Chapter 2 HW

Fabiani Rafael

Conceptual Questions

Exercise 2: Explain whether each scenario is a classification or regression prob-lem, and indicate whether we are most interested in inference or prediction. Finally, provide n := sample size and p := predictors.

(a) We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- This describes an instance of regressional inference because our y i.e the CEO salary is a continuous variable and it is an example of an inference problem since the goal is not primarily to predict a given output but more so to decipher the underlying relationships between profit, number of employees, industry and the CEO salary. Regression fits this use case because instead of primarily focusing on using the input to predict changes in the CEO salary, we want to see which factors affect the CEO salary not necessarily how they will affect CEO salary. Here we have $n = 500$, $p = 3$.

(b) We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

- This describes an instance of a classification problem geared towards prediction. This falls into the category of classification as it deals with a y (success/failure) which is categorical in nature moreover its goal is to predict the success or failure of the new product using collected data from similar products. Here $n = 20$, $p = 13$

(c) We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

- This is an example of a regressional prediction problem. It falls into the category of regression as it deals with a y which is continuous in nature and fits into the category of prediction since its primary goal is to use input data i.e the percent change in the US, British, and German markets to predict the % change in the USD/Euro exchange rate. Here $n = 52$, $p = 3$.

Exercise 4: You will now think of some real-life applications for statistical learn-ing.

(a) Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- One example of a real-life application in which classification would be useful is in sentiment analysis possibly for something like product reviews. Where the response could be a sentiment label i.e. an indicator for Positive, Negative, and possibly Neutral reception. Then some predictors could be Text data (for instance review wording), star ratings, and reviewer history to gauge the tendency of a user to

leave either highly negative or positive reviews. The goal of this example would be primarily prediction since our intent is to classify new product reviews into categories based on learned patterns from data. The model would be trained on labeled reviews where the sentiment is already known and then it would predict the sentiment of unseen reviews.

- Another example of a real-life application could be Handwritten digit recognition where in the response would be a digit label i.e. (0-9) and the predictors could be pixel values of the image, edge detection features, and shape descriptors. The primary goal would be prediction i.e. trying to predict the correct digit for new/unseen images. So the model would take an input of a handwritten digit and classify it as one of the digits 0-9 and then use that to predict the most likely represents.
- A real-life application could be in transaction classification i.e. fraudulent or legitimate. Predictors such as transaction amount, time of the transaction, location, device used, and transaction history. The goal here would again be prediction because the main intent would be to identify fraudulent transactions in real-time using past transaction data to train our model with transactions labeled as either fraudulent or legitimate. When a new transaction is made the model would predict whether it is a valid transaction i.e. non fraudulent based on its features. So the task here would be classifying new transactions rather than understanding the causes of fraud.

(b) Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

- One real-life application where regression may be found useful is in the analysis of drivers of life expectancy. A response value for this would be the person's life expectancy at birth (a continuous value. in years). Predictors in this example could be things like GDP per capita, access to healthcare, sanitation, education, and other social factors. The goal of this application would be inference since the main intent would be to understand the relationship between the predictors and the response i.e. life expectancy. The model would be used to understand how the predictors affect life expectancy and not necessarily to predict life expectancy.
- Another example of an application where regression would be of use is in forecasting retail demand. A response value for this use case could be Daily sales volume (units sold) and the predictors could be things like day of the week, time of the year, weather, promotions, competitor pricing, and maybe historical data like sales from the past 30 days. The goal of this application would be prediction since the main intent would be to predict future sales volume based on past sales data and other predictors to gauge future demand for inventory management. The model would be used to predict future sales volume based on past sales data and could help in anticipating holiday sales for a specific product to optimize stock levels.
- An example of an application that could use regression is also in predicting renewable energy production i.e. solar power, and wind farms. A response value for this could be the amount of energy produced in a given time period. Predictors could be things like weather data for instance wind speed, temperature and air pressure, time of day, season, and location. The goal of this application would be prediction since the main intent would be to predict the amount of energy produced in the future based on past data. The model would be used to predict future energy production based on trends in previous energy production and could help in optimizing energy storage and distribution.

(c) Describe three real-life applications in which cluster analysis might be useful.

- One real-life application where cluster analysis might be useful is in customer segmentation. For instance in a retail setting, where the goal would be to segment customers into distinct groups based on their purchasing behavior/history. Clustered data could be customer transaction history, customer demographics, and engagement metrics. Predictors could be things like purchase frequency, the average amount spent per order, and product preferences. Demographic data that could be used can be age, location, and income and behavioral data would be things like website click or email open rates. This could help in tailoring marketing strategies to each group. For instance, a cluster of customers who buy mostly electronics could be targeted with electronics promotions because they may be more likely to be interested in those products.

- Another real-life application where cluster analysis might be useful is anomaly detection. For instance in network security where the goal would be to detect unusual behavior in network traffic. Clustered data could be network traffic data, user activity logs, and system logs such as access attempts. Features here would be things like data transfer rates, login attempts, IP addresses, file access times, protocol type, and time of day. This could help in identifying unusual patterns in network traffic that could be indicative of a security breach because clustering may help in distinguishing normal activity from anomalous clusters and security teams could prioritize investigating outliers flagged by the model.
- Another real-life application where cluster analysis might be useful is in recommendation systems. For instance in a streaming service like Netflix where the goal would be to recommend movies or shows to users based on their viewing history. Clustered data could be user behavior for instance viewing history, user ratings, search queries, and time spent on content or potentially content attributes like genre, actors, directors, and release year. Features here could be viewing habits, with similar themes or audience appeal as content clusters. This could help in recommending movies or shows to users based on their viewing history and preferences. For instance, a user who watches a lot of action movies could be recommended more action movies or a user who watches a lot of comedies could be recommended more comedies.

Applied Questions

Exercise 8: This exercise relates to the College data set, which can be found in the file College.csv on the book website. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- Private : Public/private indicator
- Apps : Number of applications received
- Accept : Number of applicants accepted
- Enroll : Number of new students enrolled
- Top10perc : New students from top 10 % of high school class
- Top25perc : New students from top 25 % of high school class
- F.Undergrad : Number of full-time undergraduates
- P.Undergrad : Number of part-time undergraduates
- Outstate : Out-of-state tuition
- Room.Board : Room and board costs
- Books : Estimated book costs
- Personal : Estimated personal spending
- PhD : Percent of faculty with Ph.D.'s
- Terminal : Percent of faculty with terminal degree
- S.F.Ratio : Student/faculty ratio
- perc.alumni : Percent of alumni who donate
- Expend : Instructional expenditure per student
- Grad.Rate : Graduation rate

Before reading the data into R, it can be viewed in Excel or a text editor.

(a) Use the read.csv() function to read the data into R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
#check curr directory
#getwd()

#reading college into R
college <- read.csv("College.csv")
```

(b) Look at the data using the View() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later. Try the following commands:

rownames(college) <- college[, 1]

View(college)

```
# rm first col
rownames(college) <- college[, 1]

#outputs the entire contents of colege
#View(college)

#view first ten rows in college data set, with all columns
# (professor said it was okay to view just a subsection (since the data is large)
View(college[1:10, c(1:19)])
```

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

college <- college[, -1]

View(college)

```
# rm first col
college <- college[, -1]

# confirm first col is gone (note 1:19 would throw an error since theres now 18 cols)
#View(college)
View(college[1:10, c(1:18)])
```

Now you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

(c)

i. Use the summary() function to produce a numerical summary of the variables in the data set

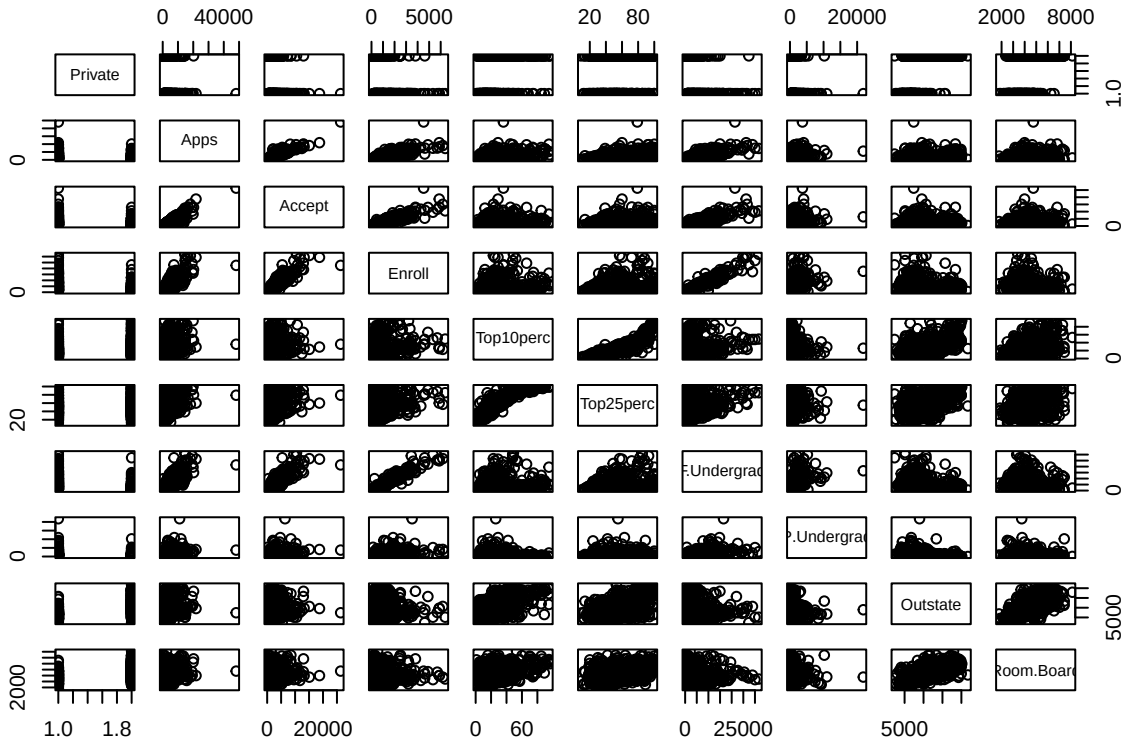
```
summary(college)
```

```
##    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
```

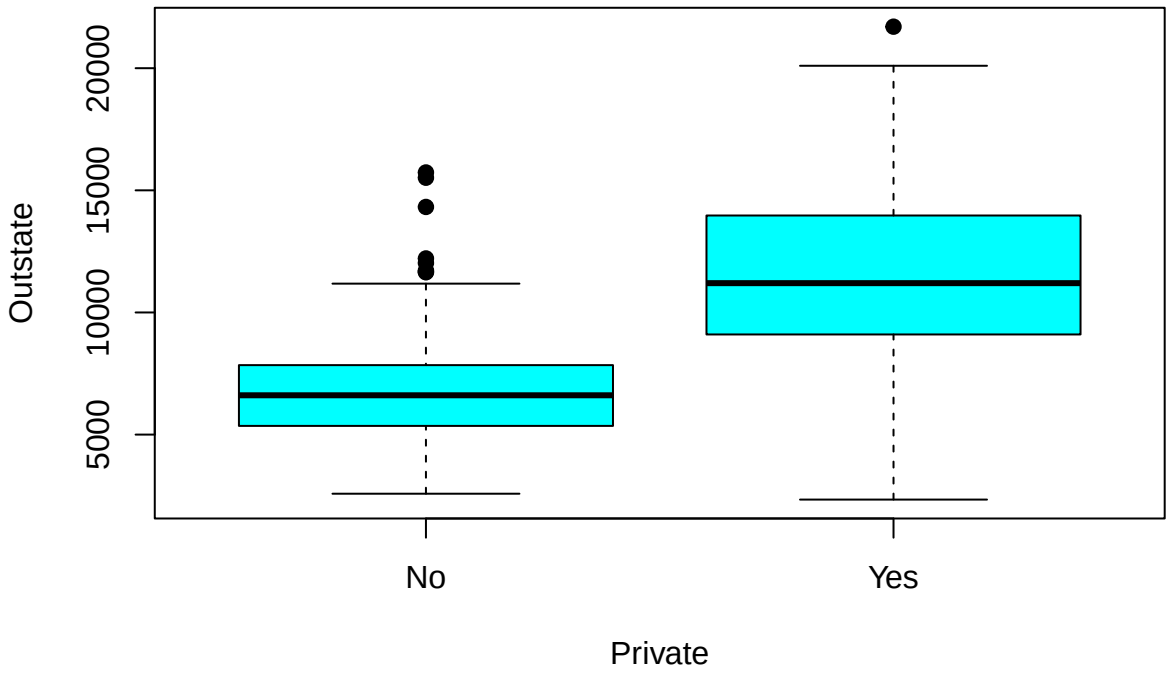
ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```r
# fix non numeric col private,
college$Private <- as.factor(college$Private)

## passing college using [,1:10] to get the first ten col
pairs(college[,1:10])
```



iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

```
# debug chk types
# str(college$Private)  # should be a factor but was char[]
# str(college$Outstate) # int[]


# fix type for plot()
#college$Private <- as.factor(college$Private) moved to ii, before i was
#using pairs with indexing to skip private like college[,2:10]
plot(college$Private, college$Outstate, xlab = "Private", ylab = "Outstate", col = "cyan", pch = 19)
```



iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

Elite <- rep("No", nrow(college))

Elite[college$Top10perc > 50] <- "Yes"

Elite <- as.factor(Elite)

college <- data.frame(college , Elite)

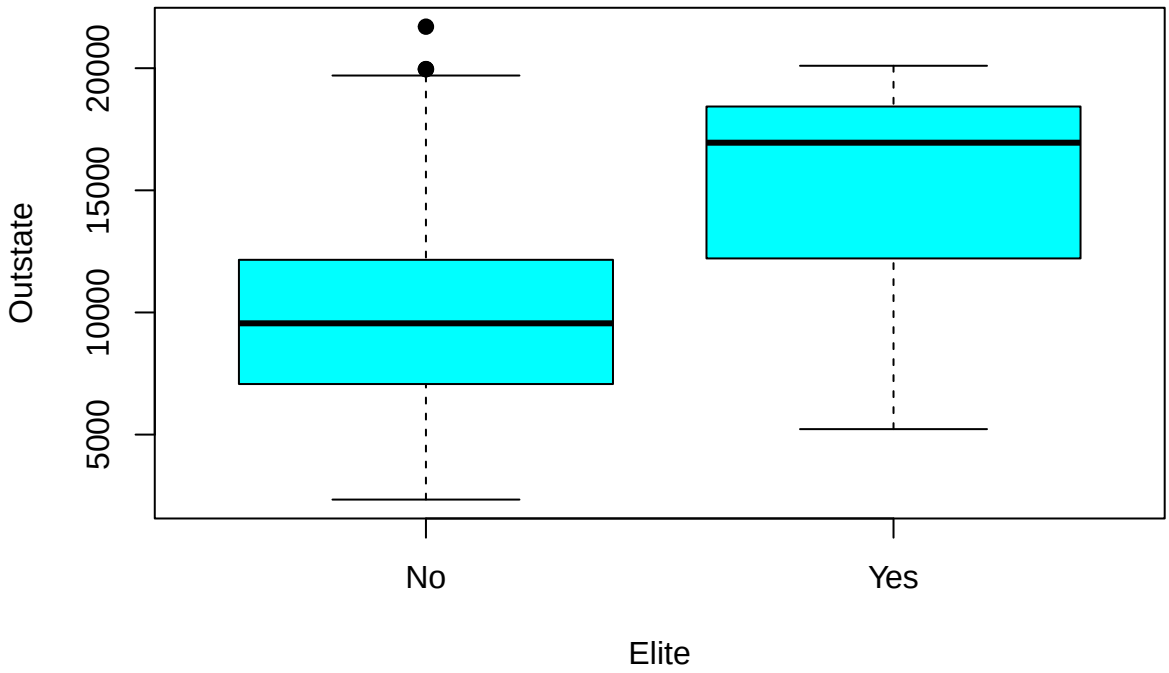
```
Elite <- rep("No", nrow(college))
Elite[college$Top10perc > 50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college , Elite)
```

Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

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

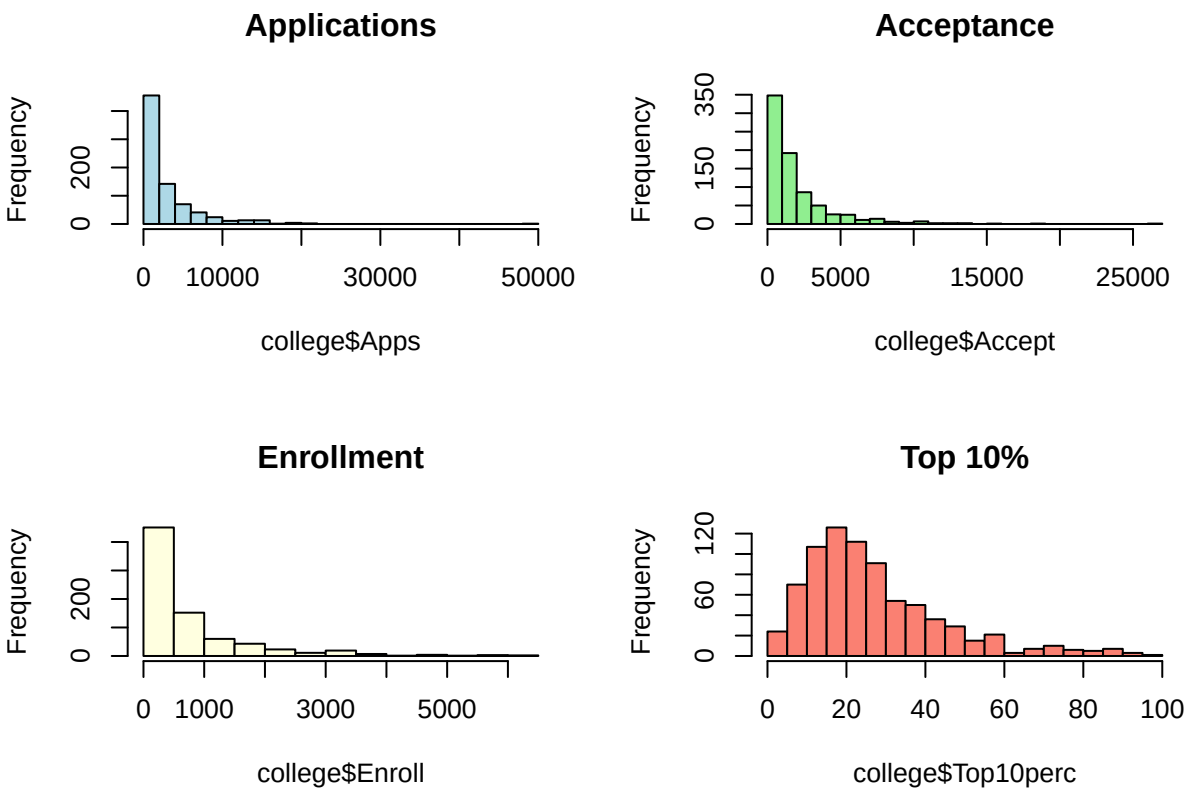
```
##  No Yes
## 699  78
```

```
plot(college$Elite, college$Outstate, xlab = "Elite", ylab = "Outstate", col = "cyan", pch = 19)
```



v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow = c(2, 2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

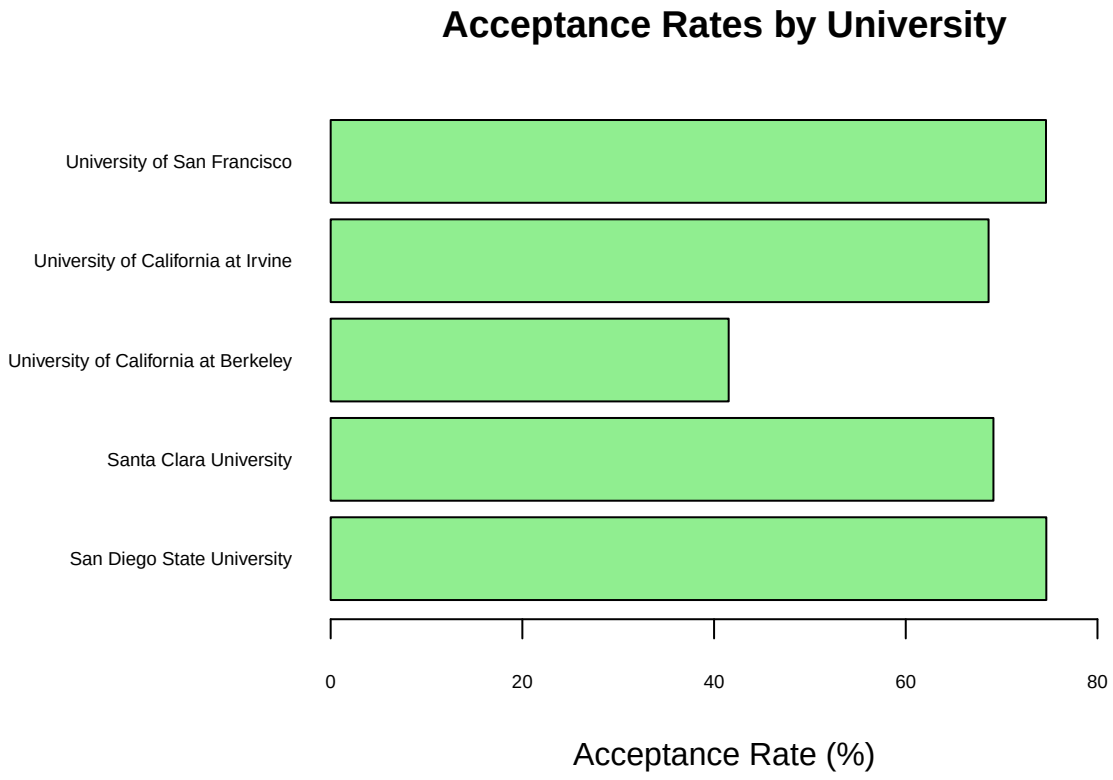
```
# divide the print window into 4 x 4 for plots
par(mfrow = c(2, 2))
hist(college$Apps, breaks = 20, col = "lightblue", main = "Applications")

hist(college$Accept, breaks = 20, col = "lightgreen", main = "Acceptance")

hist(college$Enroll, breaks = 20, col = "lightyellow", main = "Enrollment")

hist(college$Top10perc, breaks = 20, col = "salmon", main = "Top 10%")
```



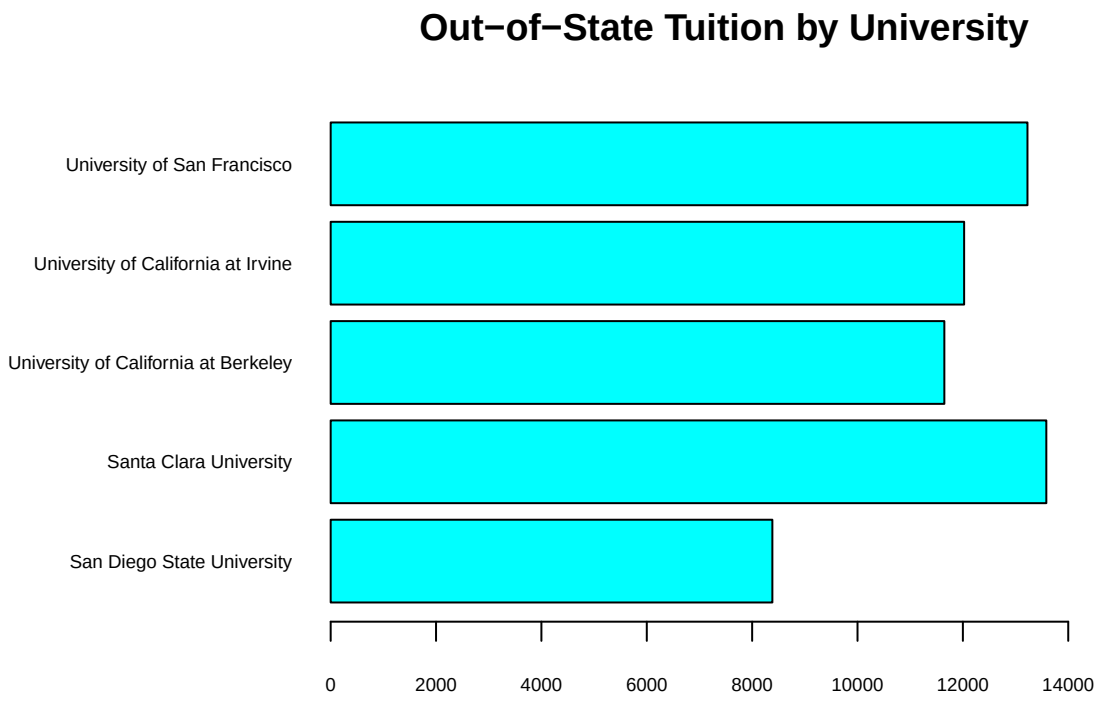
vi. Continue exploring the data, and provide a brief summary of what you discover.

```
#college_name <- "University of California at Berkeley"
#college[rownames(college) == college_name, ]


# tmp increase margins fix for names being cut off
par(mar = c(5, 10, 4, 2))

# list of colleges id like to search for
college_names <- c("University of California at Berkeley",
                   "University of California at Irvine",
                   "San Diego State University",
                   "Santa Clara University",
                   "University of San Francisco",
                   "San Francisco State University",
                   "University of California at Santa Barbara")

# conv to lower case
college_names_lower <- tolower(college_names)
rownames_lower <- tolower(rownames(college))

# collect college entries
collected_colleges<- college[rownames_lower %in% college_names_lower, ]


# get acceptance rates
collected_colleges$AcceptanceRate <- (collected_colleges$Accept / collected_colleges$Apps) * 100
```

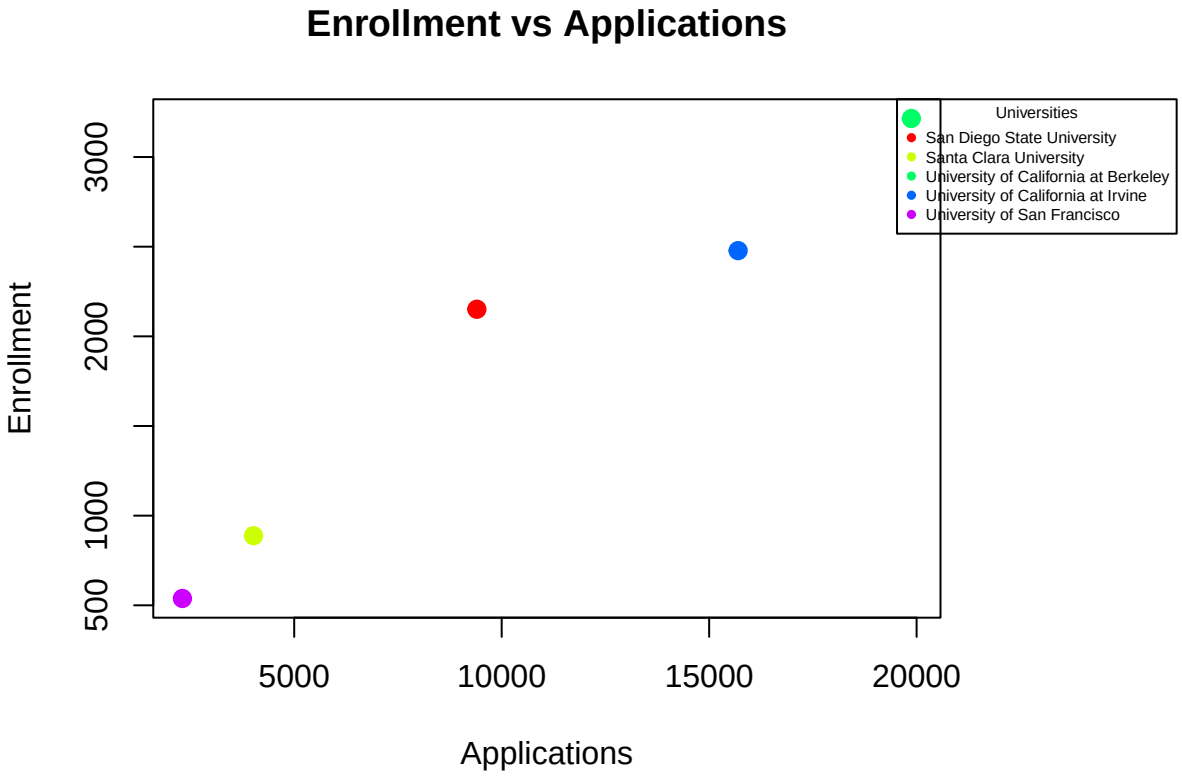
```r
# get enrollment rate
collected_colleges$EnrollmentRate <- (collected_colleges$Enroll / collected_colleges$Apps) * 100

barplot(collected_colleges$AcceptanceRate,
        names.arg = rownames(collected_colleges),
        main = "Acceptance Rates by University",
        xlab = "Acceptance Rate (%)",
        col = "lightgreen",
        horiz = TRUE,
        cex.names = 0.6,
        cex.axis = 0.6,
        las = 1,
        xlim = c(0, max(collected_colleges$AcceptanceRate) * 1.1))
```




```r
barplot(collected_colleges$Outstate,
        names.arg = rownames(collected_colleges),
        main = "Out-of-State Tuition by University",
        xlab = "Out-of-State Tuition ($)",
        col = "cyan",
        horiz = TRUE,
        cex.names = 0.6,
        cex.axis = 0.6,
        las = 1,
        xlim = c(0, max(collected_colleges$Outstate) * 1.1))
```




```r
# vector of colors (one per university)
n_colleges <- nrow(collected_colleges)
colors <- rainbow(n_colleges)

# plot margins to make space for the legend (bottom, left, top, right)
par(mar = c(5, 4, 4, 8), xpd = TRUE)

# plot the points
plot(collected_colleges$Apps, collected_colleges$Enroll,
     main = "Enrollment vs Applications",
     xlab = "Applications",
     ylab = "Enrollment",
     pch = 19,
     col = colors,
     cex = 1.2)

# legend for plot
legend("topright",
       inset = c(-0.3, 0),  # mv legend outside the plot
       legend = rownames(collected_colleges),
       col = colors,
       pch = 19,
       title = "Universities",
       cex = 0.5)
```




```
# reset to default for next ex.
par(mar = c(5, 4, 4, 2))
```

- For this portion of the exercise I wanted to compare the data for San Francisco State University , UC Berkeley, UC Santa Barbara, UC San Diego, Riverside, Santa Clara University, San Diego state and Eastbay but It seems like the data set was missing entries for San Francisco state along with other universities like UC Santa Barbara. I was able to find data for UC Berkeley, Santa Clara University, San Diego State University and University of San Francisco. I calculated the acceptance rate and enrollment rate for each of these universities and then plotted them along with the out of state tuition for each of these universities and I created a plot for enrollment vs applications. From the acceptance rate bar plot i saw that UC San Francisco and San Diego state had the highest acceptance rates. From the out of state tuition bar plot I saw that Santa Clara University had the highest out of state tuition. From the enrollment vs applications plot I saw that UC Berkeley had the highest number of applications and enrollments. I picked these colleges because I thought of applying to some of these places and had friends and professors who went to some of those schools.

Exercise 10 This exercise involves the Boston housing data set.

(a) To begin, load in the Boston data set. The Boston data set is part of the ISLR2 library.

library(ISLR2)

Now the data set is contained in the object Boston.

Boston

Read about the data set:

?Boston

How many rows are in this data set? How many columns? What do the rows and columns represent?

```r
# load the ISLR2 lib, had to install with install.packages("ISLR2")
library(ISLR2)

# load boston data set (commented out for sake of brevity)
#Boston
#?Boston
# only a segment since the entire data set is so large
Boston[1:10, c(1:13)]
```

```
##       crim   zn indus chas   nox    rm   age    dis rad tax ptratio lstat medv
## 1  0.00632 18.0  2.31    0 0.538 6.575  65.2 4.0900   1 296    15.3  4.98 24.0
## 2  0.02731  0.0  7.07    0 0.469 6.421  78.9 4.9671   2 242    17.8  9.14 21.6
## 3  0.02729  0.0  7.07    0 0.469 7.185  61.1 4.9671   2 242    17.8  4.03 34.7
## 4  0.03237  0.0  2.18    0 0.458 6.998  45.8 6.0622   3 222    18.7  2.94 33.4
## 5  0.06905  0.0  2.18    0 0.458 7.147  54.2 6.0622   3 222    18.7  5.33 36.2
## 6  0.02985  0.0  2.18    0 0.458 6.430  58.7 6.0622   3 222    18.7  5.21 28.7
## 7  0.08829 12.5  7.87    0 0.524 6.012  66.6 5.5605   5 311    15.2 12.43 22.9
## 8  0.14455 12.5  7.87    0 0.524 6.172  96.1 5.9505   5 311    15.2 19.15 27.1
## 9  0.21124 12.5  7.87    0 0.524 5.631 100.0 6.0821   5 311    15.2 29.93 16.5
## 10 0.17004 12.5  7.87    0 0.524 6.004  85.9 6.5921   5 311    15.2 17.10 18.9
```

```r
dim(Boston)
```

```
## [1] 506  13
```

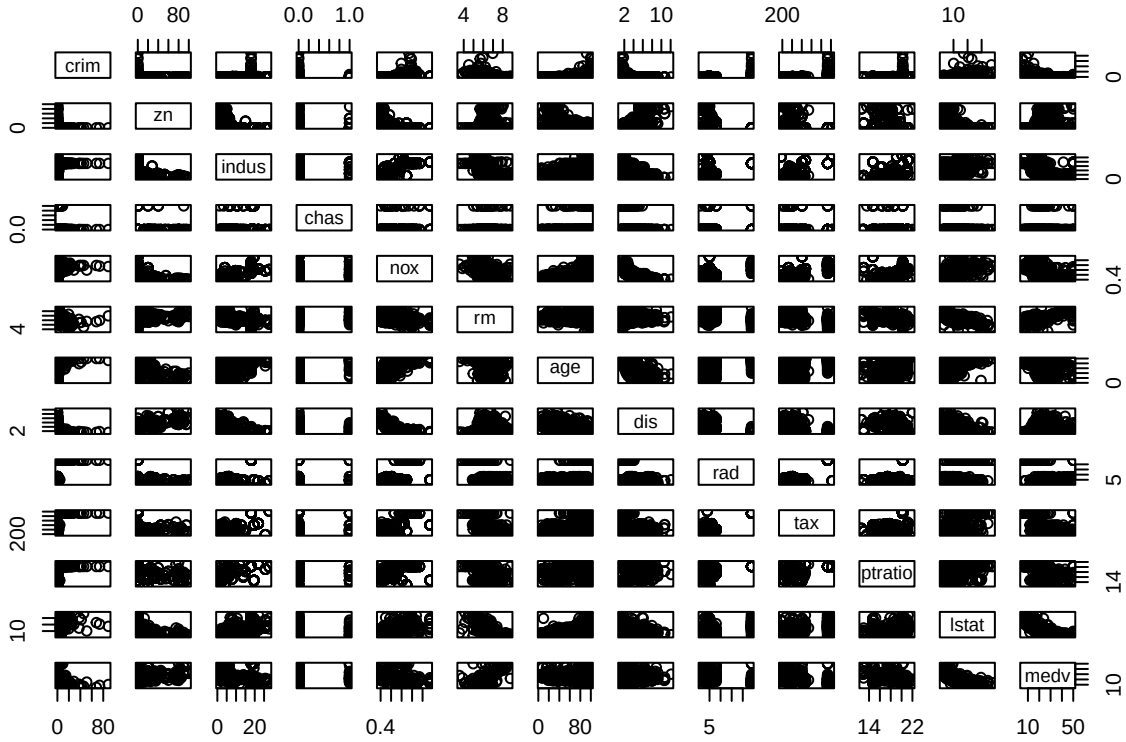
```r
colnames(Boston)
```

```
##  [1] "crim"    "zn"      "indus"   "chas"    "nox"     "rm"      "age"
##  [8] "dis"     "rad"     "tax"     "ptratio" "lstat"   "medv"
```

- The Boston data set has 506 rows and 13 columns. The rows represent the census tracts / neighborhoods in Boston. The columns represent different attributes of each neighborhood

(b) Make some pairwise scatterplots of the predictors (columns) in this data set. Describe your findings.

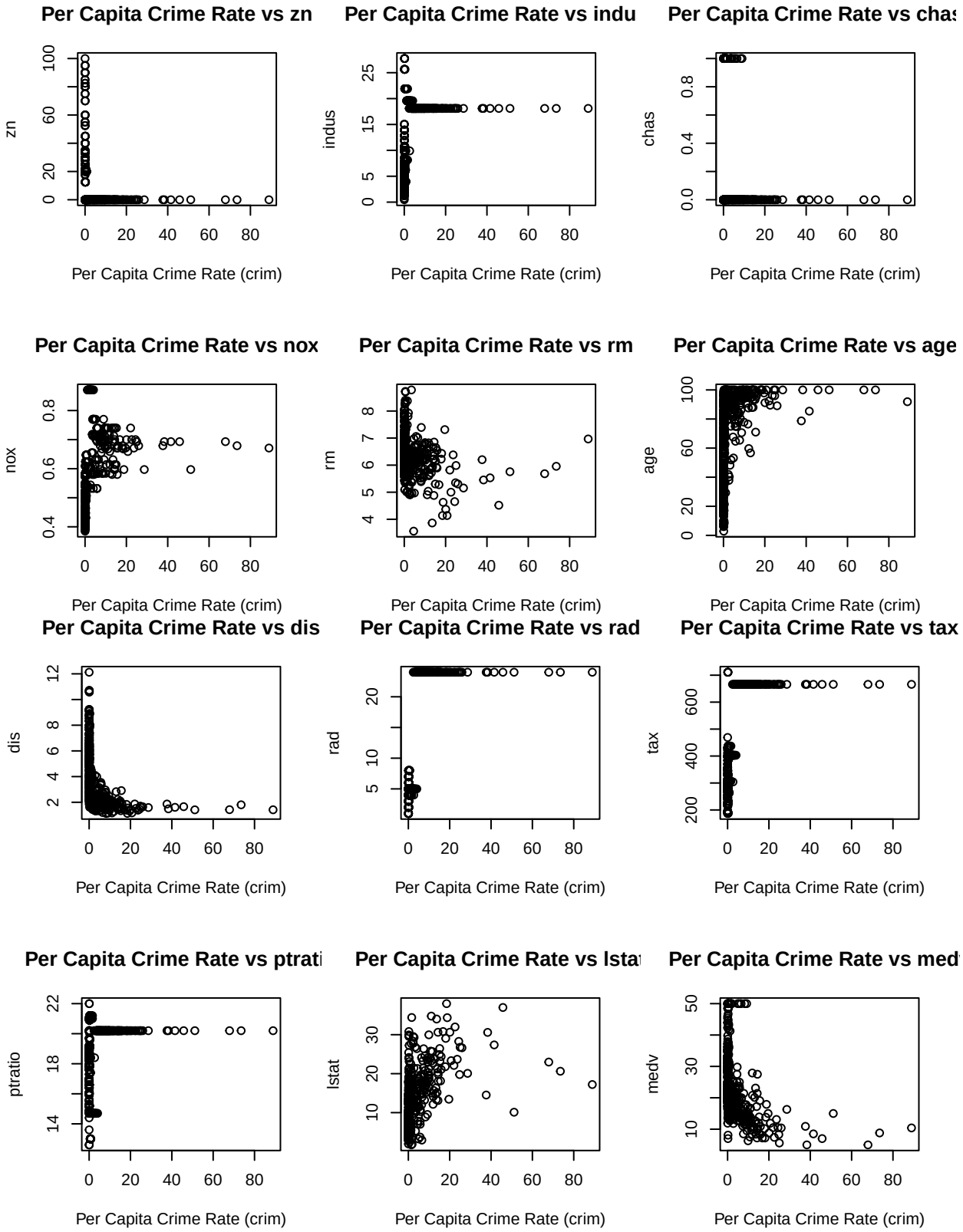
```r
pairs(Boston)
```



- From the scatterplots, we can see that some of the predictors have positive correlations for instance rm (average number of rooms per dwelling) appears to have a positive relationship with medv (median home value). This suggests that neighborhoods with larger homes tend to have higher property values. Another thing I notice is dis (distance to employment centers) seems positively correlated with medv, indicating that homes farther from the city center tend to have higher values. There are also some negative correlations for instance lstat (percent lower status of the population) seems to have a negative relationship with medv, suggesting that neighborhoods with a higher percentage of lower status residents tend to have lower property values.

(c) Are any of the predictors associated with per capita crime rate? If so, explain the relationship.

```
# plotting grid 2 x 3
par(mfrow=c(2, 3))

# plots of crime rate vs each predictor
for (i in 1:ncol(Boston)) {
    if (colnames(Boston)[i] != "crim") {
        plot(
            Boston$crim,
            Boston[, i],
            xlab = "Per Capita Crime Rate (crim)",
            ylab = colnames(Boston)[i],
            main = paste("Per Capita Crime Rate vs", colnames(Boston)[i])
        )
    }
}
```




```
cor(Boston)
```

```
##                crim          zn       indus          chas         nox
```

```
## crim     1.00000000 -0.20046922  0.40658341 -0.055891582  0.42097171
## zn      -0.20046922  1.00000000 -0.53382819 -0.042696719 -0.51660371
## indus    0.40658341 -0.53382819  1.00000000  0.062938027  0.76365145
## chas    -0.05589158 -0.04269672  0.06293803  1.000000000  0.09120281
## nox      0.42097171 -0.51660371  0.76365145  0.091202807  1.00000000
## rm      -0.21924670  0.31199059 -0.39167585  0.091251225 -0.30218819
## age      0.35273425 -0.56953734  0.64477851  0.086517774  0.73147010
## dis     -0.37967009  0.66440822 -0.70802699 -0.099175780 -0.76923011
## rad      0.62550515 -0.31194783  0.59512927 -0.007368241  0.61144056
## tax      0.58276431 -0.31456332  0.72076018 -0.035586518  0.66802320
## ptratio  0.28994558 -0.39167855  0.38324756 -0.121515174  0.18893268
## lstat    0.45562148 -0.41299457  0.60379972 -0.053929298  0.59087892
## medv    -0.38830461  0.36044534 -0.48372516  0.175260177 -0.42732077
##                 rm         age         dis          rad         tax    ptratio
## crim    -0.21924670  0.35273425 -0.37967009  0.625505145  0.58276431  0.2899456
## zn       0.31199059 -0.56953734  0.66440822 -0.311947826 -0.31456332 -0.3916785
## indus   -0.39167585  0.64477851 -0.70802699  0.595129275  0.72076018  0.3832476
## chas     0.09125123  0.08651777 -0.09917578 -0.007368241 -0.03558652 -0.1215152
## nox     -0.30218819  0.73147010 -0.76923011  0.611440563  0.66802320  0.1889327
## rm       1.00000000 -0.24026493  0.20524621 -0.209846668 -0.29204783 -0.3555015
## age     -0.24026493  1.00000000 -0.74788054  0.456022452  0.50645559  0.2615150
## dis      0.20524621 -0.74788054  1.00000000 -0.494587930 -0.53443158 -0.2324705
## rad     -0.20984667  0.45602245 -0.49458793  1.000000000  0.91022819  0.4647412
## tax     -0.29204783  0.50645559 -0.53443158  0.910228189  1.00000000  0.4608530
## ptratio -0.35550149  0.26151501 -0.23247054  0.464741179  0.46085304  1.0000000
## lstat   -0.61380827  0.60233853 -0.49699583  0.488676335  0.54399341  0.3740443
## medv     0.69535995 -0.37695457  0.24992873 -0.381626231 -0.46853593 -0.5077867
##              lstat       medv
## crim     0.4556215 -0.3883046
## zn      -0.4129946  0.3604453
## indus    0.6037997 -0.4837252
## chas    -0.0539293  0.1752602
## nox      0.5908789 -0.4273208
## rm      -0.6138083  0.6953599
## age      0.6023385 -0.3769546
## dis     -0.4969958  0.2499287
## rad      0.4886763 -0.3816262
## tax      0.5439934 -0.4685359
## ptratio  0.3740443 -0.5077867
## lstat    1.0000000 -0.7376627
## medv    -0.7376627  1.0000000
```

- Yes, from the output plots it seems that crime correlates to several predictors such as age, dis, medv,rad, indus and tax. For instance, the per capita crime rate seems to increase (have a strong positive correlation) with the age of the houses, the index of accessibility to radial highways, the proportion of non-retail business acres per town and the full-value property-tax rate per $10,000, lowe social economic stattus and nitrogen oxide concentration. There are also some strong negative correlations between crime rate and the median home value (higher values seem to correlate with lower crime), distance to employment centers and average numver of rooms per dwelling. These can be seen from inspecting their respective plots and are further affirmed from inspecting the output of **cor**(**Boston**). Wherein we see : Strongest Positive Correlations with Crime (crim):

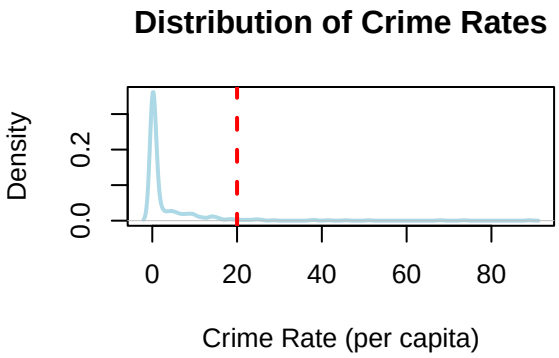
  rad (0.63) → Accessibility to radial highways tax (0.58) → Property tax rate indus (0.41) → Proportion of non-retail business acres nox (0.42) → Nitrogen oxide concentration lstat (0.46) → % lower socioeconomic status age (0.35) → Proportion of older homes

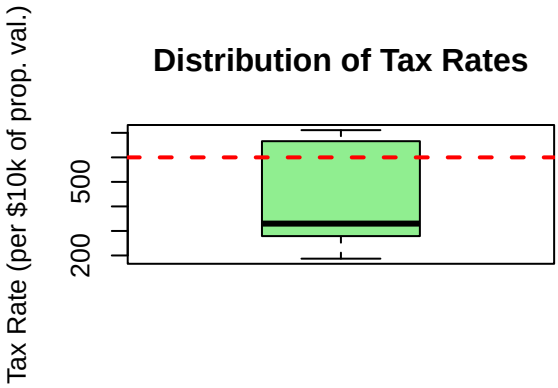
Strongest Negative Correlations with Crime (crim):

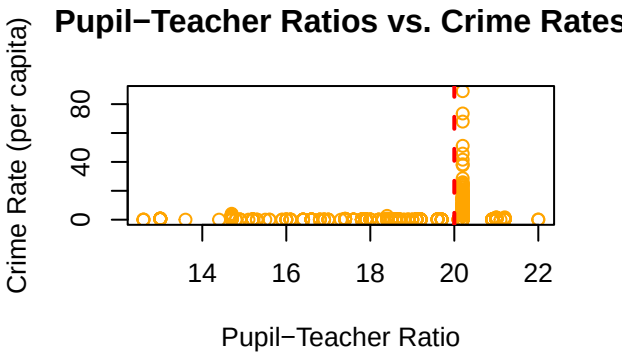
```
medv (-0.39) → Median home value (higher values = lower crime)
dis (-0.38) → Distance to employment centers (farther = less crime)
rm (-0.22) → Average number of rooms per dwelling
```

(d) Do any of the census tracts of Boston appear to have particularly high crime rates? Tax rates? Pupil-teacher ratios? Comment on the range of each predictor.

```r
# get the census tracts with the highest crime rates
high_crim <- Boston[Boston$crim > 20, ]
#high_crim

# get the census tracts with the highest tax rates
high_tax <- Boston[Boston$tax > 600, ]
#high_tax

# get the census tracts with the highest pupil-teacher ratios
high_ptratio <- Boston[Boston$ptratio > 20, ]
#high_ptratio

# get the range of each predictor
crim_range <- range(Boston$crim)
tax_range <- range(Boston$tax)
ptratio_range <- range(Boston$ptratio)

# plotting grid 2 x 3
par(mfrow=c(2, 2))

# density plot
plot(density(Boston$crim), col = "lightblue", lwd = 2,
     main = "Distribution of Crime Rates",
     xlab = "Crime Rate (per capita)", ylab = "Density")
# line to mark a threshold for high crime rate
abline(v = 20, col = "red", lwd = 2, lty = 2)

# boxplot of tax rates
boxplot(Boston$tax, col = "lightgreen",
        main = "Distribution of Tax Rates",
        cex.names = 0.7,
        ylab = "Tax Rate (per $10k of prop. val.)")
# threash hold for high tax rate
abline(h = 600, col = "red", lwd = 2, lty = 2)

# plot of pupil-teacher ratios vs. crime rates
plot(Boston$ptratio, Boston$crim, col = "orange",
     main = "Pupil-Teacher Ratios vs. Crime Rates",
     xlab = "Pupil-Teacher Ratio", ylab = "Crime Rate (per capita)")

# line to mark a threshold for high pupil-teacher ratio
abline(v = 20, col = "red", lwd = 2, lty = 2)
```







- Yes, there are census tracts with particularly high crime rates, tax rates and pupil-teacher ratios. For instance, census tracts with crime rates greater than 20 have particularly high crime rates although these appear to be outlier cases suggesting that crime might be occuring disproportionately / is concentrated in crtain tracs. The crime rate in the dataset ranges from 0.00632 to 88.98, indicating a substantial disparity. Census tracts with tax rates greater than 600 have particularly high tax burdens. The tax rate varies between 187 and 711 per $10,000 of property value. Continuing census tracts with a pupil-teacher ratio (ptratio) above 20 experience relatively high student-to-teacher ratios, which could indicate lower school resources. The pupil-teacher ratio spans from 12.6 to 22.0 across the dataset. To recap the range of each predictor is as follows: the per capita crime rate ranges from 0 to 89.0, the full-value property-tax rate per $10,000 ranges from 187 to 711, and the pupil-teacher ratio by town ranges from 12.6 to 22.0.

(e) How many of the census tracts in this data set bound the Charles river?

```
# selecting tract which bound the river , i.e chas == 1 .
#Boston[Boston$chas == 1,]

# the num of such tracts
num_tracts_river <- sum(Boston$chas == 1)
num_tracts_river
```

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

- There are 35 census tracts in this data set that bound the Charles river.

(f) What is the median pupil-teacher ratio among the towns in this data set?

```
# get the median and store it
median_ptratio <- median(Boston$ptratio)
median_ptratio
```

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

- The median pupil-teacher ratio among the towns in this data set is 19.05.

(g) Which census tract of Boston has lowest median value of owner-occupied homes? What are the values of the other predictors for that census tract, and how do those values compare to the overall ranges for those predictors? Comment on your findings.

```
# get the census tract with the lowest median value of owner-occupied homes
low_medv <- Boston[Boston$medv == min(Boston$medv), ]
# yields 19.05
low_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
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 22.98    5
```

```
# get the range of each predictor
range(Boston$medv)
```

```
## [1]  5 50
```

```
range(Boston$lstat)
```

```
## [1]  1.73 37.97
```

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

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

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

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

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

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

```
range(Boston$nox)
```

```
## [1] 0.385 0.871
```

```
range(Boston$indus)
```

```
## [1]  0.46 27.74
```

```
range(Boston$rad)
```

```
## [1]  1 24
```

```
range(Boston$age)
```

```
## [1]   2.9 100.0
```

- The census tract with the lowest median value of owner-occupied homes is row index 399. The median value of owner-occupied homes in this tract is 5.0. The values of the other predictors for this census tract are as follows: per capita crime rate is 38.3518, the proportion of lower status of the population is 30.59%, the full-value property-tax rate per $10,000 is 666, the pupil-teacher ratio by town is 20.2, the nitric oxides concentration is 0.693, the proportion of non-retail business acres per town is 18.1, the index of accessibility to radial highways is 24, the proportion of older homes is 96.1%. These values are at the extreme end of the range for each predictor. For instance, the per capita crime rate, proportion of lower status of the population, the full-value property-tax rate per $10,000, pupil-teacher ratio by town is at the upper end of the range, the nitric oxides concentration is at the upper end of the range,

the proportion of non-retail business acres per town is at the upper end of the range, the index of accessibility to radial highways is at the upper end of the range, and the proportion of older homes is at the upper end of the range. This suggests that this census tract is in a particularly disadvantaged area with high crime rates, high tax rates, and poor school resources.

(h) In this data set, how many of the census tracts average more than seven rooms per dwelling? More than eight rooms per dwelling? Comment on the census tracts that average more than eight rooms per dwelling.

```
# get tracts with rm > 7 and rm > 8
num_seven <- sum(Boston$rm > 7)   # 170 tracts
num_eight <- sum(Boston$rm > 8)   # 7 tracts

# analyze tracts with rm > 8
tracts_eight <- Boston[Boston$rm > 8, ]
summary(tracts_eight)  # Summary of these tracts
```

```
##       crim              zn            indus             chas
##  Min.   :0.02009   Min.   : 0.00   Min.   : 2.680   Min.   :0.0000
##  1st Qu.:0.33147   1st Qu.: 0.00   1st Qu.: 3.970   1st Qu.:0.0000
##  Median :0.52014   Median : 0.00   Median : 6.200   Median :0.0000
##  Mean   :0.71879   Mean   :13.62   Mean   : 7.078   Mean   :0.1538
##  3rd Qu.:0.57834   3rd Qu.:20.00   3rd Qu.: 6.200   3rd Qu.:0.0000
##  Max.   :3.47428   Max.   :95.00   Max.   :19.580   Max.   :1.0000
##       nox               rm             age              dis
##  Min.   :0.4161   Min.   :8.034   Min.   : 8.40   Min.   :1.801
##  1st Qu.:0.5040   1st Qu.:8.247   1st Qu.:70.40   1st Qu.:2.288
##  Median :0.5070   Median :8.297   Median :78.30   Median :2.894
##  Mean   :0.5392   Mean   :8.349   Mean   :71.54   Mean   :3.430
##  3rd Qu.:0.6050   3rd Qu.:8.398   3rd Qu.:86.50   3rd Qu.:3.652
##  Max.   :0.7180   Max.   :8.780   Max.   :93.90   Max.   :8.907
##       rad              tax           ptratio          lstat           medv
##  Min.   : 2.000   Min.   :224.0   Min.   :13.00   Min.   :2.47   Min.   :21.9
##  1st Qu.: 5.000   1st Qu.:264.0   1st Qu.:14.70   1st Qu.:3.32   1st Qu.:41.7
##  Median : 7.000   Median :307.0   Median :17.40   Median :4.14   Median :48.3
##  Mean   : 7.462   Mean   :325.1   Mean   :16.36   Mean   :4.31   Mean   :44.2
##  3rd Qu.: 8.000   3rd Qu.:307.0   3rd Qu.:17.40   3rd Qu.:5.12   3rd Qu.:50.0
##  Max.   :24.000   Max.   :666.0   Max.   :20.20   Max.   :7.44   Max.   :50.0
```

```
# comp medv to the overall median
median(Boston$medv)            # ovrall median: 21.2
```

```
## [1] 21.2
```

```
median(tracts_eight$medv)
```

```
## [1] 48.3
```

- These census tracts would generally correspond to affluent neighborhoods with larger homes i.e more rooms. Given that the number of rooms per dwelling (rm) is positively correlated with median home value (medv), these tracts are likely to have higher property values. They may also have lower crime rates and better school resources, as these factors are often associated with wealthier areas.