ISLR Chapter 2

Abhirup Sen

5/4/2021

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.

Obs: For this question, i did not use the csv file, the file was not found when i called it after loaded College dataset, so i had to skip some questions.

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.

```
# library(ISLR)
# detach(college)
college_newdataset = read.csv("College.csv")
dim(college_newdataset)
```

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

```
college = na.omit(college_newdataset)
dim(college)
```

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

c) Now, you should see that the first data column is Private. Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

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

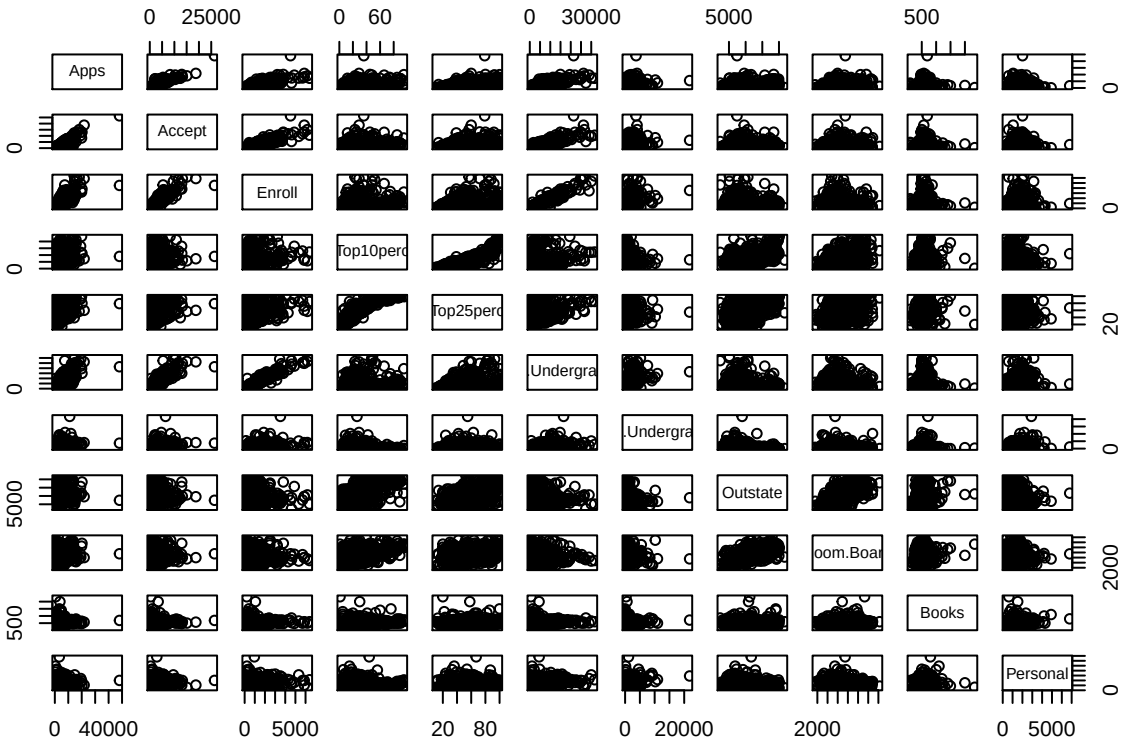
```
summary(college)
```

```
##       X                Private               Apps           Accept
##  Length:777         Length:777         Min.   :   81   Min.   :   72
##  Class :character   Class :character   1st Qu.:  776   1st Qu.:  604
##  Mode  :character   Mode  :character   Median : 1558   Median : 1110
##                                        Mean   : 3002   Mean   : 2019
##                                        3rd Qu.: 3624   3rd Qu.: 2424
##                                        Max.   :48094   Max.   :26330
##      Enroll       Top10perc       Top25perc      F.Undergrad
##  Min.   :  35   Min.   : 1.00   Min.   :  9.0   Min.   :  139
##  1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992
##  Median : 434   Median :23.00   Median : 54.0   Median : 1707
##  Mean   : 780   Mean   :27.56   Mean   : 55.8   Mean   : 3700
##  3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005
##  Max.   :6392   Max.   :96.00   Max.   :100.0   Max.   :31643
##   P.Undergrad         Outstate       Room.Board       Books
##  Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
##  1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
```

```
##  Median :  353.0   Median : 9990   Median :4200   Median : 500.0
##  Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
##  3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
##  Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
##     Personal         PhD            Terminal       S.F.Ratio
##  Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
##  1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
##  Median :1200   Median : 75.00   Median : 82.0   Median :13.60
##  Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
##  3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
##  Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80
##   perc.alumni        Expend        Grad.Rate
##  Min.   : 0.00   Min.   : 3186   Min.   : 10.00
##  1st Qu.:13.00   1st Qu.: 6751   1st Qu.: 53.00
##  Median :21.00   Median : 8377   Median : 65.00
##  Mean   :22.74   Mean   : 9660   Mean   : 65.46
##  3rd Qu.:31.00   3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :64.00   Max.   :56233   Max.   :118.00
```

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].

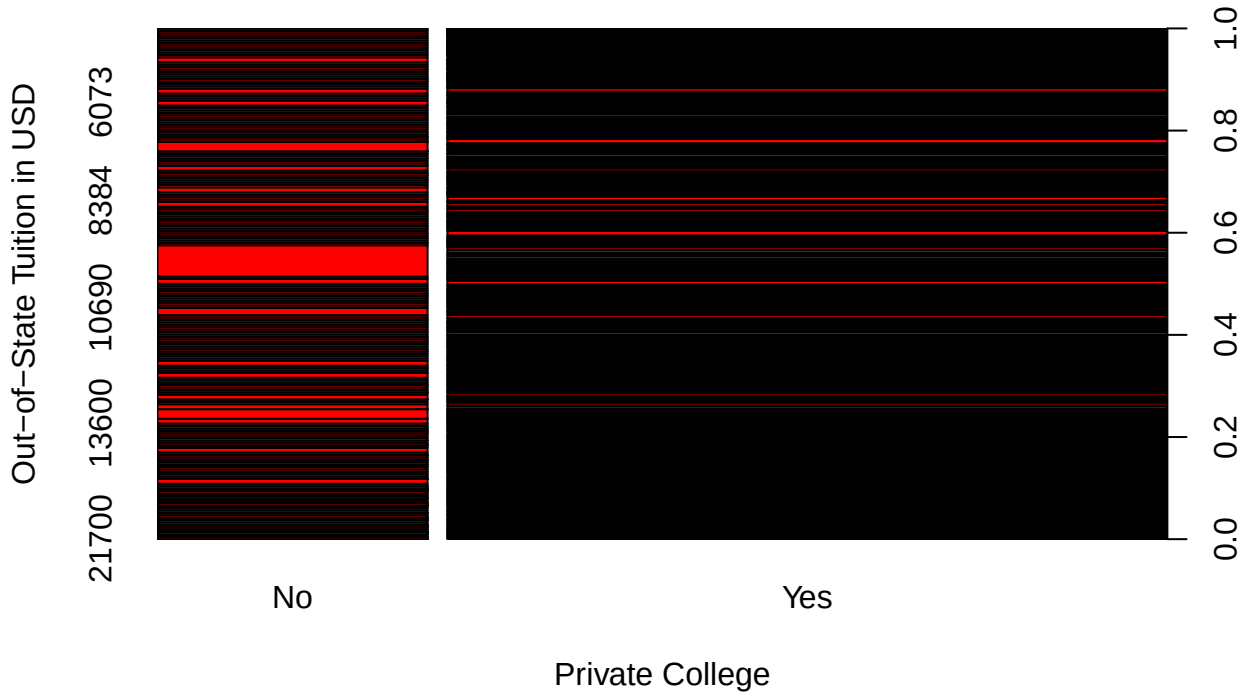
```
pairs(college[, 3:13])
```



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

```
plot(as.factor(college$Private), as.factor(college$Outstate), varwidth=T, col="red",
     xlab="Private College", ylab="Out-of-State Tuition in USD",
     main="Distribution Along the Colleges")
```

```
## Warning in rect(xleft, ybottom, xright, ytop, col = col, ...): "varwidth" is not
## a graphical parameter
```



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%.

```
Elite=rep("No", nrow(college))
Elite[college$Top10perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college, Elite)
```

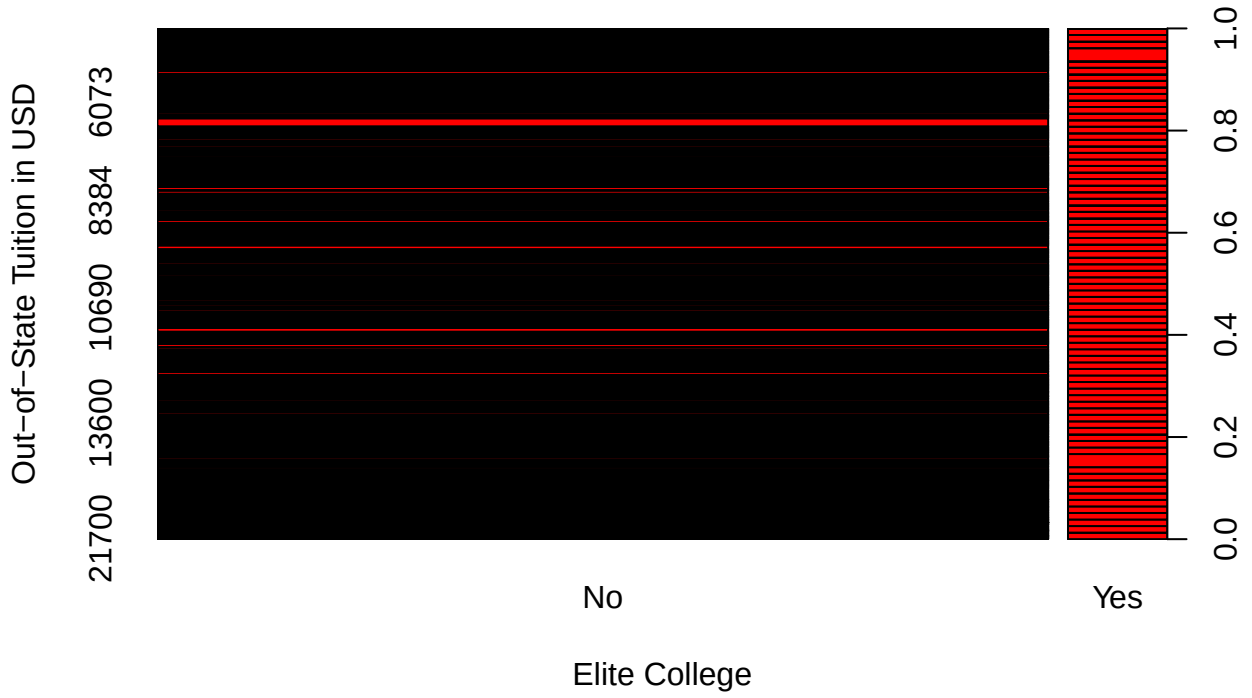
Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

```
summary(Elite)
```

```
##  No Yes
## 699  78
```

```
plot(Elite, as.factor(college$Outstate), varwidth=T, col="red",
     xlab="Elite College", ylab="Out-of-State Tuition in USD",
     main="Distribution Along the Elite Colleges")
```

```
## Warning in rect(xleft, ybottom, xright, ytop, col = col, ...): "varwidth" is not
## a graphical parameter
```

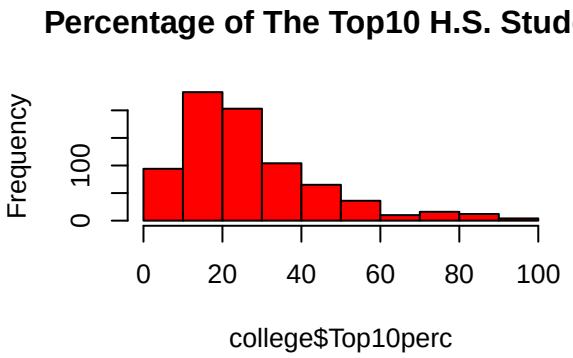


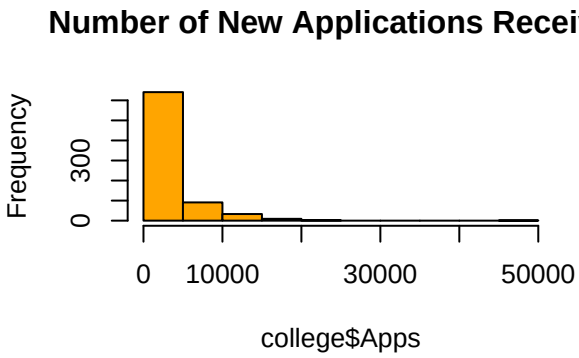
V. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to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

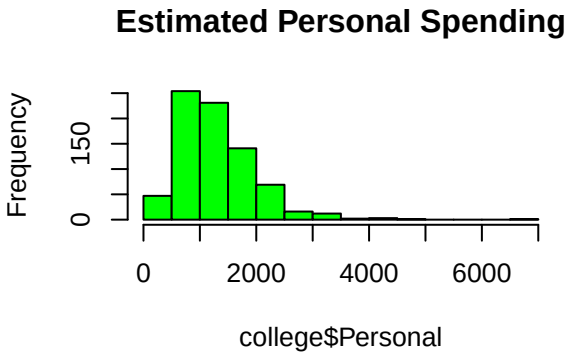
```
par(mfrow=c(2,2))
colnames(college)
```

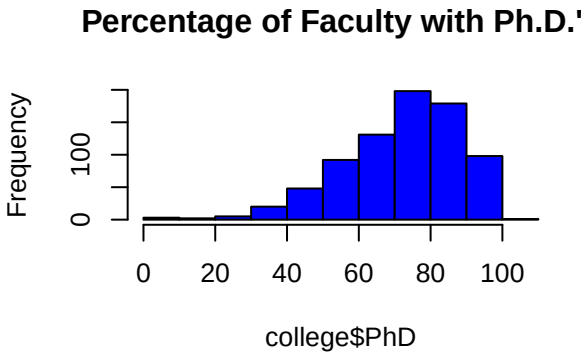
```
##  [1] "X"           "Private"     "Apps"        "Accept"      "Enroll"
##  [6] "Top10perc"   "Top25perc"   "F.Undergrad" "P.Undergrad" "Outstate"
## [11] "Room.Board"  "Books"       "Personal"    "PhD"         "Terminal"
## [16] "S.F.Ratio"   "perc.alumni" "Expend"      "Grad.Rate"   "Elite"
```

```
hist(college$Top10perc, breaks=10, col="red", main="Percentage of The Top10 H.S. Students")
hist(college$Apps, breaks=10, col="orange", main="Number of New Applications Received")
hist(college$Personal, breaks=10, col="green", main="Estimated Personal Spending")
hist(college$PhD, breaks=10, col="blue", main="Percentage of Faculty with Ph.D.'s")
```









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

```
summary(college$PhD)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

It is strange that there is a college with more than 100% of percentage, checking the college or colleges.

```
row.names(college[college$PhD>100, ])
```

```
## [1] "583"
```

Also there's an isolated university who receive a very larger number of applications than others, approximately 50 thousands.

```
summary(college$Apps)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      81     776    1558    3002    3624   48094
```

```
row.names(college[college$Apps>25000, ])
```

```
## [1] "484"
```