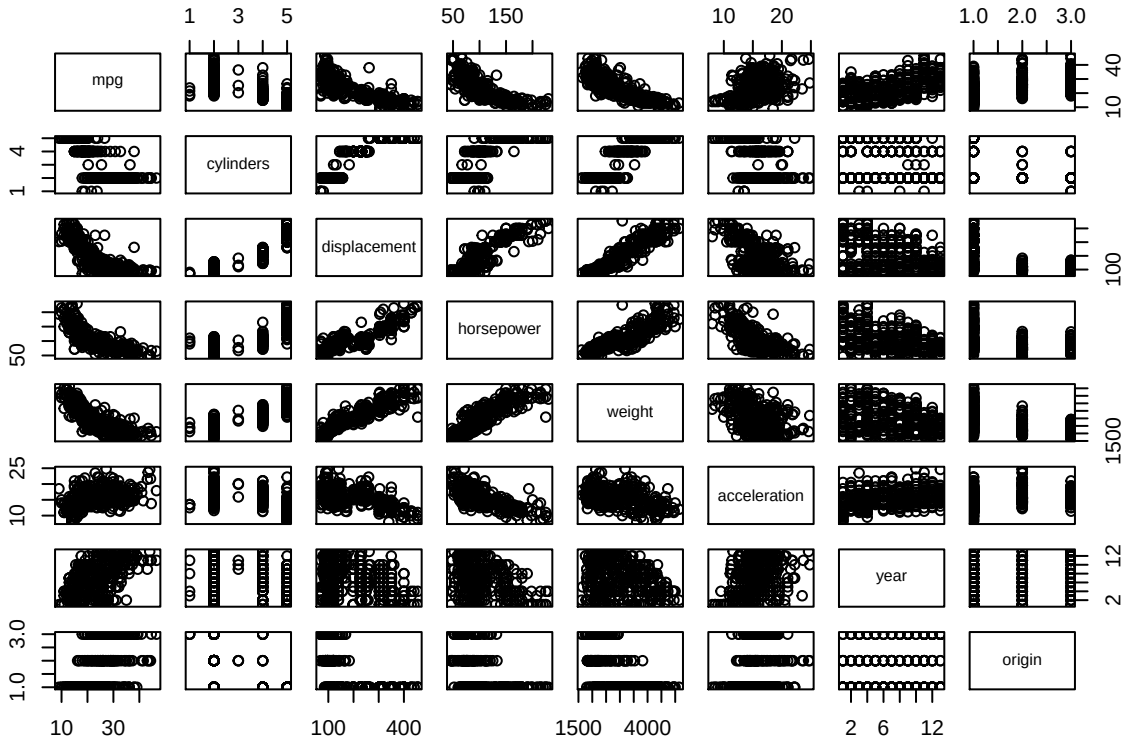
```
sapply(sub, mean)
```

```
##          mpg    cylinders displacement       weight acceleration         year
##    24.438629     5.370717   187.049844  2933.962617    15.723053    77.152648
##       origin
##     1.598131
```

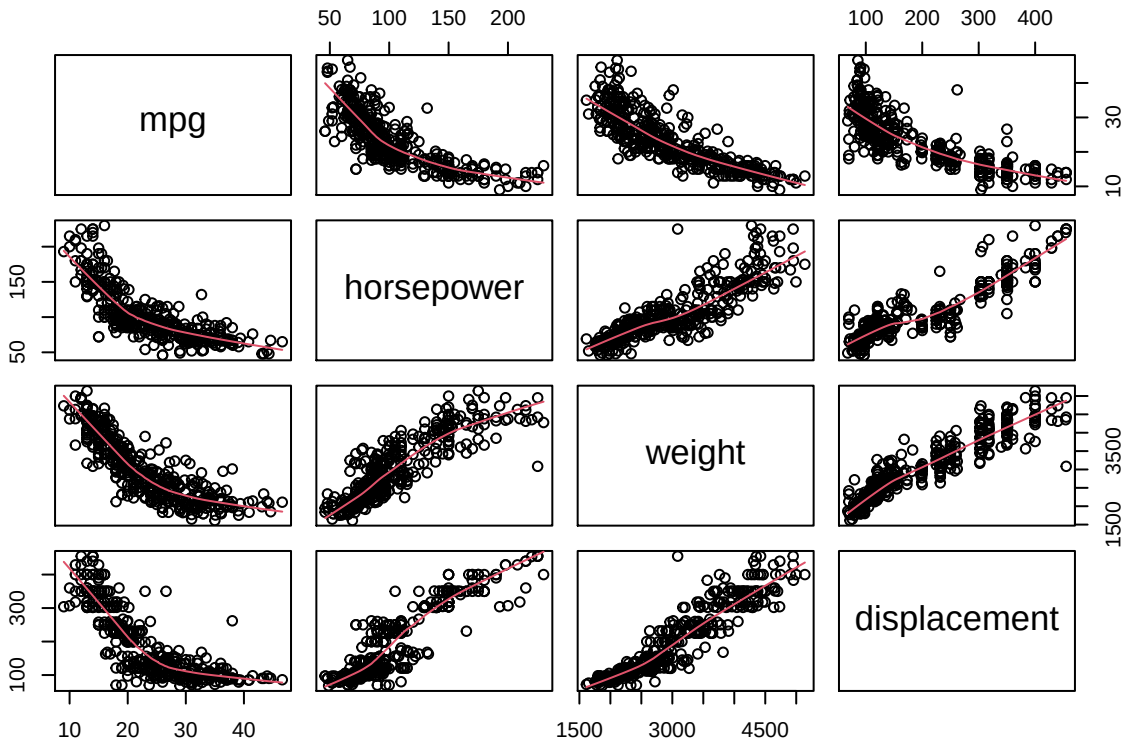
```
sapply(sub, sd)
```

```
##          mpg    cylinders displacement       weight acceleration         year
##    7.9081842    1.6534857   99.6353853  810.6429384    2.6805138    3.1112298
##       origin
##    0.8161627
```

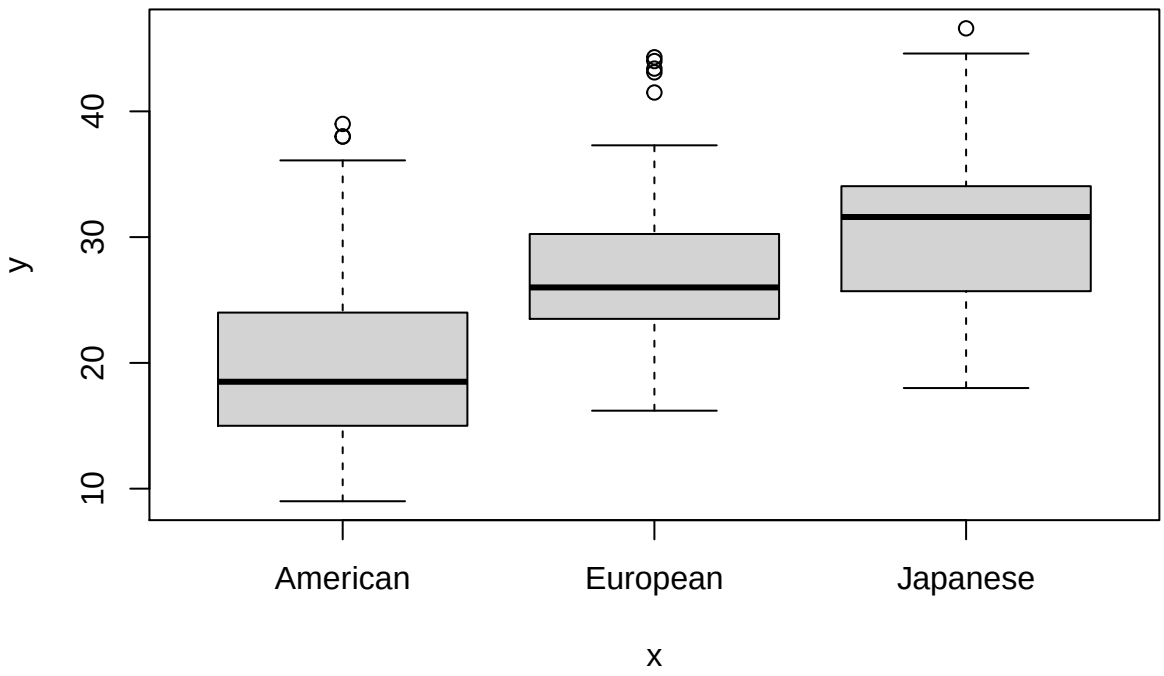
```
auto$cylinders <-as.factor(auto$cylinders)
auto$year <-as.factor(auto$year)
auto$origin <-as.factor(auto$origin)
pairs(auto[-9])
```




```
pairs(~ mpg + horsepower + weight + displacement, data = auto, panel = panel.smooth)
```




```
plot(factor(Auto$origin), Auto$mpg, names=(c("American", "European", "Japanese")))
```



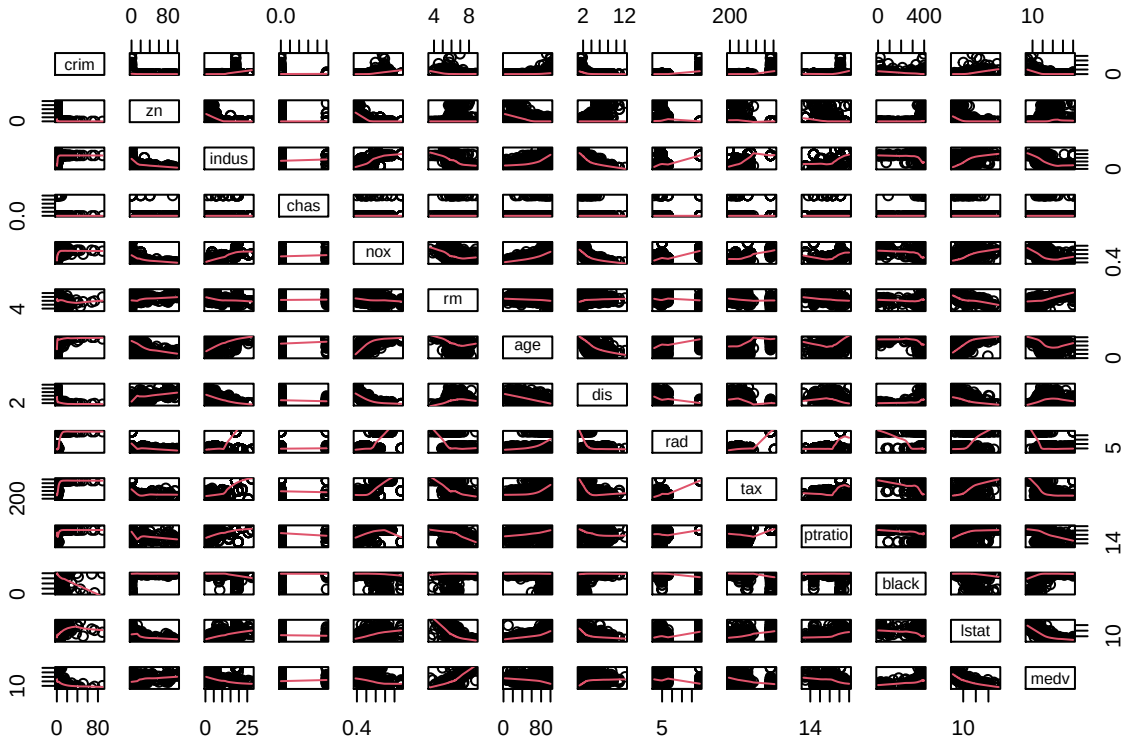
a) All variables except "name" are quantitative. e) We get better mpg on a 4 cyl vehicle compared to others. There are positive correlations between displacement/horsepower and displacement/weight as well as horsepower/weight. Also, as years have gone by, we see mpg increasing. f) Cylinder, horsepower, displacement, weight origin, and year can be used as predictors.

#10

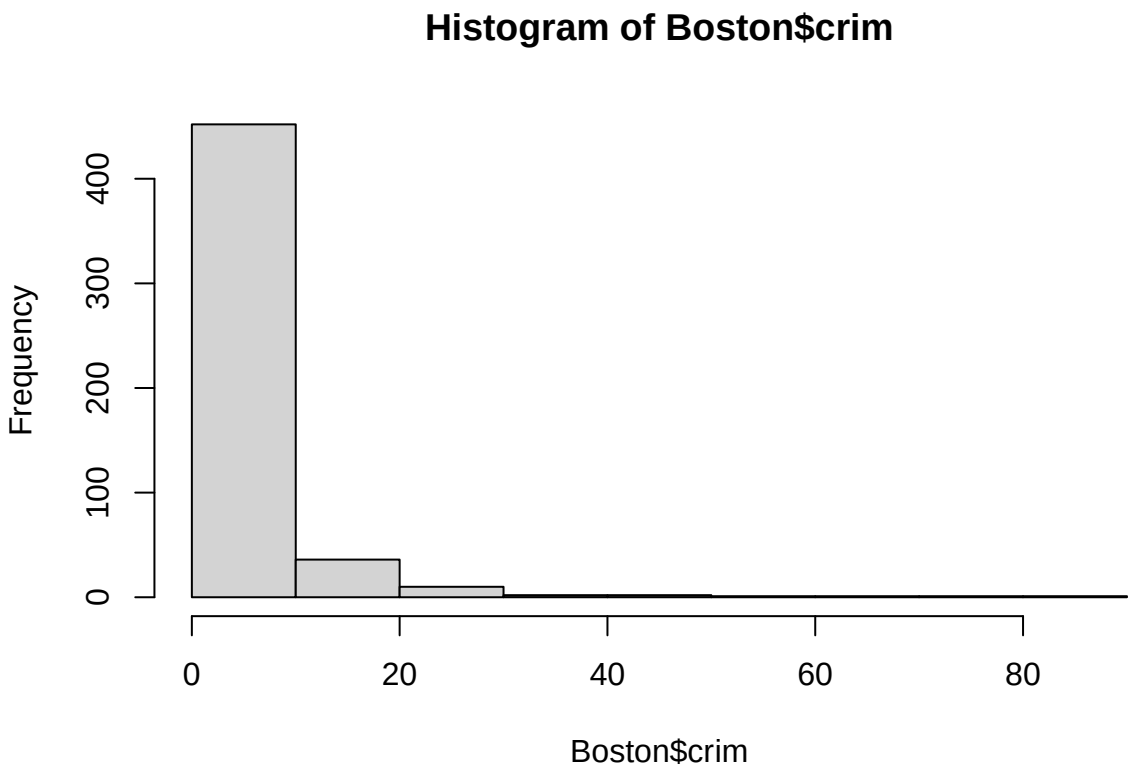
```
library(MASS)
data(Boston)
dim(Boston)
```

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

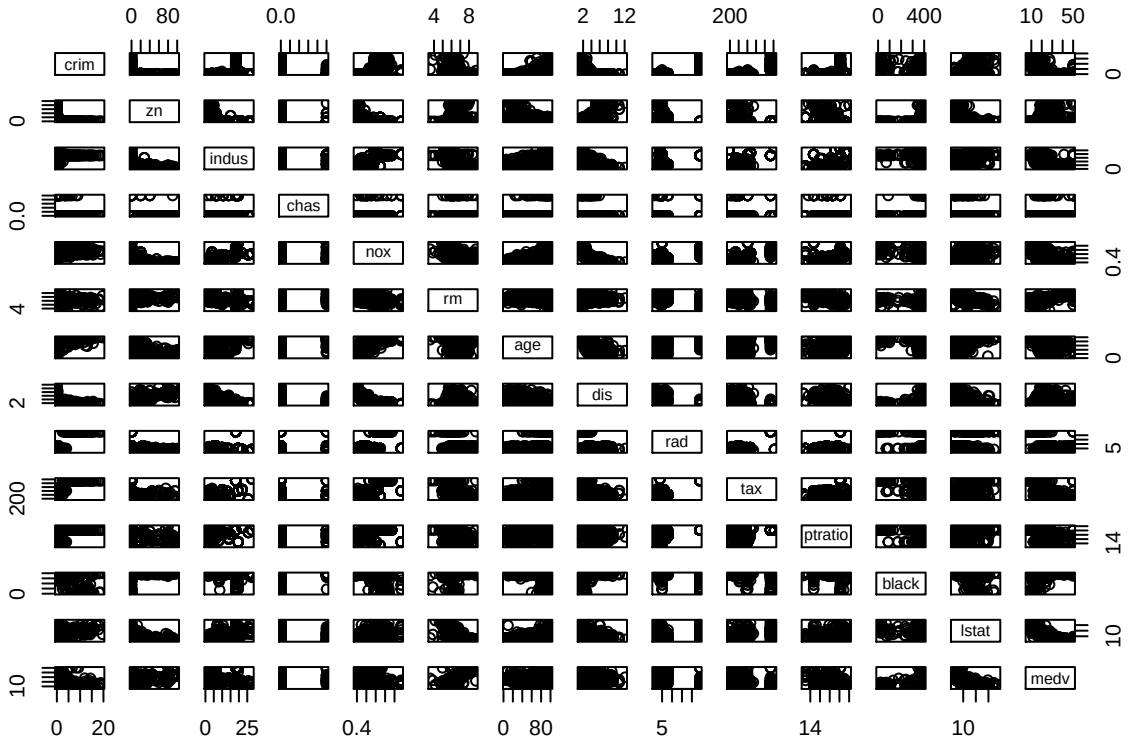
```
pairs(Boston, panel = panel.smooth)
```



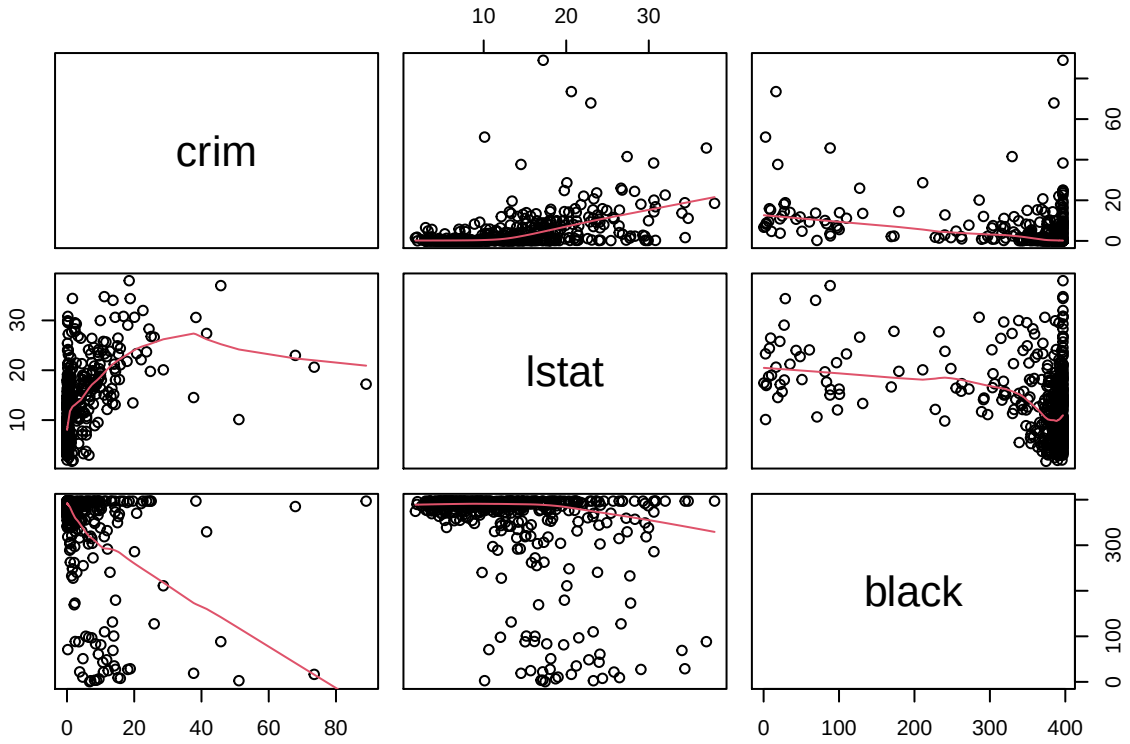

```
hist(Boston$crim)
```



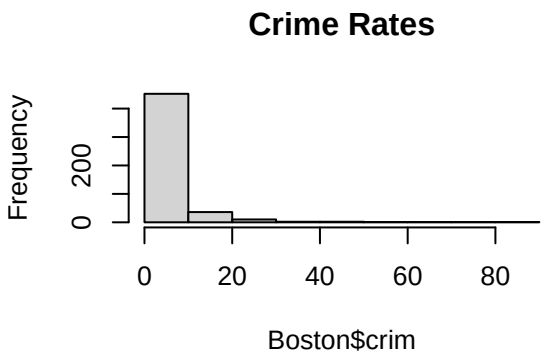

```
pairs(Boston[Boston$crim <20,])
```

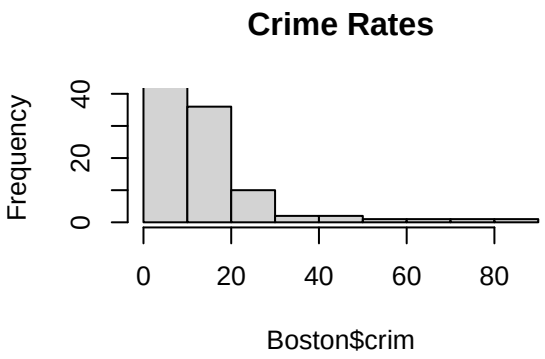


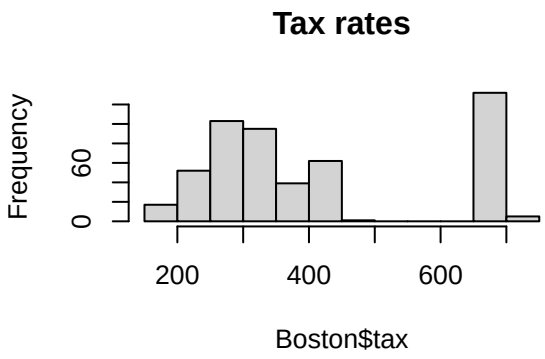

```
pairs(~ crim + lstat + black, data = Boston, panel = panel.smooth)
```

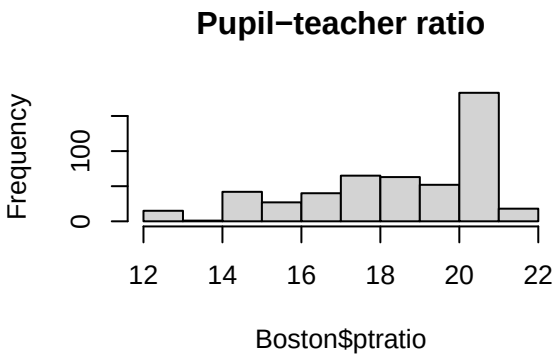



```
par(mfrow=c(2,2))
hist(Boston$crim, main="Crime Rates")
hist(Boston$crim, main="Crime Rates",  ylim=c(0, 40))
hist(Boston$tax, main="Tax rates")
hist(Boston$ptratio, main="Pupil-teacher ratio")
```










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

```
##    Mode   FALSE    TRUE
## logical     471      35
```

```
median(Boston$ptratio)
```

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

```
which.min(Boston$medv)
```

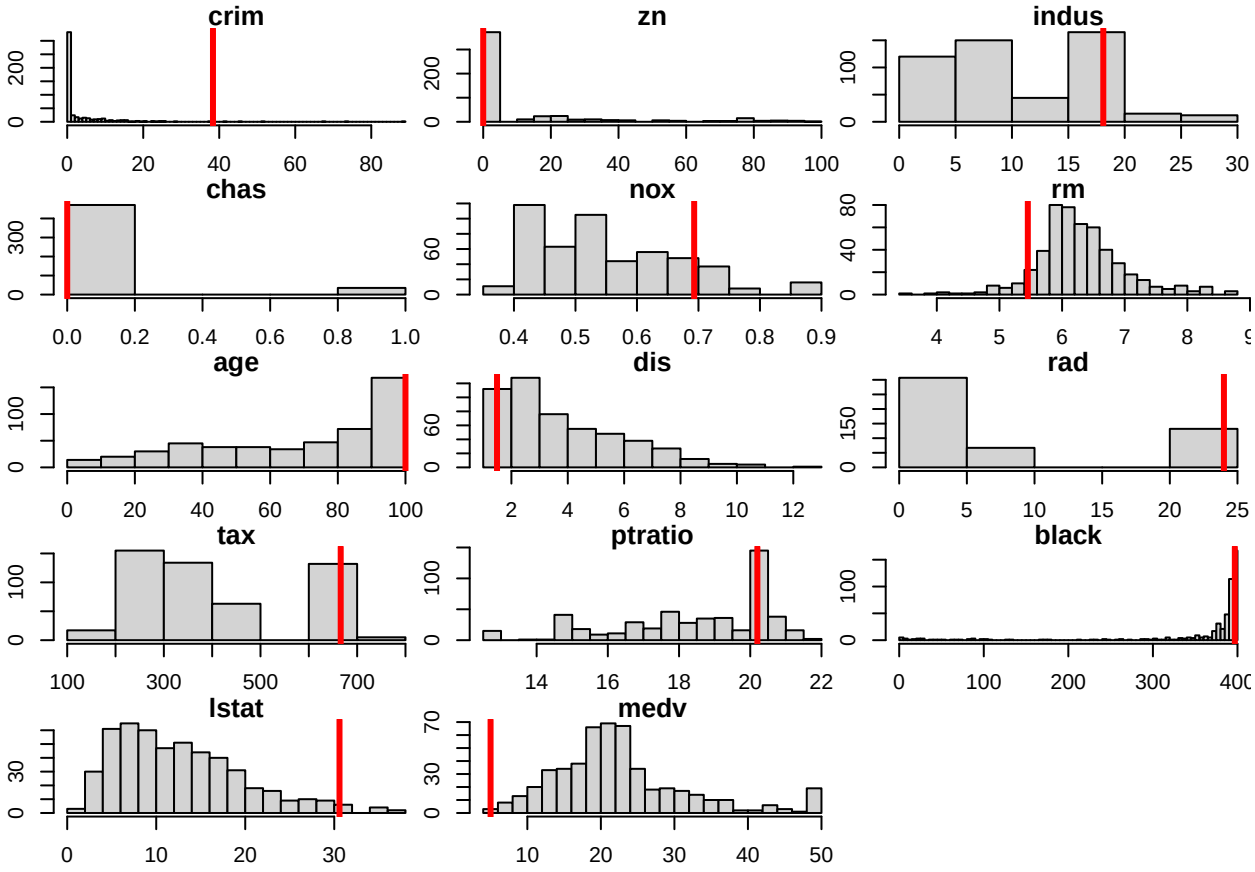
```
## [1] 399
```

```
par(mfrow=c(5,3), mar=c(2, 2, 1, 0))
for (i in 1:ncol(Boston)){
  hist(Boston[, i], main=colnames(Boston)[i], breaks="FD")
  abline(v=Boston[399, i], col="red", lw=3)
}

summary(Boston$rm >7)
```

```
##    Mode   FALSE    TRUE
## logical     442      64
```

```
summary(Boston$rm >8)
```

```
##    Mode   FALSE    TRUE
## logical     493      13
```



a) Boston has 506 rows and 14 columns. c) There may be a relationship between crime and (nox, rm, age, dis, lstat, and medv). e) 35 subarbs bound the Charles River. g) #399 has the lowest median value. h) 64 have more than 7 rooms. 13 have more than 8