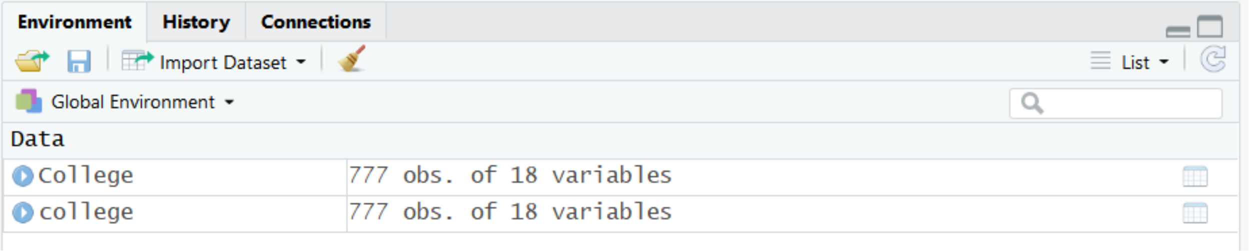
Assignment 1 Solution (Total: 20 pts)

This assignment relates to the ‘College’ data set, which can be found in the ‘ISLR’ library. It contains a number of variables for 777 different universities and colleges in the US. The variables are

- 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

1. Read the data into R. Call the loaded data “college”.
   The data college include 777 observations with 18 variables.

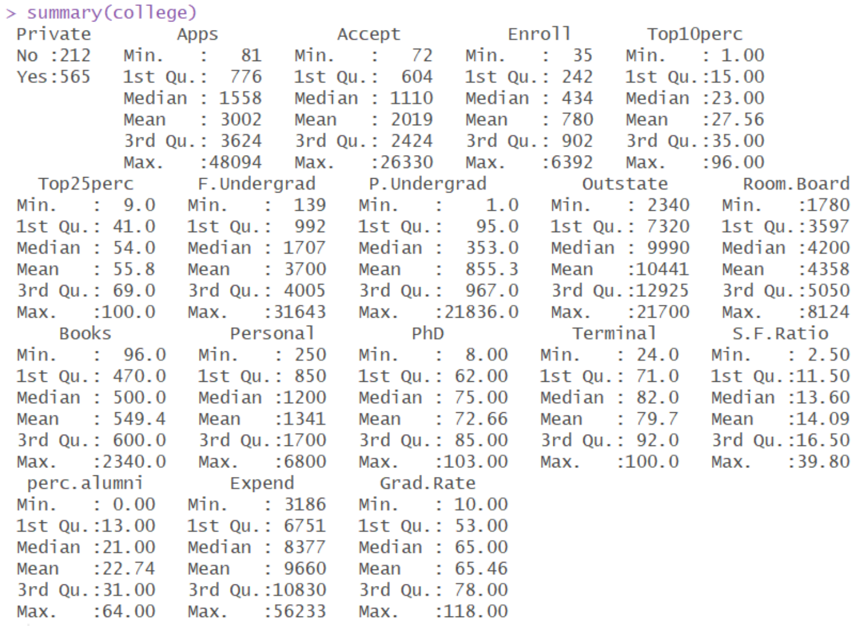


```
Data
College     777 obs. of 18 variables
college     777 obs. of 18 variables
```

2. Answer the following sub-questions
   i) (2 points; Identify the variable type) Use the “summary()” function to produce a numerical summary of the variables in the data set.

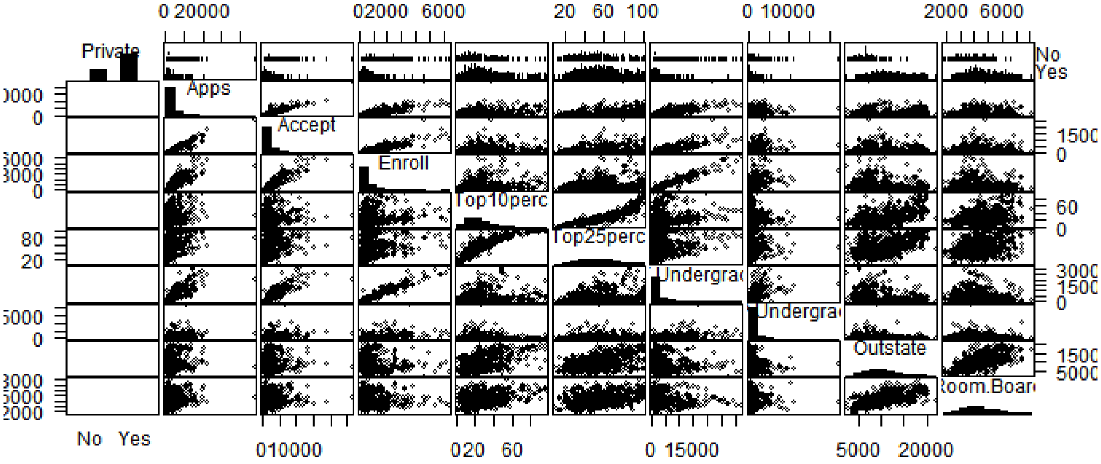
Categorical Variable	Private	2 levels
Quantitative Variables	Apps	Mean=3002, Median=1558
	Accept	Mean=1110, Median=2019
	Enroll	Mean=434, Median=780
	Top10perc	Mean=23, Median=27.56
	Top25perc	Mean=54, Median=55.8
	F.Undergrad	Mean=1707, Median=3700

P.Undergrad	Mean=353, Median=855.3
Outstate	Mean=9990, Median=10441
Rood.Board	Mean=4200, Median=4358
Books	Mean=500, Median=549.4
Personal	Mean=1200, Median=1341
Ph.D.	Mean=75, Median=72.66
Terminal	Mean=82, Median=79.7
S.F.Ratio	Mean=13.6, Median=14.09
Perc.alumi	Mean=21, Median=22.74
Expend	Mean=8377, Median=9660
Grad.Rate	Mean=65, Median=65.46

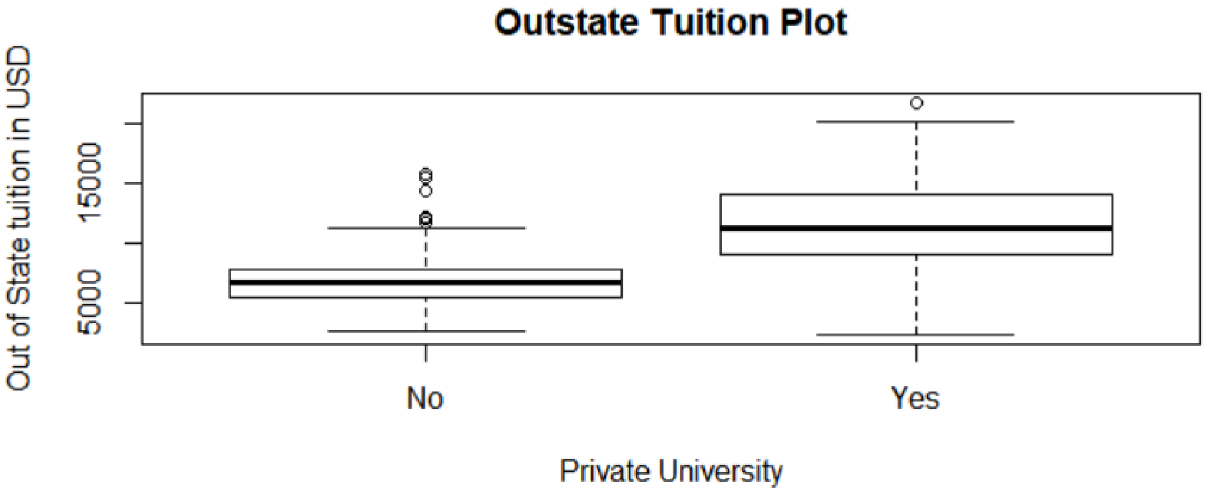


```
> summary(college)
 Private        Apps           Accept          Enroll         Top10perc
 No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
 Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
           Median : 1558   Median : 1110   Median : 434   Median :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       Room.Board
 Min.   :  9.0   Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780
 1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597
 Median : 54.0   Median : 1707   Median :  353.0   Median : 9990   Median :4200
 Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358
 3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050
 Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124
     Books           Personal         PhD            Terminal       S.F.Ratio
 Min.   :  96.0   Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50
 1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50
 Median : 500.0   Median :1200   Median : 75.00   Median : 82.0   Median :13.60
 Mean   : 549.4   Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09
 3rd Qu.: 600.0   3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50
 Max.   :2340.0   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
```

ii) (2 points) 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 then columns of a matrix A using A [,1:10].

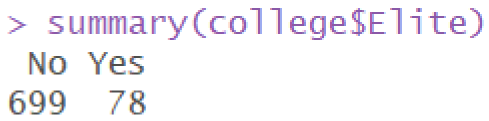


iii) (3 points) Use the “plot()” function to produce side-by-side boxplots of “Outstate” versus “Private”

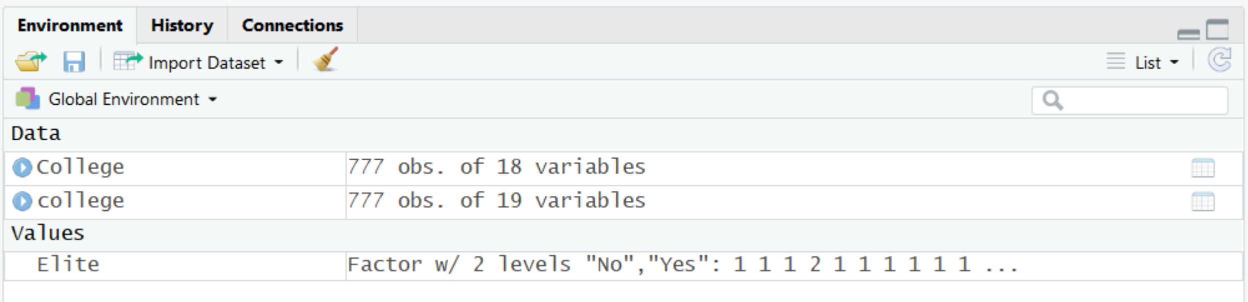


Based on the side-by-side box plot, the out of state tuitions (in $) between a private university and public university are different. The out-of-state tuitions in public university are about $5000 smaller than in public university. In addition, there are suspicious outliers for both groups.

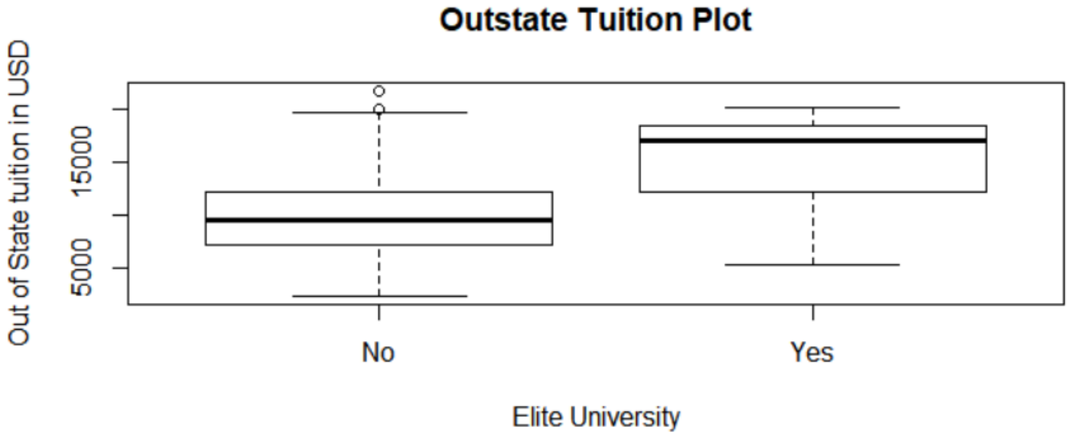
iv) (3 points) Create a new qualitative variable, called “Elite”, by binning the “Top10perc” variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.
Use the “summary()” function to see how many elite universities there are. Now use the “plot()” function to produce side-by-side boxplots of “Outstate” versus “Elite”



```
> summary(college$Elite)
 No Yes
699  78
```

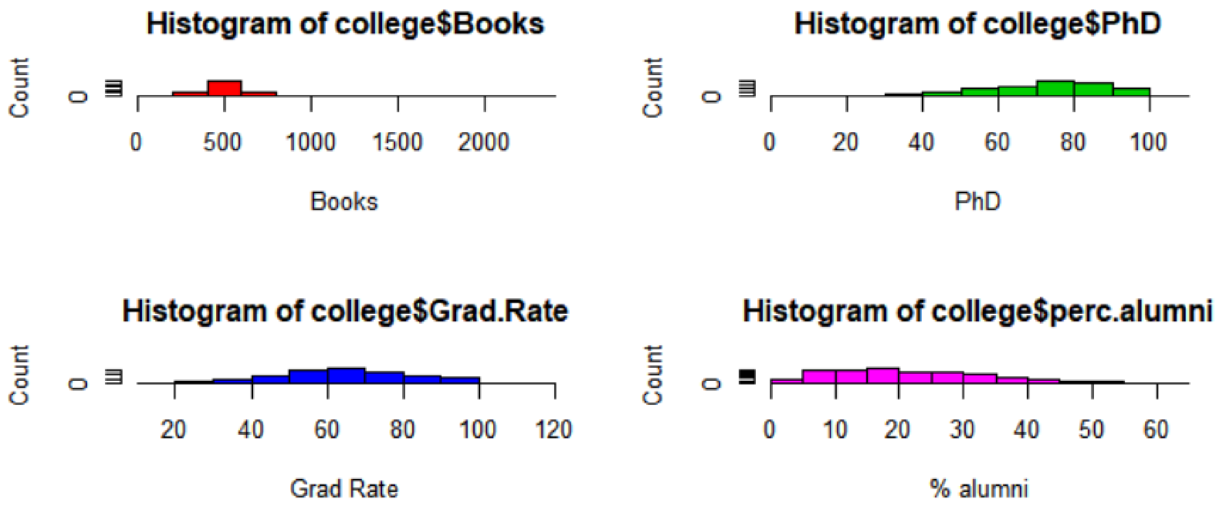


```
Data
 College     777 obs. of 18 variables
 college     777 obs. of 19 variables
Values
 Elite       Factor w/ 2 levels "No","Yes": 1 1 1 2 1 1 1 1 1 1 ...
```


Outstate Tuition Plot

A new variable, Elite University, was generated and there are 78 Elite universities and 699 nonElite universities in the updated dataset. Based on a side-by-side box plot, out of state tuition from Elite University is about $7000 larger than nonElite University. For the group of nonElite University, a couple of potential outliers are presented.

v) (3 points) Use the "hist()" function to produce some histogram 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.


Histogram of college$Books

Histogram of college$PhD

Histogram of college$Grad.Rate

Histogram of college$perc.alumni

Books: Unimodal, Symmetric, no outlier
Ph.D: Unimodal, Asymmetric (skewed to the left), no outlier
Grad. Rate: Unimodal, Symmetric, no outlier
Perc. Alumni: Unimodal, Asymmetric (skewed to the right), no outlier

vi) (2 points) Continue exploring the data, and provide a brief summary of what you discover.