STA-6543 Assignment #1

Austin Thrash - FFK221

2024-01-25

2. Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p.

(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 problem would be a *regression* problem because we are looking to see how our predictors affect CEO salary. We are most interested in *inference* as we are looking to find out how our predictors affect our outcome.
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 problem would be *classification* since we are trying to determine whether or not the new product is a success or a failure. For determining this, we are most interested in *prediction.*
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 scenario is a *regression* problem as we are looking to find a continuous numerical value. In this situation we are most interested in *prediction*, since we are attempting to predict the % change.
n = 52 p = 3

5. What are the advantages and disadvantages of a very flexible (versus a less flexible) approach for regression or classification? Under what circumstances might a more flexible approach be preferred to a less flexible approach? When might a less flexible approach be preferred?

A very flexible approach can provide the advantage of recognizing more complex patterns which can be helpful when relationships are non-linear. It can also provide the advantage of higher accuracy by reducing bias and possibly avoid making assumptions about underlying relationships in the data; This could be achieved by

using Random Forest to identify these hidden biases that simpler models potential miss. One of the main disadvantages to using a more flexible model is that the model can become over fit and capture noise in the data thus making the model less accurate.
A less flexible approach has the advantage of being more generalized which allows the model to avoid noise and focus on essential features or predictors. With generalization comes better interpretability; with fewer parameters and simpler decisions, the model becomes easier to understand. Another advantage is the reduction of overfitting, simpler models are less likely to become over fit, ignoring noise in the data. The disadvantages to a less flexible model are the opposite of a flexible model advantages, meaning its less likely to identify complex relationships and has a higher bias leading to inaccurate predictions.

6. Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a non-parametric approach)? What are its disadvantages?

Parametric methods are a two-step modeling process. The first step includes making an assumption about the functional form (or shape) of f. Such as making the assumption that f is linear in X. The second step is to fit or train the model, the most common approach is referred to as least squares; however this is just one of the many ways to fit models. Overall a parametric approach reduces the estimation of f down to estimating a set of parameters. The main difference with non-parametric methods is that they do not make assumptions about the functional form of f. Non-parametric models are more flexible and allow for more complex relationships to be identified. The advantages of a parametric approach is its simplicity and easy interpretation. However, these advantages could be interpreted as disadvantages as parametric models have limited flexibility and could possibly miss complex relationships. Incorrect assumptions about the data could also lead to inaccurate results, another disadvantage of parametric models.

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

(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.

```
setwd("~/Downloads")
college <- read.csv('College.csv', stringsAsFactors = TRUE)
```

(b) Look at the data using the View() function.

```
# The view() commands are commented out to avoid issues when knitting the file to pdf
rownames(college) <- college[,1]
#View(college)

college <- college[,-1]
#View(college)
```

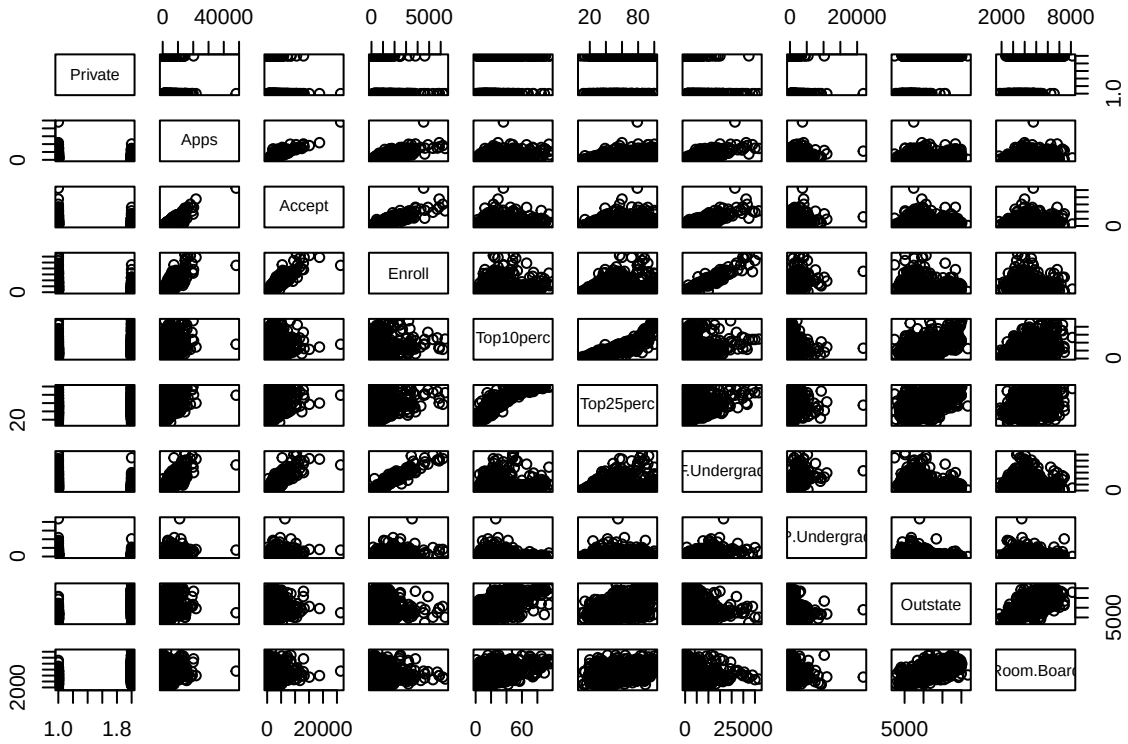
(c) i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
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
```

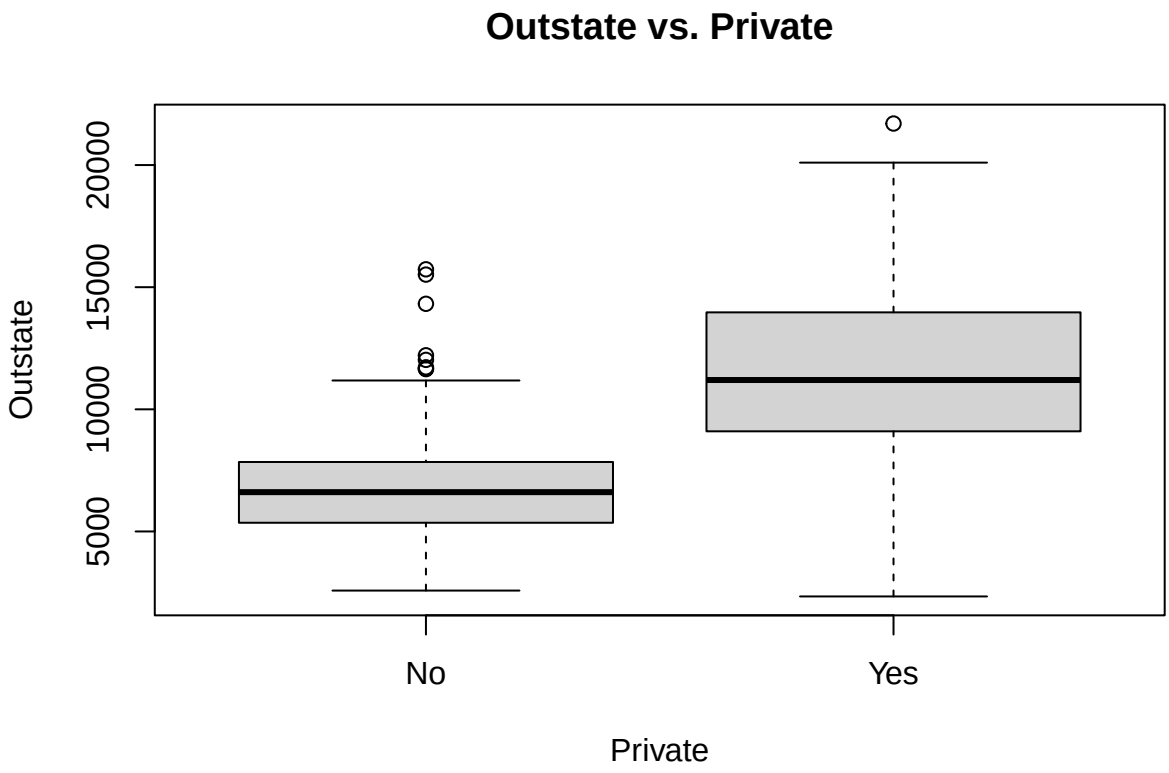
ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

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



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

```
plot(college$Private, college$Outstate,
     main = "Outstate vs. Private",
     xlab = "Private", ylab = "Outstate")
```



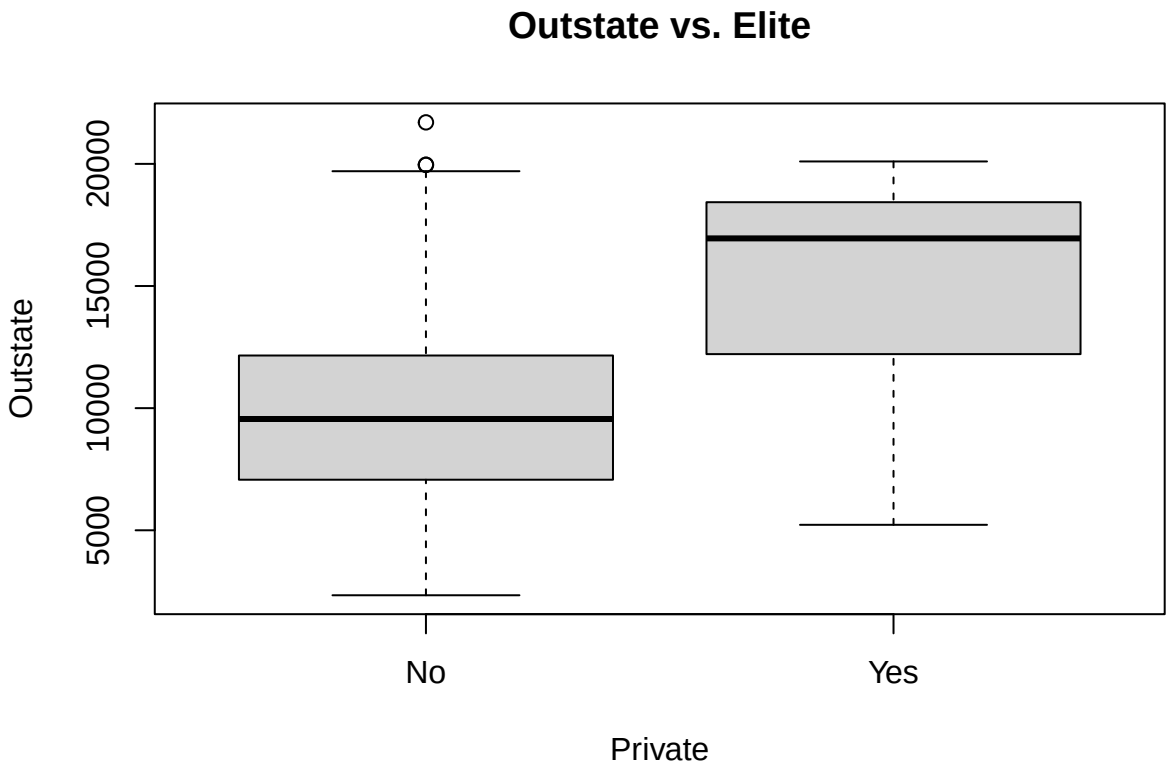
iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50 %. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

```
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$Elite, college$Outstate, main = "Outstate vs. Elite",
     xlab = "Private", ylab = "Outstate")
```

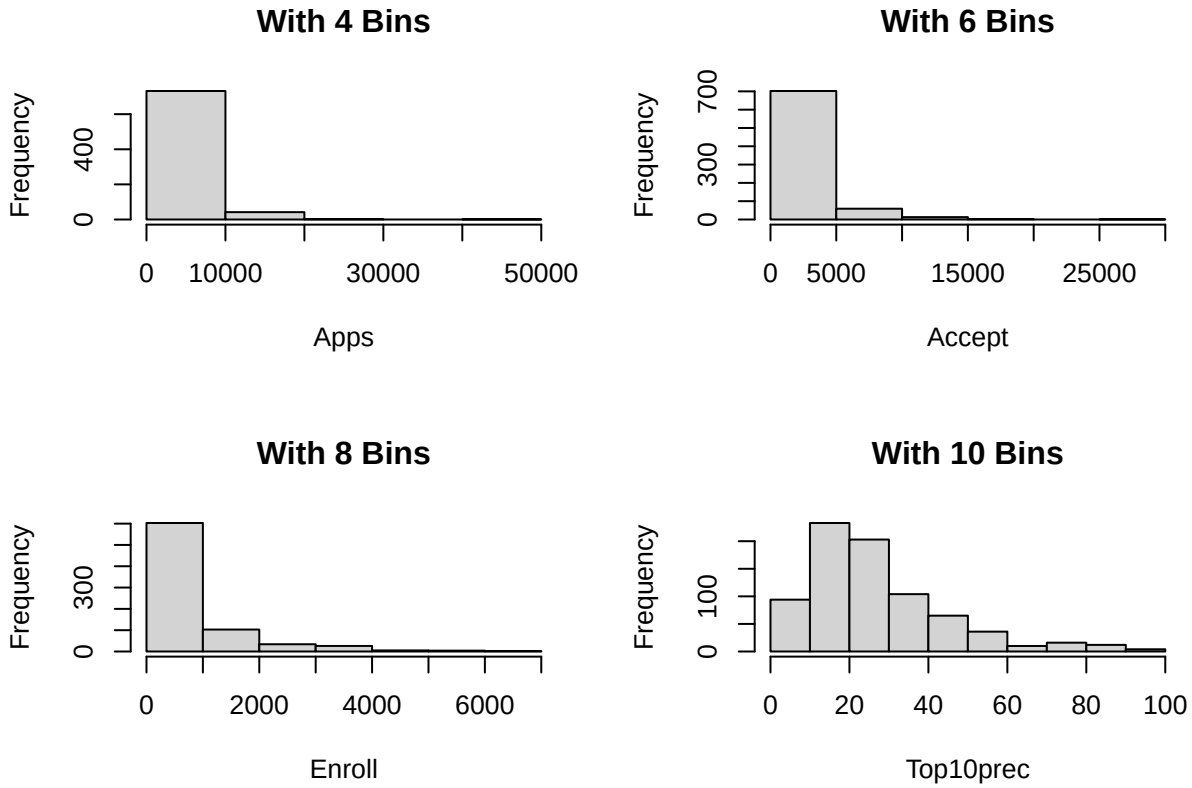


v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow = c(2, 2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments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

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

hist(college$Apps, breaks = 4, main = "With 4 Bins", xlab = "Apps")
hist(college$Accept, breaks = 6, main = "With 6 Bins", xlab = "Accept")
hist(college$Enroll, breaks = 8, main = "With 8 Bins", xlab = "Enroll")
```

```
hist(college$Top10perc, breaks = 10, main = "With 10 Bins", xlab = "Top10prec")
```



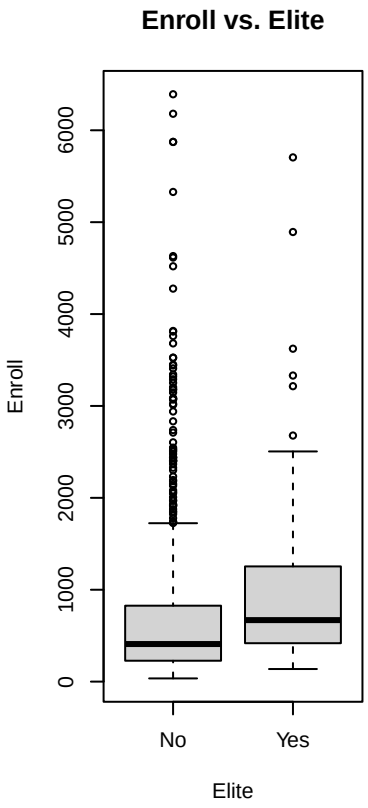
vi. Continue exploring the data, and provide a brief summary of what you discover.

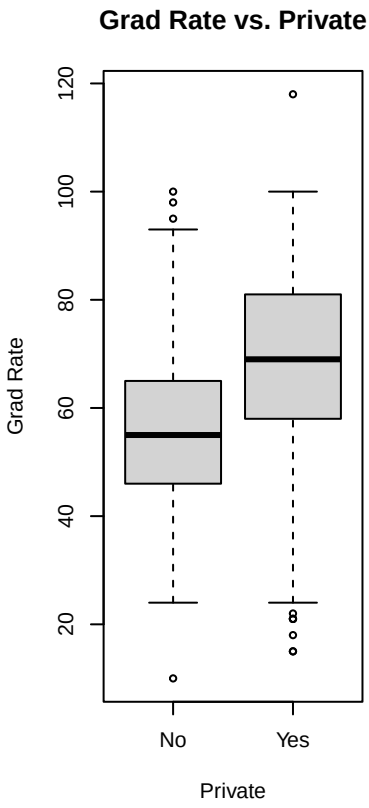
```
library(corrplot)
```

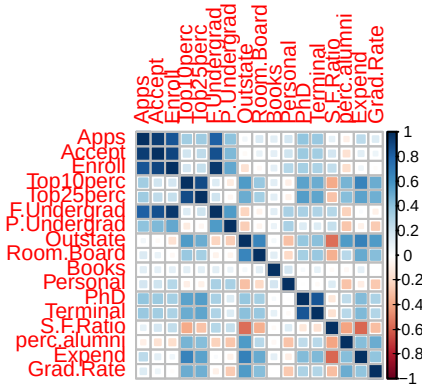
```
## corrplot 0.92 loaded
```

```
par(mfrow = c(1, 3))
plot(college$Elite, college$Enroll,
     main = "Enroll vs. Elite",
     xlab = "Elite", ylab = "Enroll")
plot(college$Private, college$Grad.Rate,
     main = "Grad Rate vs. Private",
     xlab = "Private", ylab = "Grad Rate")

correlation <- round(cor(college[,2:18]), 2)
corrplot(correlation, method = "square")
```







Upon further exploration, I made two additional box plots one being Enroll vs Elite and Grad Rate vs Private. One observation we can make about the data based on the first plot is that elite schools tend to have higher enrollment then non-elite schools and that the overall variability of elite schools is not as great as non-elite schools. Another observation can be made from the second plot, we can notice that private schools tend to have a higher graduation rate than non-private schools. We can also make a correlation plot between each of the variables and can notice a few that have high correlation such as Outstate - Expend, Outsate - Top10prec, and Expend - Top10prec.

9. This exercise involves the Auto data set studied in the lab. Make sure that the missing values have been removed from the data.

(a) Which of the predictors are quantitative, and which are qualitative?

```
library(ISLR2)

sum(is.na(Auto))
```

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

There are not any missing values in the data.

- Quantitative:
  -mpg
  -cylinders
  -displacement
  -horsepower

-weight
-acceleration
-year

- Qualitative:
  -origin
  -name

(b) What is the range of each quantitative predictor? You can answer this using the range() function.

```
variables <- c("mpg", "cylinders", "displacement", "horsepower", "weight", "acceleration", "year")

# Applying range to each of the variables in the Auto dataset and labeling the rows
ranges <- apply(Auto[, variables], 2, range)
rownames(ranges) <- c("min", "max")
print(ranges)
```

```
##      mpg cylinders displacement horsepower weight acceleration year
## min  9.0         3           68         46   1613          8.0   70
## max 46.6         8          455        230   5140         24.8   82
```

(c) What is the mean and standard deviation of each quantitative predictor?

```
# Using sapply() since we are working with a data frame, applying mean to all variables
print("Mean:")
```

```
## [1] "Mean:"
```

```
sapply(Auto[, variables], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    23.445918     5.471939   194.411990   104.469388  2977.584184    15.541327
##         year
##    75.979592
```

```
print("Standard deviation:")
```

```
## [1] "Standard deviation:"
```

```
sapply(Auto[, variables], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.805007     1.705783   104.644004    38.491160   849.402560     2.758864
##         year
##     3.683737
```

Above is the mean and standard deviation of each quantitative predictor

(d) Now remove the 10th through 85th observations. What is the range, mean, and standard deviation of each predictor in the subset of the data that remains?

```
subset_Auto <- Auto[-c(10:85),]
ranges <- apply(subset_Auto[, variables], 2, range)
rownames(ranges) <- c("min", "max")
print(ranges)
```

```
##      mpg cylinders displacement horsepower weight acceleration year
## min 11.0         3           68         46   1649          8.5   70
## max 46.6         8          455        230   4997         24.8   82
```

```
print("Mean:")
```

```
## [1] "Mean:"
```

```
sapply(subset_Auto[, variables], mean)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##    24.404430     5.373418   187.240506   100.721519  2935.971519    15.726899
##         year
##    77.145570
```

```
print("Standard deviation:")
```

```
## [1] "Standard deviation:"
```

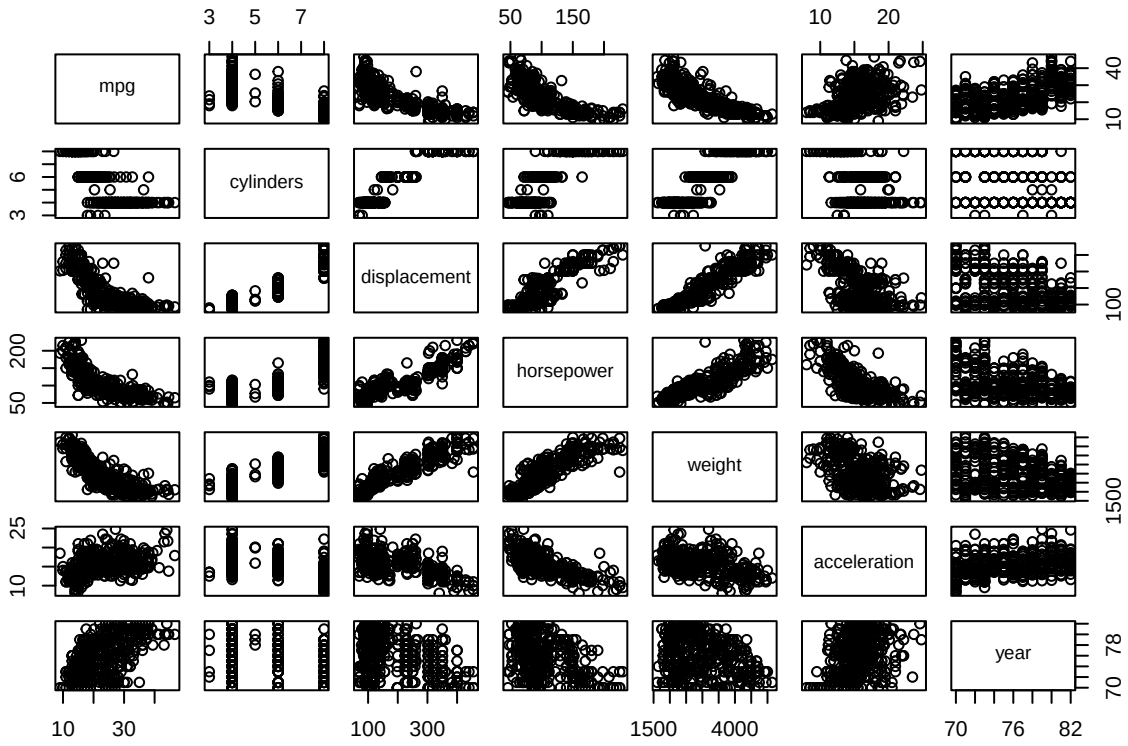
```
sapply(subset_Auto[, variables], sd)
```

```
##          mpg    cylinders displacement   horsepower       weight acceleration
##     7.867283     1.654179    99.678367    35.708853   811.300208     2.693721
##         year
##     3.106217
```

Above is the ranges, means, and standard deviations for each of the predictors.

(e) Using the full data set, investigate the predictors graphically, using scatterplots or other tools of your choice. Create some plots highlighting the relationships among the predictors. Comment on your findings.

```
pairs(Auto[,1:7])
```

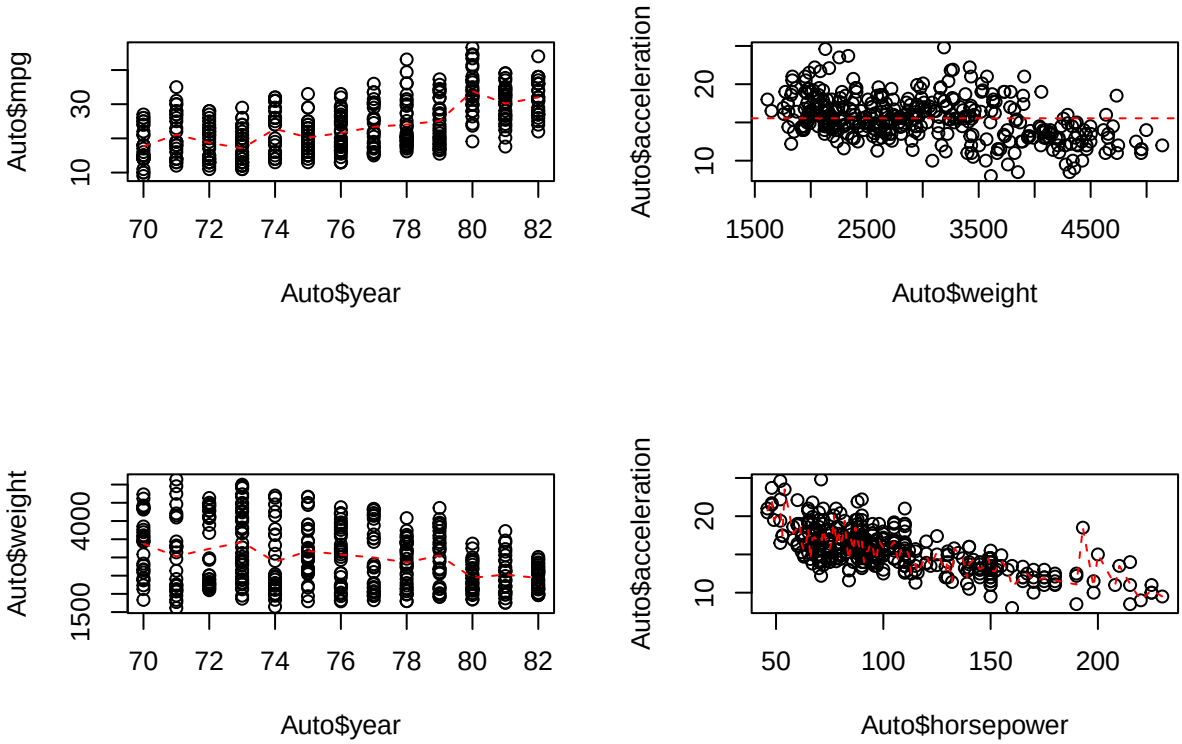



```
par(mfrow = c(2, 2))
plot(Auto$year, Auto$mpg)
mpg_means <- tapply(Auto$mpg, Auto$year, mean)
lines(names(mpg_means), mpg_means, col = "red", lty = 2)

plot(Auto$weight, Auto$acceleration)
abline(h = mean(Auto$acceleration), col = "red", lty = 2)


plot(Auto$year, Auto$weight)
weight2_means <- tapply(Auto$weight, Auto$year, mean)
lines(names(weight2_means), weight2_means, col = "red", lty = 2)

plot(Auto$horsepower, Auto$acceleration)
horsepower_means <- tapply(Auto$acceleration, Auto$horsepower, mean)
lines(names(horsepower_means), horsepower_means, col = "red", lty = 2)
```



From the pairs we can notice a few predictors that seem to have strong relationships. Some of these relationships include year vs mpg, acceleration vs weight, year vs weight, and acceleration vs horsepower.
Year vs MPG:
- In this plot we can notice that the mean MPG for vehicles increase over time indicating a positive relationship.

Acceleration vs Weight:
- This plot seems to indicate a negative relationship between weight and acceleration; as weight increases, acceleration decreases.

Year vs Weight:
- From this plot we notice a negative relationship between year and weight, as the year increases, the mean weight for vehicles decreases.

Acceleration vs horsepower:
- This plot suggests that as horsepower increases, acceleration decreases, indicating another negative relationship.

(f) Suppose that we wish to predict gas mileage (mpg) on the basis of the other variables. Do your plots suggest that any of the other variables might be useful in predicting mpg? Justify your answer.
Based on the previous plots we made, we can notice strong relationships in mpg and the following variables:
- weight
- displacement
- horsepower
- cylinders

All of these variables negatively affect mpg, for example, as the weight of the vehicle increases, the mpg decreases. These variables might prove to be useful when attempting to predict mpg. This can be visualized in the graphs with a constant downward trend when comparing mpg to the listed variables above.

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

```
library(ISLR2)
library(MASS)
```

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

```
## The following object is masked from 'package:ISLR2':
##
##     Boston
```

```
dim(Boston)
```

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

crim
per capita crime rate by town.

zn
proportion of residential land zoned for lots over 25,000 sq.ft.

indus
proportion of non-retail business acres per town.

chas
Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).

nox
nitrogen oxides concentration (parts per 10 million).

rm
average number of rooms per dwelling.

age
proportion of owner-occupied units built prior to 1940.

dis
weighted mean of distances to five Boston employment centres.

rad
index of accessibility to radial highways.

tax
full-value property-tax rate per $10,000.

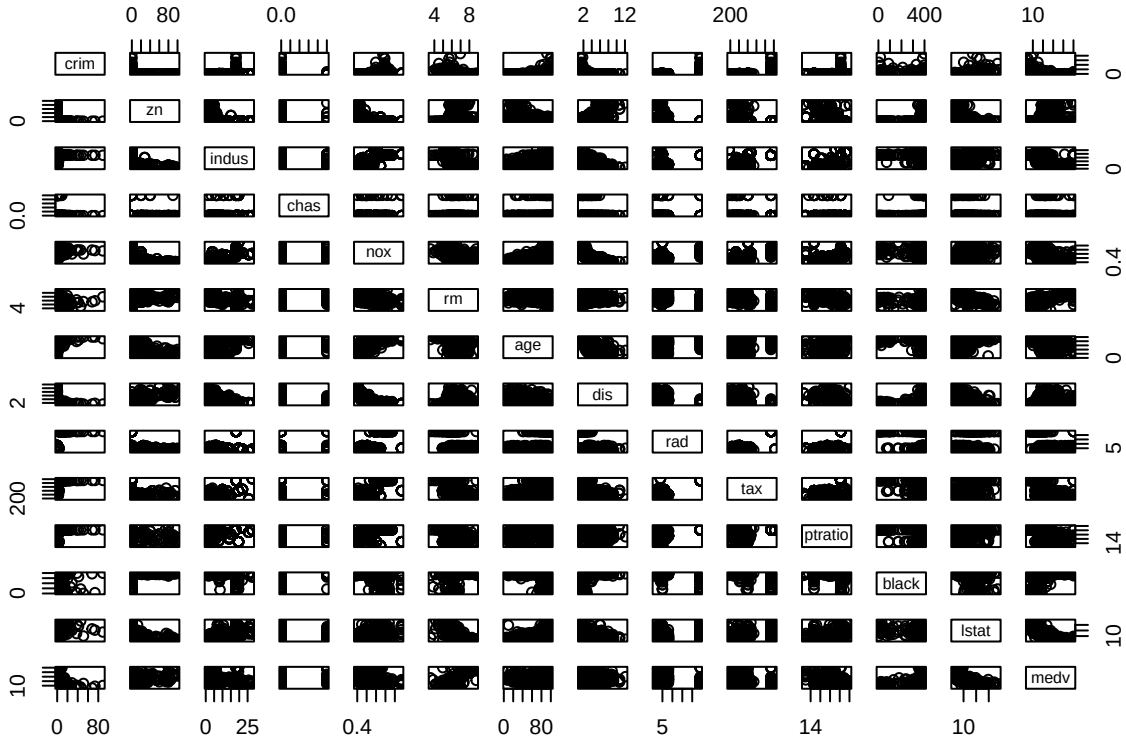
ptratio
pupil-teacher ratio by town.

lstat
lower status of the population (percent).

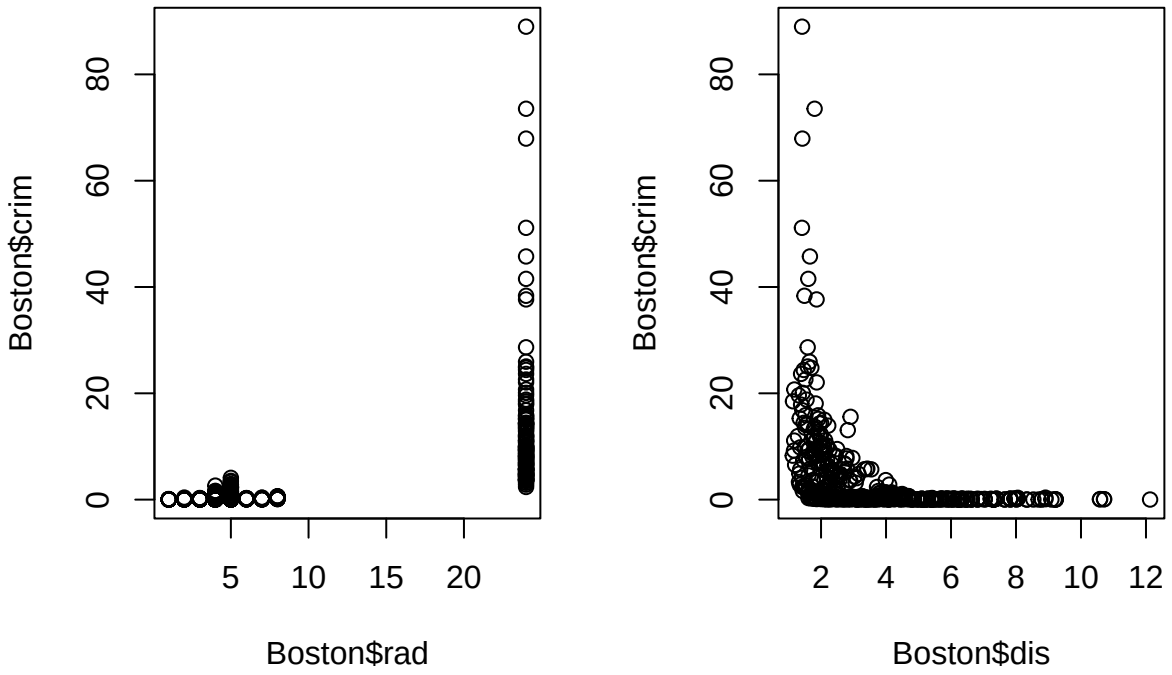
medv
median value of owner-occupied homes in $1000s.

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

```
pairs(Boston)
```




```
par(mfrow = c(1, 2))
plot(Boston$rad, Boston$crim)
plot(Boston$dis, Boston$crim)
```



Based on our findings, the plots do not seem to indicate much relationship wise, however we can look at the associations closer in the next question as using some sort of selection method may provide more clear results.

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

```
full.model <- lm(crim ~ ., data = Boston)

step.model <- step(full.model, direction = "both",
                   trace = FALSE)
summary(step.model)
```

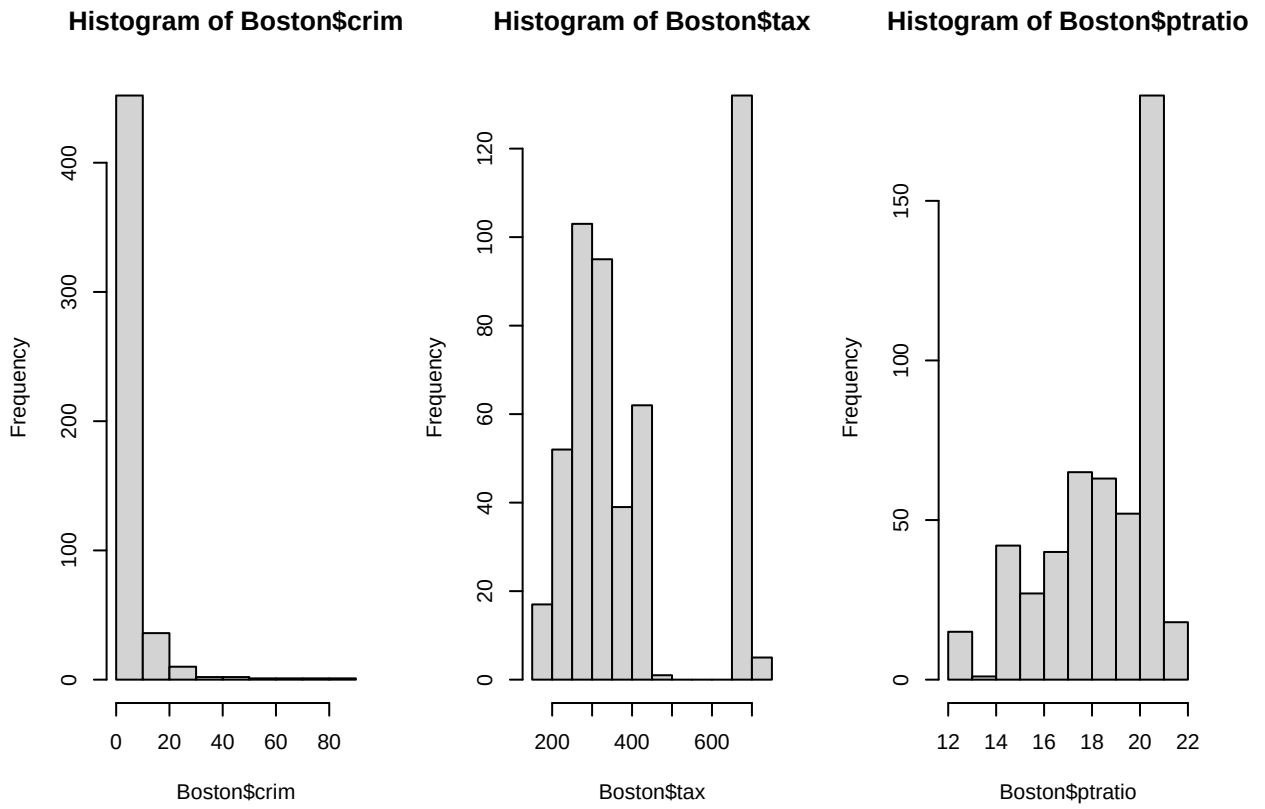
```
##
## Call:
## lm(formula = crim ~ zn + nox + dis + rad + ptratio + black +
##     lstat + medv, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.860 -2.102 -0.363  0.895 75.702
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.683128   6.086010   3.234 0.001301 **
## zn            0.043293   0.017977   2.408 0.016394 *
## nox         -12.753708   4.760157  -2.679 0.007623 **
## dis          -0.918318   0.261932  -3.506 0.000496 ***
## rad           0.532617   0.049727  10.711  < 2e-16 ***
## ptratio      -0.310541   0.182941  -1.697 0.090229 .
## black        -0.007922   0.003615  -2.191 0.028897 *
## lstat         0.110173   0.069219   1.592 0.112097
## medv         -0.174207   0.053988  -3.227 0.001334 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.428 on 497 degrees of freedom
## Multiple R-squared:  0.4505, Adjusted R-squared:  0.4416
## F-statistic: 50.92 on 8 and 497 DF,  p-value: < 2.2e-16
```

Since the pair() function gave us a lot to look at, making it harder to interpret, we can utilize step wise model selection to identify which predictors might be the best when predicting the 'crim' variable. From our results we can notice that 'rad', 'dis', 'nox', and 'medv' seem to be the strongest predictors. The results suggest that 'rad' has a strong positive association with 'crim' while 'dis' has a strong negative association, this can be seen when looking at the estimate value for both of these variables. The association with 'crim' and the remaining variables is not as prevalent indicated by having estimate values closer to zero.

(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.

```
par(mfrow = c(1, 3))
hist(Boston$crim)
hist(Boston$tax)
hist(Boston$ptratio)
```




```
vars <- c("crim", "tax", "ptratio")
ranges <- apply(Boston[,vars], 2, range)
rownames(ranges) <- c("min", "max")
print(ranges)
```

```
##         crim tax ptratio
## min  0.00632 187    12.6
## max 88.97620 711    22.0
```

These ranges seem to have large variability which could possibly indicate considerable inequality in terms of crime rate, tax rates, and pupil-teacher ratios. In the histograms we do seem to notice considerable spikes in tax rates near 600 and pupil-teacher ratio around 21. These large spikes could indicate that some tracts do indeed have significantly higher tax rates and potential inequalities involving educational resources in some parts of Boston.

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

```
nrow(Boston[Boston$chas == 1,])
```

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

We can see that 35 census tracts in this data set bound the Charles river.

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

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

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

From our results, the median pupil-teacher ratio among the towns is 19.05

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

(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.

```
##        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
```

Census tract 399 and 406 have the lowest median value of owner-occupied homes with a value of 5. Comparing the predictors of these two tracts seem to be pretty similar, when compared to the overall ranges of these predictors we can see that they sit close to the average of these ranges.

```
dim(subset(Boston,rm>7))[1]
```

(h) In this dataset, 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.

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

```
dim(subset(Boston,rm>8))[1]
```

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

64 census tracts average more than seven rooms per dwelling and 13 census tracts average more than eight rooms per dwelling.

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

```
##         crim           zn        indus         chas          nox           rm
##   3.61352356  11.36363636  11.13677866   0.06916996   0.55469506   6.28463439
##          age          dis          rad          tax      ptratio        black
##  68.57490119   3.79504269   9.54940711 408.23715415  18.45553360 356.67403162
##        lstat         medv
##  12.65306324  22.53280632
```

```
sapply(Boston_8rooms, mean)
```

```
##        crim          zn       indus        chas         nox          rm
##   0.7187954  13.6153846   7.0784615   0.1538462   0.5392385   8.3485385
##         age         dis         rad         tax     ptratio       black
##  71.5384615   3.4301923   7.4615385 325.0769231  16.3615385 385.2107692
##       lstat        medv
##   4.3100000  44.2000000
```

When compared to the overall average, census tracts that average more than 8 rooms per dwelling averages do not vary much. However some interesting observations can be made with crim, tax, and lstat. We can see that with 8 rooms per dwelling crime and tax rate average is lower than that of the overall average, however the lower status of the population increases in these settings.