i.Exercise2.4.8

Kaarthik, Sahil and Vidhi

5/24/2020

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.

```
getwd()
```

```
## [1] "C:/Users/Sahil Shah/Documents/GitHub/Data-Driven-Knowledge-Discovery/KDD"
```

```
college <- read.csv("College.csv")
```

(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:

```
#fix(college)
rownames ( college )= college [,1]
college = college [,-1]
##fix ( college )
```

```
library(knitr)
library(kableExtra)
kable(head(college), format = "latex", booktabs=T) %>%
  kable_styling(latex_options="scale_down")
```

	Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad	P.Undergrad	Outstate	Room.Board	Books	Personal	PhD	Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Abilene Christian University	Yes	1660	1232	721	23	52	2885	537	7440	3300	450	2200	70	78	18.1	12	7041	60
Adelphi University	Yes	2186	1924	512	16	29	2683	1227	12280	6450	750	1500	29	30	12.2	16	10527	56
Adrian College	Yes	1428	1097	336	22	50	1036	99	11250	3750	400	1165	53	66	12.9	30	8735	54
Agnes Scott College	Yes	417	349	137	60	89	510	63	12960	5450	450	875	92	97	7.7	37	19016	59
Alaska Pacific University	Yes	193	146	55	16	44	249	869	7560	4120	800	1500	76	72	11.9	2	10922	15
Albertson College	Yes	587	479	158	38	62	678	41	13500	3335	500	675	67	73	9.4	11	9727	55

```
# head(college)
# library()
```

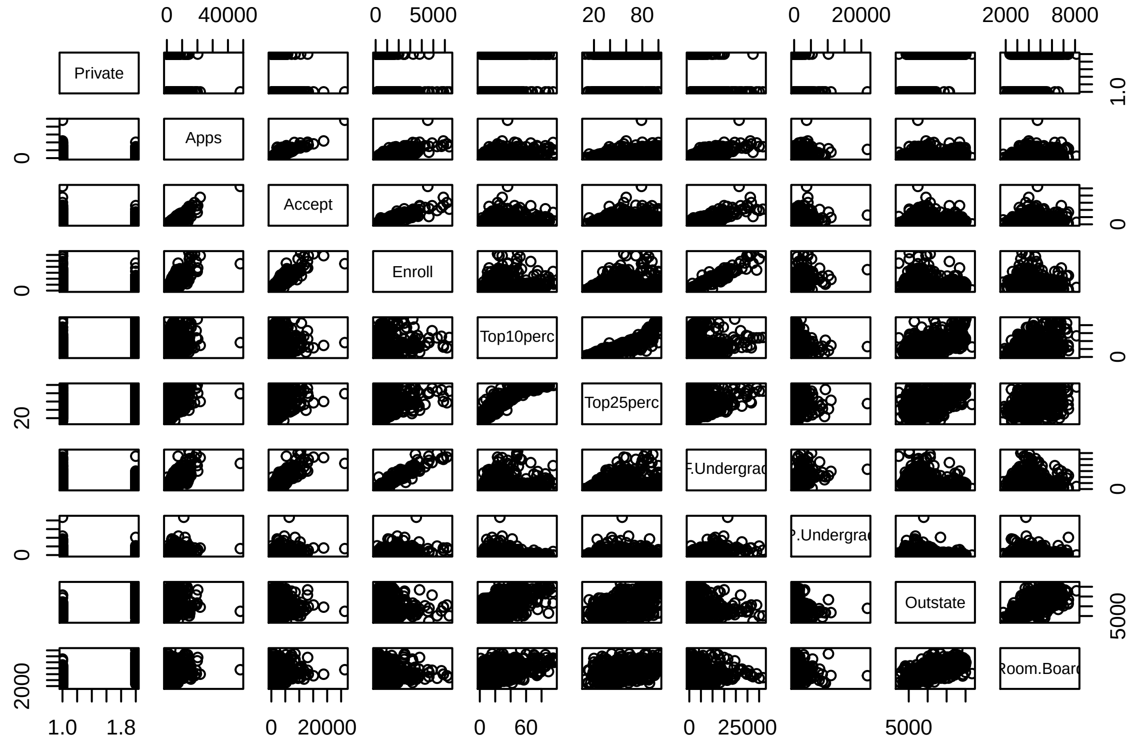
(c) i. Use the `summary()` function to produce a numerical summary of the variables in the data set.

```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
```

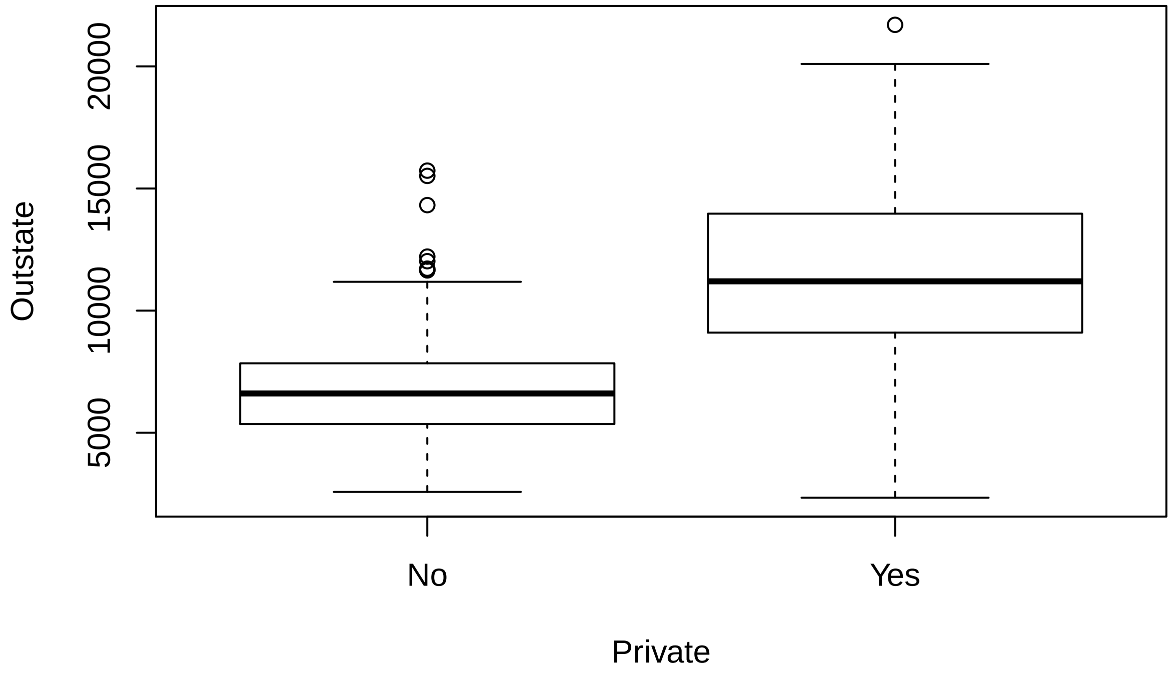
(c) ii. Use the `pairs()` function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

```r
pairs(college[,1:10])
```



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

```r
attach(college)
plot(Outstate~Private, data=college, xlab = "Private", ylab = "Outstate")
```

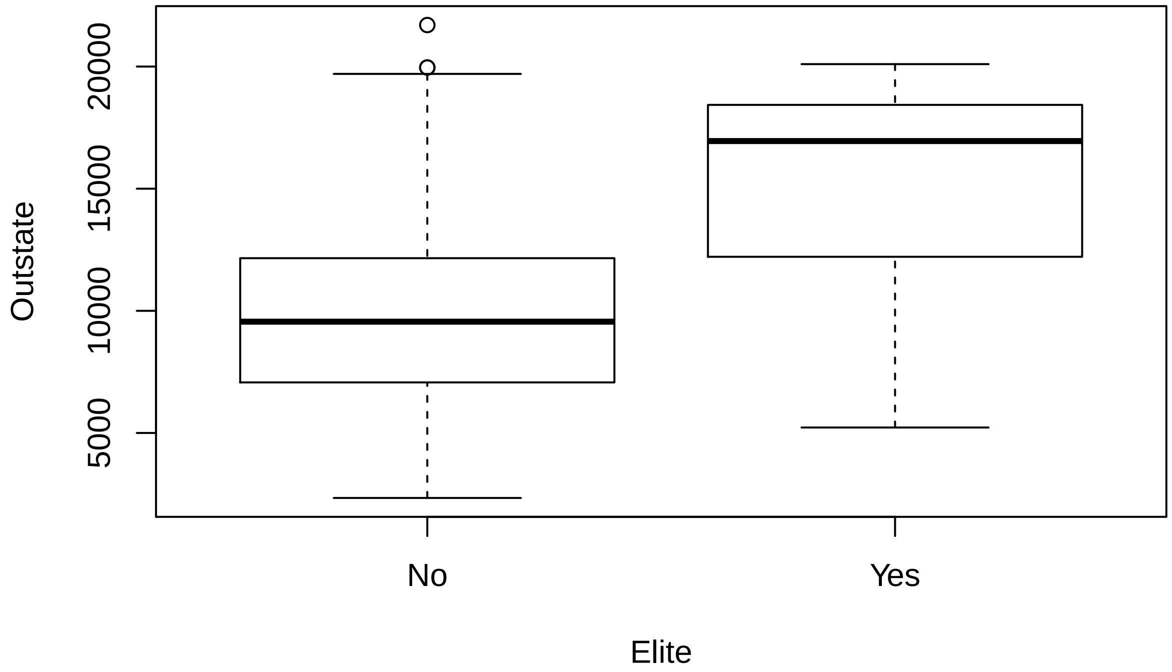


(c) iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

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

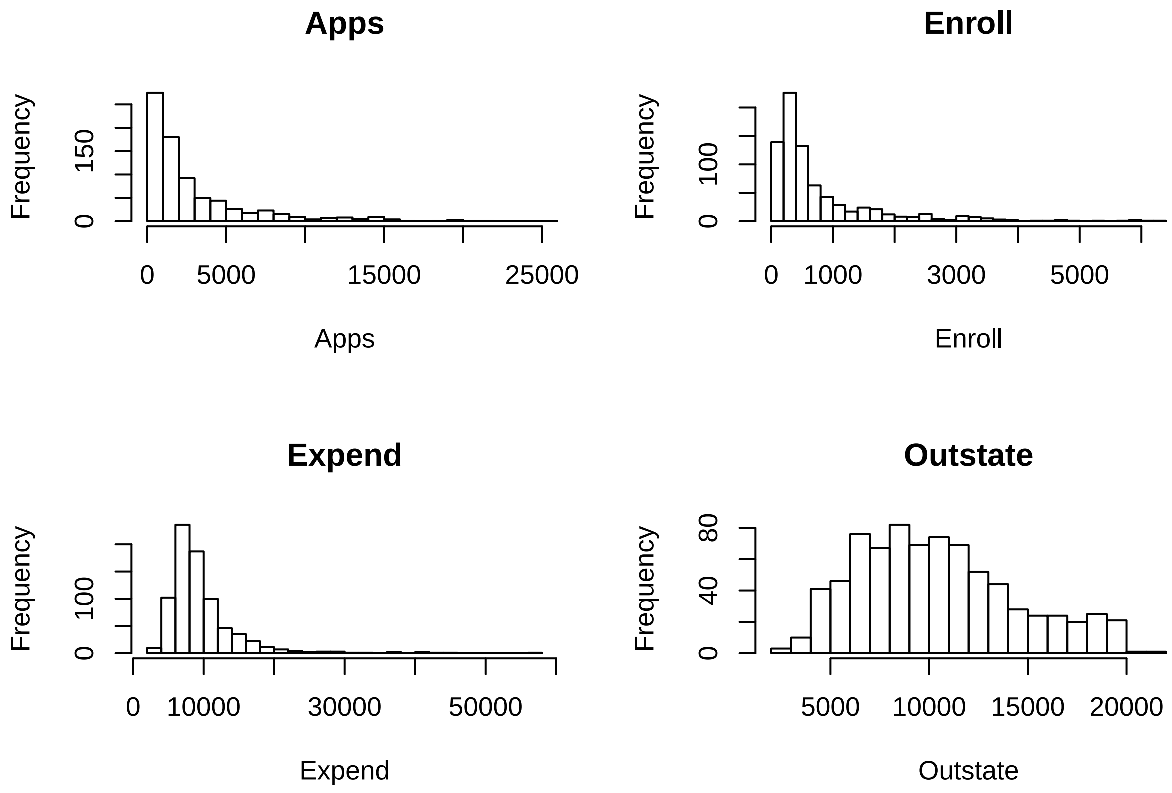
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            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   No :699  
##  1st Qu.: 53.00   Yes: 78  
##  Median : 65.00            
##  Mean   : 65.46            
##  3rd Qu.: 78.00            
##  Max.   :118.00            
```

```r
plot(Outstate~Elite, xlab = "Elite", ylab = "Outstate")
```



(c) 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.

```r
par(mfrow=c(2,2))
hist(Apps, breaks=50, xlim=c(0,25000), main="Apps")
hist(Enroll, breaks = 25, main = "Enroll")
hist(Expend, breaks = 25, main = "Expend")
hist(Outstate, breaks = 25, main = "Outstate")
```



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

```r
table(Private, Elite)
```

```
##        Elite
## Private  No Yes
##     No  199  13
##     Yes 500  65
```

```r
library(gmodels)
CrossTable(Private, Elite)
```

```
## 
## 
##    Cell Contents
## |-------------------------|
## |                       N |
## | Chi-square contribution |
## |           N / Row Total |
## |           N / Col Total |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  777
##
##
##              | Elite
##      Private |        No |       Yes | Row Total |
## -------------|-----------|-----------|-----------|
##           No |       199 |        13 |       212 |
##              |     0.360 |     3.223 |           |
##              |     0.939 |     0.061 |     0.273 |
##              |     0.285 |     0.167 |           |
##              |     0.256 |     0.017 |           |
## -------------|-----------|-----------|-----------|
##          Yes |       500 |        65 |       565 |
##              |     0.135 |     1.209 |           |
##              |     0.885 |     0.115 |     0.727 |
##              |     0.715 |     0.833 |           |
##              |     0.644 |     0.084 |           |
## -------------|-----------|-----------|-----------|
## Column Total |       699 |        78 |       777 |
##              |     0.900 |     0.100 |           |
## -------------|-----------|-----------|-----------|
##
##
```

According to the scatter plot from the **College** dataset, it is really hard to find the correlation in the small screen as the laptop. Although we can zoom in and find the relationship between the variables or plot the single graph with providing two variables as x and y using the `plot` function. From the variable Elite that we created that only hold the top 10 percent more than 50 as in yes. From the table shown above it generates the confusion matrix between the Elite and Private University.