HW2_Estrada-Rand_Noah

Noah Estrada-Rand

9/8/2019

Problem Set 2

Problem 8

a) Basic Setup

```
setwd("C:/Users/noahe/Desktop/MGSC310")
college <-read.csv("College.csv")
dim(college)
```

```
## [1] 777  19
```

b) First Few Columns of New Data Frame

```
View(college)
rownames(college) = college[,1]
View(college)
college <- college[,-1]
View(college)
head(college)
```

```
##                              Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
## Agnes Scott College              Yes  417    349    137        60
## Alaska Pacific University        Yes  193    146     55        16
## Albertson College                Yes  587    479    158        38
##                              Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University        52        2885         537     7440
## Adelphi University                  29        2683        1227    12280
## Adrian College                      50        1036          99    11250
## Agnes Scott College                 89         510          63    12960
## Alaska Pacific University           44         249         869     7560
## Albertson College                   62         678          41    13500
##                              Room.Board Books Personal PhD Terminal
## Abilene Christian University       3300   450     2200  70       78
## Adelphi University                 6450   750     1500  29       30
## Adrian College                     3750   400     1165  53       66
## Agnes Scott College                5450   450      875  92       97
## Alaska Pacific University          4120   800     1500  76       72
## Albertson College                  3335   500      675  67       73
##                              S.F.Ratio perc.alumni Expend Grad.Rate
## Abilene Christian University      18.1          12   7041        60
## Adelphi University                12.2          16  10527        56
## Adrian College                    12.9          30   8735        54
```

```
## Agnes Scott College                7.7          37  19016        59
## Alaska Pacific University         11.9           2  10922        15
## Albertson College                  9.4          11   9727        55
```

c)

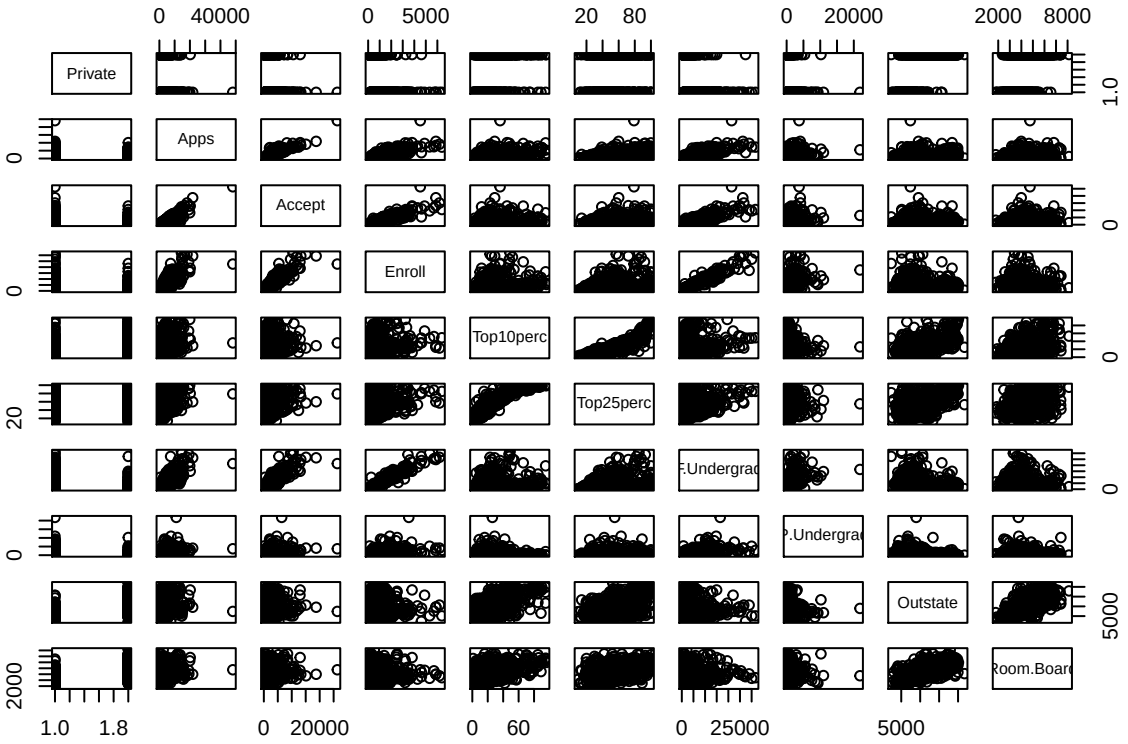
i) Summary Statistics

```
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
```

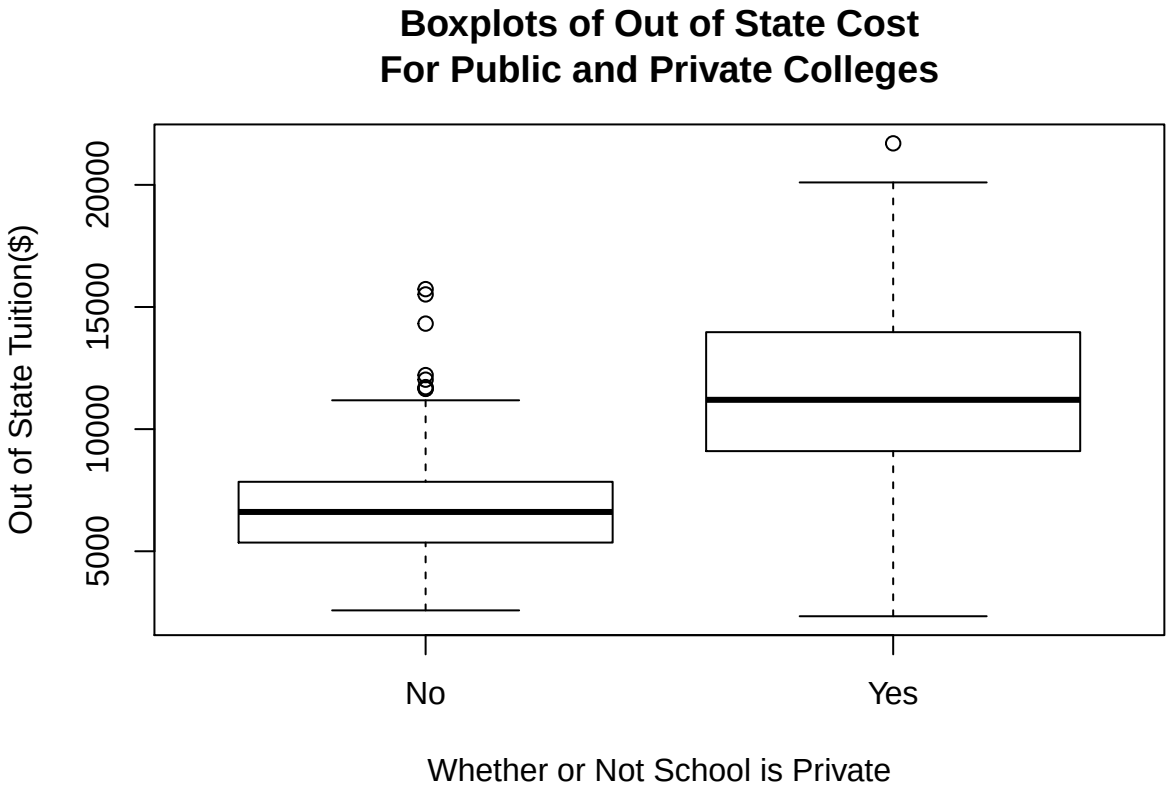
ii) Scatterplot Matrix

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



iii)

```
plot(college$Outstate~college$Private,
     main = "Boxplots of Out of State Cost\nFor Public and Private Colleges",
     ylab = "Out of State Tuition($)",xlab = "Whether or Not School is Private")
```



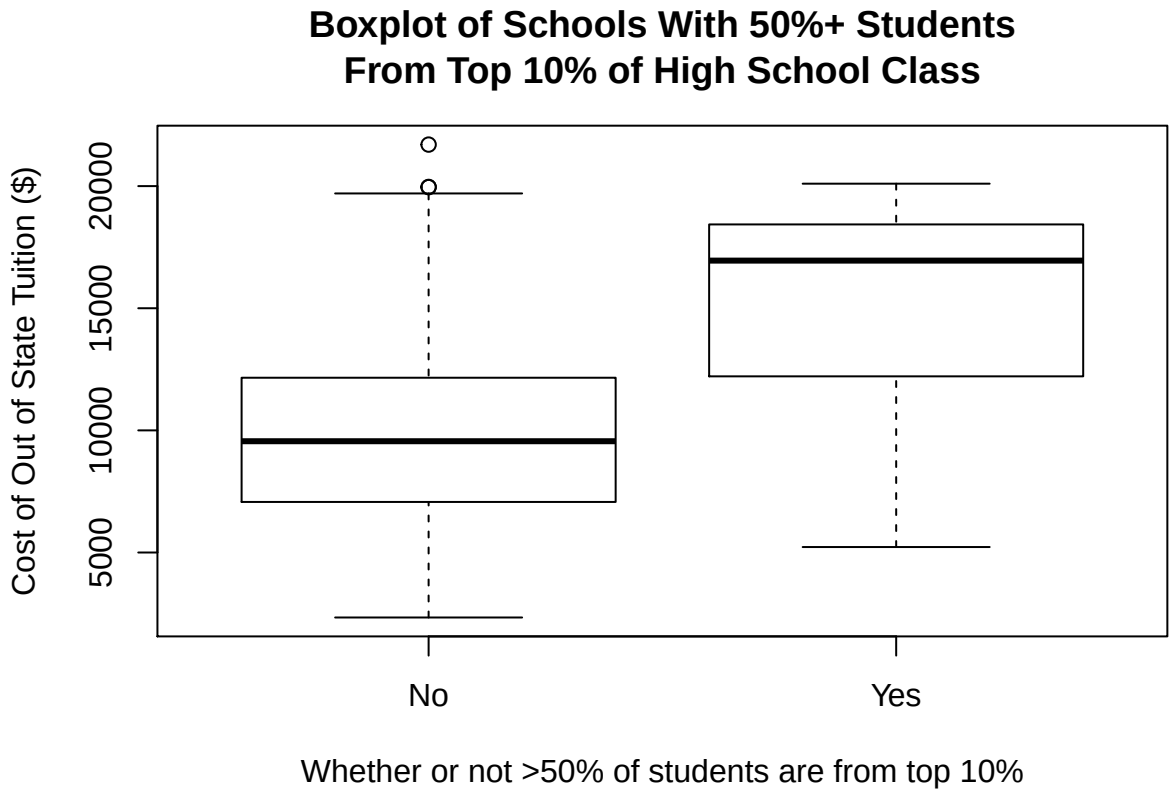
iv) Making New Variable for Levels of Elite Students

```
Elite <- rep("No",nrow(college))
Elite[college$Top10perc >50] <- "Yes"
Elite <- as.factor(Elite)
college <- data.frame(college,Elite)
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 :699
##  1st Qu.: 53.00   Yes: 78
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

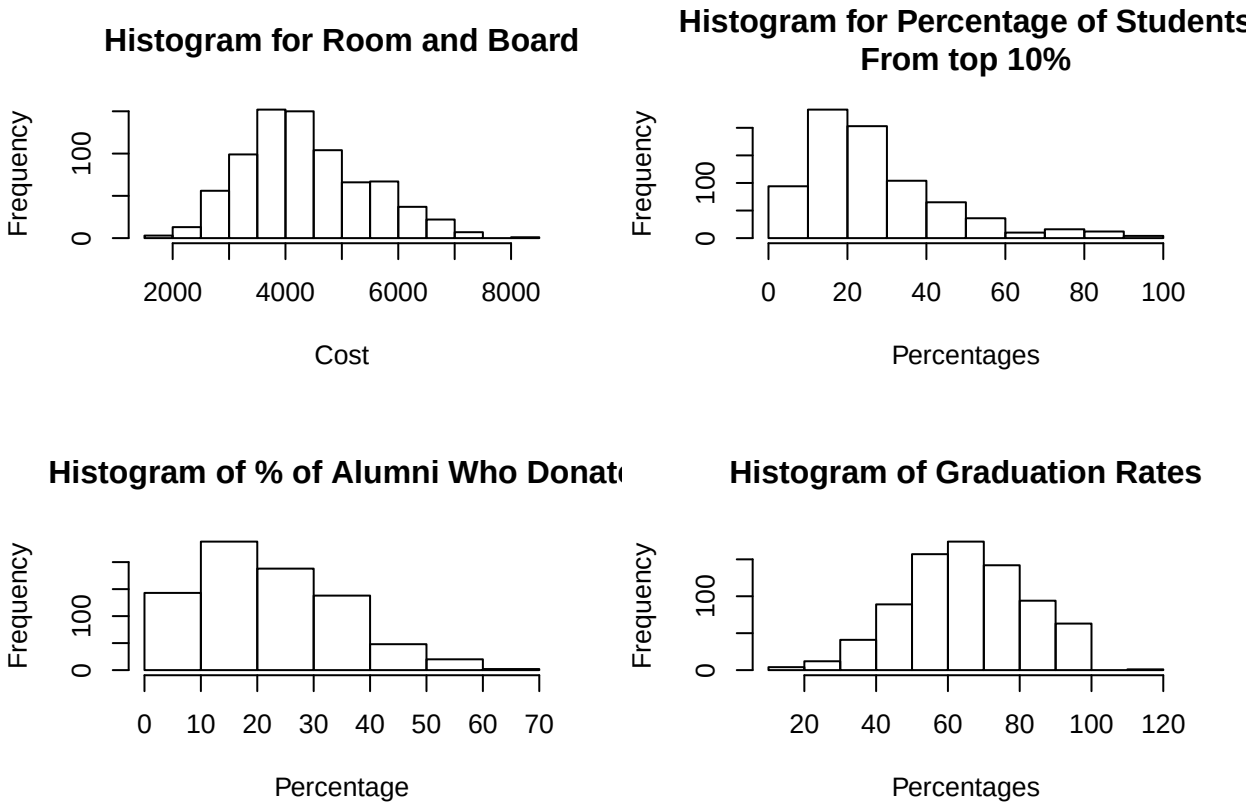
```
plot(college$Outstate~college$Elite, main = "Boxplot of Schools With 50%+ Students\nFrom Top 10% of High
     xlab = "Whether or not >50% of students are from top 10%",
     ylab = "Cost of Out of State Tuition ($)")
```



v)

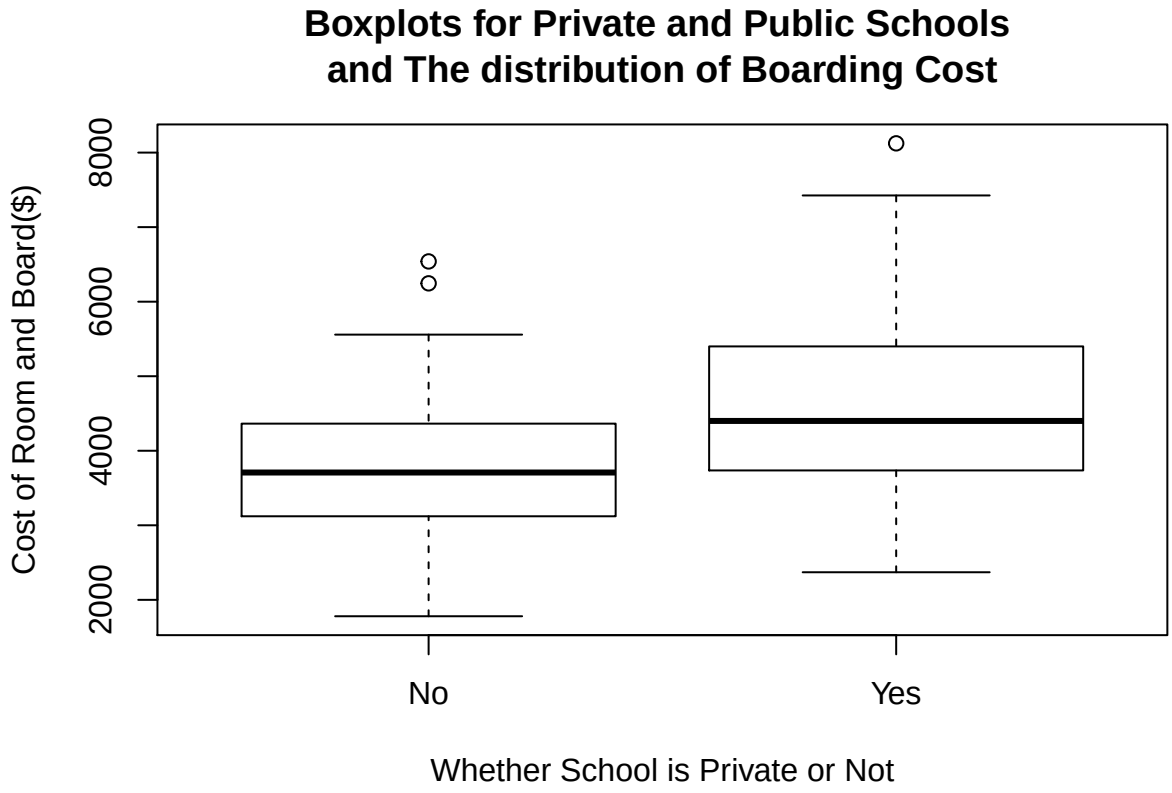
```
par(mfrow = c(2,2))
hist(college$Room.Board, main = "Histogram for Room and Board",
```

```
     xlab = "Cost",ylab = "Frequency",breaks = 15)
hist(college$Top10perc,main = "Histogram for Percentage of Students\nFrom top 10%",
     xlab = "Percentages",ylab = "Frequency", breaks = 10)
hist(college$perc.alumni, main = "Histogram of % of Alumni Who Donate",
     xlab = "Percentage",ylab = "Frequency", breaks = 5)
hist(college$Grad.Rate,main = "Histogram of Graduation Rates",
     xlab = "Percentages",ylab = "Frequency",breaks = 12)
```



vi)

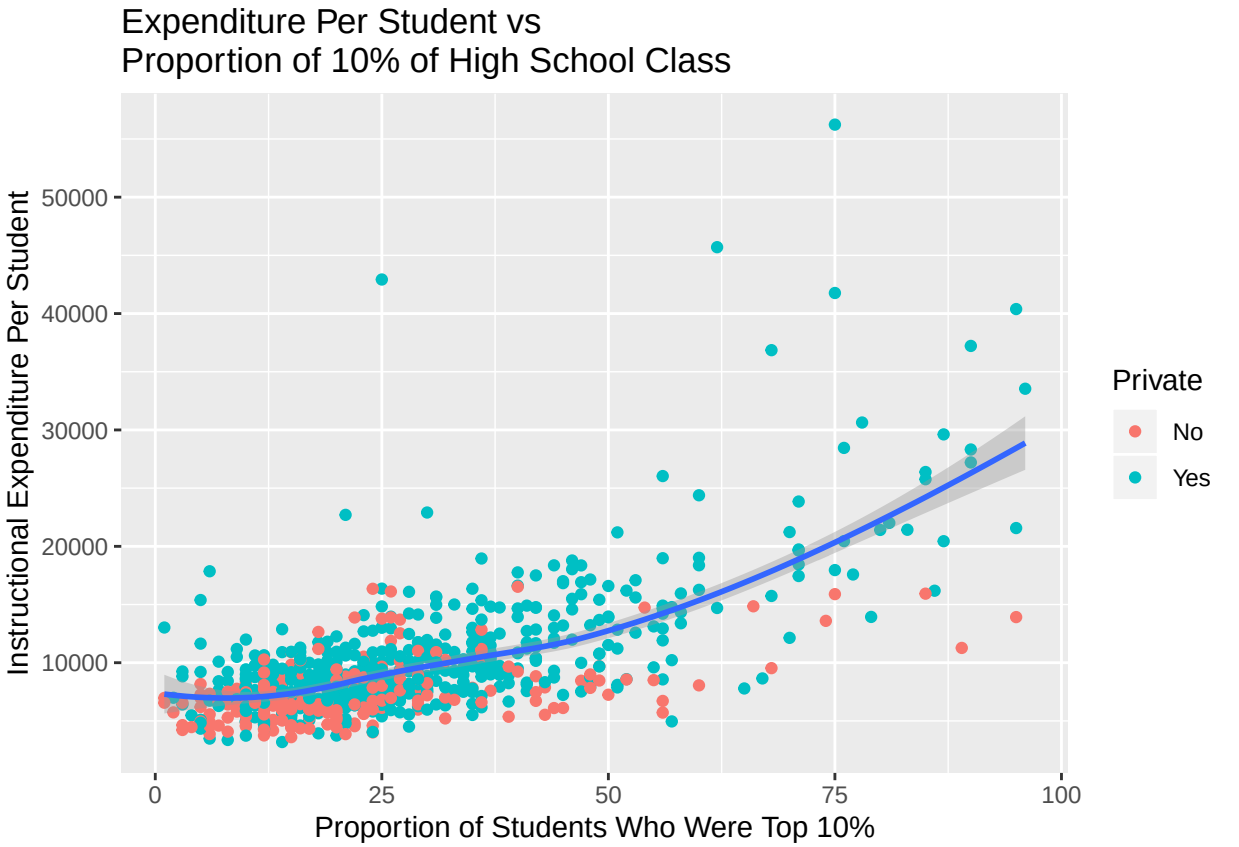
```
boxplot(college$Room.Board~college$Private, xlab = "Whether School is Private or Not",
        ylab = "Cost of Room and Board($)",
        main = "Boxplots for Private and Public Schools \nand The distribution of Boarding Cost")
```



Further analysis of college data highlighted the fact that private schools have a higher room and board cost on average. Moreover, it appears that although both distributions overlap significantly, the inner quartile range of private colleges is significantly higher than that of public schools. Thus, when one attends a private school they should expect higher room and board costs.

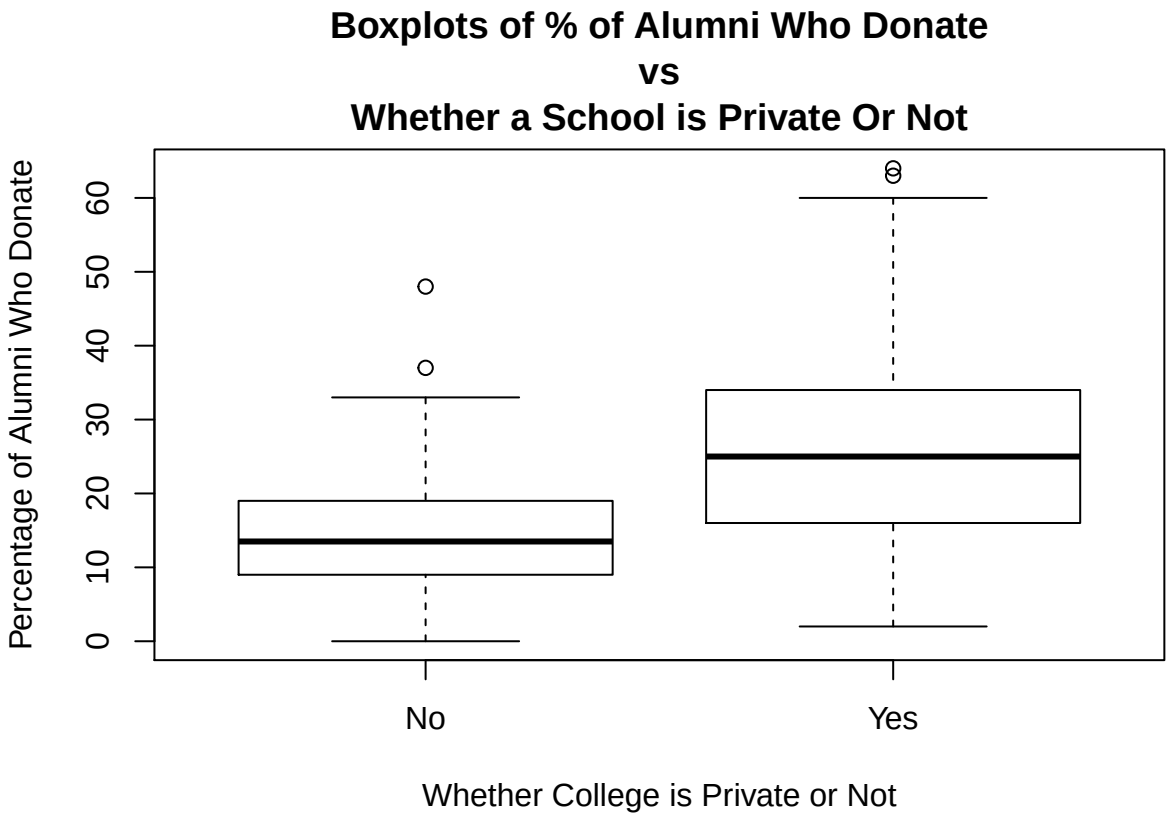
```
ggplot(college,aes(Top10perc,Expend)) + geom_point(aes(Top10perc,Expend,color = Private)) +
  geom_smooth() + labs(title = "Expenditure Per Student vs \nProportion of 10% of High School Class")+
  xlab("Proportion of Students Who Were Top 10%") + ylab("Instructional Expenditure Per Student")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In looking at the percentage of new students taken from the top 10% of their high school class, it was found that this particular variable held a strong positive correlation with instructional expenditure per student. This points to a positive linear relationship in which the higher the percetage of elite students, the higher the instructional expenditure per student. Thus, schools with higher achieving individuals spend more on their instruction than schools with lower achieving students. Furthermore this plot serves to illustrate that the majority of the higher spending schools are private institutions.

```
boxplot(college$perc.alumni~college$Private, xlab = "Whether College is Private or Not",
        ylab = "Percentage of Alumni Who Donate",
        main = "Boxplots of % of Alumni Who Donate\nvs\nWhether a School is Private Or Not")
```



In once again considering the variable of private versus public instituions, it becomes apparent that private schools garner much higher amounts of alumni donations than do public institutions. However, this does not necessarily mean that private instutions dominate this metric. Looking at the boxplots one can observe that private institutions also have a much larger range of proportions of alumni who donate in comparison to public institutions. Yet, it still remains true that private institutions garner a higher percentage of alumni donations on average.

Problem 10

a)

The boston data set has 506 rows and 14 columns. In this particular dataset, each row delineates a different town while each column is a different metric of each town.

```
library(MASS)
?Boston
```

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

```
bos <- Boston
nrow(bos)
```

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

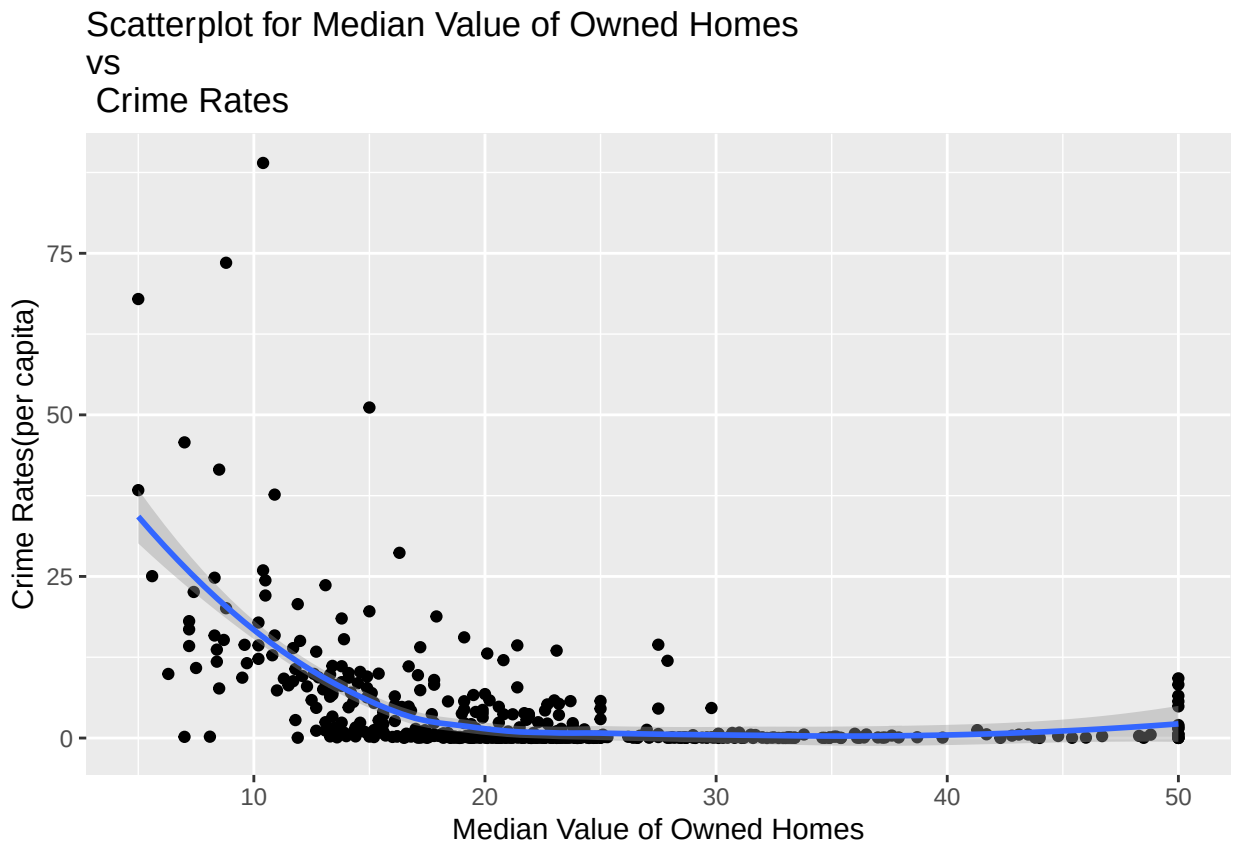
```
ncol(bos)
```

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

b)

```
ggplot(bos,aes(medv,crim)) + geom_point(aes(medv,crim)) + xlab("Median Value of Owned Homes")+
  ylab("Crime Rates(per capita)") + geom_smooth()+
  labs(title = "Scatterplot for Median Value of Owned Homes\nvs\n Crime Rates")
```

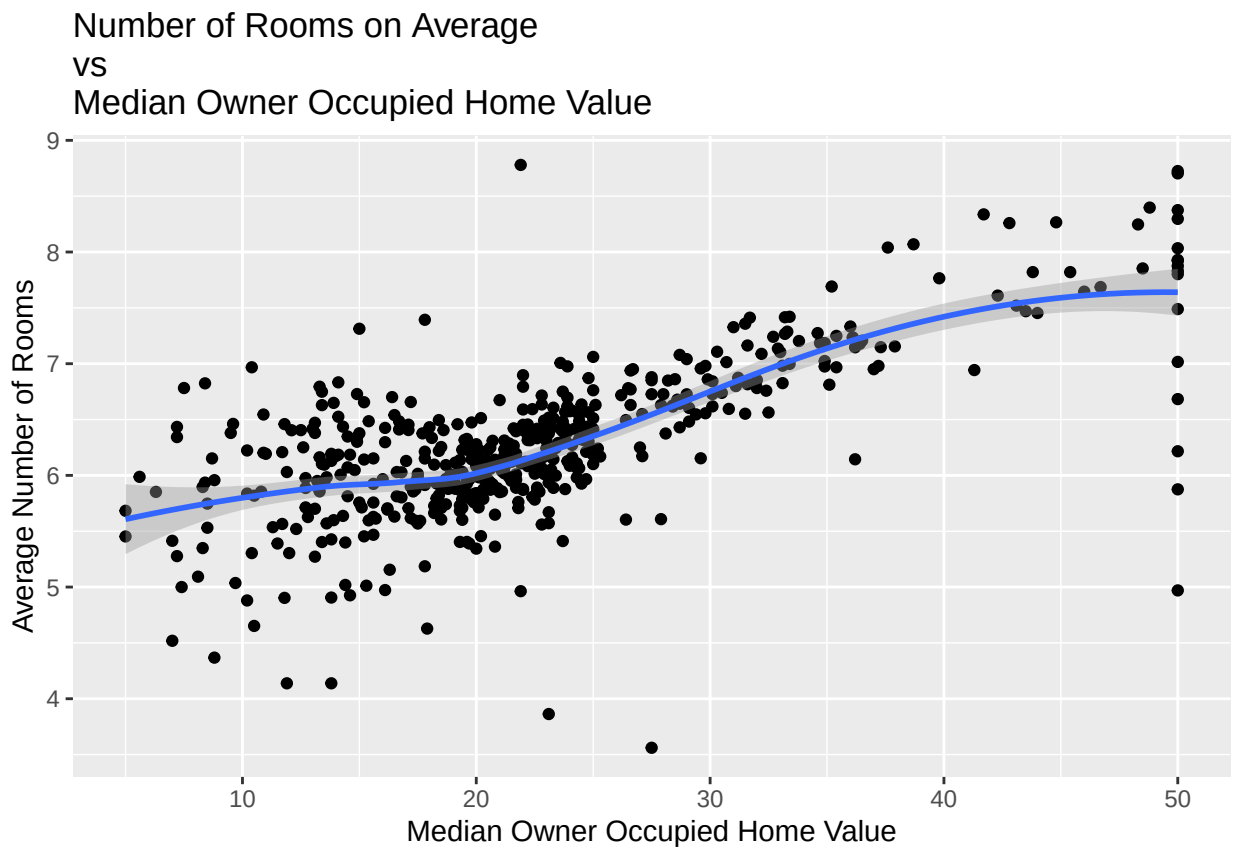
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



From the above graph there appears to a weak negative correlation between the median value of owned homes in the town and the crime rate in the town. When looking at the plot it becomes apparent that the highest crime rates exist in the lower range of median values of owned homes, indicating that wealthier neighborhoods have lower crime rates.

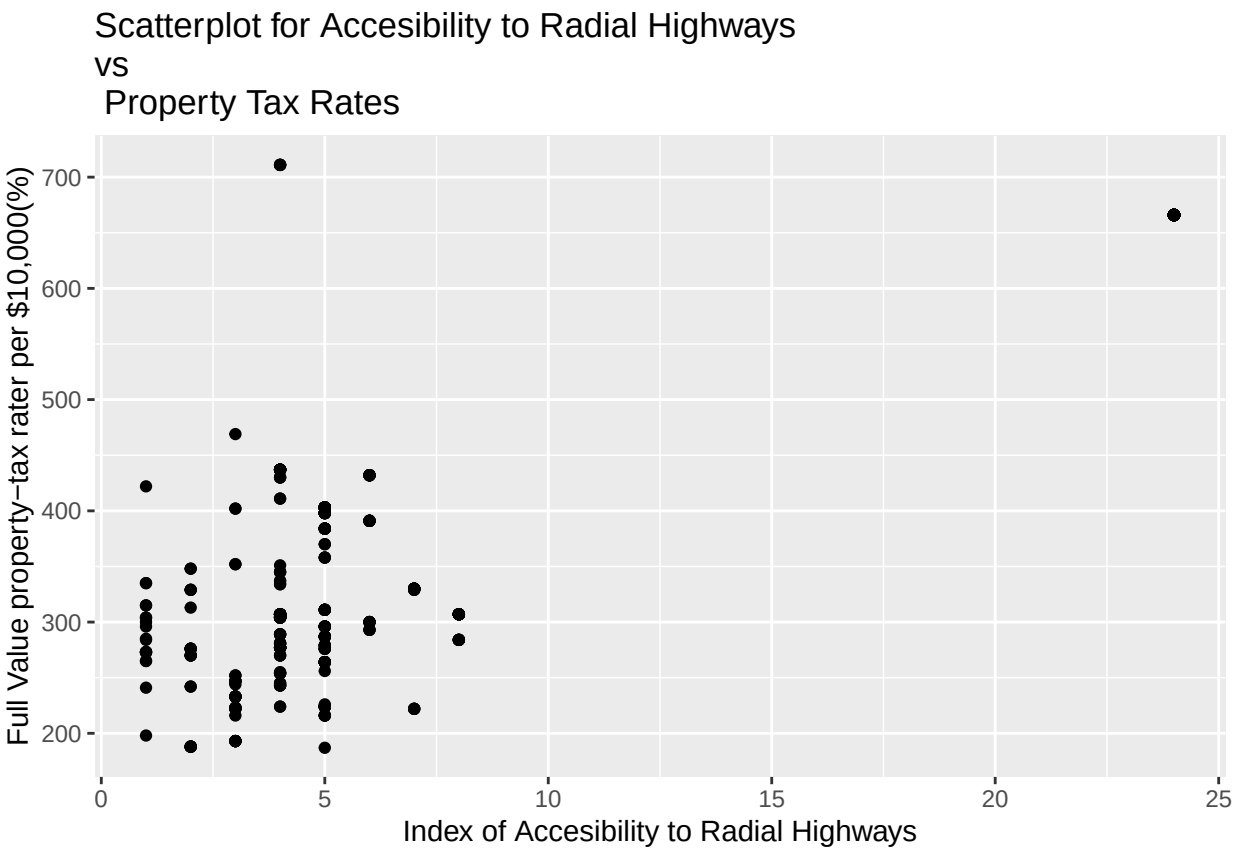
```
ggplot(bos,aes(medv,rm)) + geom_point(aes(medv,rm)) + geom_smooth() +
  labs(title = "Number of Rooms on Average\nvs\nMedian Owner Occupied Home Value") +
  xlab("Median Owner Occupied Home Value") + ylab("Average Number of Rooms")
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



As illustrated by the above scatterplot, it becomes apparent that there is a strong positive correlation between median home values and the number of rooms per dwelling. This poinst to the idea that higher valued homes tend to have more rooms.

```
ggplot(bos,aes(rad,tax)) + geom_point(aes(rad,tax)) + xlab("Index of Accesibility to Radial Highways") +
  ylab("Full Value property-tax rater per $10,000(%)") +
  labs(title = "Scatterplot for Accesibility to Radial Highways\nvs\n Property Tax Rates")
```



When considering possible correlations between accesibility to radial highways and property taxes, there appears to be little correlation if any. While calculating the pearson's correlation coefficient for the two variables indicates a strong positive linear relationship between the two, graphical analysis indicates that the relationship may not necessarily be significant in relation to the real world.

c)

```
cor(bos$crim,bos)
```

```
##      crim         zn     indus        chas       nox         rm       age
## [1,]    1 -0.2004692 0.4065834 -0.05589158 0.4209717 -0.2192467 0.3527343
##             dis       rad       tax   ptratio      black     lstat
## [1,] -0.3796701 0.6255051 0.5827643 0.2899456 -0.3850639 0.4556215
##            medv
## [1,] -0.3883046
```

Looking at the correlation between crime rates and all other variables in the dataset, a few interesting trends emerge. First and foremost, it appears that crime rates have a moderate strength correlation with tax rates in each town. Furthermore, there is a similar correlational relationship between crime rates and percent of the lower status of the population. Yet, the strongest correlation concerning crime rates is found between crime and accesibility to radial highways. A decently strong positive correlation is found between these two variables indicating a positive linear relationship between accesibility to highways and crime. Thus, it would be somewhat reasonable to assume that the closer a town is to highways, the higher the level of crime.
When Looking at negative correlations, median value of owner occupied homes and the proportion of blacks per town are negatively correlated to the level of crime in each town. Thus, crime rate has a negative linear relationship with both the level of black citizens and median home value in each city. This implies that the wealthier the neighborhood, the less crime, as highlighted earlier in this report. These findings present

significant implications when looking at predictors of crime throughout Boston.

d)

```
range(bos$crim)
```

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

```
boxplot.stats(bos$crim)
```

```
## $stats
## [1] 0.00632 0.08199 0.25651 3.67822 8.98296
##
## $n
## [1] 506
##
## $conf
## [1] 0.00391236 0.50910764
##
## $out
##  [1] 13.52220  9.23230 11.10810 18.49820 19.60910 15.28800  9.82349
##  [8] 23.64820 17.86670 88.97620 15.87440  9.18702 20.08490 16.81180
## [15] 24.39380 22.59710 14.33370 11.57790 13.35980 38.35180  9.91655
## [22] 25.04610 14.23620  9.59571 24.80170 41.52920 67.92080 20.71620
## [29] 11.95110 14.43830 51.13580 14.05070 18.81100 28.65580 45.74610
## [36] 18.08460 10.83420 25.94060 73.53410 11.81230 11.08740 12.04820
## [43] 15.86030 12.24720 37.66190  9.33889 10.06230 13.91340 11.16040
## [50] 14.42080 15.17720 13.67810  9.39063 22.05110  9.72418  9.96654
## [57] 12.80230 10.67180  9.92485  9.32909  9.51363 15.57570 13.07510
## [64] 15.02340 10.23300 14.33370
```

When weighing the range of crime rates in suburbs throughout Boston, it becomes glaringly apparent that there are numerous outliers in this dataset. In looking at the list of outliers, the list is extensive, detailing that numerous suburbs have abnormally high levels of crime when compared to the bulk of the remaining data. Furthermore, the range of this metric is large, ranging from .6% to 88.97% per capita indicative of a very large range of crime rates throughout Boston.

```
range(bos$tax)
```

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

```
boxplot.stats(bos$tax)
```

```
## $stats
## [1] 187 279 330 666 711
##
## $n
## [1] 506
##
## $conf
## [1] 302.8173 357.1827
##
## $out
## numeric(0)
```

When considering the range of values for property taxes, there do not appear to be any outliers in the given data. In this case, the lowest tax rate per $10,000 is $187 while the highest is $711.

```
range(bos$ptratio)
```

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

```
boxplot.stats(bos$ptratio)
```

```
## $stats
## [1] 13.60 17.40 19.05 20.20 22.00
##
## $n
## [1] 506
##
## $conf
## [1] 18.85333 19.24667
##
## $out
##  [1] 12.6 12.6 12.6 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0 13.0
## [15] 13.0
```

Now, in consideration of pupil-teacher ratios a few outliers appear when alayzing the range of the predictor. In this case, however, the outliers exist on the lower end of the metric, with only abnormally low values existing. Overall, however, the pupil teacher ratio values range from 12.6 to 22.0 students per teacher in Boston.

e)

There are 35 tracts that bound the river.

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

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

f)

The median pupil to teacher ratio across all towns is 19.05 pupils per teacher.

```
median(bos$ptratio)
```

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

g)

There are two suburbs in Boston with the lowest median value of owner occupied homes is a median value of $5,000. These entries are entries 399 and 406.

```
min(bos$medv)
```

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

The values of the remaining predictors in the suburbs with the lowest median value of owner occupied homes are shown below:

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

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

```
## 406 22.98    5
```

The corresponding ranges are also below:

When comparing the crime rate of the suburbs with the two lowest median value of owner occupied homes to the entire range of crime rates, it becomes clear that suburb 399 is in the lower half of the range while suburb 406 is in the upper half of the range. This is curious as this is the only variable where the two suburbs differ significantly, while all other variables are nearly identical.

```
range(bos$crim)
```

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

When looking at proportion of residential lan zoned for lots over 25,00 square feet, the values of suburbs 399 and 406 score the lowest score possible, 0. This indicates that no lots of residential land are zoned for over 25,000 square feet, indicating that homes and housing may only comprise a small portion of the land in these two suburbs.

```
range(bos$zn)
```

```
## [1]   0 100
```

Regarding the proportion of non-reatail business acres per town, suburbs 399 and 406 fall in the middle of the overall range with both scoring values of 18.1% of land belonging to non-retail businesses. This indicates that the majority of businesses in these areas are retail based businesses.

```
range(bos$indus)
```

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

When analyzing the average number of rooms per dwelling, the suburbs with the lowest median value of owner occupied homes fall close to the middle of the range of values, both around an average of 5 rooms per dwelling.

```
range(bos$rm)
```

```
## [1] 3.561 8.780
```

When considering the predictor of age, suburbs 399 and 406 have 100% of their owner-occupied dwellings built before 1940, indicating that both are older suburbs in the Boston area.

```
range(bos$age)
```

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

In regards to the weighted mean of distance to five Boston employment centers, suburbs 399 and 406 score close to the minimum value, indicating that they are close to places of employment.

```
range(bos$dis)
```

```
## [1]  1.1296 12.1265
```

However, when considering accesibility from these suburbs to raidal highways, they score the highest value of 24, indicating that they are not conveniently located near radial highways.

```
range(bos$rad)
```

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

When considering the full-value property tax rate, suburbs 399 and 406 have among the highest tax rate per $10000. This indicates that while the neighborhood may be older and not as highly appraised as other suburbs, they still experience relatively high rates of property tax.

```
range(bos$tax)
```

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

When weighing the pupil to teacher ratio in these suburbs, they score among the highest ratios both town having a pupil to teacher ratio of 20.2 students per teacher.

```
range(bos$ptratio)
```

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

Further analyzing other predictor variables for suburbs 399 and 406, it appears that these two suburbs score clsoe to the highest score for the proportion of blacks per town.

```
range(bos$black)
```

```
## [1]   0.32 396.90
```

Overall these findings present curious implications. While the two suburbs analyzed above have the lowest median value of owner-occupied homes, they are among one of the highest taxed within the data set. Furthermore, the fact that zero percent of both suburbs are zone for residential lots over 25,000 square feet, it becomes apparent that the homes there are smaller.

H)

From the dataset, there are 64 total suburbs which average more than 7 rooms per dwelling.

```
nrow(bos[bos$rm >7,])
```

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

Further analyses revealed that only 13 suburbs average more than 8 rooms per dwelling. This is only a small proportion of the total neighborhoods, indicating that far fewer rooms is the norm. However, when looking at other predictors of these neigborhoods, it becomes clear that these neighborhoods also have other distinct characteristics. For instance, their crime rate are nearly zero, indicating a safer environment overall. Furthermore, the median value of owner occupied homes ranges from $21,000 to $50,000 in value. Lastly, the pupil teacher ratios are also in the lower end of that predictors values, indicating a more personalized education for the children therein. Observing the trends in this data regarding these suburbs with homes averaging more than 8 rooms, it becomes clear that these suburbs are much wealthier overall with many of the variables supporting this conclusion.

```
nrow(bos[bos$rm >8,])
```

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