Analytics 512: Solution Key for Homework 0

02/04/19

ISLR 2.4.2

(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 is a regression problem. Since we want to understand factors, we are interested in inference. The number of observations is $n = 500$ and the number of predictors is $p = 3$.

(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 is a classification problem, with $n = 20$ observations and $p = 13$ predictors. We are interested in prediction.

(c) *We are interest in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.* This is a regression problem. There are $n = 52$ observations and $p = 3$ predictors. The goal is prediction.

ISLR 2.4.5

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 is expected to have few systematic errors (low bias), but it may be unstable (have large variance). A more flexible approach might be preferred when there is a highly nonlinear or otherwise complicated relation between predictors and response. A less flexible approach is better when there are too many predictors.

Xtra #4

Load the MNIST data:

```
load("../../Data/mnist_all.RData")
```

a) Modify the function **myclosest()** so that it uses exactly **k** neighbors instead of 100 to classify a test digit. The new function should have two arguments, namely **mydigit** and **k**.

Here is the modified code.

```
myclosest = function(mydigit,k){
  digit.dist = function(j){
    return(sqrt(mean((test$x[mydigit,] - train$x[j,])^2) ) )
  }
  mnist.distances = sapply(1:60000,FUN = digit.dist)
  myclosest = head(order(mnist.distances),k) # replace 100 with k
  mytable <- table(train$y[myclosest])
  myindex = which.max(mytable)
```

```
  return(as.numeric(names(mytable[myindex])))
}
```

b) Demonstrate the modified function by trying to classify a test digit of your choice. Find a value of k such that the classification is correct and another value of $k < 1000$ such that the classification of the same test digit is incorrect.

We choose digit 1432.This is a 2.

```
mydigit = 1432
test$y[mydigit]
```

```
## [1] 2
```

For $k = 10$ the digit is classified correctly. For $k = 900$, the digit is misclassified as a 1.

```
myclosest(mydigit,10)
```

```
## [1] 2
```

```
myclosest(mydigit,900)
```

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

ISLR 2.4.8 (5)

Read the data, introduce row names, delete the first column, and look at the result.

```
college <- read.csv("../../Data/College.csv")
rownames(college) <- college[,1]
college <- college[,-1]
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
```

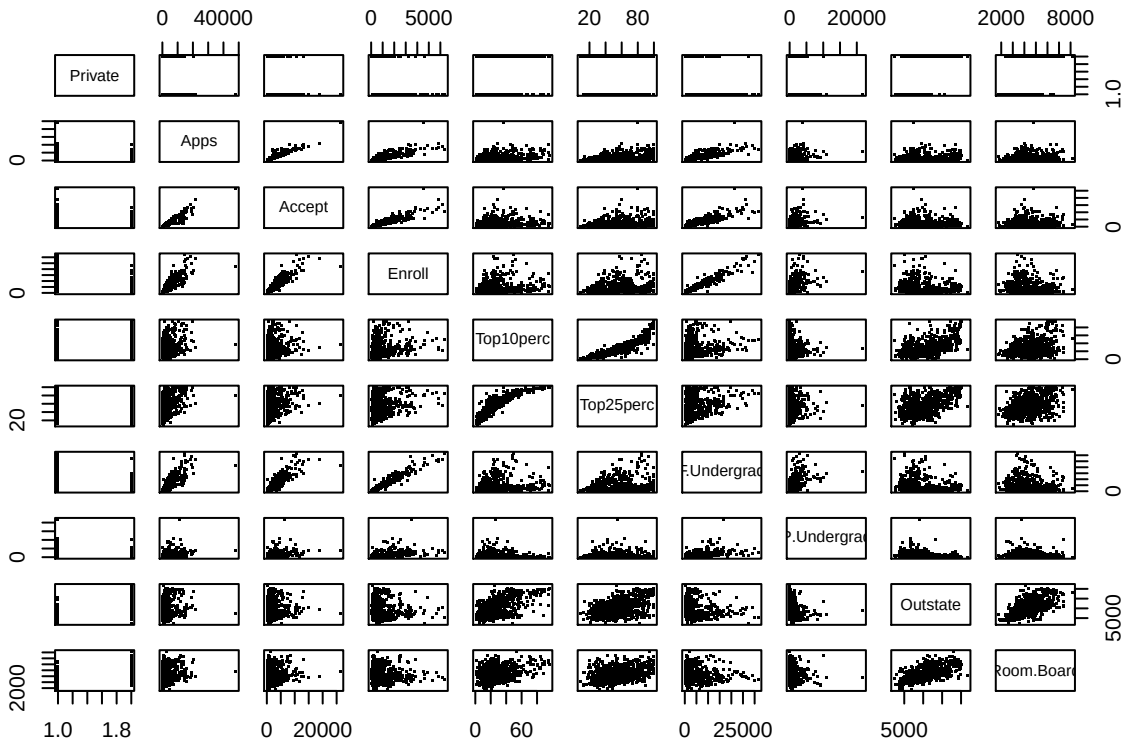
```
## Adrian College                     12.9          30   8735        54
## Agnes Scott College                 7.7          37  19016        59
## Alaska Pacific University          11.9           2  10922        15
## Albertson College                   9.4          11   9727        55
```

Summaries and plots. Use **pairs()** with the plot symbol **pch = 46**.

```
summary(college)
```

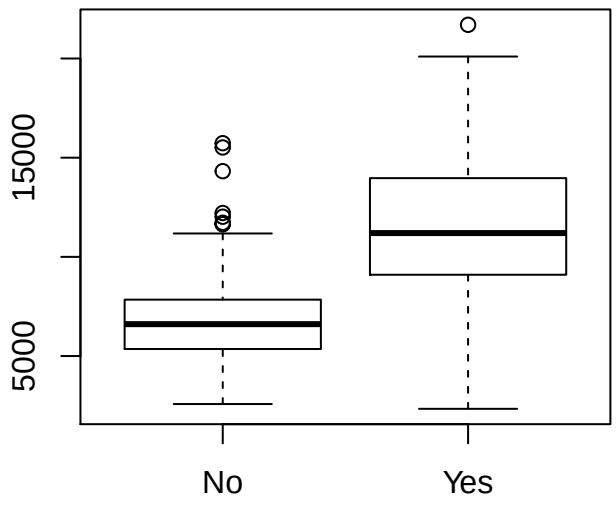
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00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

```
pairs(college[,1:10], pch = 46)
```



The "pairs" plot reveals many associations.

```
boxplot(Outstate ~ Private, data = college)
```



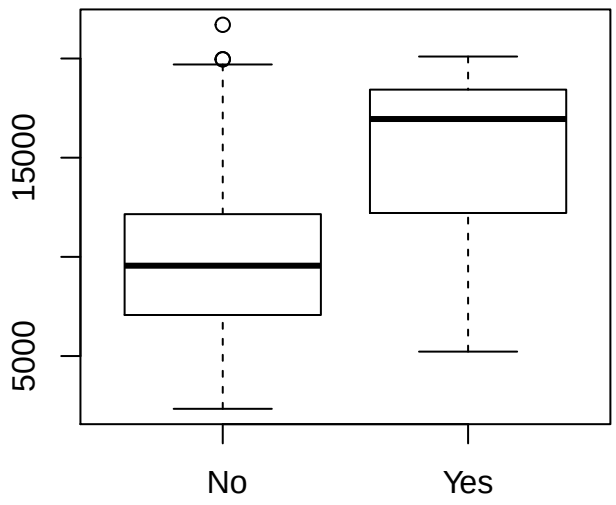
The side-by-side box plot shows that private colleges tend to have higher out-of-state tuition, no doubt because their overall tuition is higher.

Looking at "elite" colleges:

```
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
```

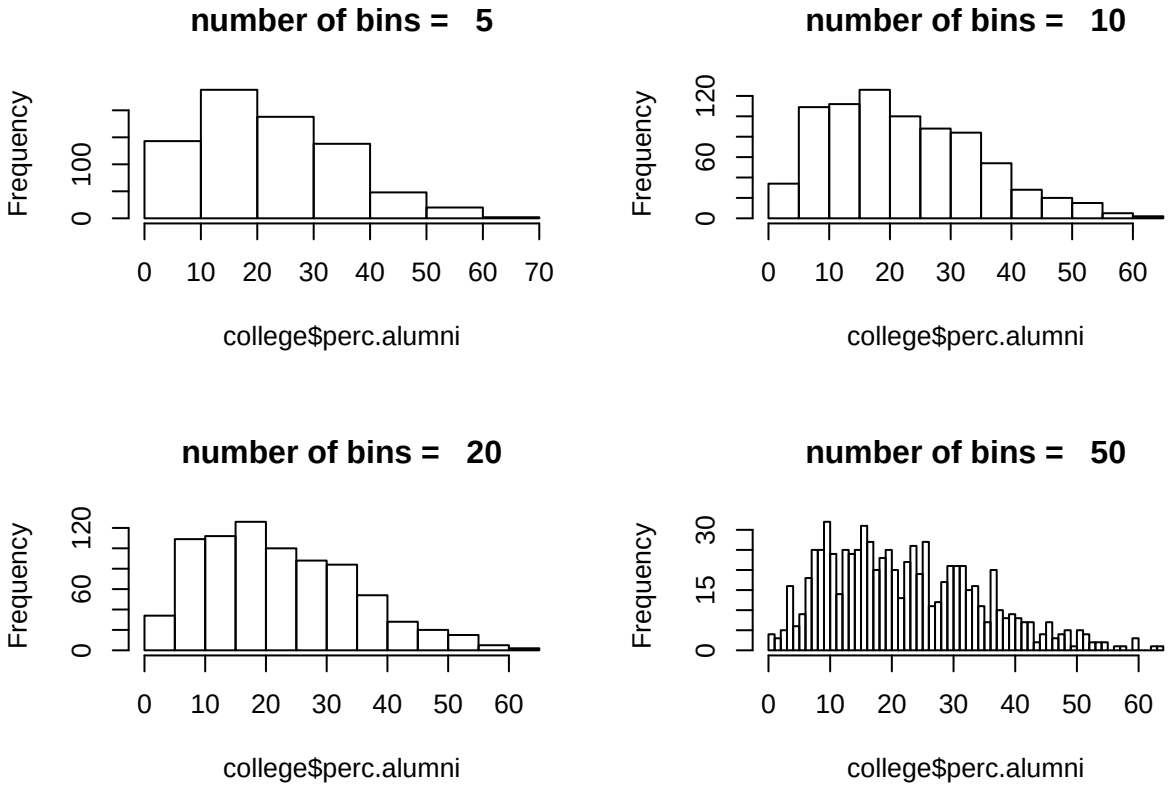
```
boxplot(Outstate ~ Elite, data = college)
```



There are altogether 78 colleges in this group. Their out-of-state tuition tends to be substantially higher than that of non-elite colleges.

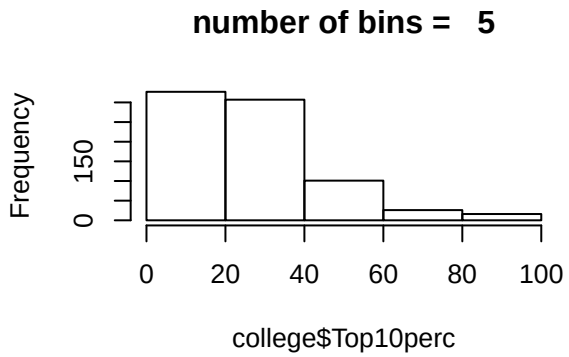
Make histograms with various bin sizes for a number of quantitative variables.

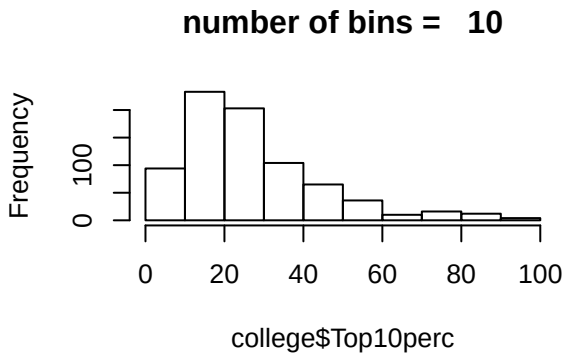
```
par(mfrow = c(2,2))
for (n in c(5,10,20,50)){
  hist(college$perc.alumni, breaks = n, main = paste("number of bins =  ",n))
}
```

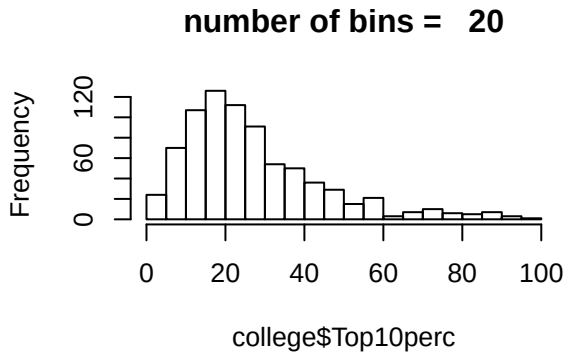


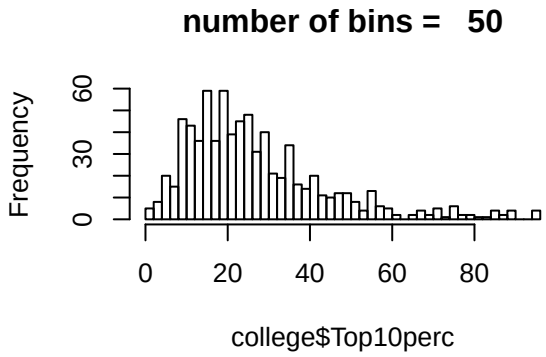
Lower been size (n = 5) does not give enough detail, high bin size (n=40) shows too much extraneous variability. Ten or 20 bins are just right in all cases.

```
par(mfrow = c(2,2))
for (n in c(5,10,20,50)){
  hist(college$Top10perc, breaks = n, main = paste("number of bins =  ",n))
}
```

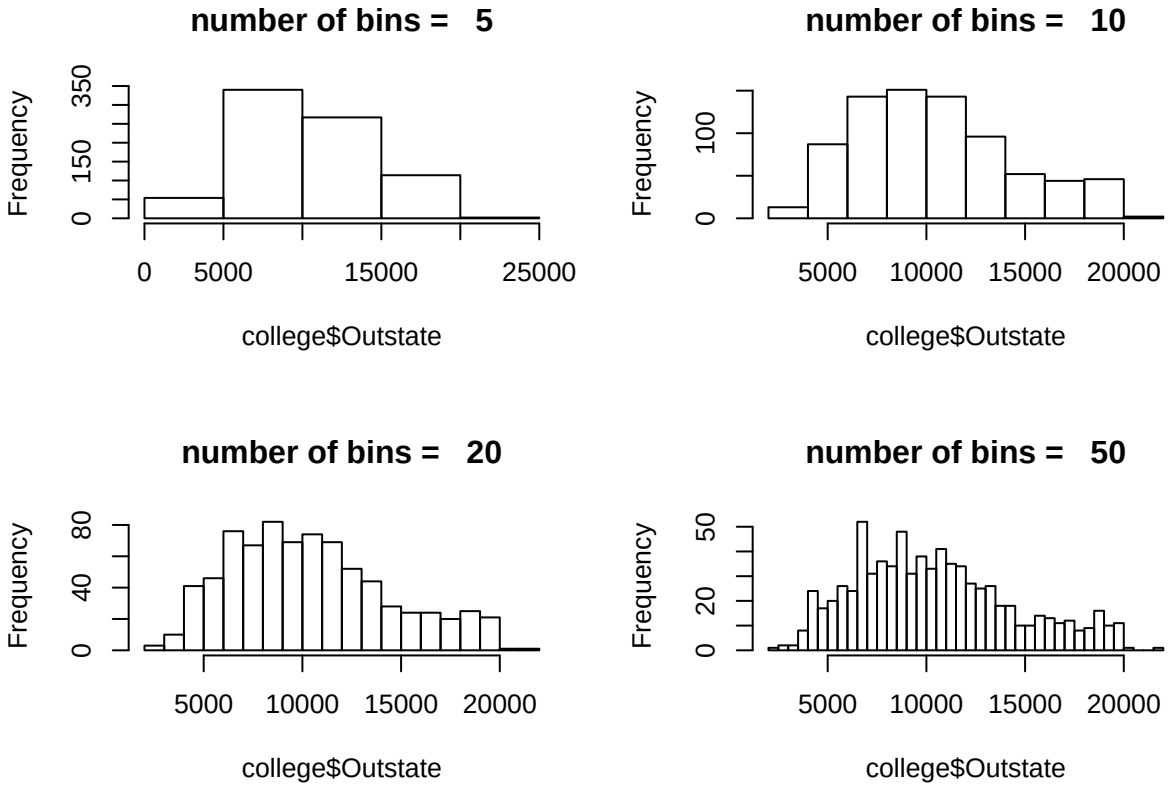






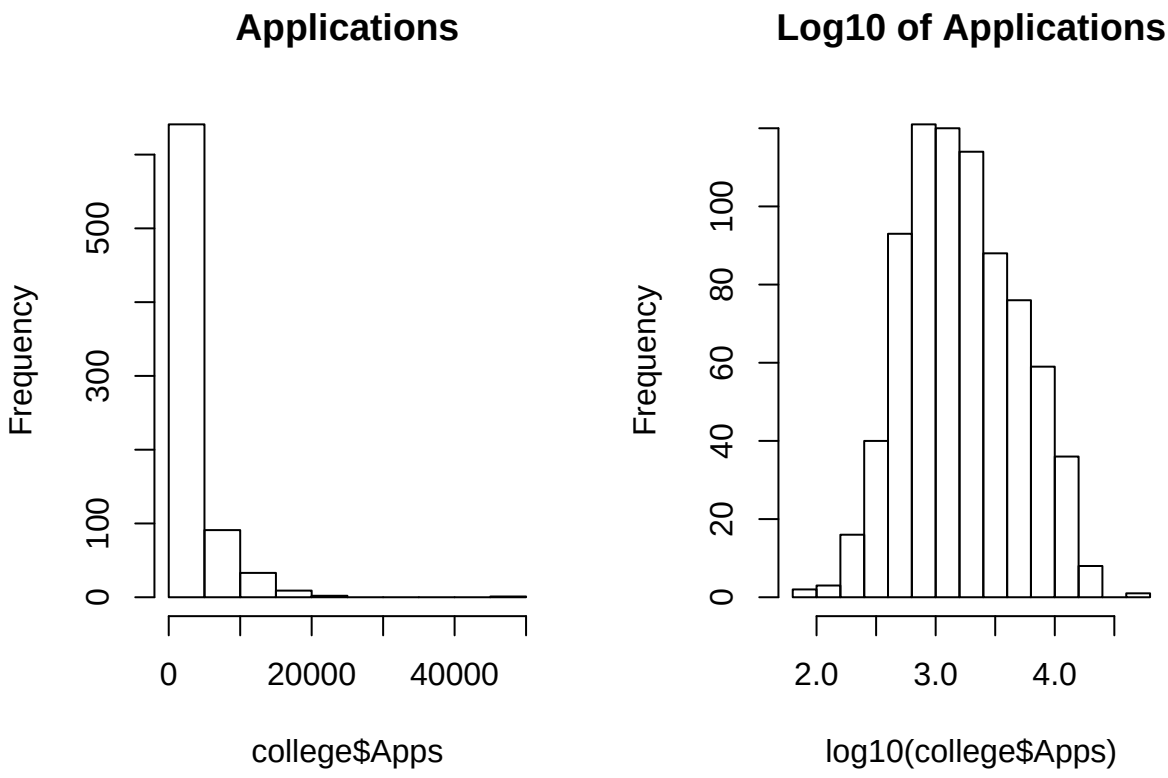



```
par(mfrow = c(2,2))
for (n in c(5,10,20,50)){
  hist(college$Outstate, breaks = n, main = paste("number of bins =  ",n))
}
```



There are many ways to explore the data further. Make a histogram of the number of applications. This is a highly skewed plot, since there are some colleges with very large numbers of applications. Therefore, a histogram of the logarithms of applications gives a better picture. Use n = 10 bins in both cases.

```
par(mfrow = c(1,2))
  hist(college$Apps, breaks = 10, main = "Applications")
  hist(log10(college$Apps), breaks = 10, main = "Log10 of Applications")
```



Xtra #6

This problem uses the Shiny app at https://keeganhines.shinyapps.io/bias_variance/ . Before working on this problem, load the app, read the explanation, play with the slider and the "Generate New Data" button, and answer the questions at the bottom of the page ("Check your understanding") for yourself or discuss them with others.

Model complexity = degree of the polynomial that is being fitted.

a) Make 10 different simulations with model complexity = 1. Compute the average Residual SSE and find the approximate range of the highest order coefficient for these 10 simulations. This is a measure for the baseline variance for a low complexity model.

Solution. The average residual SSE is about 81 and the highest order coefficient ranges from about-5 to about -3.

b) Make 10 different simulations with model complexity = 10. Compute the average Residual SSE. Which coefficient has the largest range in this case? What is that range? This is a measure for the variance for a high complexity model.

Solution. The mean residual SSE is about 19. Coefficients 5 and 6 have the largest range. For example, the range of coefficient 5 is from about -15,000 to about 12,000 in ten simulations.

c) How do your results illustrate the bias - variance trade-off? The answer should be a short paragraph.

Solution. As the model complexity increases from 1 (the most rigid model) to 10 (very flexible), the bias goes down (the residual SSE decreases), while the variance increases (the coefficients of the model show more variability and therefore are less reliable).

d) For which model complexity between 1 and 15 do you typically obtain a curve which is most similar and overall close to the unknown curve that is to be estimated? Try multiple simulation for several different model complexities, summarize what you see, and explain your answer. Pictures or numerical results are not required.

Solution. This happens most frequently for model complexity 2 and sometimes for model complexity 3. In these cases, one obtains a curve that is concave down, just like the original curve. However, the residual sum squares is not the smallest in that case. A model with this flexibility is capable of following the general curve without overfitting.