chapter 2

StatisticalLearning

2024-10-21

`

#chapter-2 (8)

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.

```
library (ISLR)
```

```
## Warning: package 'ISLR' was built under R version 4.3.3
```

b. Look at the data using the fix() 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:

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 4.3.3
```

```
## ── Attaching core tidyverse packages ──────────────────────── tidyverse 2.0.0 ──
## ✔ dplyr     1.1.4     ✔ readr     2.1.5
## ✔ forcats   1.0.0     ✔ stringr   1.5.1
## ✔ ggplot2   3.4.4     ✔ tibble    3.2.1
## ✔ lubridate 1.9.3     ✔ tidyr     1.3.1
## ✔ purrr     1.0.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflicts() ──
## ✖ dplyr::filter() masks stats::filter()
## ✖ dplyr::lag()    masks stats::lag()
## ℹ Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to become errors
```

```
rownames(College) <- College$X
College$X <- NULL

glimpse(College)
```

```
## Rows: 777
## Columns: 18
## $ Private     <fct> Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes, Yes…
## $ Apps        <dbl> 1660, 2186, 1428, 417, 193, 587, 353, 1899, 1038, 582, 173…
## $ Accept      <dbl> 1232, 1924, 1097, 349, 146, 479, 340, 1720, 839, 498, 1425…
## $ Enroll      <dbl> 721, 512, 336, 137, 55, 158, 103, 489, 227, 172, 472, 484,…
## $ Top10perc   <dbl> 23, 16, 22, 60, 16, 38, 17, 37, 30, 21, 37, 44, 38, 44, 23…
## $ Top25perc   <dbl> 52, 29, 50, 89, 44, 62, 45, 68, 63, 44, 75, 77, 64, 73, 46…
## $ F.Undergrad <dbl> 2885, 2683, 1036, 510, 249, 678, 416, 1594, 973, 799, 1830…
## $ P.Undergrad <dbl> 537, 1227, 99, 63, 869, 41, 230, 32, 306, 78, 110, 44, 638…
## $ Outstate    <dbl> 7440, 12280, 11250, 12960, 7560, 13500, 13290, 13868, 1559…
## $ Room.Board  <dbl> 3300, 6450, 3750, 5450, 4120, 3335, 5720, 4826, 4400, 3380…
## $ Books       <dbl> 450, 750, 400, 450, 800, 500, 500, 450, 300, 660, 500, 400…
## $ Personal    <dbl> 2200, 1500, 1165, 875, 1500, 675, 1500, 850, 500, 1800, 60…
## $ PhD         <dbl> 70, 29, 53, 92, 76, 67, 90, 89, 79, 40, 82, 73, 60, 79, 36…
## $ Terminal    <dbl> 78, 30, 66, 97, 72, 73, 93, 100, 84, 41, 88, 91, 84, 87, 6…
## $ S.F.Ratio   <dbl> 18.1, 12.2, 12.9, 7.7, 11.9, 9.4, 11.5, 13.7, 11.3, 11.5, …
## $ perc.alumni <dbl> 12, 16, 30, 37, 2, 11, 26, 37, 23, 15, 31, 41, 21, 32, 26,…
## $ Expend      <dbl> 7041, 10527, 8735, 19016, 10922, 9727, 8861, 11487, 11644,…
## $ Grad.Rate   <dbl> 60, 56, 54, 59, 15, 55, 63, 73, 80, 52, 73, 76, 74, 68, 55…
```

C.

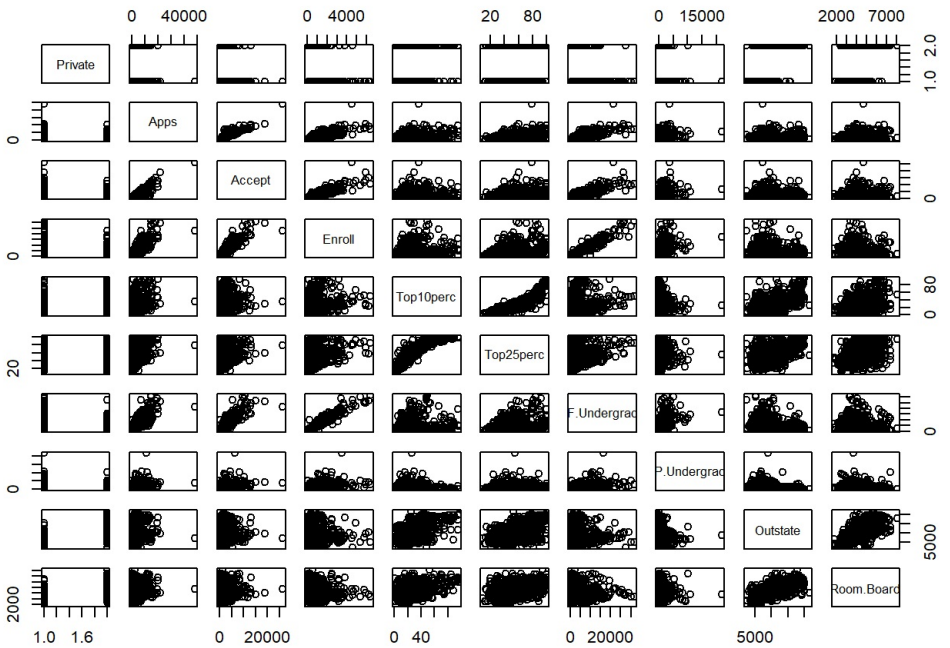
i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
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
```

INTERPRETATION #Out of 777 colleges, 565 are private, and 212 are public. #The number of applications received by colleges ranges from 81 to 48,094, with a median of 1,558. 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].

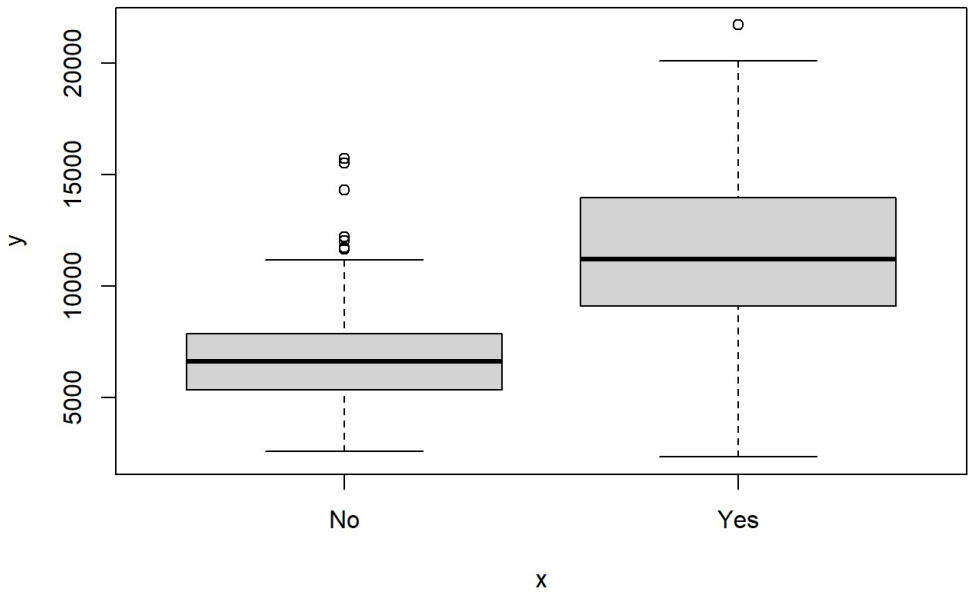
```
pairs(College [,1:10])
```



intrepertation #Positive correlations, like between Apps and Accept, indicate that more applications generally lead to more acceptances. # such as Outstate vs. Enroll, may show that higher tuition can negatively impact enrollment. #the relationship between enrollment and student quality can help identify whether colleges with more selective admissions have higher or lower student bodies.

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

```
plot(College$Private,College$Outstate)
```



INTERPRETATION #private

colleges(yes) show a higher median out-of-state tuition #public colleges(no) will show a lower median tuition it has a outliers #Private colleges typically offer higher-priced education, whereas public colleges tend to be more cost-effective

iv. Create a new qualitative variable, called Elite,bybinning 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$Top25perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(College,Elite)
```

Use the summary() function to see how many elite universities there are.

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

```
##  No Yes
## 328 449
```

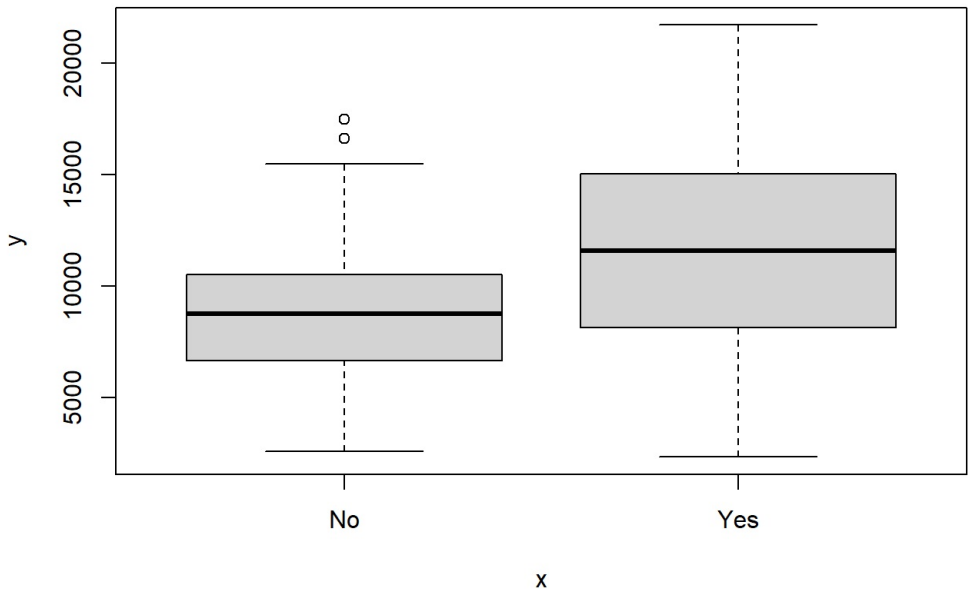
INTERPRETATION #one category (“Yes”) dominates, representing about 58% of the dataset

```
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   No :328
##  1st Qu.: 53.00   Yes:449
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

INTERPRETATION #acceptances ranging from 72 to 26,330 and a median of 1,110. #enrollments spanning from 35 to 6,392, and a median of 434. #wide range of costs, academic profiles, and institutional resources, highlighting the diversity in the landscape of higher education Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

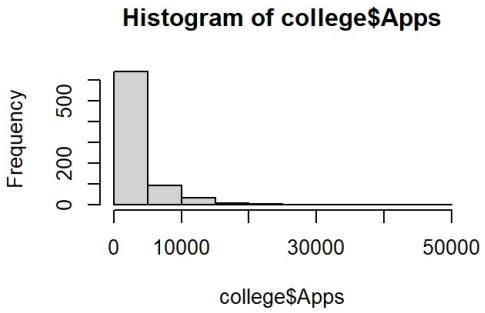
```
plot(college$Elite,college$Outstate)
```

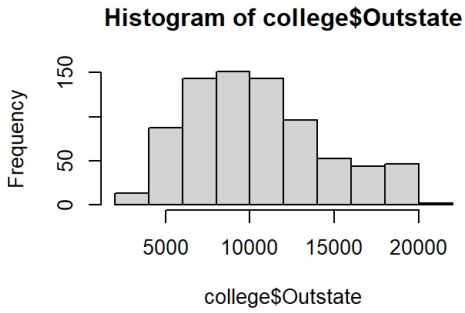


intrepertation #If the out-of-state tuition for Elite colleges is significantly higher #Elite status doesn't have a strong relationship with out-of-state tuition.

v. Use the hist() function to produce some histograms withdiffering 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 fourplots can be made simultaneously. Modifying the arguments to this function will divide the screen in other ways.

```
par(mfrow=c(2,2))
hist(college$Apps)
hist(college$Outstate)
```

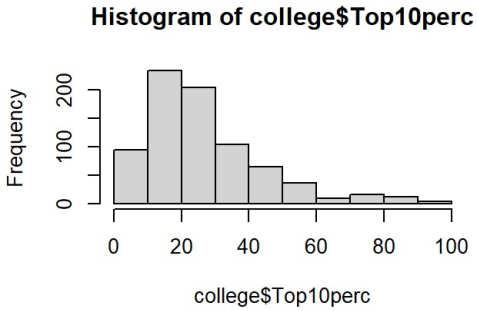


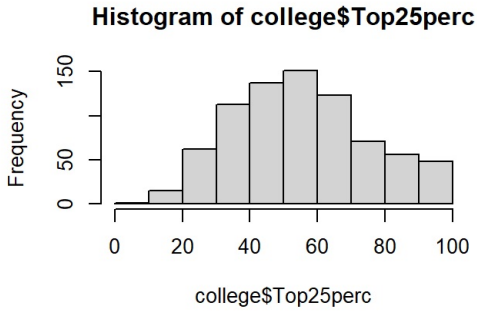


INTERPRETAION #The histogram

shows a right-skewed distribution, indicating that most colleges receive a relatively low number of applications, with a few outliers #The histogram for out-of-state tuition is also be right-skewed, reflecting that most colleges charge a moderate amount for out-of-state tuition, while a few institutions have much higher fees. vi.Continue exploring the data, and provide a brief summary of what you discover

```
par(mfrow=c(2,2))
hist(college$Top10perc)
hist(college$Top25perc)
```





INTERPRETATION #The histogram

for Top10perc shoes a distribution that is slightly right-skewed, indicating that a moderate number of colleges have a high percentage of students from the top 10% of their high school classes. #the histogram for Top25perc shows the median percentage, indicating that many institutions enroll a considerable proportion of students from the top 25% of their high school class. #with a few elite institutions attracting high-achieving students, while the majority fall within a more moderate range

```
fit = glm(Grad.Rate~., data = college)
fit
```

```
##
## Call:  glm(formula = Grad.Rate ~ ., data = college)
##
## Coefficients:
## (Intercept)   PrivateYes         Apps       Accept       Enroll    Top10perc
##  35.6911902    3.2753438    0.0013583   -0.0008132    0.0022046    0.0771177
##   Top25perc  F.Undergrad  P.Undergrad     Outstate   Room.Board        Books
##   0.0722750   -0.0004704   -0.0015117    0.0010057    0.0019374   -0.0024534
##    Personal          PhD     Terminal    S.F.Ratio  perc.alumni       Expend
##  -0.0016671    0.0776580   -0.0691915    0.0877899    0.2761555   -0.0004404
##    EliteYes
##   2.3526158
##
## Degrees of Freedom: 776 Total (i.e. Null);  758 Residual
## Null Deviance:       229000
## Residual Deviance: 123000    AIC: 6180
```

INTERPRETATION # a private institution (PrivateYes) is associated with a 3.28% higher graduation rate compared to public colleges # as a higher percentage of students from the top 10% and 25% of their high school class with coefficients of 0.077 and 0.072 #he decrease in deviance from 229,000 (null) to 123,000 (residual) indicates a substantial portion of the variation in graduation rates

#chapter-2(10) (a) To begin, load in the Boston data set. The Boston data set is part of the MASS library in R.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
?Boston
```

```
## starting httpd help server ...
```

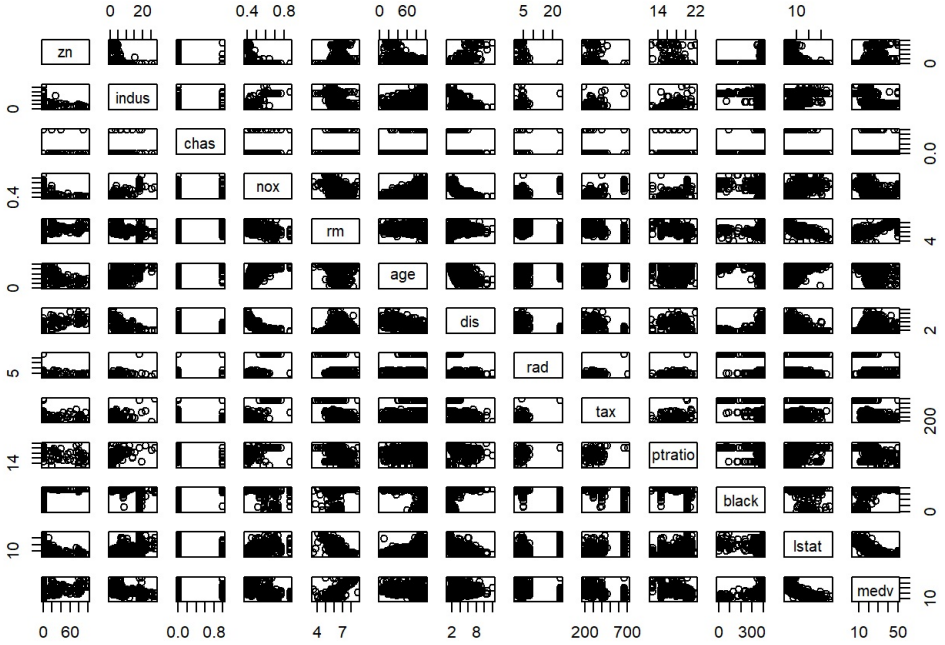
```
##  done
```

```
dim(Boston)
```

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

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

```
 pairs(Boston[,2:14])
```

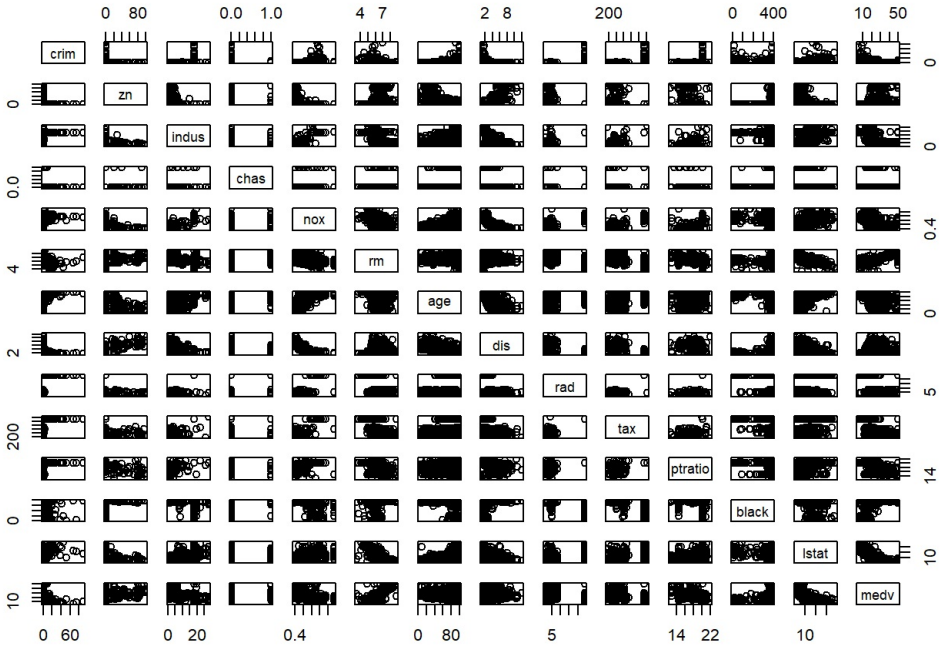


INTERPRETATiON #strong positive

correlation between RM and MEDV indicates that larger homes #a negative correlation between the percentage of lower-status residents (LSTAT) and MEDV suggests that neighborhoods with more lower-status residents #such as NOXand INDUS, exhibit strong correlations

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

```
pairs(Boston)
```



INTERPRETATION # a negative

correlation might be observed between the percentage of lower-status residents #Positive relationship between the average number of rooms per dwelling (RM) and the median value of homes (MEDV), indicating that more rooms typically correspond to higher home values.

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

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

INTERPRETATION # The first quartile suggests that 25% of the data falls below (0.08204) #The median (0.25651) serves as a measure of central tendency #the maximum value of 88.97620 highlights the presence of extreme outliers

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

```
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
```

INTERPRETATION #the minimum value is 0.00632, indicating that some areas experience very low crime rates #the first quartile (0.08205) suggests that 25% of neighborhoods have low crime rates #The first quartile (279.0) and median (330.0) show that the majority of neighborhoods have relatively low to moderate tax rates, with the mean (408.2) indicating a slight skew towards higher values #The first quartile (17.40) and median (19.05) indicate that a majority of neighborhoods, while the mean (18.46) suggests a slight skew toward lower ratios E) How many of the suburbs in this data set bound the Charles river?

```
table(Boston$chas)
```

```
##
##   0   1
## 471  35
```

INTERPRETATION #out of 506 total observations in the Boston dataset, 471 neighborhoods do not border the Charles River (chas = 0), while only 35 neighborhoods are located next to the river (chas = 1). #This significant disparity suggests that only a small proportion (approximately 6.9%) of the neighborhoods in the dataset are near the Charles River (f) What is the median pupil-teacher ratio among the towns in this data set?

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

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

INTERPRETATION #The median is a robust measure of central tendency that represents the middle value when all observations are sorted, meaning that 50% of the neighborhoods have a pupil-teacher ratio below this value and 50% have a ratio
(g) Which suburb of Boston has lowest median value of owner occupied homes? What are the values of the other predictors for that suburb

```
Boston[Boston$medv == min(Boston$medv), ]
```

```
##        crim zn indus chas   nox    rm age    dis rad tax ptratio  black lstat
## 399 38.3518  0  18.1    0 0.693 5.453 100 1.4896  24 666    20.2 396.90 30.59
## 406 67.9208  0  18.1    0 0.693 5.683 100 1.4254  24 666    20.2 384.97 22.98
##     medv
## 399    5
## 406    5
```

#the low median home price, such as a high crime rate (crim), which may reduce the attractiveness of the area Boston has lowest median value of owneroccupied homes?

```
Boston_percentiles <- sapply(Boston[ ,-4], function(x) rank(x)/length(x)) %>%
  as.data.frame()

Boston_percentiles[c(399, 406),]
```

```
##          crim        zn     indus       nox        rm      age        dis
## 399 0.9881423 0.3685771 0.7579051 0.8448617 0.0770751 0.958498 0.05731225
## 406 0.9960474 0.3685771 0.7579051 0.8448617 0.1363636 0.958498 0.04150198
##           rad       tax   ptratio     black     lstat        medv
## 399 0.8705534 0.8606719 0.7519763 0.8814229 0.9782609 0.002964427
## 406 0.8705534 0.8606719 0.7519763 0.3498024 0.8992095 0.002964427
```

INTERPRETATION # rows 399 and 406, the percentile rankings of two specific neighborhoods for all variables in the dataset. These percentile values will indicate how these two neighborhoods rank in terms of features like crime rate (crim), tax rate (tax), median home value (medv) (h) In this data set, how many of the suburbs average more than seven rooms per dwelling? More than eight rooms per dwelling?Comment on the suburbs that average more than eight rooms per dwelling.

#More than seven rooms per dwelling:

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

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

More than eight rooms per dwelling:

```
Boston_8rooms <- Boston[Boston$rm > 8, ]
nrow(Boston_8rooms)
```

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

the percentile rank for these 13 suburbs

```
Boston_8rooms_perc <- Boston_percentiles[as.numeric(rownames(Boston_8rooms)), ]
glimpse(Boston_8rooms_perc)
```

```
## Rows: 13
## Columns: 13
## $ crim    <dbl> 0.34387352, 0.69367589, 0.03557312, 0.52766798, 0.58893281, 0.…
## $ zn      <dbl> 0.3685771, 0.3685771, 0.9950593, 0.3685771, 0.3685771, 0.36857…
## $ indus   <dbl> 0.09683794, 0.91798419, 0.08992095, 0.34090909, 0.34090909, 0.…
## $ nox     <dbl> 0.21541502, 0.68873518, 0.08893281, 0.38833992, 0.38833992, 0.…
## $ rm      <dbl> 0.9802372, 0.9920949, 0.9762846, 0.9861660, 0.9980237, 0.97826…
## $ age     <dbl> 0.48418972, 0.74604743, 0.14130435, 0.50988142, 0.55830040, 0.…
## $ dis     <dbl> 0.5454545, 0.2766798, 0.7480237, 0.4634387, 0.4634387, 0.50296…
## $ rad     <dbl> 0.06422925, 0.49407115, 0.27173913, 0.71640316, 0.71640316, 0.…
## $ tax     <dbl> 0.21541502, 0.63932806, 0.07806324, 0.41600791, 0.41600791, 0.…
## $ ptratio <dbl> 0.37154150, 0.06818182, 0.06818182, 0.26778656, 0.26778656, 0.…
## $ black   <dbl> 0.8814229, 0.4100791, 0.4624506, 0.3537549, 0.3201581, 0.39328…
## $ lstat   <dbl> 0.075098814, 0.035573123, 0.011857708, 0.071146245, 0.09881422…
## $ medv    <dbl> 0.9367589, 0.9851779, 0.9851779, 0.9565217, 0.9851779, 0.93280…
```

the mean of each column to simplify:

```
sapply(Boston_8rooms_perc, mean)
```

```
##       crim         zn      indus        nox         rm        age        dis
## 0.53815749 0.54651870 0.33612040 0.47514442 0.98814229 0.48008513 0.47202797
##        rad        tax    ptratio      black      lstat       medv
## 0.57236242 0.37473396 0.24407115 0.43987534 0.09007297 0.93227425
```

the average percentage of students from lower socio-economic backgrounds (lstat)

#a lower mean lstat that the neighborhoods with larger homes generally have a wealthier population