An Introduction to Statistical Learning 2.4 Exercise 8

Soodabeh

September 25

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. Before reading the data into R, it can be viewed in Excel or a text editor.

- `Private` : Public/private indicator
- `Apps` : Number of applications received
- `Accept` : Number of applicants accepted
- `Enroll` : Number of new students enrolled
- `Top10perc` : New students from top 10 % of high school class
- `Top25perc` : New students from top 25 % of high school class
- `F.Undergrad` : Number of full-time undergraduates
- `P.Undergrad` : Number of part-time undergraduates
- `Outstate` : Out-of-state tuition
- `Room.Board` : Room and board costs
- `Books` : Estimated book costs
- `Personal` : Estimated personal spending
- `PhD` : Percent of faculty with Ph.D.'s
- `Terminal` : Percent of faculty with terminal degree
- `S.F.Ratio` : Student/faculty ratio
- `perc.alumni` : Percent of alumni who donate
- `Expend` : Instructional expenditure per student
- `Grad.Rate` : Graduation rate

(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.

(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. Try the following commands:

```
>rownames (college )=college [,1]
>fix(college)
```

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate

university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
>college =college [,-1]
>fix(college)
```

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 thatR is giving to each row.

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

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].

iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private. iv. Create a new qualitative variable, called Elite, bybinning 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.

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.

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

Answer:

According to the scatter plots in step (ii), there is a strong correlation between column Accept as the output variable and Apps as the input variable. There is also a correlation (with high variability) between column Outstate (as dependent variable) and column Top10perc (as independent variable) which means that besides Top10perc there are additional independent variables need to be considered to enable an accurate prediction of Outstate.

However, it is also evident that there is a strong correlation between columns Enroll and F.Undergrad whereas the correlation between Outstate and Room. Board is quite weak

with high range of variability implying that these correlations are not meaningful as an effective predictive tool. The comparison of two boxplots in step (iii) implies that the out-of-state tuition fees in private colleges are substantially higher than those for public colleges due to higher 1Q, 2Q, and 3Q.

The boxplots of out-of-state tuition versus Elite in (iv) indicate that the elite colleges have much higher out-of-state tuition fees compared to normal colleges. In step (v), I demonstrated four histograms. The top left panel displays a left-skewed histogram for percentage of faculty with Ph.D. degree showing that most of the colleges and universities have hired more than 60% of their faculties from Ph.D. degree holders. The top right panel displays the histogram of student/faculty ratio for all the colleges showing a normal distribution with a mean of 14. The bottom left panel shows the histogram of students from top 25% of their respective high school classes describing a normal distribution with an average of 55 percentage.

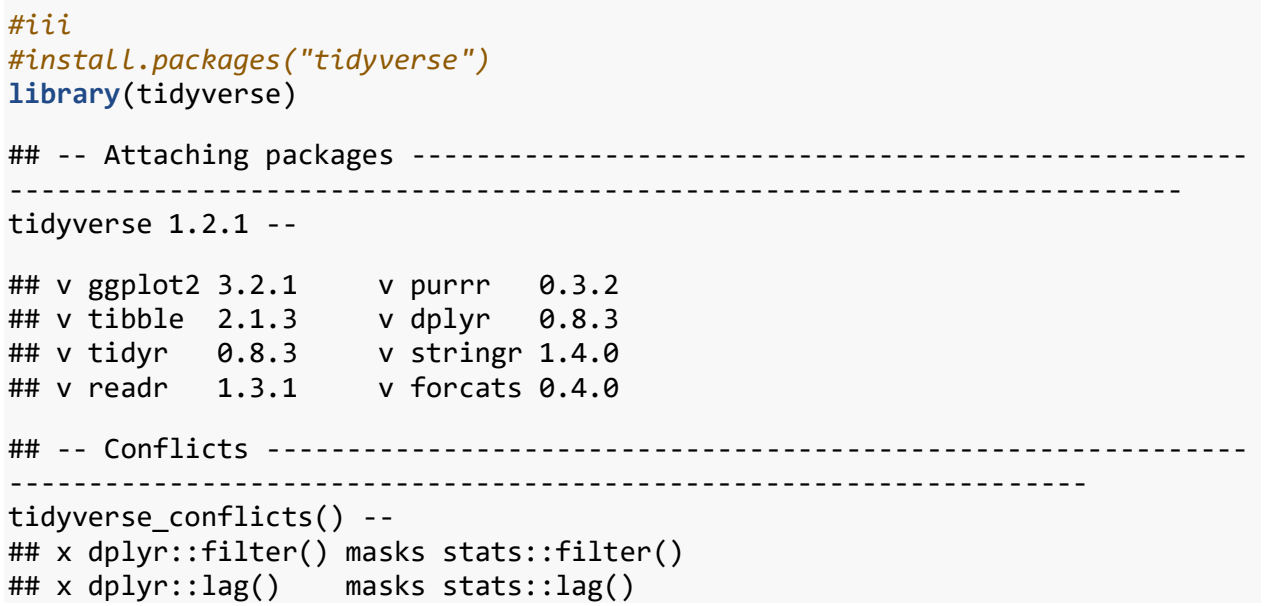
The histogram of new students accepted from top 10% of their high school classes, as shown in the bottom right panel, is right-skewed indicating that only a small fraction of schools admit more than 60% of their students from top 10% of their respective high schools.

```
#(a) read the data
college<- data.frame(read.csv("C:/Users/", stringsAsFactors = FALSE))
#View(college)
#rownames(college)

#(b) change the row names
fix(college)
rownames(college)=college [,1]
fix(college)

#View(college)
college =college [,-1]
fix(college)

#(c)
#i.
summary(college)
```

```
##    Private                Apps           Accept          Enroll
##  Length:777         Min.   :   81   Min.   :   72   Min.   :  35
##  Class :character   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##  Mode  :character   Median : 1558   Median : 1110   Median : 434
##                     Mean   : 3002   Mean   : 2019   Mean   : 780
##                     3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                     Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
```

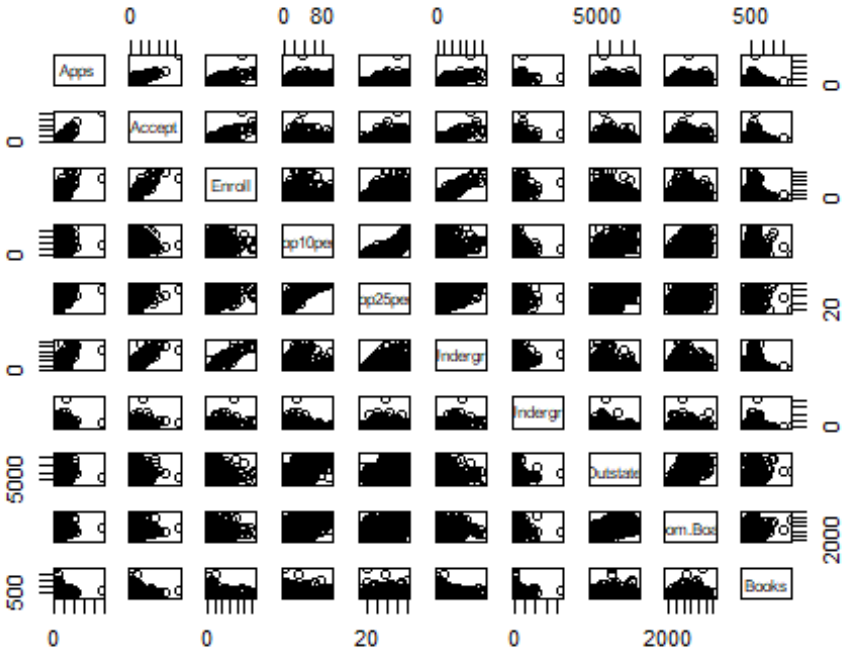
```
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend         Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

```
#ii.
pairs(college[,2:11])
```



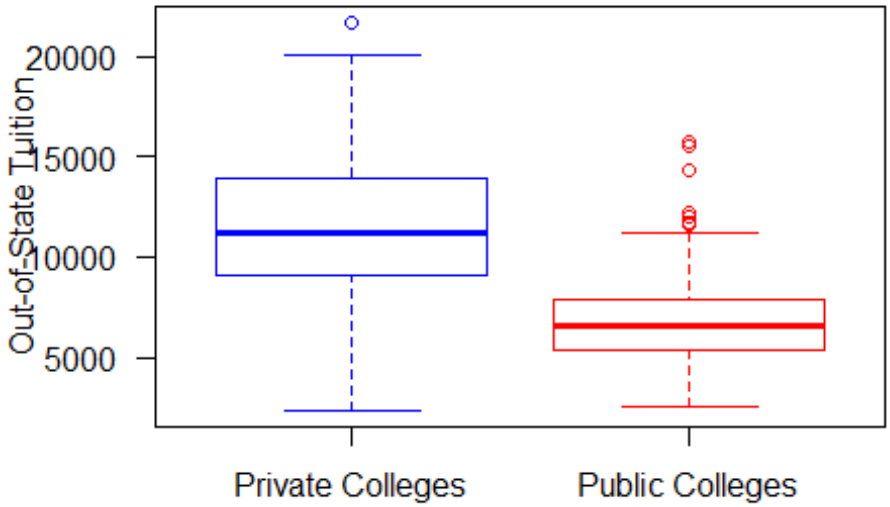
```
#iii
#install.packages("tidyverse")
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------
-------------------------------------------------------------------------
tidyverse 1.2.1 --

## v ggplot2 3.2.1     v purrr   0.3.2
## v tibble  2.1.3     v dplyr   0.8.3
## v tidyr   0.8.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.4.0

## -- Conflicts ------------------------------------------------------------
-----------------------------------------------------------------
tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

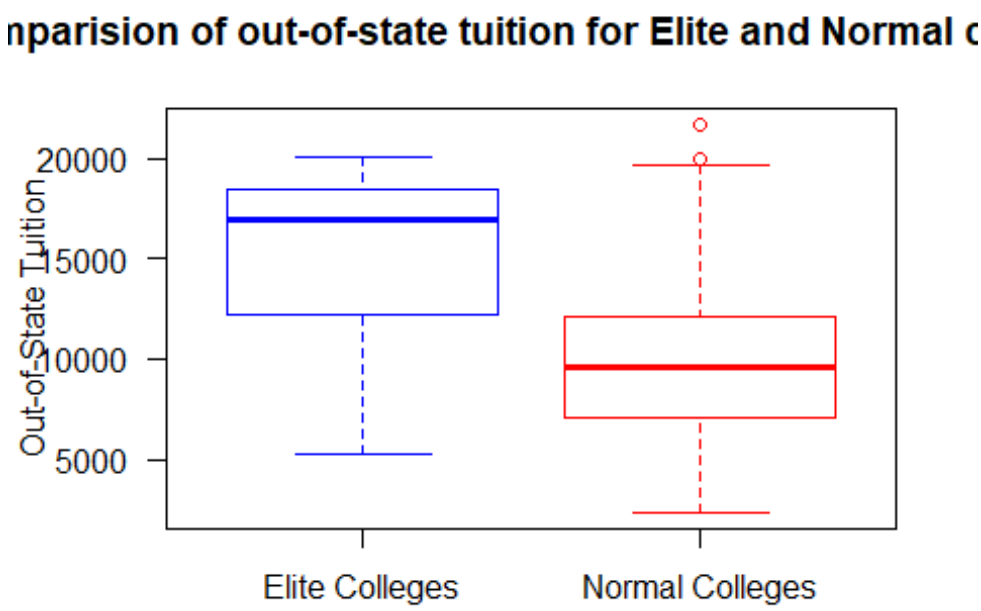



```
Private=filter(college,college$Private == "Yes")$Outstate
Public=filter(college,college$Private == "No")$Outstate

par(mfrow=c(1,1))
boxplot(Private,Public,
        main="Comparision of out-of-state tuition for Private and Public
colleges",
        ylab= "Out-of-State Tuition",
        at = c(1,2),
        names = c("Private Colleges", "Public Colleges"),
        las=1,
        col=c("white","white"),
        border =c("blue","red"),
        horizontal = F,
        notch = FALSE
)
```

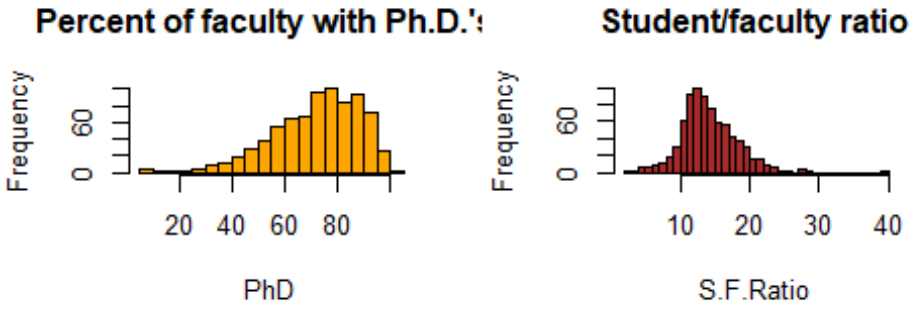
parision of out-of-state tuition for Private and Public

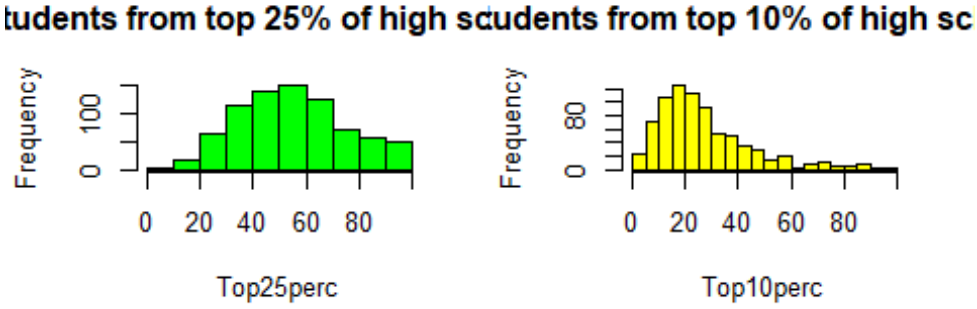



```
#iv.
Elite=rep("No",nrow(college ))
Elite[college$Top10perc >50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college ,Elite)

summary(college$Elite)

##  No Yes
## 699  78

Non_Elite=filter(college,college$Elite == "No")$Outstate
Elite=filter(college,college$Elite == "Yes")$Outstate

boxplot(Elite,Non_Elite,
        main="Comparision of out-of-state tuition for Elite and Normal
colleges",
        ylab= "Out-of-State Tuition",
        at = c(1,2),
        names = c("Elite Colleges", "Normal Colleges"),
        las=1,
        col=c("white","white"),
        border =c("blue","red"),
        horizontal = F,
        notch = FALSE
)
```




```r
#v.
par(mfrow=c(2,2))
hist(college$PhD,15, col="orange", xlab = "PhD", main="Percent of faculty
with Ph.D.'s ")
hist(college$S.F.Ratio,30, col="brown", xlab = "S.F.Ratio",
main="Student/faculty ratio")
hist(college$Top25perc,10, col="green", xlab = "Top25perc", main="New
students from top 25% of high school class")
hist(college$Top10perc,20,col="yellow", xlab = "Top10perc", main="New
students from top 10% of high school class ")
```






```
#hist(college$Grad.Rate,25, col="red", xlab = "Grad.Rate", main="Graduation
rate ")
#hist(college$Outstate,20, col="blue", xlab = "Outstate", main="Out-of-state
tuition")
mean(college$S.F.Ratio)

## [1] 14.0897

mean(college$Top25perc)

## [1] 55.79665

#vi.
```