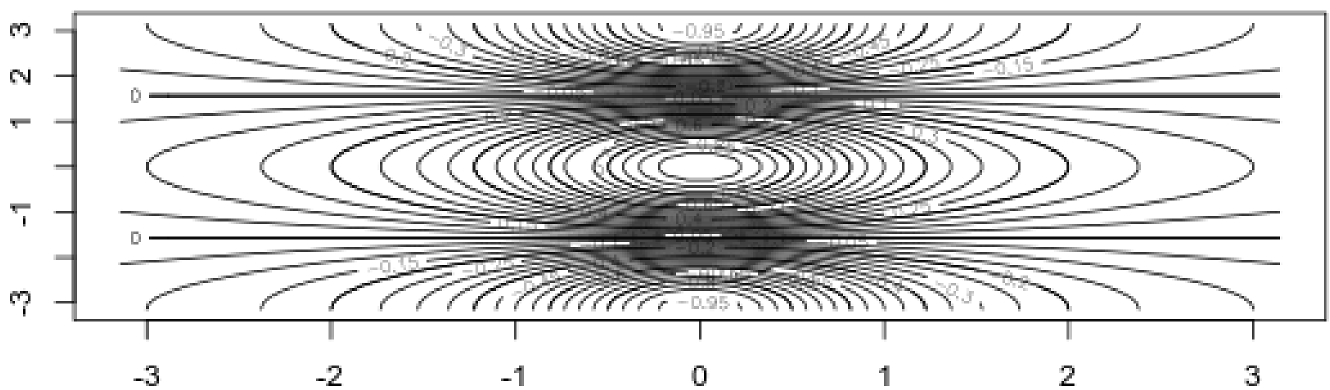
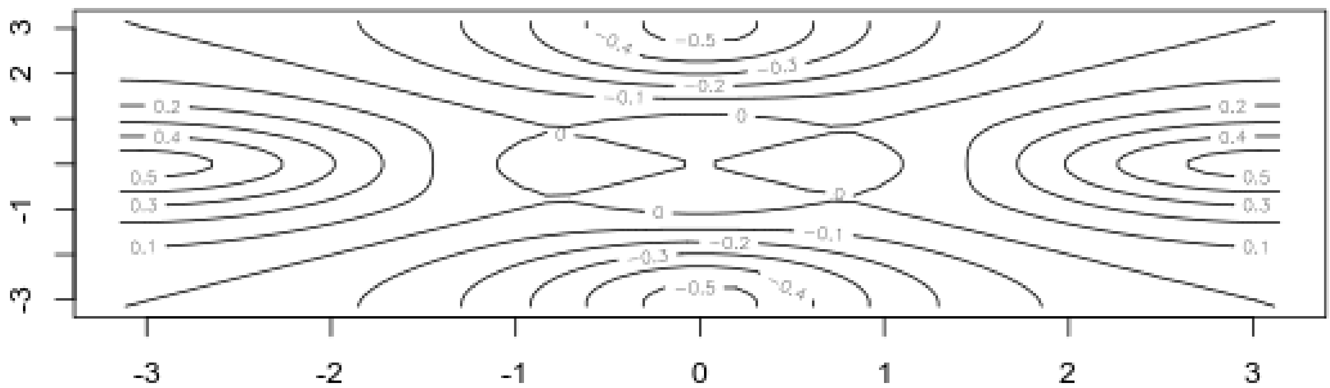
Chapter 2

R Lab

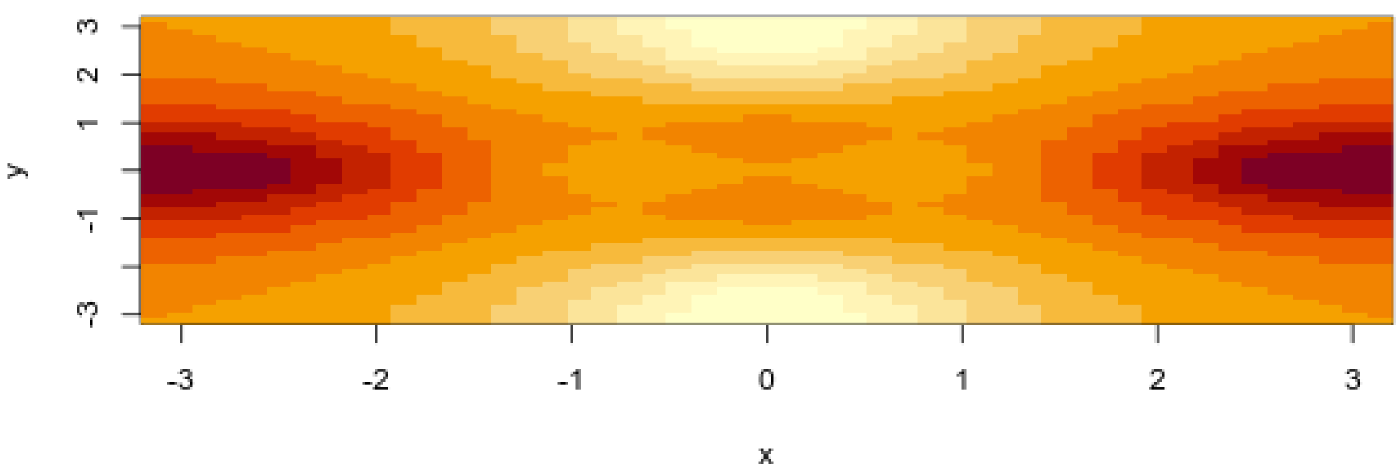
```
x <- seq(-pi, pi, length.out = 50); y <- x

f <- outer(x, y, function(x, y) cos(y)/(1+x^2))

contour(x, y, f)

contour(x, y, f, nlevels = 45, add = T)
```



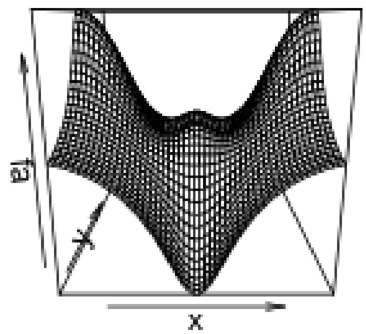

```
fa <- (f - t(f)) / 2

contour(x, y, fa, nlevels = 15)
```



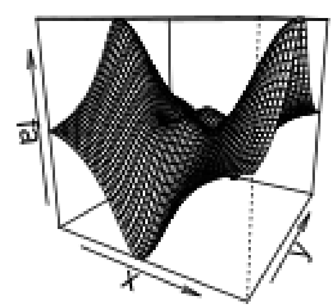

```r
image(x, y, fa)
```



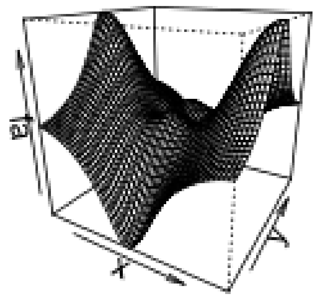

```r
persp(x, y, fa)
```



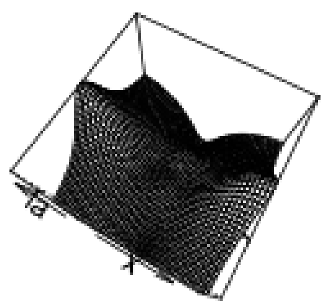

```r
persp(x, y, fa, theta = 30)
```



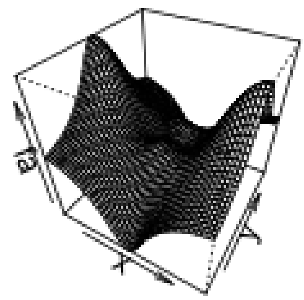

```r
persp(x, y, fa, theta = 30, phi = 20)
```



```r
persp(x, y, fa, theta = 30, phi = 70)
```



```r
persp(x, y, fa, theta = 30, phi = 40)
```



Conceptual

1.)

For each of parts (a) through (d), indicate whether i. or ii. is correct, and explain your answer. In general, do we expect the performance of a flexible statistical learning method to perform better or worse than an inflexible method when:

a.)The sample size n is extremely large, and the number of predictors p is small ?

Better. A flexible method will fit the data closer and with the large sample size, would perform better than an inflexible approach.

b.) The number of predictors p is extremely large, and the number of observations n is small ?

Worse. A flexible method would overfit the small number of observations.

c.) The relationship between the predictors and response is highly non-linear ?

Better. With more degrees of freedom, a flexible method would fit better than an inflexible one.

d.) The variance of the error terms, i.e. $\sigma 2=Var(\varepsilon)$, is extremely high ?

Worse. A flexible method would fit to the noise in the error terms and increase variance.

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.

Regression. 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.

Classification. n = 20, p = 13

c.) We are interesting in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

Regression. n = 52, p = 3

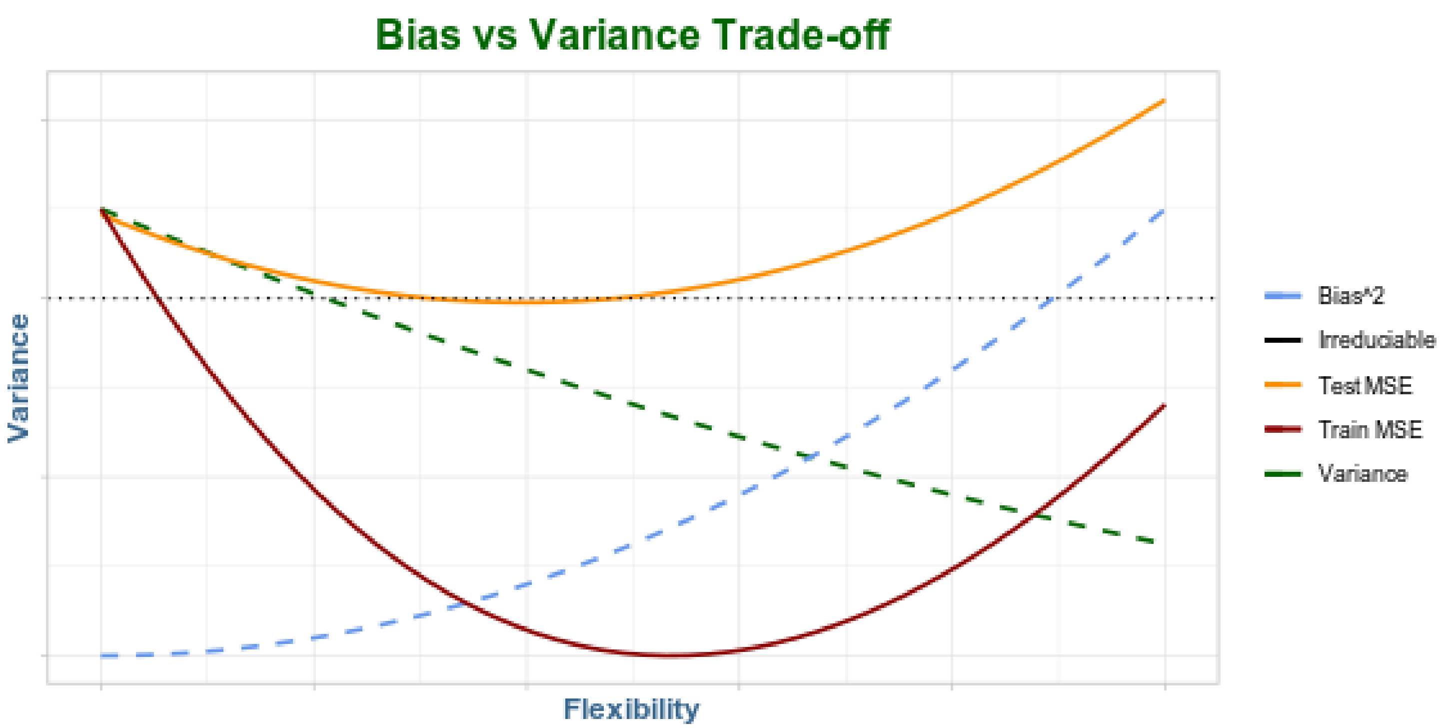
3.)

We now revisit the bias-variance decomposition.

a.) Provide a sketch of typical (squared) bias, variance, training error, test error, and Bayes (or irreducible) error curves, on a single plot, as we go from less flexible statistical learning methods towards more flexible approaches. The x-axis should represent the amount of flexibility in the method, and the y-axis should represent the values for each curve. There should be five curves. Make sure to label each one.

```
mu <- 2

Z <- rnorm(20000, mu)

MSE <- function(estimate, mu) {
   return(sum((estimate - mu)^2) / length(estimate))
}

n <- 50
shrink <- seq(0,0.5, length=n)
test.mse <- numeric(n)
train.mse <- numeric(n)
bias <- numeric(n)
variance <- numeric(n)

for (i in 1:n) {
   test.mse[i] <- MSE((1 - shrink[i]) * Z, mu)
   bias[i] <- mu * shrink[i]
   variance[i] <- (1 - shrink[i])^2
   train.mse[i] <- (variance[i] - bias[i] ) ^2
}

data.table(x = shrink, var = variance, bias = bias^2, test.mse = test.mse, train.mse = train.ms
   ggplot(data = .) +
```

```
  geom_line(aes(x, var, col = "Variance"), lwd = .8, lty = 2,) +
  geom_line(aes(x, bias, col = "Bias^2"), lwd = .8, lty = 2) +
  geom_line(aes(x, test.mse, col = "Test MSE"), lwd = .8) +
  geom_line(aes(x, train.mse, col = "Train MSE"), lwd = .8) +
  geom_hline(aes(yintercept = .8, col = "Irreducible"), lty = 3) +
  scale_colour_manual(values=c("cornflowerblue", "black", "darkorange", "darkred", "darkgreen'
  labs(title = "Bias vs Variance Trade-off", y = "Variance", x = "Flexibility") +
  theme(legend.title = element_blank(),
     axis.text.x=element_blank(), axis.text.y=element_blank())
```



b.) Explain why each of the five curves has the shape displayed in part (a

4.)

You will now think of some real-life applications for statistical learning.

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.

- Infering if an e-mail is spam/ham. Y = Spam {Y|No}, Y = {words in email, subject, from addr}.
- Predicting if a customer will redeemd a coupon. Y = Redeem {Y||No}, X = {purchase history, coupon value, frequency of store visit}.
- Predicting if an inmate will recidivate. Y = Recid {Y|N}, X = {crime type, age, release date, time served}

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.

- Predicting the sale price of a home. Y = sale price, X = {year built, sq footage, quality}.
- Predicting the next day return of a stock. Y = log(Return), X = {prior returns}
- Predicting the customer annual spend on clothing. Y = {Spend? $}, X = {num of items purchased last 1 year, inventory}

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 ?

_The advantages of a very flexible approach are that it may give a better fit for non-linear models and it decreases the bias.

The disadvantages of a very flexible approach are that it requires estimating a greater number of parameters, it follows the noise too closely (overfit) and it increases the variance.

A more flexible approach would be preferred to a less flexible approach when we are interested in prediction and not the interpretability of the results.

A less flexible approach would be preferred to a more flexible approach when we are interested in inference and the interpretability of the results._

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 nonparametric approach) ? What are its disadvantages ?

_A parametric approach reduces the problem of estimating f down to one of estimating a set of parameters because it assumes a form for f.

A non-parametric approach does not assume a patricular form of f and so requires a very large sample to accurately estimate f.

The advantages of a parametric approach to regression or classification are the simplifying of modeling f to a few parameters and not as many observations are required compared to a non-parametric approach.

The disadvantages of a parametric approach to regression or classification are a potentially inaccurate estimate f if the form of f assumed is wrong or to overfit the observations if more flexible models are used._

7.)

The table below provides a training data set containing 6 observations, 3 predictors, and 1 qualitative response variable. Suppose we wish to use this data set to make a prediction for Y when X1 = X2 = X3 = 0 using K-nearest neighbors.

```r
dat <- data.table(Obs = 1:6,
           X1 = c(0, 2, 0, 0, -1, 1),
           X2 = c(3, 0, 1, 1, 0, 1),
           X3 = c(0, 0, 3, 2, 1, 1),
           Y = c("Red", "Red", "Red", "Green", "Green", "Red"))

dat
```

```
   Obs X1 X2 X3     Y
1:   1  0  3  0   Red
2:   2  2  0  0   Red
3:   3  0  1  3   Red
4:   4  0  1  2 Green
5:   5 -1  0  1 Green
6:   6  1  1  1   Red
```

a.) Compute the Euclidean distance between each observation and the test point, X1 = X2 = X3 = 0.

```r
dat[, Distance := ( (X1 - 0)^2 + (X2 - 0)^2 + (X3 - 0)^2 )^.5 ]
dat
```

```
   Obs X1 X2 X3     Y Distance
1:   1  0  3  0   Red 3.000000
2:   2  2  0  0   Red 2.000000
3:   3  0  1  3   Red 3.162278
4:   4  0  1  2 Green 2.236068
5:   5 -1  0  1 Green 1.414214
6:   6  1  1  1   Red 1.732051
```

b.) What is our prediction with K=1 ? Why ?

If $K = 1$ then $X_5 \in N_0$, so that:

$P(Y = Red|X = x_0) = \frac{1}{1} \sum_{i \in N_0} I(y_i = Red) = 0$

and

$P(Y = Green|X = x_0) = \frac{1}{1} \sum_{i \in N_0} I(y_i = Greeen) = 1$

or:

```r
setorder(dat, Distance)

dat[1]$Y
```

```
[1] "Green"
```

Our prediction is *green.*

c.) What is our prediction with K=3 ? Why ?

```r
dat[1:3]
```

```
   Obs X1 X2 X3     Y Distance
1:   5 -1  0  1 Green 1.414214
2:   6  1  1  1   Red 1.732051
3:   2  2  0  0   Red 2.000000
```

2 out of the closest 3 points are red, so we would predict red.

d.) If the Bayes decision boundary in this problem is highly nonlinear, then would we expect the best value for K to be large or small ? Why ?

As K becomes larger, the boundary becomes inflexible (linear). So in this case we would expect the best value for K to be small.

Applied

8.)

This exercise relates to the "College" data set, which can be found in the file "College.csv". 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.

```r
data(College)
```

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.

```r
# fix(College)

head(College)
```

```
                             Private Apps Accept Enroll Top10perc Top25perc
Abilene Christian University     Yes 1660   1232    721        23        52
Adelphi University               Yes 2186   1924    512        16        29
Adrian College                   Yes 1428   1097    336        22        50
Agnes Scott College              Yes  417    349    137        60        89
Alaska Pacific University        Yes  193    146     55        16        44
Albertson College                Yes  587    479    158        38        62
                             F.Undergrad P.Undergrad Outstate Room.Board Books
Abilene Christian University        2885         537     7440       3300   450
Adelphi University                  2683        1227    12280       6450   750
Adrian College                      1036          99    11250       3750   400
Agnes Scott College                  510          63    12960       5450   450
Alaska Pacific University            249         869     7560       4120   800
Albertson College                    678          41    13500       3335   500
```

```
                             Personal PhD Terminal S.F.Ratio perc.alumni Expend
Abilene Christian University     2200  70       78      18.1          12   7041
Adelphi University               1500  29       30      12.2          16  10527
Adrian College                   1165  53       66      12.9          30   8735
Agnes Scott College               875  92       97       7.7          37  19016
Alaska Pacific University        1500  76       72      11.9           2  10922
Albertson College                 675  67       73       9.4          11   9727
                             Grad.Rate
Abilene Christian University        60
Adelphi University                  56
Adrian College                      54
Agnes Scott College                 59
Alaska Pacific University           15
Albertson College                   55
```

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

```r
summary(College)
```

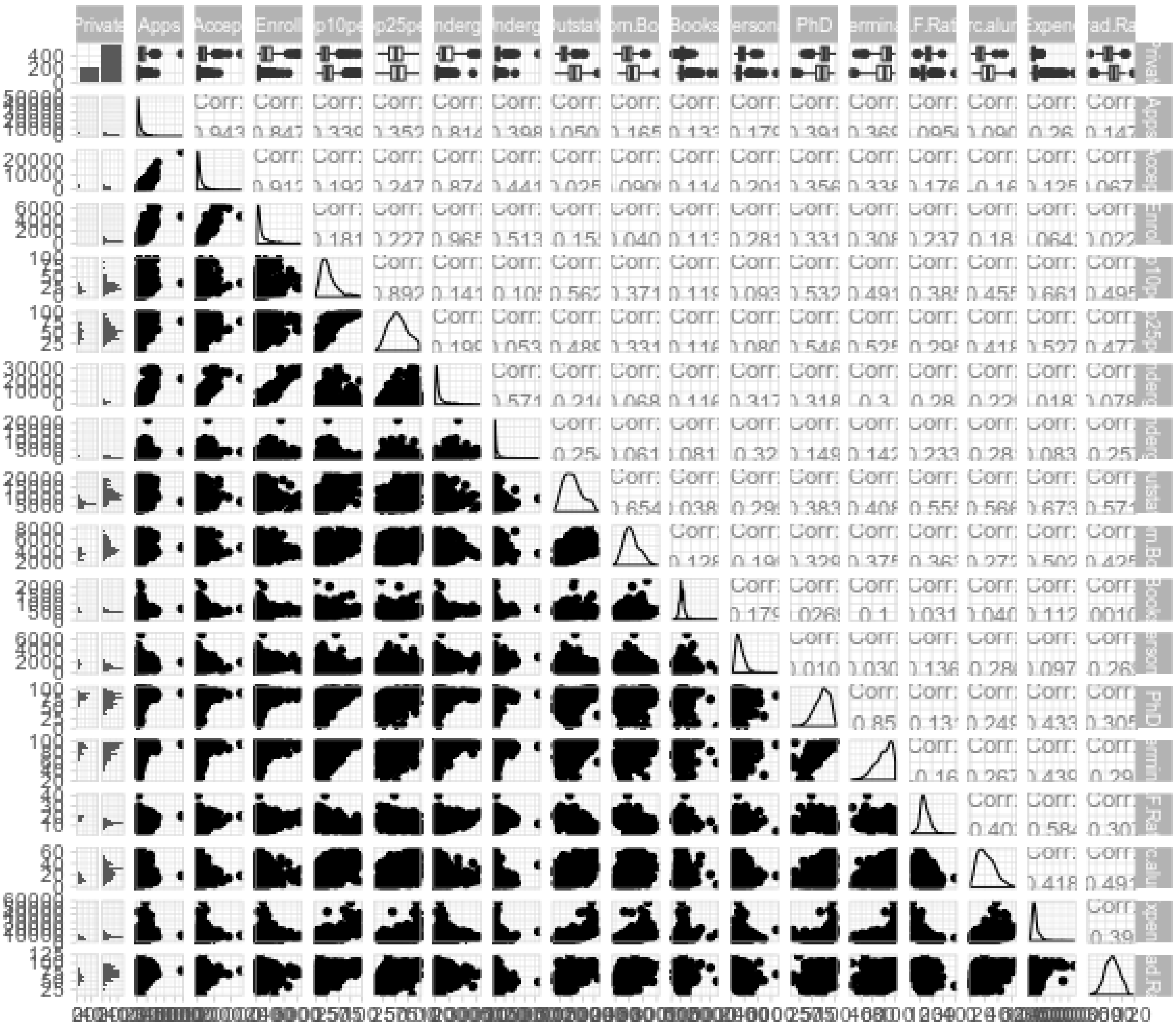
```
 Private        Apps           Accept          Enroll       Top10perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
           Median : 1558   Median : 1110   Median : 434   Median :23.00
           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  
```

Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data.

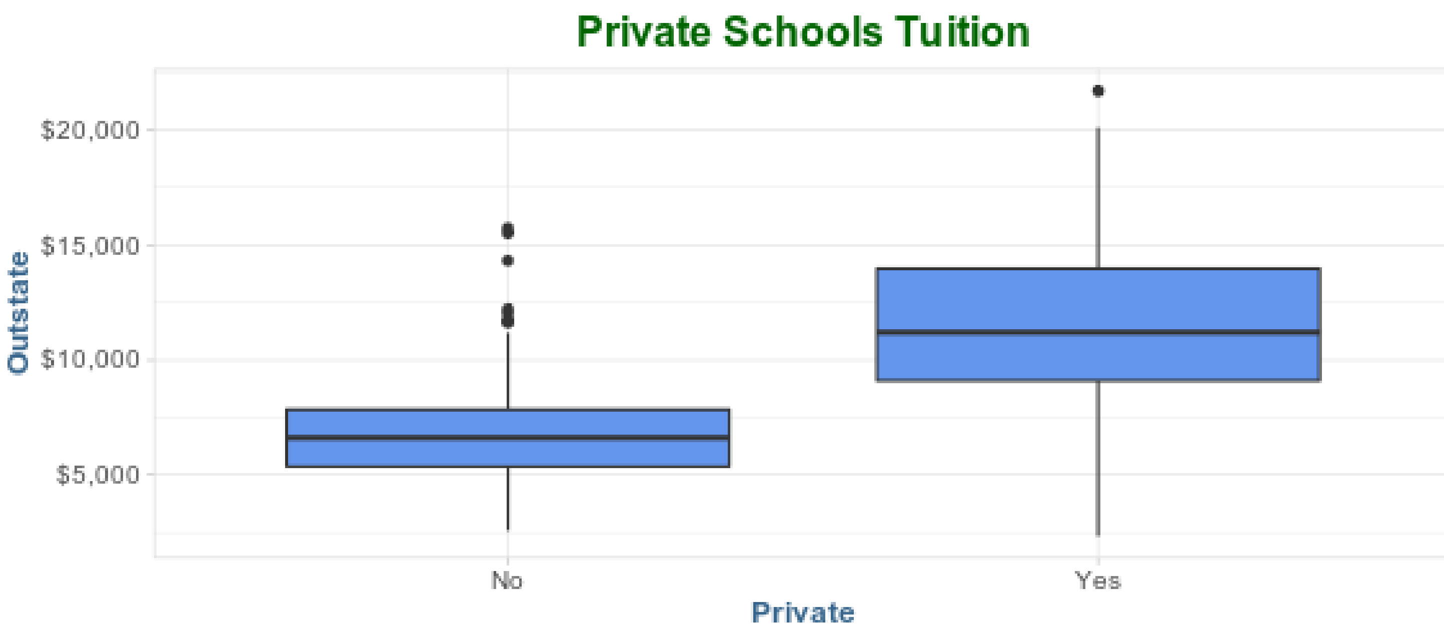
```r
ggpairs(College)
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



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

```r
ggplot(College, aes(Private, Outstate)) +
   geom_boxplot(fill = "cornflowerblue") +
   scale_y_continuous(label = dollar) +
   labs(title = "Private Schools Tuition")
```



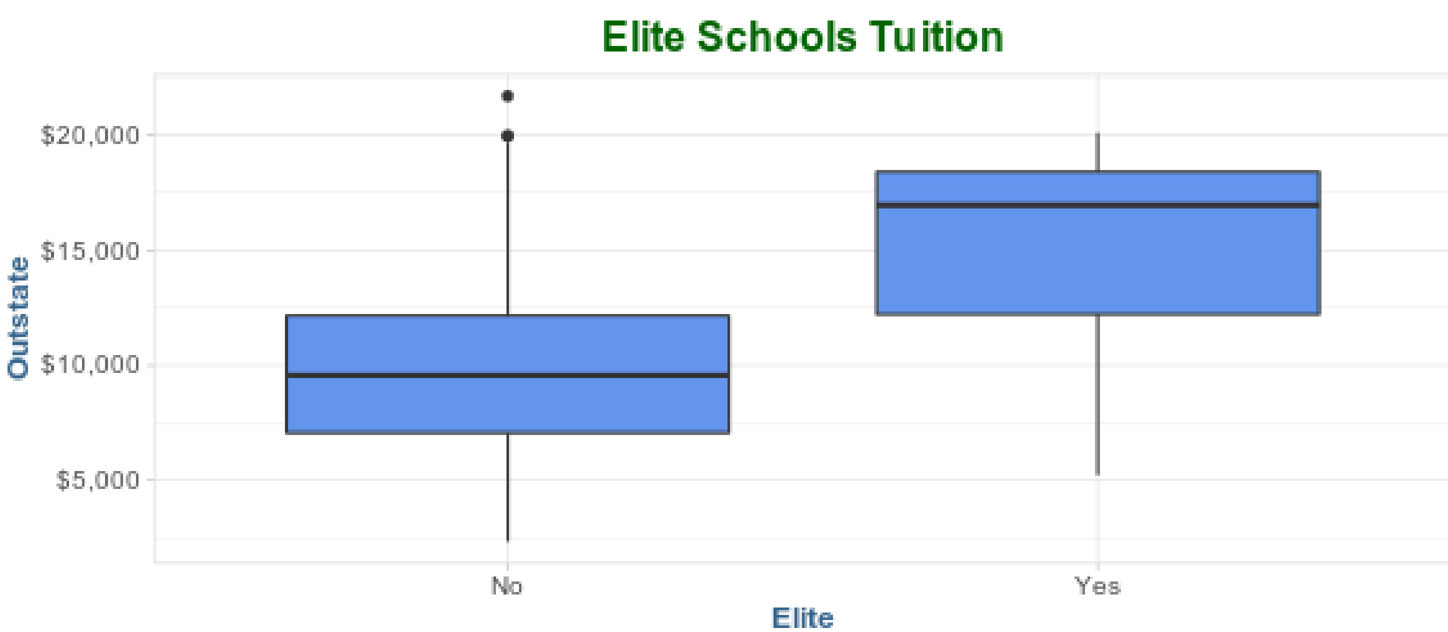
Create a new qualitative variable, called “Elite”, by binning the “Top10perc” variable. Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of “Outstate” versus “Elite”.

```
college <- as.data.table(College)
college[, Elite := ifelse(Top10perc > 50, "Yes", "No")]

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   Length:777        
 1st Qu.: 53.00   Class :character  
 Median : 65.00   Mode  :character  
 Mean   : 65.46                     
 3rd Qu.: 78.00                     
 Max.   :118.00                     
```

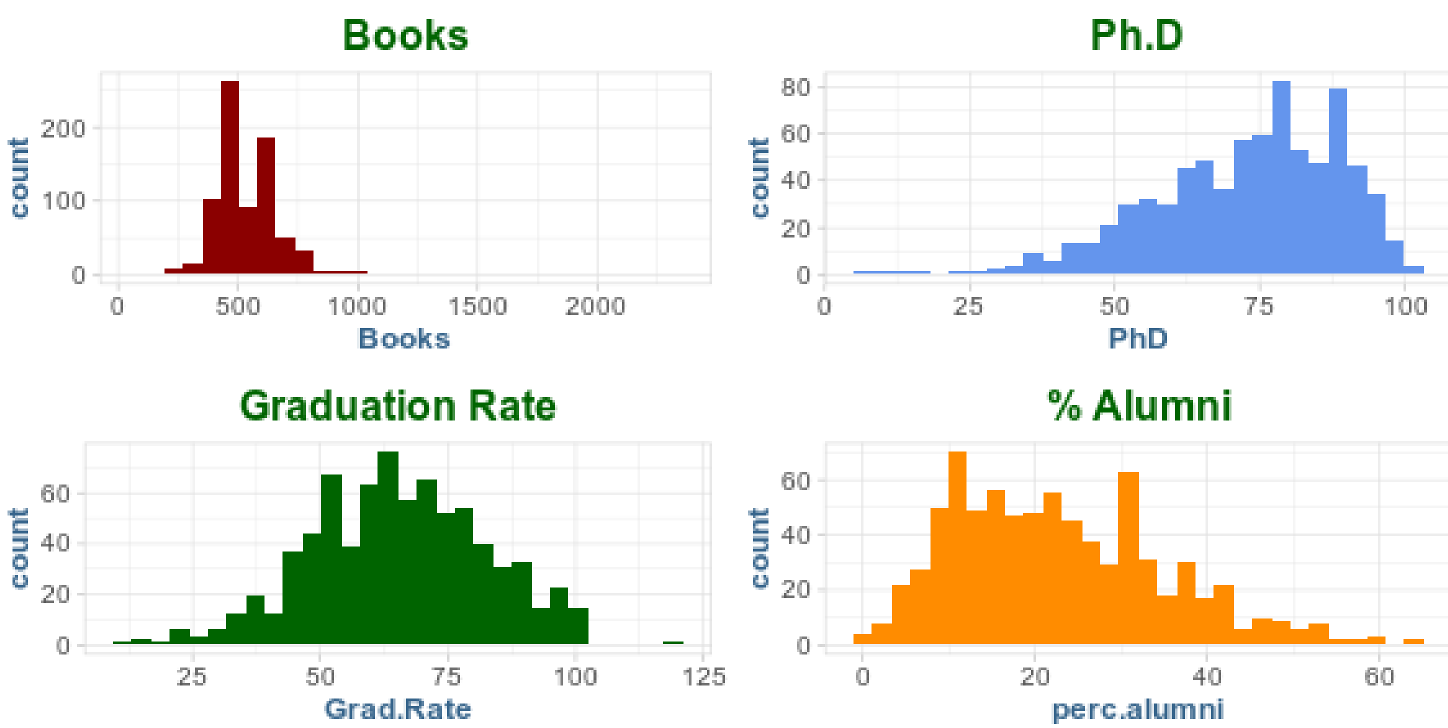
```r
ggplot(college, aes(Elite, Outstate)) +
   geom_boxplot(fill = "cornflowerblue") +
   scale_y_continuous(label = dollar) +
   labs(title = "Elite Schools Tuition")
```



Use the hist() function to produce some histograms with numbers of bins for a few of the quantitative variables.

```r
p1 <- ggplot(college) +
   geom_histogram(aes(Books), fill = "darkred", bins = 30) +
   labs(title = "Books")

p2 <-ggplot(college) +
   geom_histogram(aes(PhD), fill = "cornflowerblue", bins = 30) +
```

```
  labs(title = "Ph.D")

p3 <-ggplot(college) +
  geom_histogram(aes(Grad.Rate), fill = "darkgreen", bins = 30) +
  labs(title = "Graduation Rate")

p4 <- ggplot(college) +
  geom_histogram(aes(perc.alumni), fill = "darkorange", bins = 30) +
  labs(title = "% Alumni")

grid.arrange(p1, p2, p3, p4, nrow = 2)
```



9.)

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

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

```
auto <- ISLR::Auto
str(auto)
```

```
'data.frame':   392 obs. of  9 variables:
 $ mpg         : num  18 15 18 16 17 15 14 14 14 15 ...
 $ cylinders   : num  8 8 8 8 8 8 8 8 8 8 ...
 $ displacement: num  307 350 318 304 302 429 454 440 455 390 ...
```

```
 $ horsepower  : num  130 165 150 150 140 198 220 215 225 190 ...
 $ weight      : num  3504 3693 3436 3433 3449 ...
 $ acceleration: num  12 11.5 11 12 10.5 10 9 8.5 10 8.5 ...
 $ year        : num  70 70 70 70 70 70 70 70 70 70 ...
 $ origin      : num  1 1 1 1 1 1 1 1 1 1 ...
 $ name        : Factor w/ 304 levels "amc ambassador brougham",..: 49 36 231 14 161 141 54 223
```

All variables except "horsepower" and "name" are quantitative.

b.) What is the range of each quantitative predictor ?

```r
summary(auto[, -c(4, 9)])
```

```
      mpg          cylinders      displacement       weight      acceleration  
 Min.   : 9.00   Min.   :3.000   Min.   : 68.0   Min.   :1613   Min.   : 8.00  
 1st Qu.:17.00   1st Qu.:4.000   1st Qu.:105.0   1st Qu.:2225   1st Qu.:13.78  
 Median :22.75   Median :4.000   Median :151.0   Median :2804   Median :15.50  
 Mean   :23.45   Mean   :5.472   Mean   :194.4   Mean   :2978   Mean   :15.54  
 3rd Qu.:29.00   3rd Qu.:8.000   3rd Qu.:275.8   3rd Qu.:3615   3rd Qu.:17.02  
 Max.   :46.60   Max.   :8.000   Max.   :455.0   Max.   :5140   Max.   :24.80  
      year           origin     
 Min.   :70.00   Min.   :1.000  
 1st Qu.:73.00   1st Qu.:1.000  
 Median :76.00   Median :1.000  
 Mean   :75.98   Mean   :1.577  
 3rd Qu.:79.00   3rd Qu.:2.000  
 Max.   :82.00   Max.   :3.000  
```

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

```r
sapply(auto[, -c(4, 9)], mean)
```

```
         mpg    cylinders displacement       weight acceleration         year 
   23.445918     5.471939   194.411990  2977.584184    15.541327    75.979592 
      origin 
    1.576531 
```

```r
sapply(auto[, -c(4, 9)], sd)
```

```
         mpg    cylinders displacement       weight acceleration         year 
   7.8050075    1.7057832  104.6440039  849.4025600    2.7588641    3.6837365 
      origin 
   0.8055182 
```

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 ?

```r
subset <- auto[-c(10:85), -c(4,9)]
sapply(subset, range)
```

```
     mpg cylinders displacement weight acceleration year origin
[1,] 11.0         3           68   1649          8.5   70      1
[2,] 46.6         8          455   4997         24.8   82      3
```

```r
sapply(subset, mean)
```

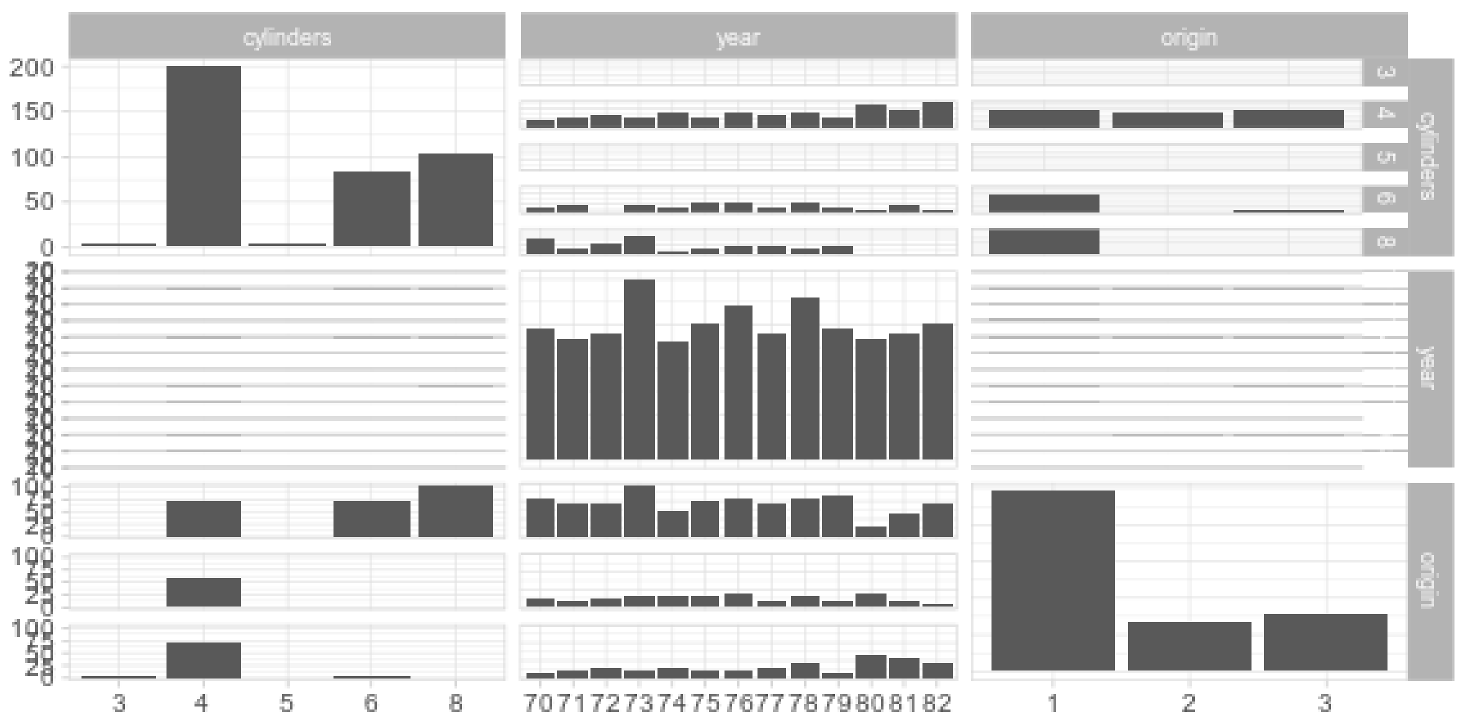
```
         mpg    cylinders displacement       weight acceleration         year
   24.404430     5.373418   187.240506  2935.971519    15.726899    77.145570
      origin
    1.601266
```

```r
sapply(subset, sd)
```

```
         mpg    cylinders displacement       weight acceleration         year
    7.867283     1.654179    99.678367   811.300208     2.693721     3.106217
      origin
    0.819910
```

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.

```r
auto$cylinders <- as.factor(auto$cylinders)
auto$year <- as.factor(auto$year)
auto$origin <- as.factor(auto$origin)

ggpairs(auto[, c("cylinders", "year", "origin")])
```



f.) Suppose that we wish to predict gas mileage (“mpg”) on the basis of other variables. Do your plots suggest that any of the other variables might be useful in predicting "mpg" ?

From the plots above, the cylinders, horsepower, year and origin can be used as predictors. Displacement and weight were not used because they are highly correlated with horespower and with each other.

```
auto$horsepower <- as.numeric(auto$horsepower)
cor(auto$weight, auto$horsepower)
```

```
[1] 0.8645377
```

```
cor(auto$weight, auto$displacement)
```

```
[1] 0.9329944
```

```
cor(auto$displacement, auto$horsepower)
```

```
[1] 0.897257
```

10.)

This exercise involves the "Boston" housing data set.

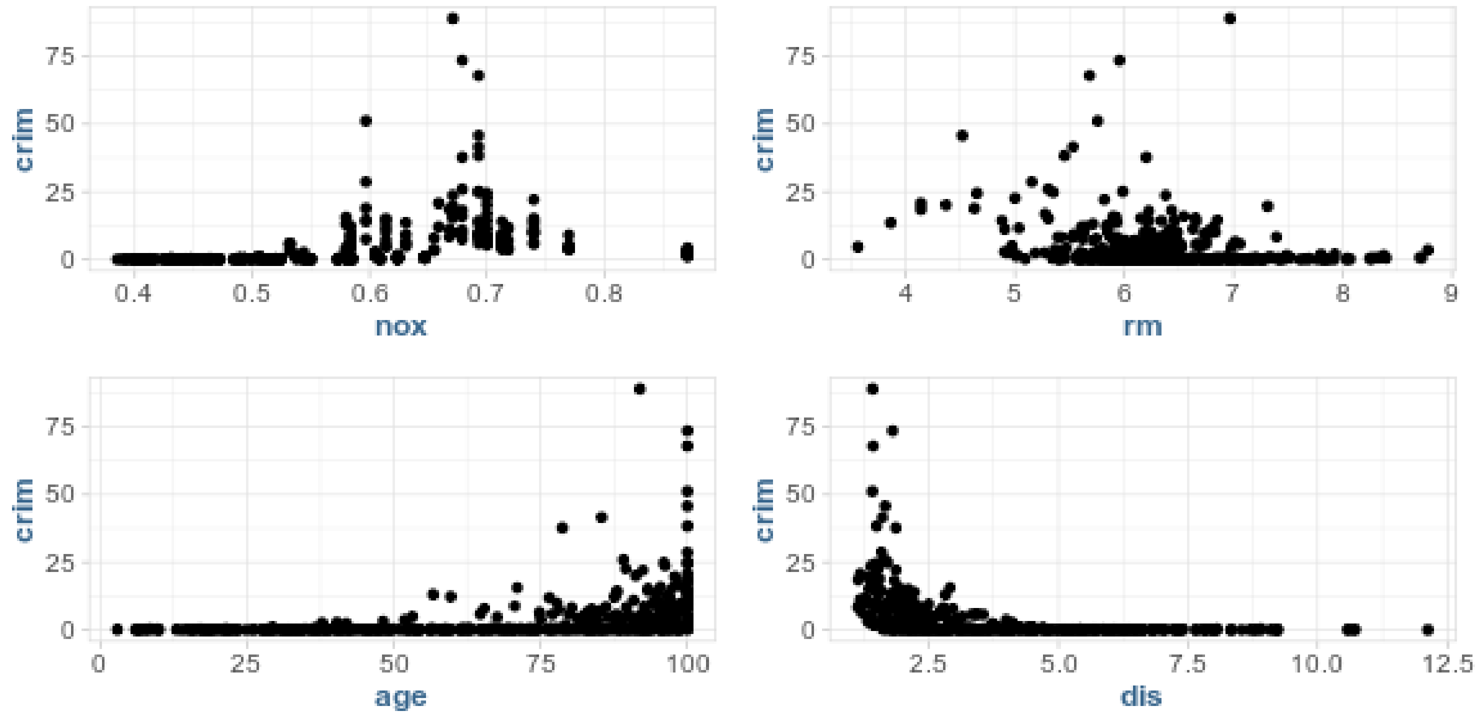
a.) To begin, load in the "Boston" data set.

```
boston <- MASS::Boston

boston$chas <- as.factor(boston$chas)
dim(Boston)
```

```
[1] 506  14
```

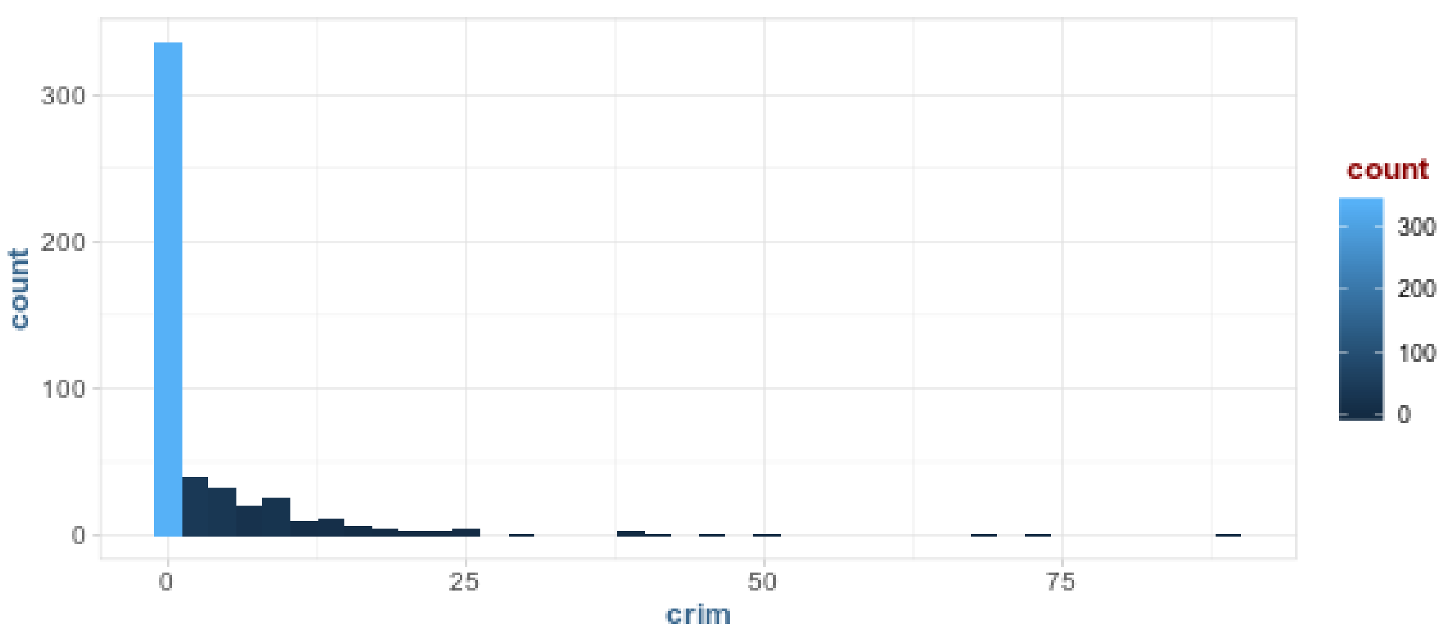
b.) Make some pairwise scatterplots of the predictors in this data set.

```
p1 <- ggplot(boston) +
   geom_point(aes(nox, crim))

p2 <- ggplot(boston) +
   geom_point(aes(rm, crim))

p3 <- ggplot(boston) +
   geom_point(aes(age, crim))

p4 <- ggplot(boston) +
   geom_point(aes(dis, crim))

grid.arrange(p1, p2, p3, p4, nrow = 2)
```



c.) Are any of the predictors associated with per capita crime rate ?

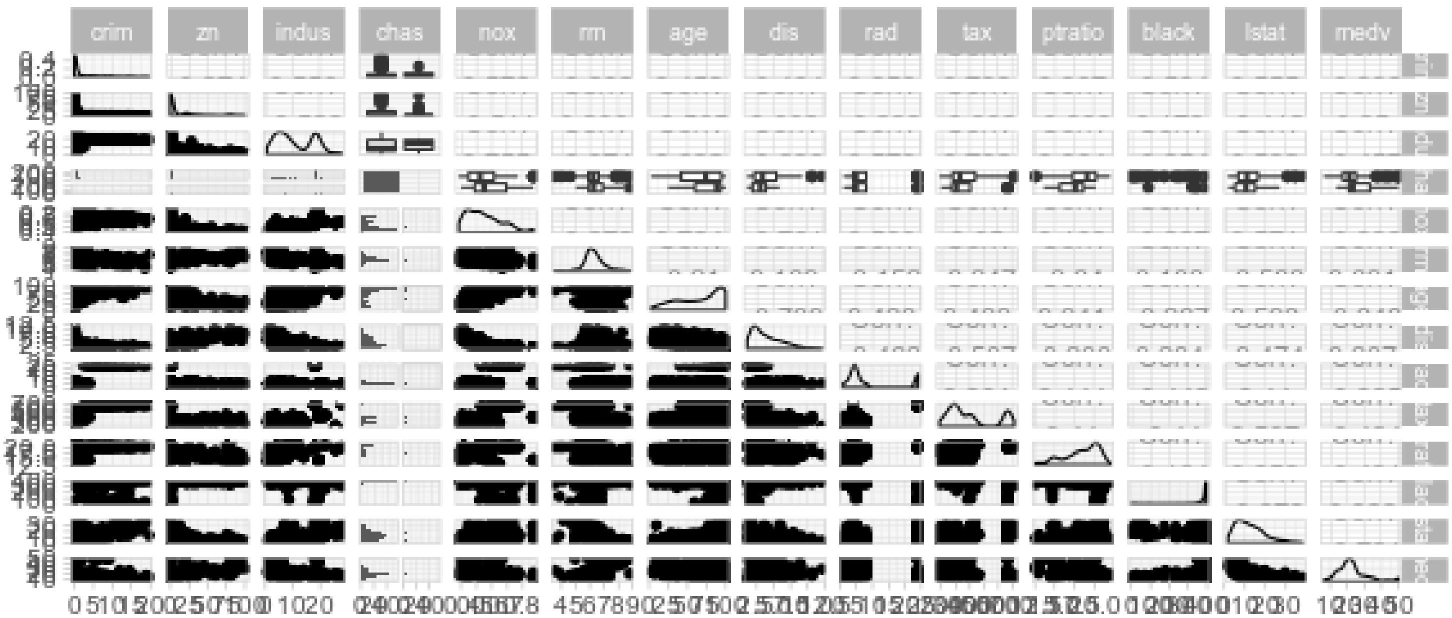
```r
ggplot(boston) +
   geom_histogram(aes(crim, fill = ..count..), bins = 40)
```



Most suburbs do not have any crime (80% of data falls in crim < 20).

```r
ggpairs(boston[boston$crim < 20,])
```

```
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



d.) Do any of the suburbs of Boston appear to have particularly high crime rates ? Tax rates ? Pupil-teacher ratios ?

```r
nrow(Boston[Boston$tax == 666, ])
```

```
[1] 132
```

e.) How many of the suburbs in this data set bound the Charles river ?

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

```
[1] 35
```

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

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

```
[1] 19.05
```

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

```
row.names(Boston[min(Boston$medv), ])
```

```
[1] "5"
```

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

```
[1] 187 711
```

```
boston[min(boston$medv), ]$tax
```

```
[1] 222
```

h.) In this data set, how many of the suburbs average more than seven rooms per dwelling ? More than eight rooms per dwelling ?

```
row.names(Boston[min(boston$medv), ])
```

```
[1] "5"
```

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

```
[1] 64
```

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

```
[1] 13
```