STAT 641 HW 1

Jack Cunningham (jgavc@tamu.edu)

2/6/24

Homework 1

```r
require(tidyverse)
```

```
Loading required package: tidyverse
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.2     v readr     2.1.4
v forcats   1.0.0     v stringr   1.5.0
v ggplot2   3.4.2     v tibble    3.2.1
v lubridate 1.9.2     v tidyr     1.3.0
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to becom
```

```r
require(ISLR2)
```

```
Loading required package: ISLR2
```

```
Warning: package 'ISLR2' was built under R version 4.3.2
```

1)

a. Since n is extremely large and p is small we would expect a more flexible approach to be better than an inflexible method. The high amount of observations help alleviate the risk of following the random error too closely. The small amount of predictors allows to avoid the curse of dimensionality, it is easier to gather nearby points that share properties in order to inform future predictions.

b. We would expect the performance of a flexible method to do worse than an inflexible method in this case. With a small amount of observations combined with the fact we have many predictors we are likely to see over fitting become a problem in the flexible method.

c. A flexible method will do better than an inflexible method in this case, it is more capable of capturing the true shape of f. An inflexible method would introduce a significant amount of bias.

d. A flexible method will suffer and do worse than an inflexible method in this case. The more flexible a method gets the more it tries to fit the data, in an application where the variance of error is high a flexible method will be picking up a lot of noise from the data.

2)

a. This is a regression problem, our response variable is CEO salary which is a quantitative value. We state we are interested in what factors impact salary thus this is an inference problem. n = 500 and p = 3.

b. This is a classification problem, our response variable is the result of the product which is qualitative with two possibilities, success or failure. We are interested in whether our product will be a success therefore this is a prediction problem. n = 20 and p = 13.

c. This is a regression problem, our response variable is the percentage change in USD/Euro rates which is a quantitative value. This is a prediction problem. n = 52 and p = 3.

3)

a.
```r
college <- read.csv("College.csv")
```

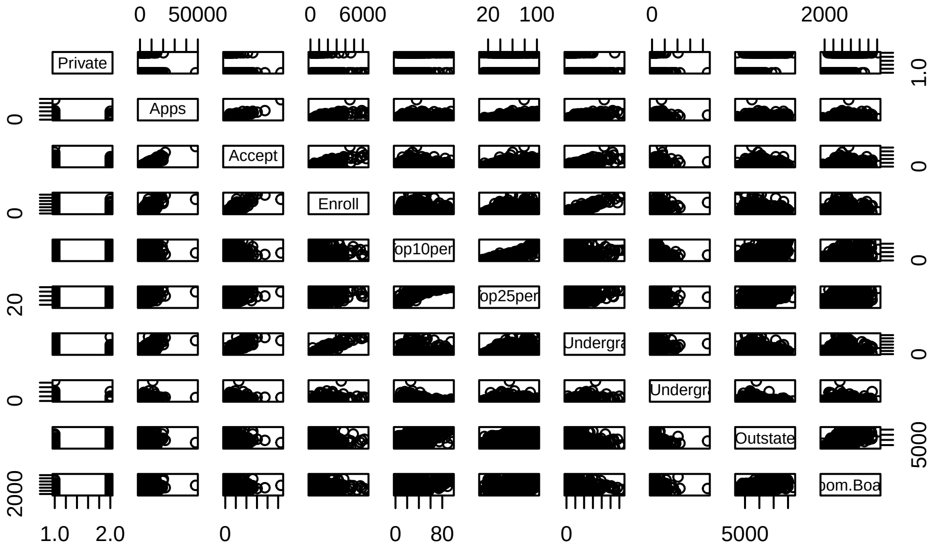
b.
```r
rownames(college) <- college[, 1]
college <- college[, -1]
```

c.
```r
summary(college)
```

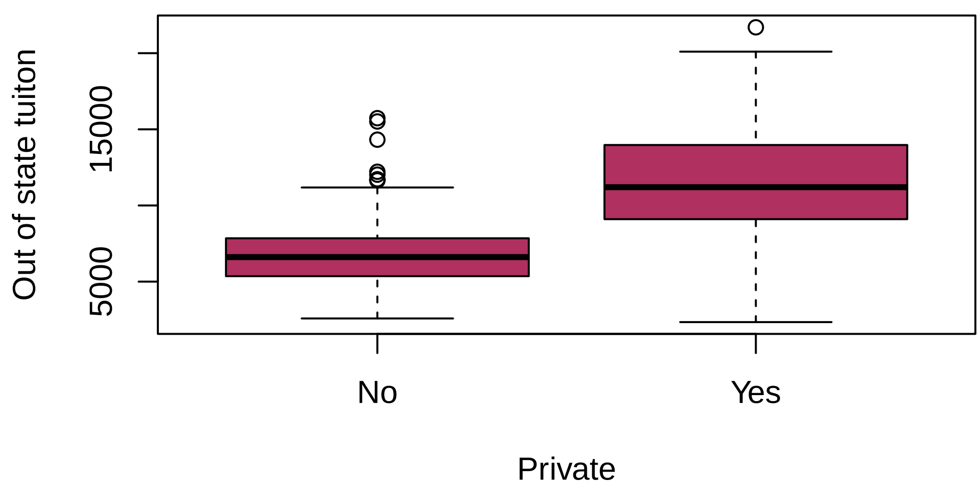
```
   Private                Apps           Accept          Enroll
 Length:777         Min.   :   81   Min.   :   72   Min.   :  35
 Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
 Mode  :character   Median : 1558   Median : 1110   Median : 434
```

```
                      Mean   : 3002   Mean   : 2019   Mean   : 780  
                      3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902  
                      Max.   :48094   Max.   :26330   Max.   :6392  
   Top10perc       Top25perc      F.Undergrad     P.Undergrad    
 Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0  
 1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0  
 Median :23.00   Median : 54.0   Median : 1707   Median :  353.0  
 Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3  
 3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0  
 Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0  
    Outstate       Room.Board       Books           Personal   
 Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250  
 1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850  
 Median : 9990   Median :4200   Median : 500.0   Median :1200  
 Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341  
 3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700  
 Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800  
      PhD            Terminal       S.F.Ratio      perc.alumni   
 Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00  
 1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00  
 Median : 75.00   Median : 82.0   Median :13.60   Median :21.00  
 Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74  
 3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00  
 Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00  
     Expend        Grad.Rate     
 Min.   : 3186   Min.   : 10.00  
 1st Qu.: 6751   1st Qu.: 53.00  
 Median : 8377   Median : 65.00  
 Mean   : 9660   Mean   : 65.46  
 3rd Qu.:10830   3rd Qu.: 78.00  
 Max.   :56233   Max.   :118.00  
```

```r
college$Private <- as.factor(college$Private)
pairs(college[,1:10])
```




```r
plot(college$Private,college$Outstate, col = "maroon", xlab = "Private", ylab = "Out of st
```

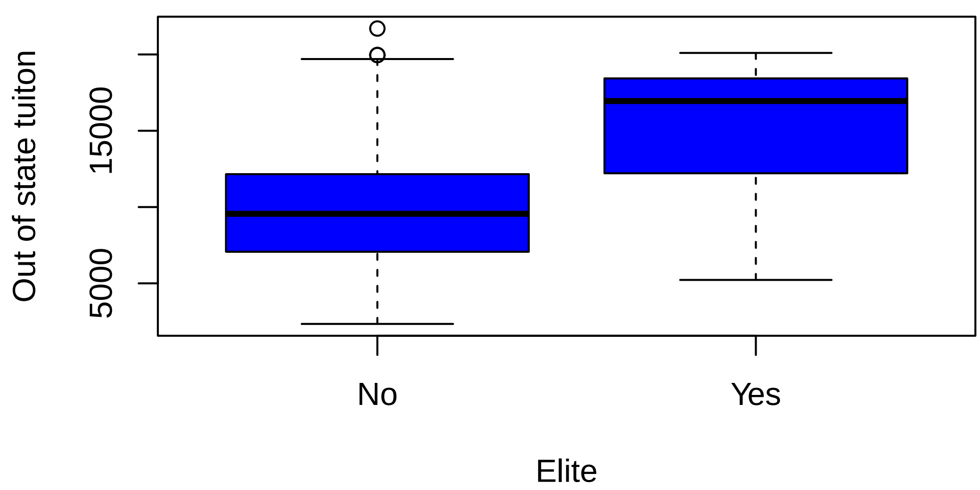


```r
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college, Elite)
```

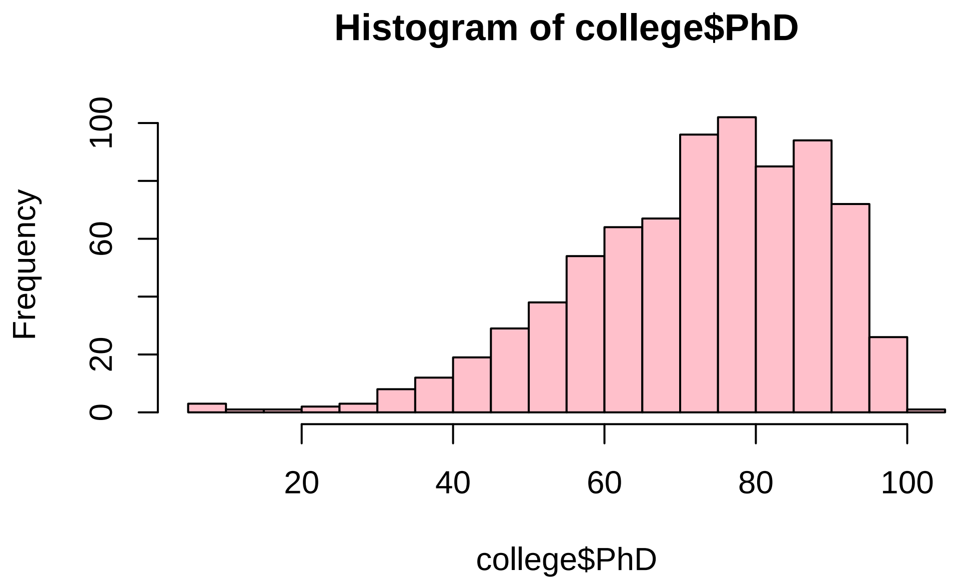
```r
summary(college$Elite)
```

```
 No Yes
699  78
```

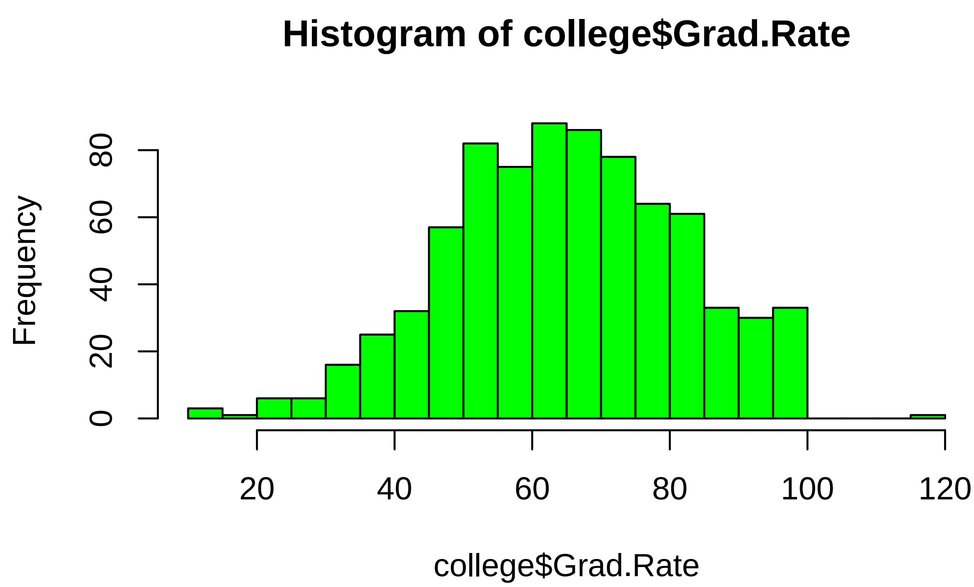
```r
plot(x = college$Elite, y = college$Outstate, xlab = "Elite", ylab = "Out of state tuiton"
```



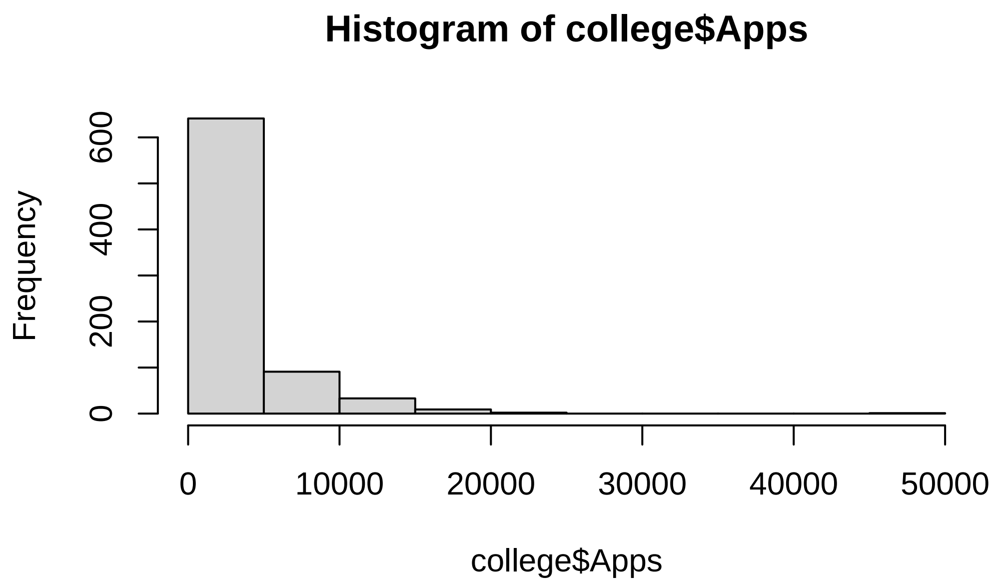

```r
hist(college$PhD, breaks = 15, col = "pink")
```




```r
hist(college$Grad.Rate, breaks = 20, col = "green")
```

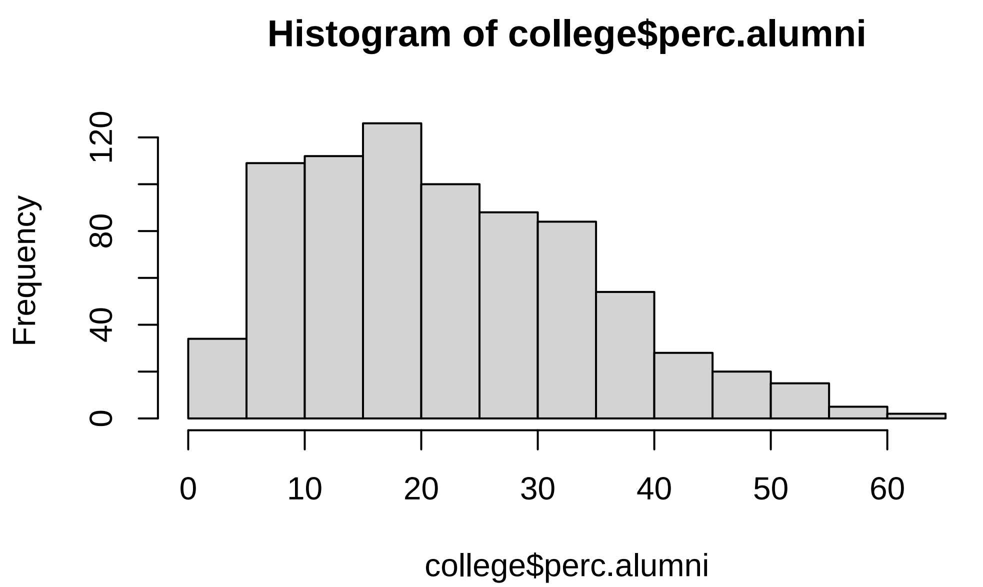


```
hist(college$Apps)
```

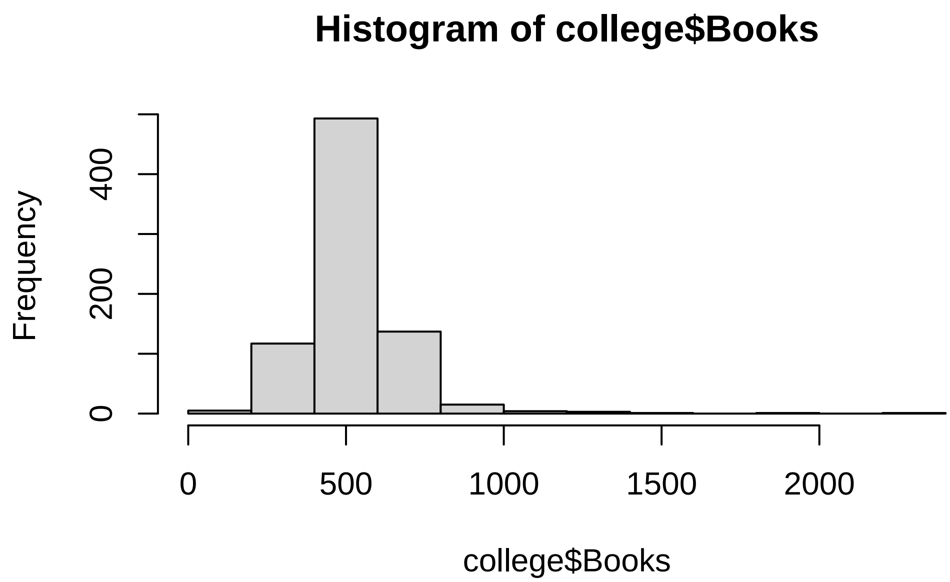



```
hist(college$Expend)
```


```r
hist(college$perc.alumni)
```



```r
hist(college$Books)
```




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

4.

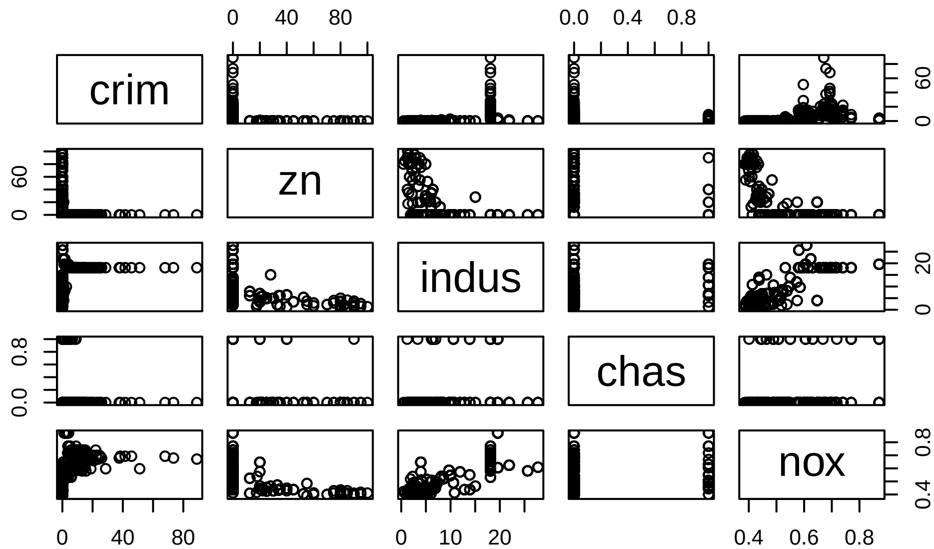
```r
?Boston
```

```
starting httpd help server ... done
```

a) There are 506 rows and 13 columns. Each row is an observation of a Boston Suburb and each column is a measurement of the suburb (i.e. crime rate).

b)

```r
pairs(Boston[,1:5])
```



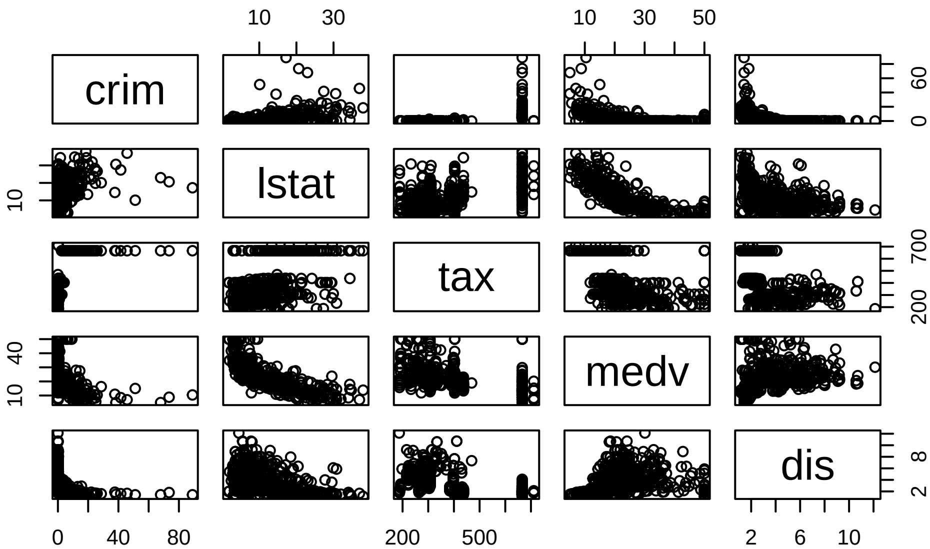
A few findings we can point out:

In areas with low industry have high amounts of residential land are zoned. This makes sense, I'd imagine the majority of industry are in areas close to the city. The further you get the more space there is to set aside large lots of land to live in.

We can also see that in areas with low residential zoning there appears to be high amounts of crime, furthering the idea those areas are closer to the city.

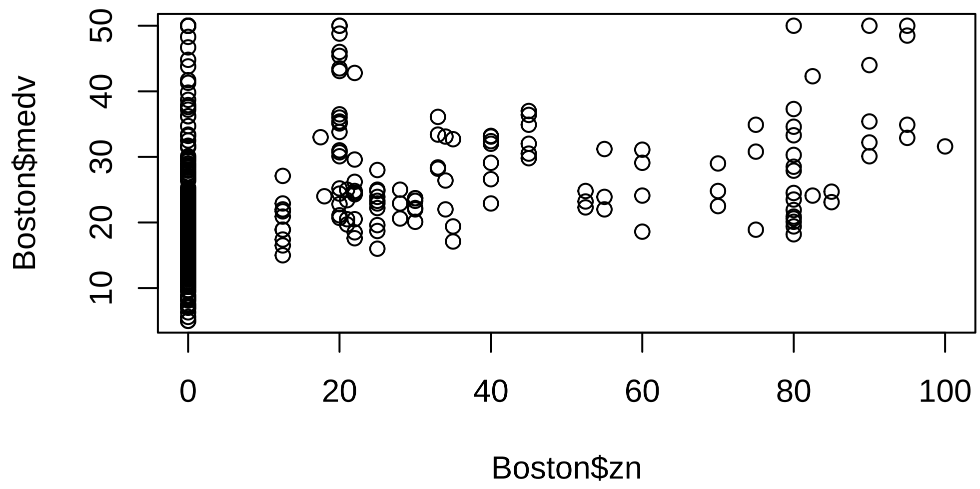
c)

```
pairs(crim ~ lstat + tax + medv + dis, data = Boston)
```



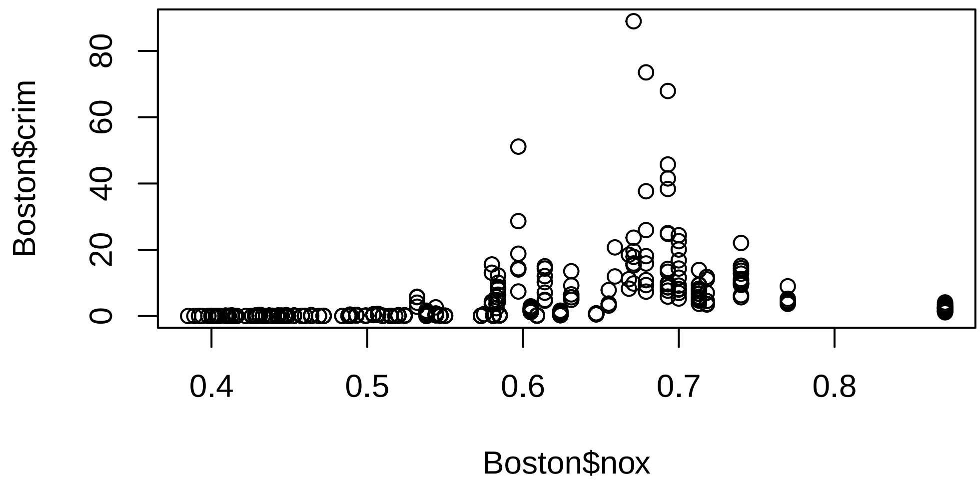
Crime appears to be lower in areas with higher median value homes. More affluent areas seem to not have the high amount of crime that less desirable areas do. But this could be due to theses houses being on larger plots of land and naturally people being further apart.

```
plot(Boston$zn, Boston$medv)
```



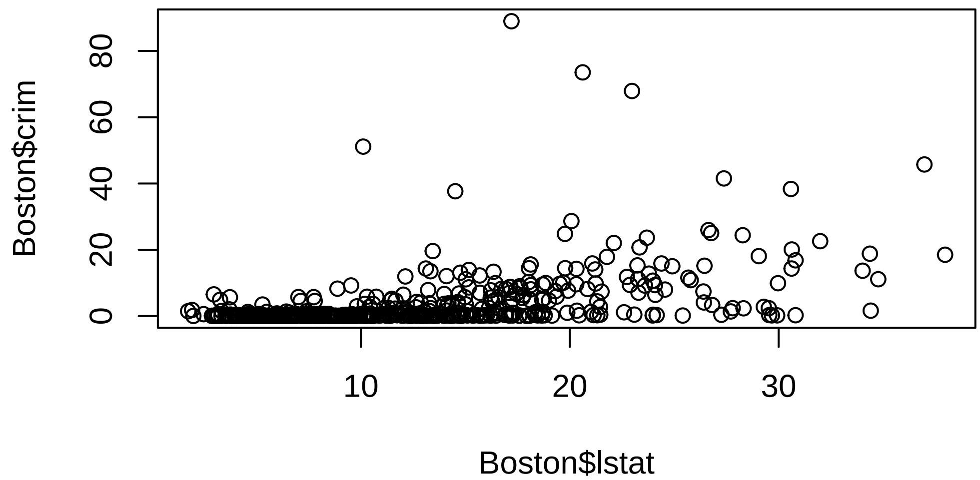
It is hard to tell due to the high density of suburbs having a proportion of 0 for this residential land but there doesn't appear to be a strong relationship between median value of houses in a suburb and proportion of large plots of residential land. So it seems the more affluent the area the less crime there is.

```
plot(Boston$nox, Boston$crim)
```



We can see here that crime appears to be more prevalent in areas with high nitrogen oxide concentration, indicating that crime occurs more in areas where high amounts of people/industry is.

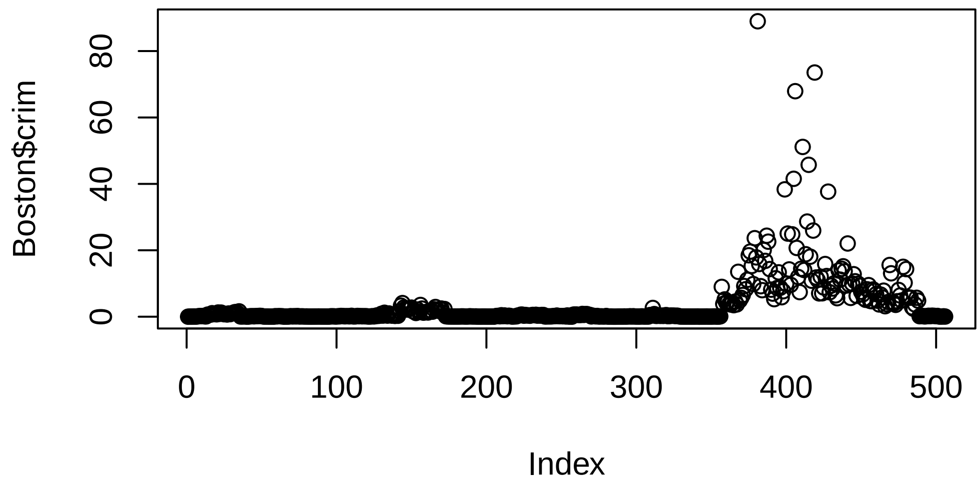
```
plot(Boston$lstat, Boston$crim)
```



In areas with very few low status rates crime is low, in poorer areas crime begins to become a problem.

d)

```
plot(Boston$crim)
```



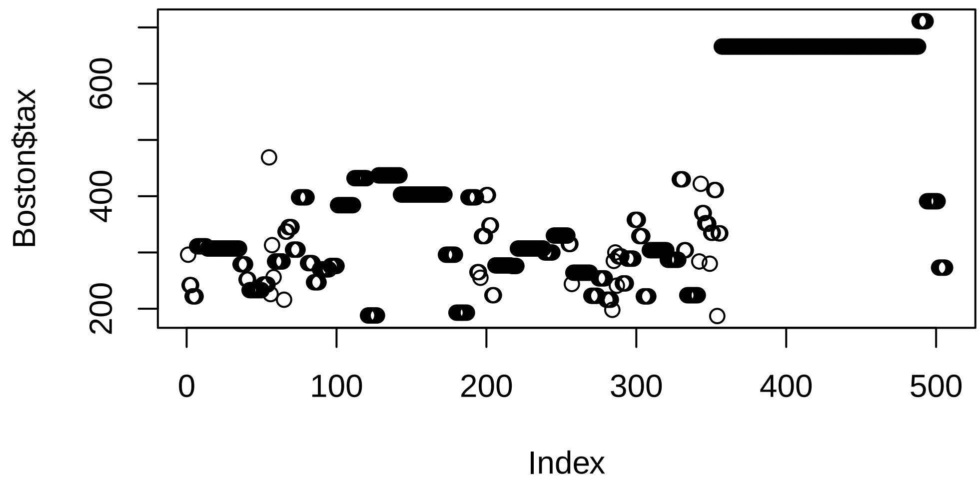
We can see that the observations from 350-490 have higher crime than the rest of the observations.

```
summary(Boston$crim)
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
 0.00632  0.08204  0.25651  3.61352  3.67708 88.97620
```

We can see the range of crime rate goes from nearly 0 to 89. We can also see that the vast amount of suburbs have relatively low crime rates.

```
plot(Boston$tax)
```



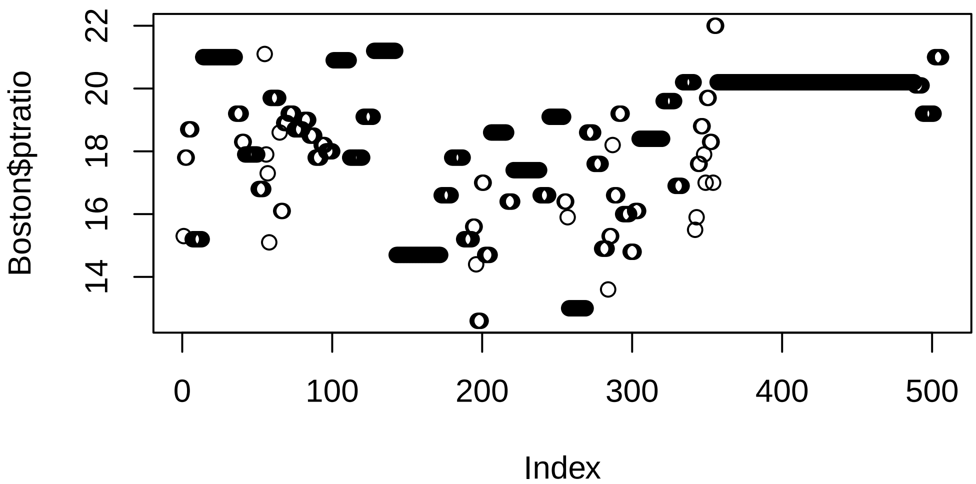
That same range that had high crime appear to also have the highest property taxes.

```
summary(Boston$tax)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  187.0   279.0   330.0   408.2   666.0   711.0 
```

The range of property tax rates is 187 per $10,000 to 711 per $10,000.

```
plot(Boston$ptratio)
```



We see here that same range have the same pupil teacher ratio which is relatively high, perhaps this set of suburbs have strict rules on their pupil-teacher ratio.

There is also a small set with a rather low ratio of 13.

```
summary(Boston$ptratio)
```

```
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  12.60   17.40   19.05   18.46   20.20   22.00 
```

The range of the pupil teacher ratio is 12.60 to 22.00.

e)

```
sum(Boston$chas)
```

```
[1] 35
```

There are 35 suburbs who border the Charles river.

f)

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

```
[1] 19.05
```

The median pupil teacher ratio is 19.05.

g)

```r
Boston[order(Boston$medv)[1:5], ]
```

```
        crim zn indus chas   nox    rm   age    dis rad tax ptratio lstat medv
399 38.35180  0  18.1    0 0.693 5.453 100.0 1.4896  24 666    20.2 30.59  5.0
406 67.92080  0  18.1    0 0.693 5.683 100.0 1.4254  24 666    20.2 22.98  5.0
401 25.04610  0  18.1    0 0.693 5.987 100.0 1.5888  24 666    20.2 26.77  5.6
400  9.91655  0  18.1    0 0.693 5.852  77.8 1.5004  24 666    20.2 29.97  6.3
415 45.74610  0  18.1    0 0.693 4.519 100.0 1.6582  24 666    20.2 36.98  7.0
```

We can see that the two suburbs that share the lowest median value of owner-occupied homes are index #399 and 406.

```r
Boston[c(399, 406),]
```

```
       crim zn indus chas   nox    rm age    dis rad tax ptratio lstat medv
399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 30.59    5
406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 22.98    5
```

Quickly taking mean of all the other suburbs to compare values and calculating the percentile of each feature for our two standout suburbs.

```r
Boston|>
  slice(-c(399,406)) |>
  summarise(across(where(is.numeric), mean))
```

```
      crim       zn    indus       chas       nox       rm     age      dis
1 3.417005 11.40873 11.10915 0.06944444 0.5541462 6.287478 68.4502 3.804319
       rad      tax  ptratio    lstat     medv
1 9.492063 407.2143 18.44861 12.59698 22.60238
```

```r
Boston |>
  summarise(across(where(is.numeric), percent_rank)) |>
  slice(c(399, 406))
```

```
Warning: Returning more (or less) than 1 row per `summarise()` group was deprecated in
dplyr 1.1.0.
i Please use `reframe()` instead.
i When switching from `summarise()` to `reframe()`, remember that `reframe()`
  always returns an ungrouped data frame and adjust accordingly.
```

```
       crim zn     indus chas       nox         rm       age        dis
1 0.9881188  0 0.6277228    0 0.8316832 0.07524752 0.9168317 0.05544554
2 0.9960396  0 0.6277228    0 0.8316832 0.13465347 0.9168317 0.03960396
        rad       tax   ptratio     lstat medv
1 0.7405941 0.7306931 0.6138614 0.9782178    0
2 0.7405941 0.7306931 0.6138614 0.8990099    0
```

We can see that the crime of these two suburbs are much higher than the average of others, in the top two percentile of suburbs.

There is no zoning for large residential land.

There is slightly more industry than average.

They do not border Charles river.

The nitrogen level is quite a bit higher than others.

The average rooms per dwelling is a lot lower than the mean.

All owner-occupied units were built before 1940.

They are closer to Boston employment centers.

They are closer to highways.

Property tax value is higher.

The pupil-teacher ratio is higher.

They have an extremely high rate of lower status individuals.

h)

```r
Boston |>
  filter(rm > 7) |>
  summarize(n = n())
```

```
   n
1 64
```

There are 64 rows with an average of 7 rooms or more.

```r
Boston |>
  filter(rm > 8) |>
  summarize(n = n())
```

```
   n
1 13
```

There are 13 suburbs with an average of 8 rooms or more.

```r
Boston |>
  mutate(many_rooms = rm > 8) |>
  group_by(many_rooms) |>
  summarise(across(where(is.numeric), mean))
```

```
# A tibble: 2 x 14
  many_rooms  crim    zn indus   chas   nox    rm   age   dis   rad   tax
  <lgl>      <dbl> <dbl> <dbl>  <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 FALSE      3.69   11.3 11.2  0.0669 0.555  6.23  68.5  3.80  9.60  410.
2 TRUE       0.719  13.6  7.08 0.154  0.539  8.35  71.5  3.43  7.46  325.
# i 3 more variables: ptratio <dbl>, lstat <dbl>, medv <dbl>
```

One can investigate any particular differences they find interesting, a few of mine:

We can see the extremely low crime rate of these suburbs, the low pupil-teacher ratio, the extremely low ratio of lower-class individuals and the extremely valuable houses.