ComS 573: Homework #1

Due on February 7, 2014

Professor De Brabanter at 10am

Josh Davis

Problem 1

Answer the following questions using the table below.

Observation	X_1	X_2	X_3	Y
1	0	3	0	Red
2	2	0	0	Red
3	0	1	3	Red
4	0	1	2	Green
5	-1	0	1	Green
6	1	1	1	Red

Part A

Compute the Euclidean distance between each observation and the test point, $X_1 = X_2 = X_3 = 0$.

Solution

The equation for Euclidean distance is: $\text{dist} = \sqrt{(x_1 - 0)^2 + (x_2 - 0)^2 + (x_3 - 0)^2}$ Thus giving us:

Observation	Equation	Result
1	$\sqrt{0^2 + 3^2 + 0^2}$	3
2	$\sqrt{2^2 + 0^2 + 0^2}$	2
3	$\sqrt{0^2 + 1^2 + 3^2}$	3.16
4	$\sqrt{0^2 + 1^2 + 2^2}$	2.24
5	$\sqrt{-1^2 + 0^2 + 1^2}$	1.41
6	$\sqrt{1^2 + 1^2 + 1^2}$	1.73

Part B

Prediction with $k = 1$.

Solution

For $k = 1$, the prediction for our test point includes a single neighbor, thus it includes Observation 5 which is Green. Since the probability of being Green is 1, our test point should be Green as well.

Part C

Prediction with $k = 3$.

Solution

For $k = 3$, the prediction for our test includes three neighbors: Observation 5 (Green), Observation 6 (Red), and Observation 2 (Red). The probability for Green is then 1/3 and the probability of Red is 2/3. The test point should then be Red.

Part D

If the Bayes decision boundary (gold standard) is highly nonlinear in this problem, then would we expect the best value for k to be large or small?

Solution

If the Bayes decision boundary is highly nonlinear, then we would expect the best value for k to be small. This is because the larger the value of k, the less flexible our model becomes. The less flexible that it is, the more linear it gets.

Problem 2

Suppose we would like to fit a straight line through the origin, i.e., $Y_i = \beta_1 x_i + e_i$ with $i = 1, \ldots, n$, $\mathrm{E}[e_i] = 0$, and $\mathrm{Var}[e_i] = \sigma_e^2$ and $\mathrm{Cov}[e_i, e_j] = 0, \forall i \neq j$.

Part A

Find the least squares esimator for $\hat{\beta}_1$ for the slope β_1.

Solution

To find the least squares estimator, we should minimize our Residual Sum of Squares, RSS:

$$\begin{aligned} RSS &= \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 x_i\right)^2 \end{aligned}$$

By taking the partial derivative in respect to $\hat{\beta}_1$, we get:

$$\frac{\partial}{\partial \hat{\beta}_1}(RSS) = -2 \sum_{i=1}^{n} x_i (Y_i - \hat{\beta}_1 x_i) = 0$$

This gives us:

$$\begin{aligned} \sum_{i=1}^{n} x_i (Y_i - \hat{\beta}_1 x_i) &= \sum_{i=1}^{n} x_i Y_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 \\ &= \sum_{i=1}^{n} x_i Y_i - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 \end{aligned}$$

Solving for $\hat{\beta}_1$ gives the final estimator for β_1:

$$\hat{\beta}_1 = \frac{\sum x_i Y_i}{\sum x_i^2}$$

Problem 2 continued on next page...

Part B

Calculate the bias and the variance for the estimated slope $\hat{\beta}_1$.

Solution

For the bias, we need to calculate the expected value $\mathrm{E}[\hat{\beta}_1]$:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_1] &= \mathrm{E}\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{\sum x_i \mathrm{E}[Y_i]}{\sum x_i^2} \\
&= \frac{\sum x_i(\beta_1 x_i)}{\sum x_i^2} \\
&= \frac{\sum x_i^2 \beta_1}{\sum x_i^2} \\
&= \beta_1 \frac{\sum x_i^2 \beta_1}{\sum x_i^2} \\
&= \beta_1
\end{aligned}$$

Thus since our estimator's expected value is β_1, we can conclude that the bias of our estimator is 0.

For the variance:

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_1] &= \mathrm{Var}\left[\frac{\sum x_i Y_i}{\sum x_i^2}\right] \\
&= \frac{\sum x_i^2}{\sum x_i^2 \sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{\sum x_i^2}{\sum x_i^2 \sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{1}{\sum x_i^2} \mathrm{Var}[Y_i] \\
&= \frac{1}{\sum x_i^2} \sigma^2 \\
&= \frac{\sigma^2}{\sum x_i^2}
\end{aligned}$$

Problem 3

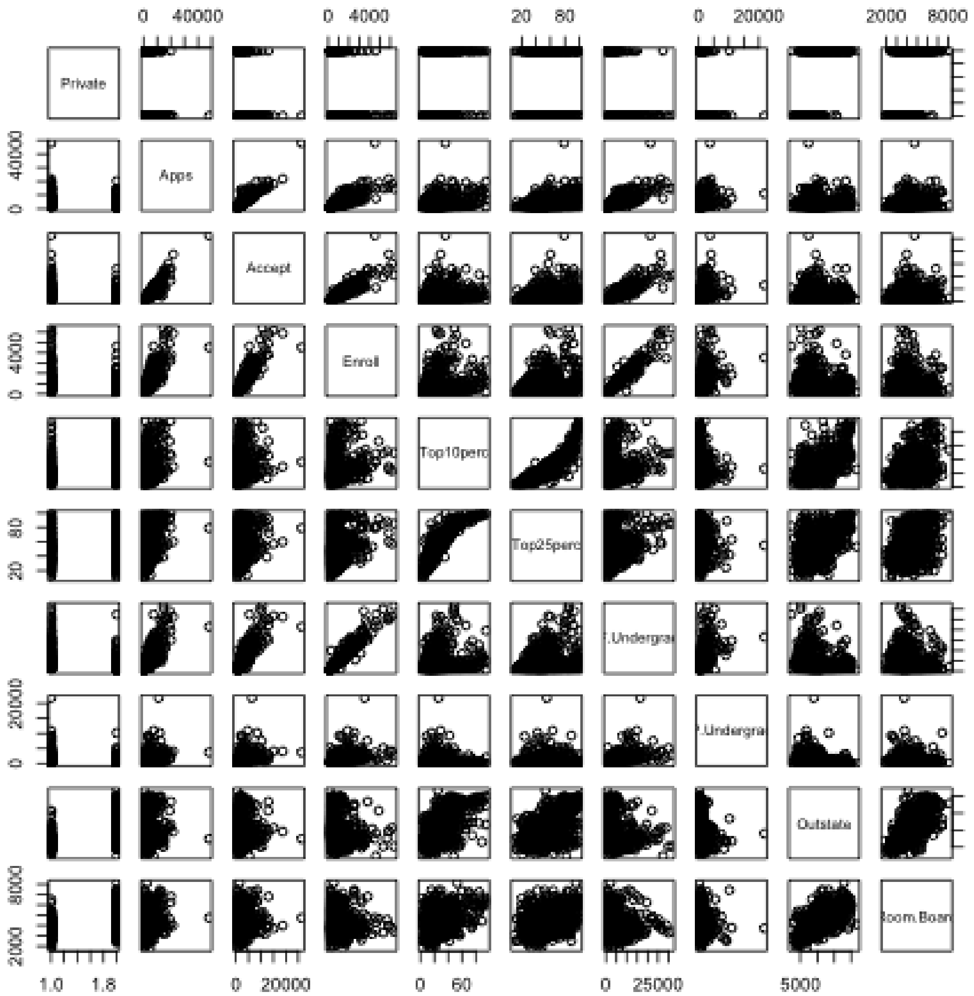
The code has been appended to the end of this PDF.

Figure 1: Summary from Part I

```
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  
```

Problem 3 continued on next page...

Figure 2: Pairs plot from Part II



Problem 3 continued on next page. . .

Figure 3: Boxplot of Outstate vs Private in Part III

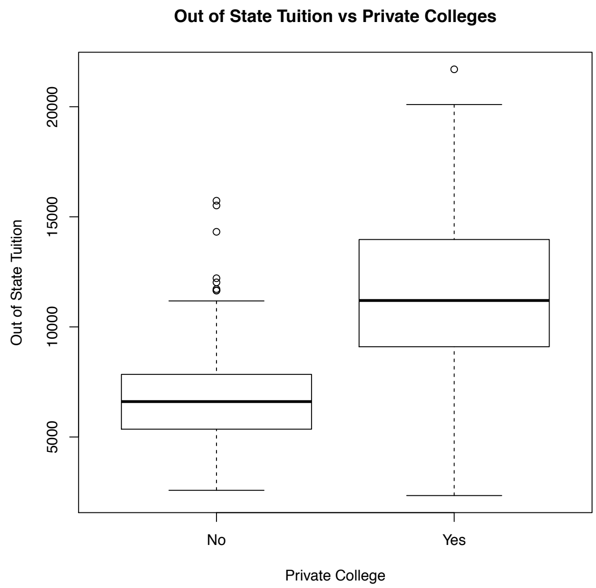
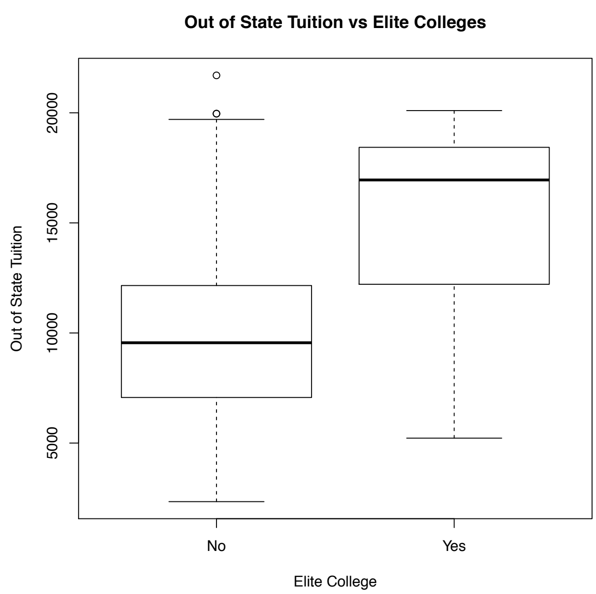
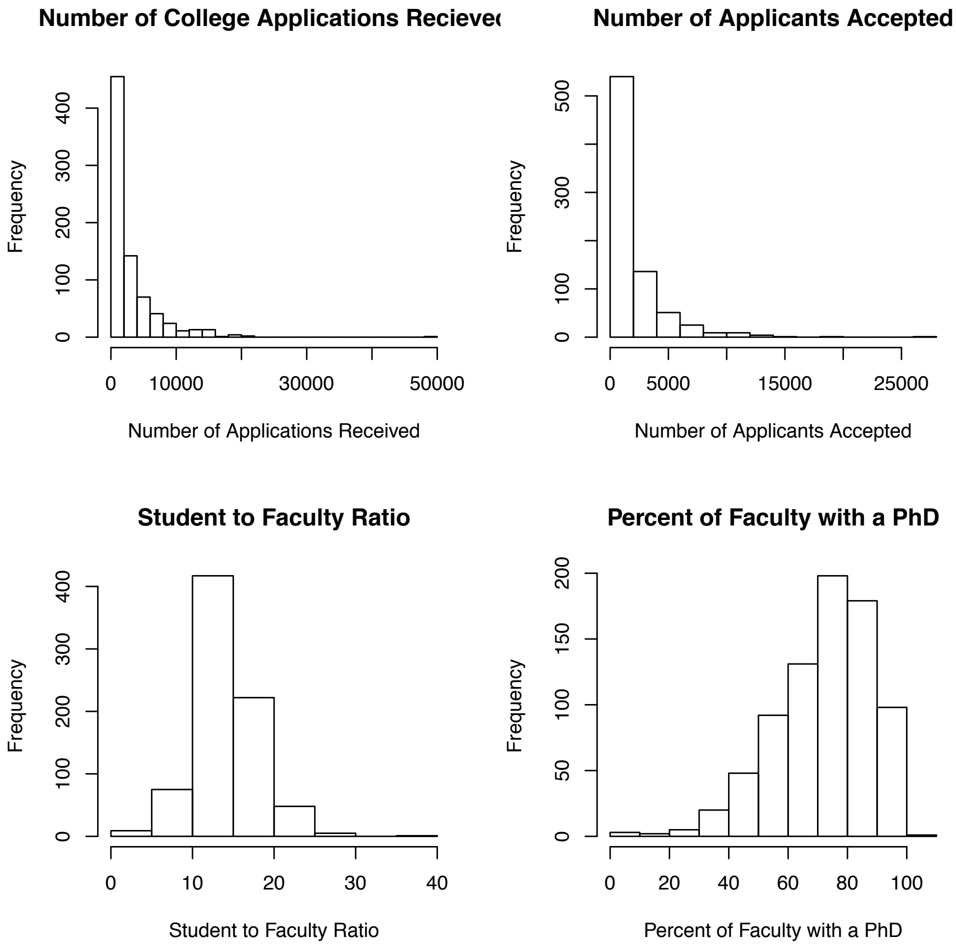


Figure 4: Boxplot of Elite colleges vs Outstate in Part IV



Problem 3 continued on next page. . .

Figure 5: Various histograms from Part V



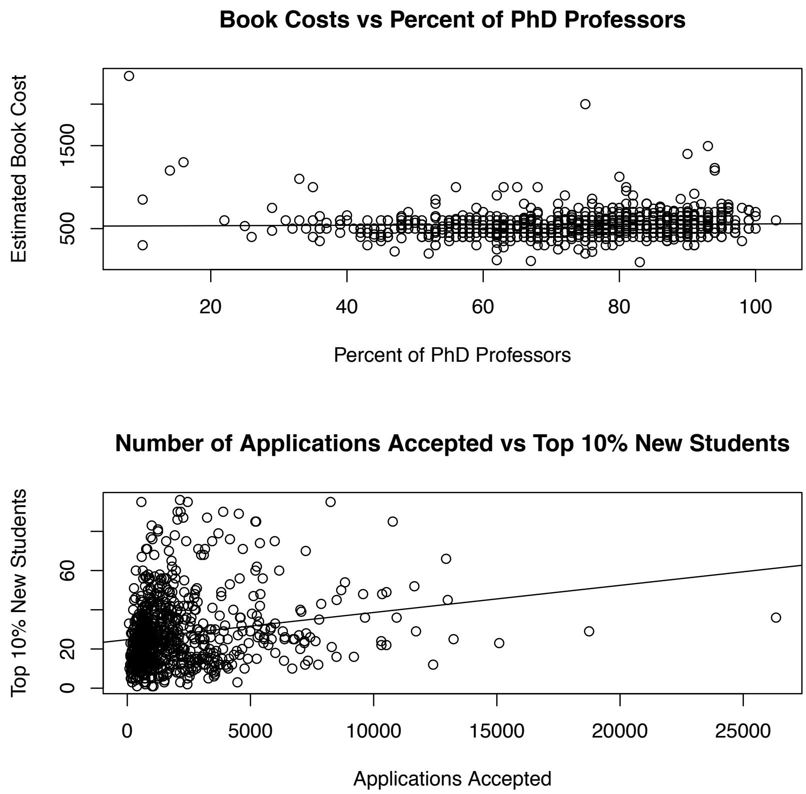
Below is my exploration of the data. I wanted to see if there were a few different relationships between a few different sets of the data.

As you can see, there seems to be a relationship between number of applications accepted and the percent of students coming from the top 10% of high school class.

However, there doesn't seem to be a relationship between book costs vs the percent of professors with a PhD.

Problem 3 continued on next page...

Figure 6: Part VI Exploration



Problem 4

Consider the following equation of a straight line $Y_i = \beta_0 + \beta_1 x_i + e_i$ with $i = 1, \dots, n$, $\mathrm{E}[e_i] = 0$, $\mathrm{Var}[e_i] = \sigma_e^2$, and $\mathrm{Cov}[e_i, e_j] = 0, \forall i \neq j$.

As in class, our estimator for β_1 is:

$$\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

which gives us:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1$$

as the two estimators for our line as given in the book and in lecture.

Part A

Calculate the bias for the estimator of the intercept $\hat{\beta}_0$.

Solution

In class, we determined that $\hat{\beta}_1$ is unbiased and thus $\mathrm{E}[\hat{\beta}_1] = \beta_1$.

Our expectation for $\hat{\beta}_0$ is thus:

$$\begin{aligned}
\mathrm{E}[\hat{\beta}_0] &= \mathrm{E}[\bar{y} - \hat{\beta}_1 \bar{x}] \\
&= \mathrm{E}[\bar{y}] - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \frac{1}{n} \sum \mathrm{E}[y_i] - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \frac{1}{n} \sum (\beta_0 + \beta_1 x_i) - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \frac{1}{n} \sum (\beta_1 x_i) - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \beta_1 \frac{1}{n} \sum (x_i) - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \beta_1 \bar{x} - \mathrm{E}[\hat{\beta}_1 \bar{x}] \\
&= \beta_0 + \beta_1 \bar{x} - \beta_1 \bar{x} \\
&= \beta_0
\end{aligned}$$

which shows that our estimator $\hat{\beta}_1$ is unbiased.

Problem 4 continued on next page. . .

Part B

Calculate the variance for the estimator of the intercept $\hat{\beta}_0$.

Solution

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_0] &= \mathrm{Var}[\bar{y} - \hat{\beta}_1 \bar{x}] \\
&= \mathrm{Var}[\bar{y}] + \mathrm{Var}[-\hat{\beta}_1 \bar{x}] + 2\mathrm{Cov}[\bar{y}, -\hat{\beta}_1 \bar{x}]
\end{aligned}$$

but by our assumption 3:

$$\begin{aligned}
\mathrm{Var}[\hat{\beta}_0] &= \mathrm{Var}[\bar{y}] + \mathrm{Var}[-\hat{\beta}_1 \bar{x}] + 0 \\
&= \frac{1}{n^2} \sum (\mathrm{Var}[y_i]) + \mathrm{Var}[-\hat{\beta}_1 \bar{x}] \\
&= \frac{n\sigma^2}{n^2} + \mathrm{Var}[-\hat{\beta}_1 \bar{x}] \\
&= \frac{\sigma^2}{n} + \mathrm{Var}[\hat{\beta}_1 \bar{x}] \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \mathrm{Var}[\hat{\beta}_1] \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \mathrm{Var}\left[\frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}\right] \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \mathrm{Var}\left[\sum (x_i - \bar{x})(y_i - \bar{y})\right] \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \left(\sum (x_i - \bar{x})^2\right) (\mathrm{Var}[y_i - \bar{y}]) \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \left(\frac{1}{\sum (x_i - \bar{x})^2}\right)^2 \left(\sum (x_i - \bar{x})^2\right) \sigma^2 \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \left(\frac{1}{\sum (x_i - \bar{x})^2}\right) \sigma^2 \\
&= \frac{\sigma^2}{n} + \frac{\bar{x}^2 \sigma^2}{\sum (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)
\end{aligned}$$

which agrees with equations 3.8 on page 66 of the textbook.

```r
# Problem 3
#

# Download data if it doesn't exist
data <- function() {
    if (!file.exists('./College.csv')) {
        download.file('http://www-bcf.usc.edu/~gareth/ISL/College.csv', destfile =
'./College.csv')
    }
}

#
# Part A:
# Read in the data from the CSV file.
#

# Read in Data
college <- read.csv('./College.csv', header = TRUE)

#
# Part B:
# Look at the data using various functions.
#

# View/edit the data
# fix(college)

# View the data
# View(college)

# Remove first column according to page 55
college <- college[,-1]

#
# Part C:
#

#
# Part I:
# Show a summary of the college data.
#

partI <- function() {
    college.summary <- capture.output(summary(college))
    cat(college.summary,
        file = 'partI.txt',
        sep = '\n')
```

```r
#
# Part II:
# Show a scatterplot of the first 10 columns of college.
#

partII <- function() {
    png('partII.png')
    pairs(college[,1:10])
    dev.off()
}

#
# Part III:
# Produce side-by-side boxplots of Outstate vs Private.
#

partIII <- function() {
    pdf('partIII.pdf')
    plot(college$Private, college$Outstate,
         main = 'Out of State Tuition vs Private Colleges',
         xlab = 'Private College',
         ylab = 'Out of State Tuition')
    dev.off()
}

#
# Create a new qualitative variable for Elite colleges. Show various statistics
# for the Elite colleges.
#

partIV <- function() {
    pdf('partIV.pdf')
    Elite <- rep('No', nrow(college))
    Elite[college$Top10perc > 50] <- 'Yes'
    Elite <- as.factor(Elite)
    college <- data.frame(college, Elite)

    # Show number of elite vs non-elite colleges
    summary(college)

    # Show boxplot for Outstate vs Elite
    plot(college$Elite, college$Outstate,
         main = 'Out of State Tuition vs Elite Colleges',
         xlab = 'Elite College',
         ylab = 'Out of State Tuition')
    dev.off()
```

```r
#
# Part V:
# Show some histograms with differing numbers of bins for a few quantitative
# variables.
#

partV  <- function() {
    pdf('partV.pdf')
    par(mfrow=c(2,2))
    hist(college$Apps, 20,
         main = 'Number of College Applications Recieved',
         xlab = 'Number of Applications Received')
    hist(college$Accept, 10,
         main = 'Number of Applicants Accepted',
         xlab = 'Number of Applicants Accepted')
    hist(college$S.F.Ratio, 10,
         main = 'Student to Faculty Ratio',
         xlab = 'Student to Faculty Ratio')
    hist(college$PhD, 10,
         main = 'Percent of Faculty with a PhD',
         xlab = 'Percent of Faculty with a PhD')
    dev.off()
}

#
# Part VI:
# Continue exploring the data and report what you find.
#

partVI <- function() {
    pdf('partVI.pdf')

    par(mfrow=c(2,1))
    df <- data.frame(x = college$PhD)
    df$y <- college$Books

    T <- lm(y~x,data=df)

    plot(college$PhD, college$Books,
         main = 'Book Costs vs Percent of PhD Professors',
         xlab = 'Percent of PhD Professors',
         ylab = 'Estimated Book Cost')
    abline(T)

    df <- data.frame(x = college$Accept)
    df$y <- college$Top10perc
```

```
    T <- lm(y~x,data=df)

    plot(college$Accept, college$Top10perc,
         main = 'Number of Applications Accepted vs Top 10% New Students',
         xlab = 'Applications Accepted',
         ylab = 'Top 10% New Students')
    abline(T)

    dev.off()
}

# Runs all of the parts of the homework
run <- function() {
    data()
    partI()
    partII()
    partIII()
    partIV()
    partV()
    partVI()
}
```