Assignment1

Yue Zhang

2025-09-08

```r
setwd("/Users/yuezhang/Documents/Biostat/PH1976")
getwd()
library(ISLR2)
library(ggplot2)
library(tidyr)
library(dtplyr)
library(tidyverse)
```

```
## -- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
## v dplyr     1.1.4     v readr     2.1.5
## v forcats   1.0.0     v stringr   1.5.1
## v lubridate 1.9.3     v tibble    3.2.1
## v purrr     1.1.0
## -- 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 become error
```

```r
library(ggcorrplot)
```

1.

(a). Flexible methods will be better. With extremely large dataset and very few predictors, we have enough data to estimate the relationship without overfitting.
(b). Flexible methods will perform worse. Flexible models will overfit and have high variance.
(c). Flexible methods will perform better. Because the relationship tends to be non-linear and inflexible models may miss the true complexity and have high bias.

2.
(a). n = 500, p = 3 This is a regression problem and we're interested in inference.
(b). n = 20, p = 13 This is a classification problem and we're interested in prediction.
(c). n = 52, p = 3 This is a regression problem and we're interested in prediction.

5.

A flexible model can fit complex relationships between predictors and response. It can capture non linear relationship and has low bias. However, it will suffer from high variance and when noise is high or sample size is small, it tends to overfit. Flexible models are less interpretable. A less flexible model is easier to use and interpret and has lower variance. It won't capture a lot noise. But it may have higher bias if the relationship is complicated. A flexible approach is preferred when the relationship is highly non-linear or complex with large sample size. A less flexible approach is preferred when the sample size is small relative to the number of predictors.

7.

```
X1 = c(0, 2, 0, 0, -1, 1)
X2 = c(3, 0, 1, 1, 0, 1)
X3 = c(0, 0, 3, 2, 1, 1)
Y = c("red", "red", "red", "green", "green", "red")
df1 = cbind.data.frame(X1, X2, X3, Y)
```

(a).

```
df1$distance = sqrt((df1$X1 - 0)^2 + (df1$X2 - 0)^2 + (df1$X3 - 0)^2)
print(round(df1$distance, 3))
```

```
## [1] 3.000 2.000 3.162 2.236 1.414 1.732
```

```
#Ordered distance from small to large: 5, 6, 2, 4, 1, 3
```

(b).

```
knn = function(k) {
  names(which.max(table(df1[["Y"]][order(df1$distance)[1:k]])))
}
knn(1)
```

```
## [1] "green"
```

```
#For k = 1, it's only based on point 5.
```

(c).

```
knn(3)
```

```
## [1] "red"
```

```
#For k = 3, it's based on point 2, 5, 6
```

(d). We would expect the best value of k to be small as it yields to low bias and higher variance, which is better to capture complex structure while large k oversmooths and increases bias.

8.

(a).

```
college = read.csv("../Data/College.csv")
```

(b).

```
rownames(college) = college[, 1]
View(college)

college = college[, -1]
View(college)
```

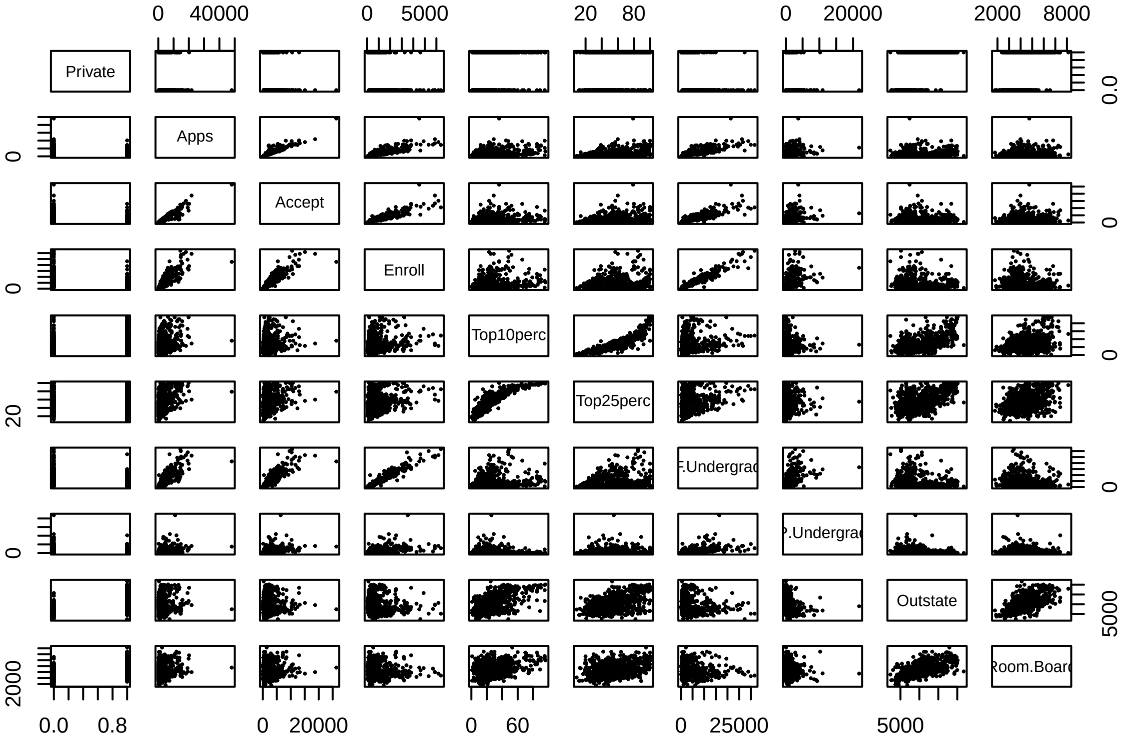
(c). *(i)*.

```
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
```

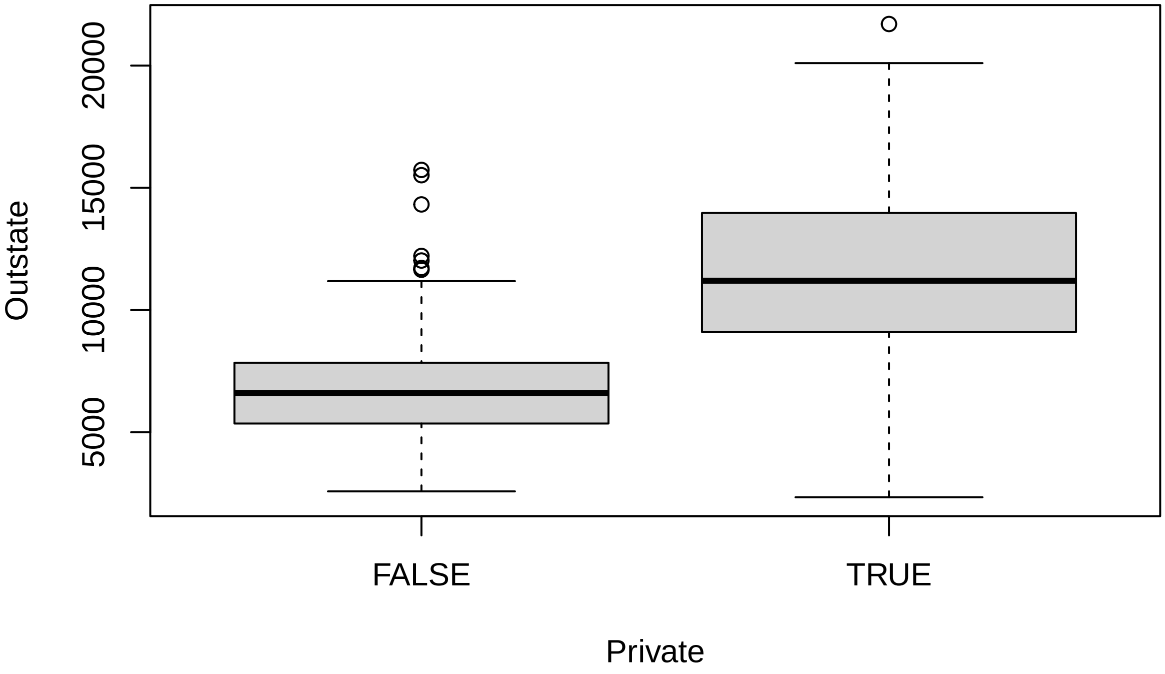
(ii).

```
college$Private = college$Private == "Yes"
pairs(college[, 1:10], pch = 19, cex = 0.2)
```



(iii).

```r
plot(college$Outstate ~ factor(college$Private), xlab = "Private", ylab = "Outstate")
```



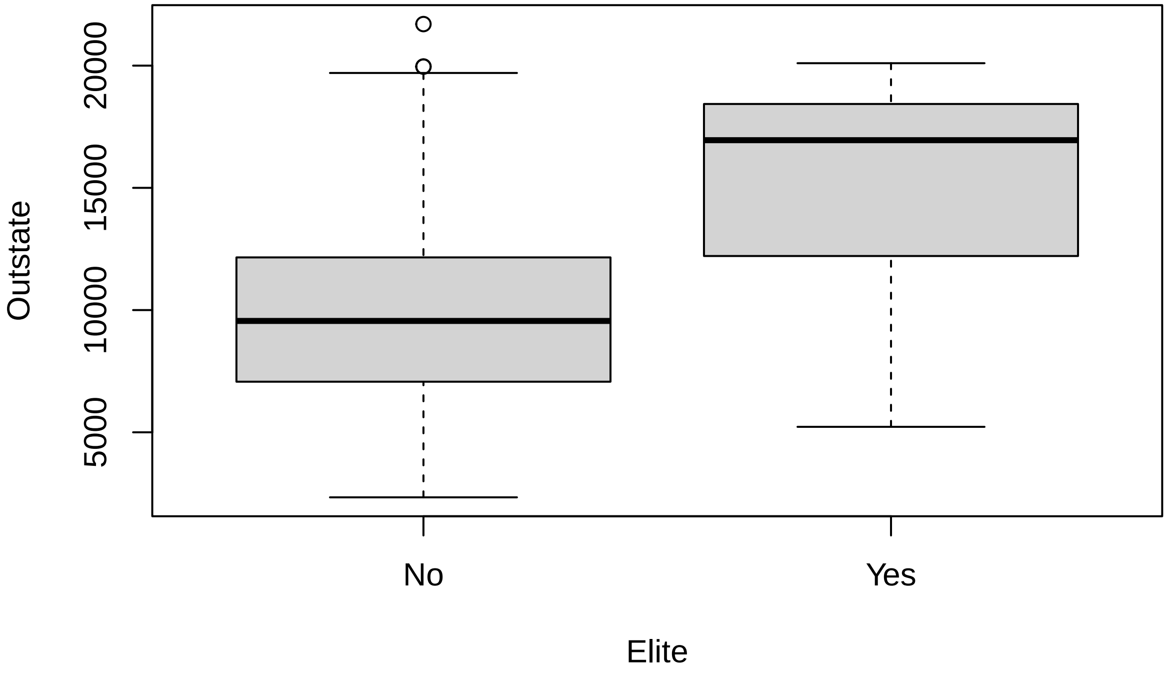
(iv).

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

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

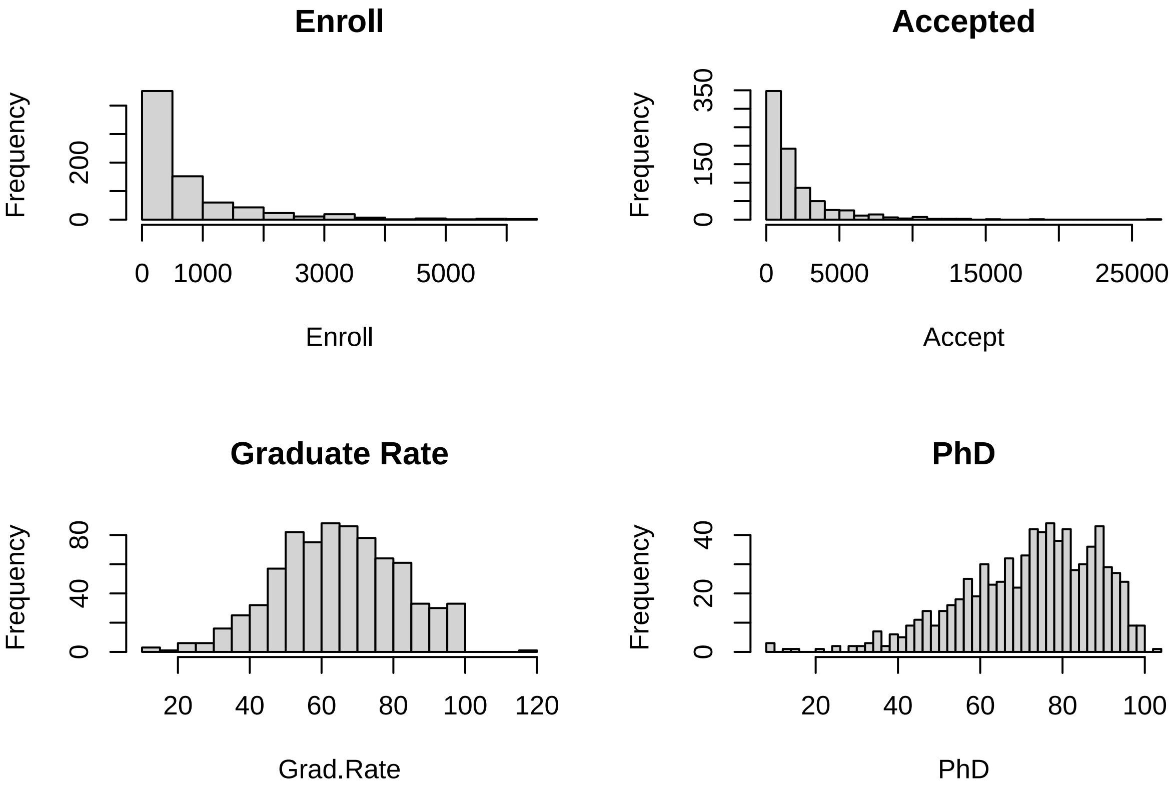
```
##  No Yes
## 699  78
```

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



(v).

```
par(mfrow = c(2, 2))
hist(college$Enroll, breaks = 10, main = "Enroll", xlab = "Enroll")
hist(college$Accept, breaks = 20, main = "Accepted", xlab = "Accept")
hist(college$Grad.Rate, breaks = 30, main = "Graduate Rate", xlab = "Grad.Rate")
hist(college$PhD, breaks = 50, main = "PhD", xlab = "PhD")
```

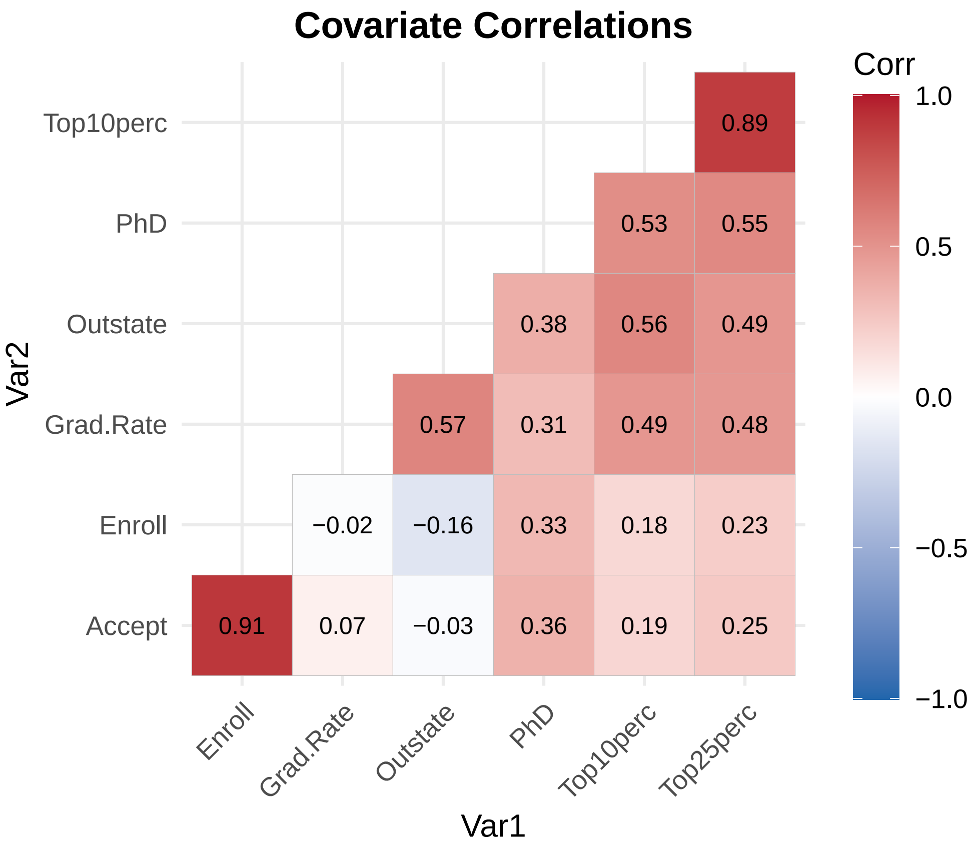


```r
par(mfrow = c(1, 1))
```

(vi).

```r
cont_vars = college %>%
  select(Grad.Rate, Accept, Enroll, Top10perc, Top25perc, PhD, Outstate)

cor_mat = cor(cont_vars, use = "pairwise.complete.obs")
ggcorrplot(
  cor_mat,
  method = "square",
  type = "lower",
  hc.order = TRUE,
  lab = TRUE,
  lab_size = 3,
  colors = c("#2166ac", "white", "#b2182b"),
  hc.method = "complete",
  tl.srt = 45,
  show.diag = NULL
) +
  ggplot2::labs(title = "Covariate Correlations") +
  ggplot2::theme_minimal(base_size = 12) +
  ggplot2::theme(
    plot.title = element_text(face = "bold", hjust = 0.5),
    axis.text.x = element_text(angle = 45, vjust = 1, hjust = 1)
  ) +
  guides(fill = guide_colorbar(barheight = unit(8, "cm")))
```



From the correlation matrix, we can see that the number of students accepted and the number of students enrolled are highly correlated. The proportion of students from the top 10% high schools is strongly correlated with the proportion from the top 25% high schools. Both enrollment and acceptance to the university are negatively correlated with outstate which makes sense as out of state tuition is higher than in-state tuition.

10.

(a).

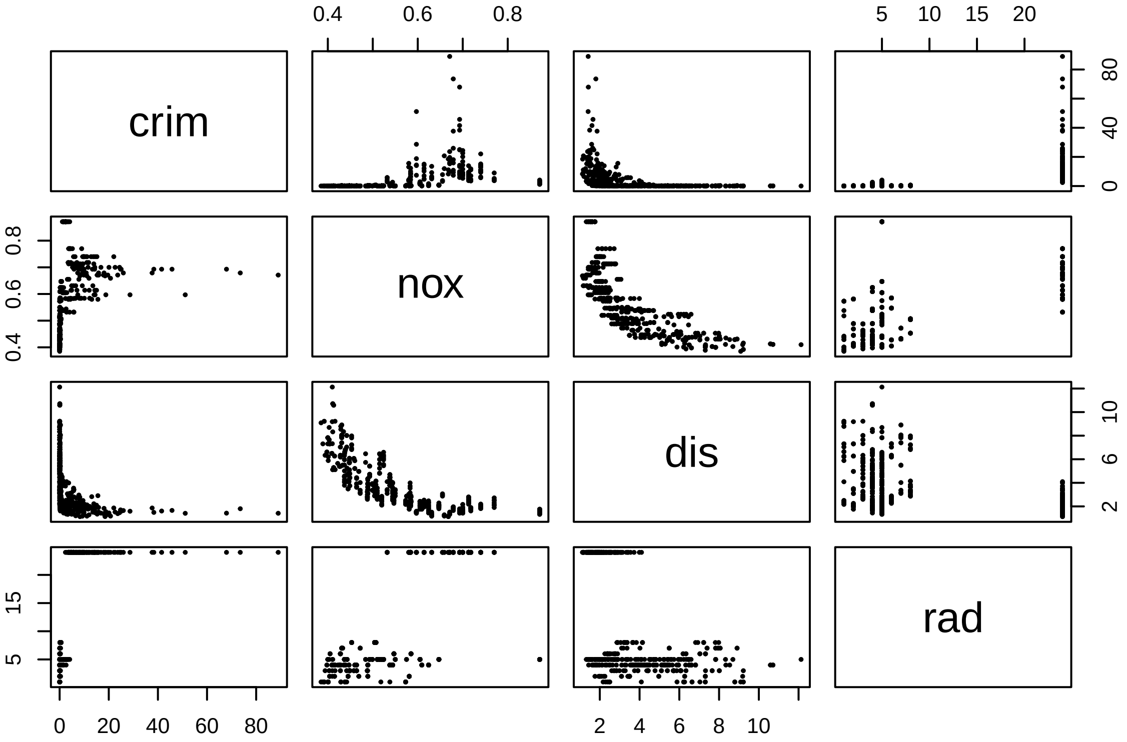
```
?Boston
View(Boston)
str(Boston)
```

```
## 'data.frame':    506 obs. of  13 variables:
##  $ crim   : num  0.00632 0.02731 0.02729 0.03237 0.06905 ...
##  $ zn     : num  18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...
##  $ indus  : num  2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...
##  $ chas   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ nox    : num  0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...
##  $ rm     : num  6.58 6.42 7.18 7 7.15 ...
##  $ age    : num  65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...
##  $ dis    : num  4.09 4.97 4.97 6.06 6.06 ...
##  $ rad    : int  1 2 2 3 3 3 5 5 5 5 ...
##  $ tax    : num  296 242 242 222 222 222 311 311 311 311 ...
##  $ ptratio: num  15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...
##  $ lstat  : num  4.98 9.14 4.03 2.94 5.33 ...
##  $ medv   : num  24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...
```

There are 506 rows and 13 columns in this dataset. Each row represents a suburb in Boston. Each column describes the characteristics of the suburb (crime rate, proportion of residential land, proportion of non-retail business, Charles River, nitrogen oxide concentration, average number of rooms per dwelling, proportion of owner-occupied units, weighted mean of distances to five employment centers, accessibility to radial highways, full-value property tax rate, pupil-teacher ratio, lower status of the population, and median value of owner-occupied homes)

(b).

```
pairs(Boston[, c(1, 5, 8, 9)], pch = 19, cex = 0.3)
```

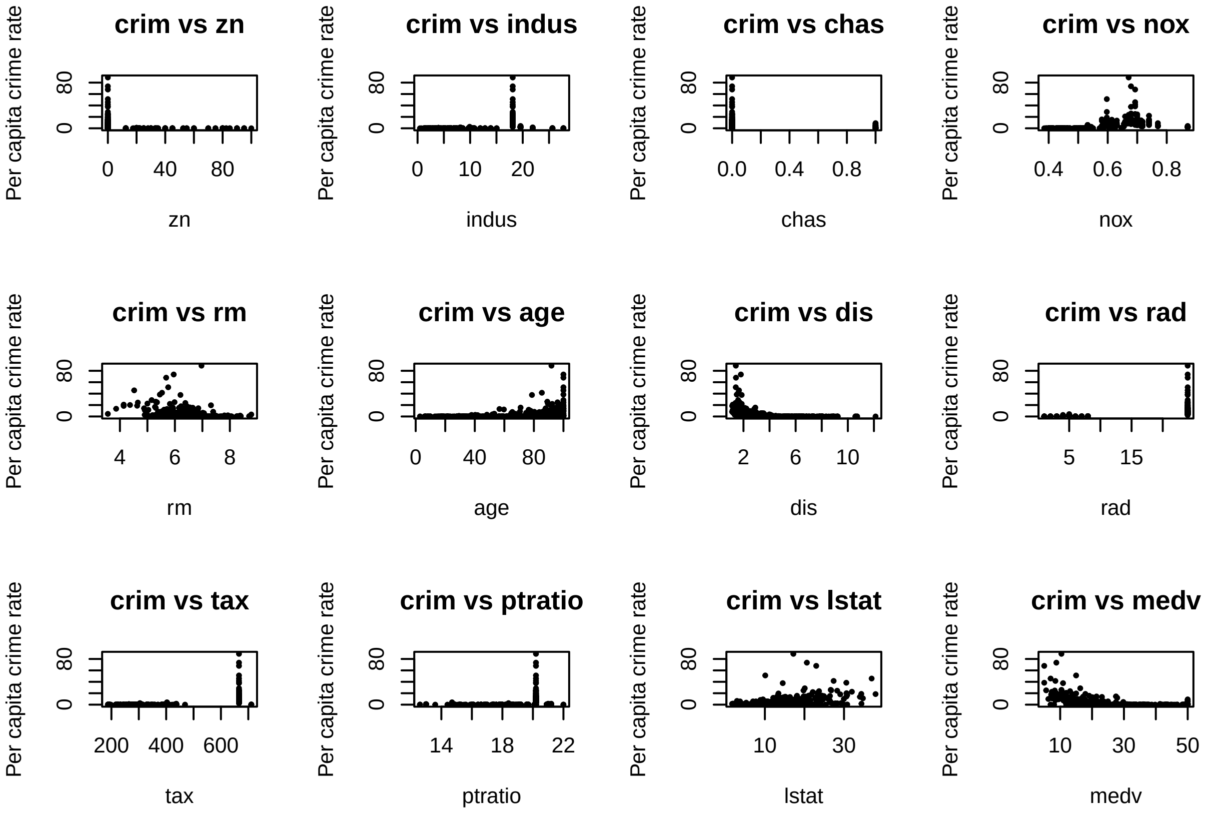


The pairwise scatterplot shows that higher crime rates tend to happen in areas that are close to employment centers and close to highways. Areas closer to employment centers also suffer from high nitrogen oxide concentration.

(c).

```
par(mfrow = c(3, 4))

for (var in names(Boston)[-1]) {
  plot(Boston[[var]], Boston$crim,
       xlab = var, ylab = "Per capita crime rate",
       main = paste("crim vs", var), pch = 19, cex = 0.4)
}
```



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

Crime rate is higher in areas with more industrial land use, server air pollution, closer to employment centers, smaller homes, lower median housing values, closer to highway, lower property tax, older housing and larger proportion of lower-status population.

(d).

```
range(Boston$crim)
```

```
## [1]  0.00632 88.97620
```

```
range(Boston$tax)
```

```
## [1] 187 711
```

```
range(Boston$ptratio)
```

```
## [1] 12.6 22.0
```

```
which.max(Boston$crim)
```

```
## [1] 381
```

```
which.max(Boston$tax)
```

```
## [1] 489
```

```
which.max(Boston$ptratio)
```

```
## [1] 355
```

```
summary(Boston[, c("crim", "tax", "ptratio")])
```

```
##       crim               tax           ptratio     
##  Min.   : 0.00632   Min.   :187.0   Min.   :12.60  
##  1st Qu.: 0.08205   1st Qu.:279.0   1st Qu.:17.40  
##  Median : 0.25651   Median :330.0   Median :19.05  
##  Mean   : 3.61352   Mean   :408.2   Mean   :18.46  
##  3rd Qu.: 3.67708   3rd Qu.:666.0   3rd Qu.:20.20  
##  Max.   :88.97620   Max.   :711.0   Max.   :22.00  
```

The range of crime rate is [0.00632, 88.97620], census tract 381 has the highest crime rate. The range of property tax rate is [187, 711], census tract 489 has the highest property tax rate. The range of pupil-teacher ratio is [12.6, 22.0], census tract 355 has the highest pupil-teacher ratio.

(e).

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

```
## [1] 35
```

There are 35 census tracts set bound the Charles river.

(f).

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

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

The median pupil-teacher ratio is 19.05

(g).

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

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

```
Boston[which.min(Boston$medv), ]
```

```
##        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
```

```
summary(Boston)
```

```
##       crim                zn             indus            chas        
##  Min.   : 0.00632   Min.   :  0.00   Min.   : 0.46   Min.   :0.00000  
##  1st Qu.: 0.08205   1st Qu.:  0.00   1st Qu.: 5.19   1st Qu.:0.00000  
##  Median : 0.25651   Median :  0.00   Median : 9.69   Median :0.00000  
##  Mean   : 3.61352   Mean   : 11.36   Mean   :11.14   Mean   :0.06917  
##  3rd Qu.: 3.67708   3rd Qu.: 12.50   3rd Qu.:18.10   3rd Qu.:0.00000  
##  Max.   :88.97620   Max.   :100.00   Max.   :27.74   Max.   :1.00000  
##       nox               rm             age              dis        
##  Min.   :0.3850   Min.   :3.561   Min.   :  2.90   Min.   : 1.130  
##  1st Qu.:0.4490   1st Qu.:5.886   1st Qu.: 45.02   1st Qu.: 2.100  
##  Median :0.5380   Median :6.208   Median : 77.50   Median : 3.207  
##  Mean   :0.5547   Mean   :6.285   Mean   : 68.57   Mean   : 3.795  
##  3rd Qu.:0.6240   3rd Qu.:6.623   3rd Qu.: 94.08   3rd Qu.: 5.188  
##  Max.   :0.8710   Max.   :8.780   Max.   :100.00   Max.   :12.127  
##       rad              tax           ptratio          lstat      
##  Min.   : 1.000   Min.   :187.0   Min.   :12.60   Min.   : 1.73  
##  1st Qu.: 4.000   1st Qu.:279.0   1st Qu.:17.40   1st Qu.: 6.95  
##  Median : 5.000   Median :330.0   Median :19.05   Median :11.36  
##  Mean   : 9.549   Mean   :408.2   Mean   :18.46   Mean   :12.65  
##  3rd Qu.:24.000   3rd Qu.:666.0   3rd Qu.:20.20   3rd Qu.:16.95  
##  Max.   :24.000   Max.   :711.0   Max.   :22.00   Max.   :37.97  
##       medv      
##  Min.   : 5.00  
##  1st Qu.:17.02  
##  Median :21.20  
##  Mean   :22.53  
##  3rd Qu.:25.00  
##  Max.   :50.00
```

Census tract 399 has the lowest median value of owner-occupied homes. Compared to the overall ranges for predictors, we can see that this census tract has high crime, high pollution, high proportion of lower-status population, old housing, and high taxes, along with small dwellings.

(h).

```
sum(Boston$rm > 7)
```

```
## [1] 64
```

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

```
## [1] 13
```

```
Boston[Boston$rm > 8, ]
```

```
##        crim zn indus chas    nox    rm  age    dis rad tax ptratio lstat medv
## 98  0.12083  0  2.89    0 0.4450 8.069 76.0 3.4952   2 276    18.0  4.21 38.7
## 164 1.51902  0 19.58    1 0.6050 8.375 93.9 2.1620   5 403    14.7  3.32 50.0
## 205 0.02009 95  2.68    0 0.4161 8.034 31.9 5.1180   4 224    14.7  2.88 50.0
## 225 0.31533  0  6.20    0 0.5040 8.266 78.3 2.8944   8 307    17.4  4.14 44.8
## 226 0.52693  0  6.20    0 0.5040 8.725 83.0 2.8944   8 307    17.4  4.63 50.0
## 227 0.38214  0  6.20    0 0.5040 8.040 86.5 3.2157   8 307    17.4  3.13 37.6
## 233 0.57529  0  6.20    0 0.5070 8.337 73.3 3.8384   8 307    17.4  2.47 41.7
## 234 0.33147  0  6.20    0 0.5070 8.247 70.4 3.6519   8 307    17.4  3.95 48.3
## 254 0.36894 22  5.86    0 0.4310 8.259  8.4 8.9067   7 330    19.1  3.54 42.8
## 258 0.61154 20  3.97    0 0.6470 8.704 86.9 1.8010   5 264    13.0  5.12 50.0
## 263 0.52014 20  3.97    0 0.6470 8.398 91.5 2.2885   5 264    13.0  5.91 48.8
## 268 0.57834 20  3.97    0 0.5750 8.297 67.0 2.4216   5 264    13.0  7.44 50.0
## 365 3.47428  0 18.10    1 0.7180 8.780 82.9 1.9047  24 666    20.2  5.29 21.9
```

There are 64 tracts that have more than 7 rooms per dwelling and 13 tracts that have more than 8 rooms per dwelling. Those 13 tracts tend to associate with high median housing values, low crime rate, lower status of the population, reflecting the most affluent neighborhoods in the dataset.