Assignment 2

1(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

```
college <- read.csv("https://scads.eecs.wsu.edu/wp-content/uploads/2017/09/College.csv")
```

1(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

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

However, we still need to eliminate the first column in the data where the names are stored. Try

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

1(c).i. Use the summary() function to produce a numerical summary of the variables in the data set. (Respond to this question with the mean graduation rate included in the summary result).

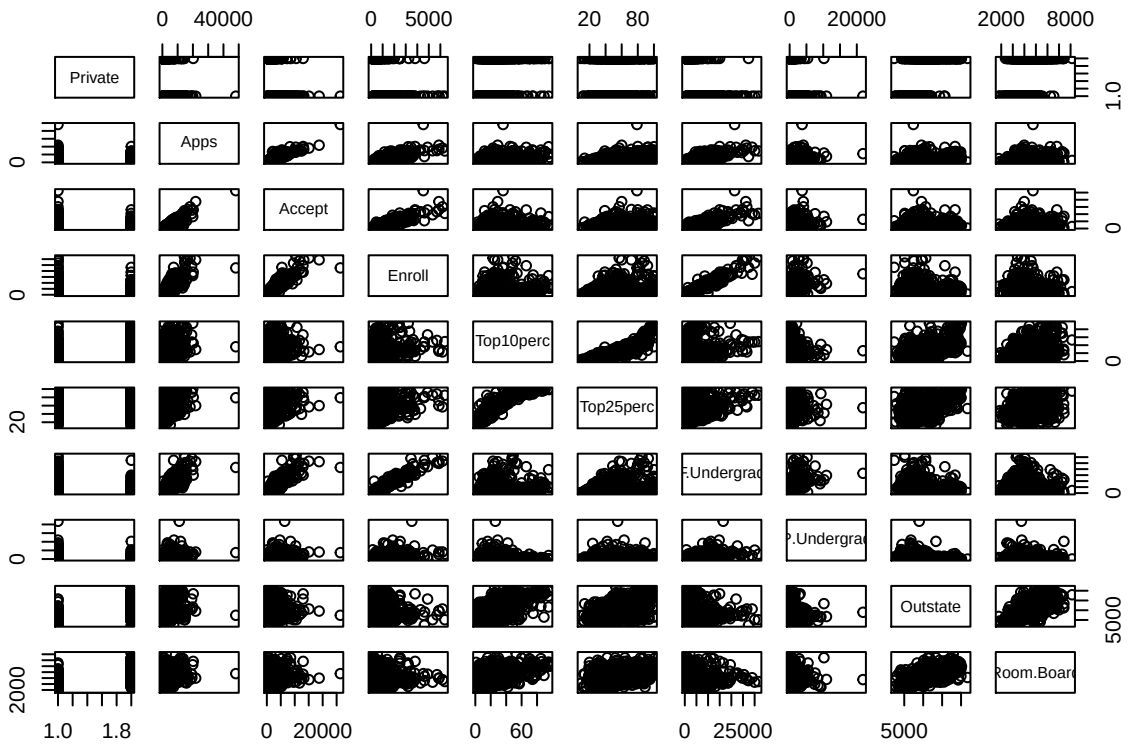
```
summary(college)
```

```
##  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
```

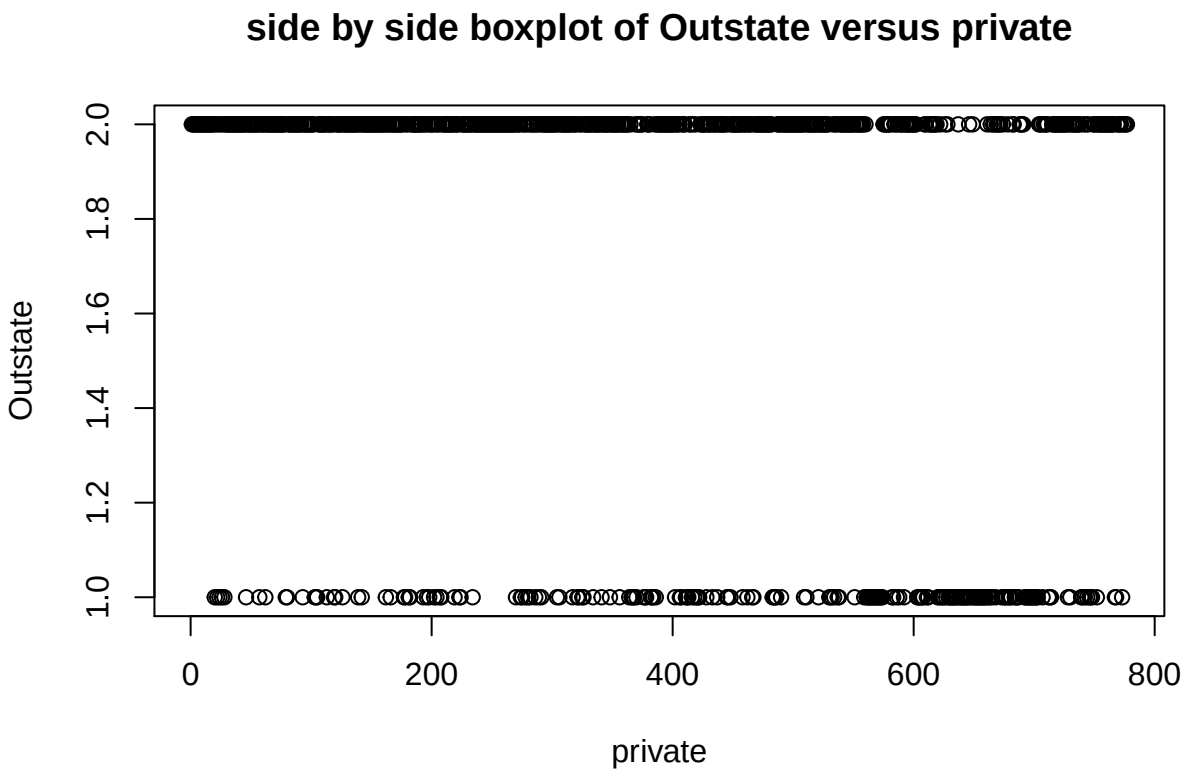
ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].

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

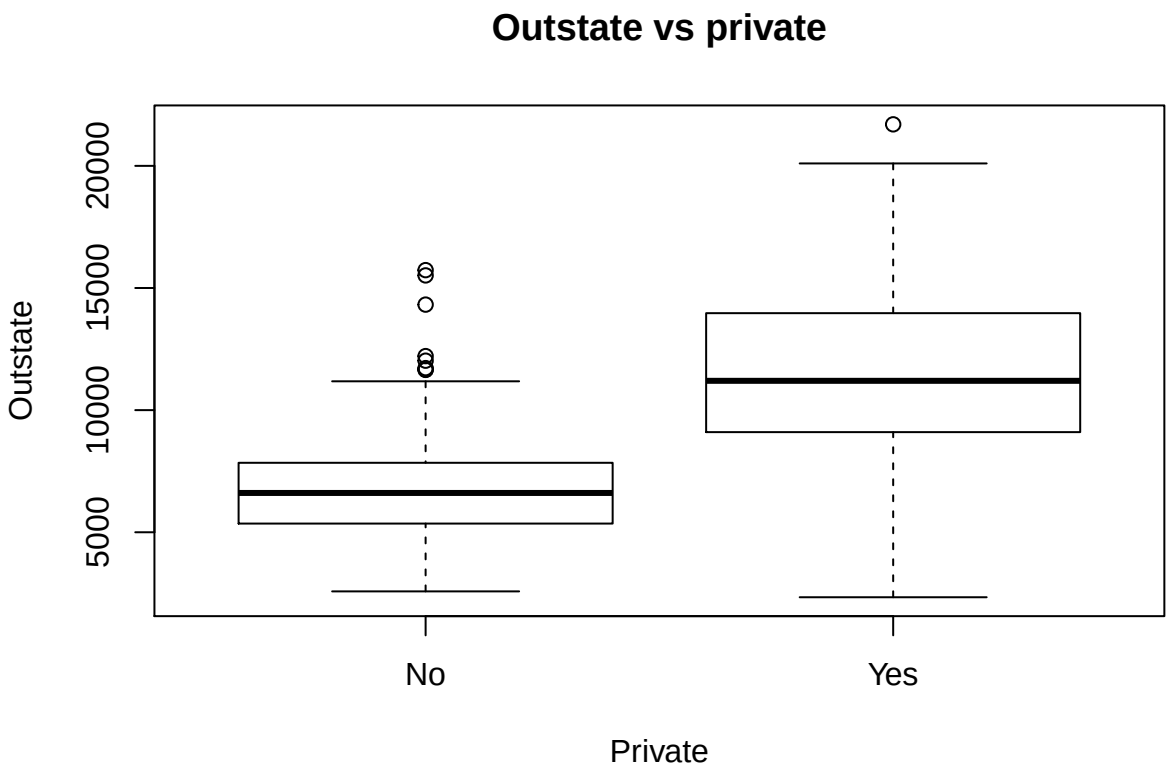


iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private.

```
plot(y=college$Oustate, x=college$Private, main="side by side boxplot of Outstate versus private", xlal
```




```
boxplot(Outstate ~ Private, data = college, xlab = "Private", ylab = "Outstate", main = "Outstate vs pri
```



iv. Create a new qualitative variable, called Top, by binning the Top25perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 25% of their high school classes exceeds 50%.

```
Top=rep("No",nrow(college ))
 Top[college$Top25perc >50]=" Yes"
   Top=as.factor(Top)
 college=data.frame(college, Top)
```

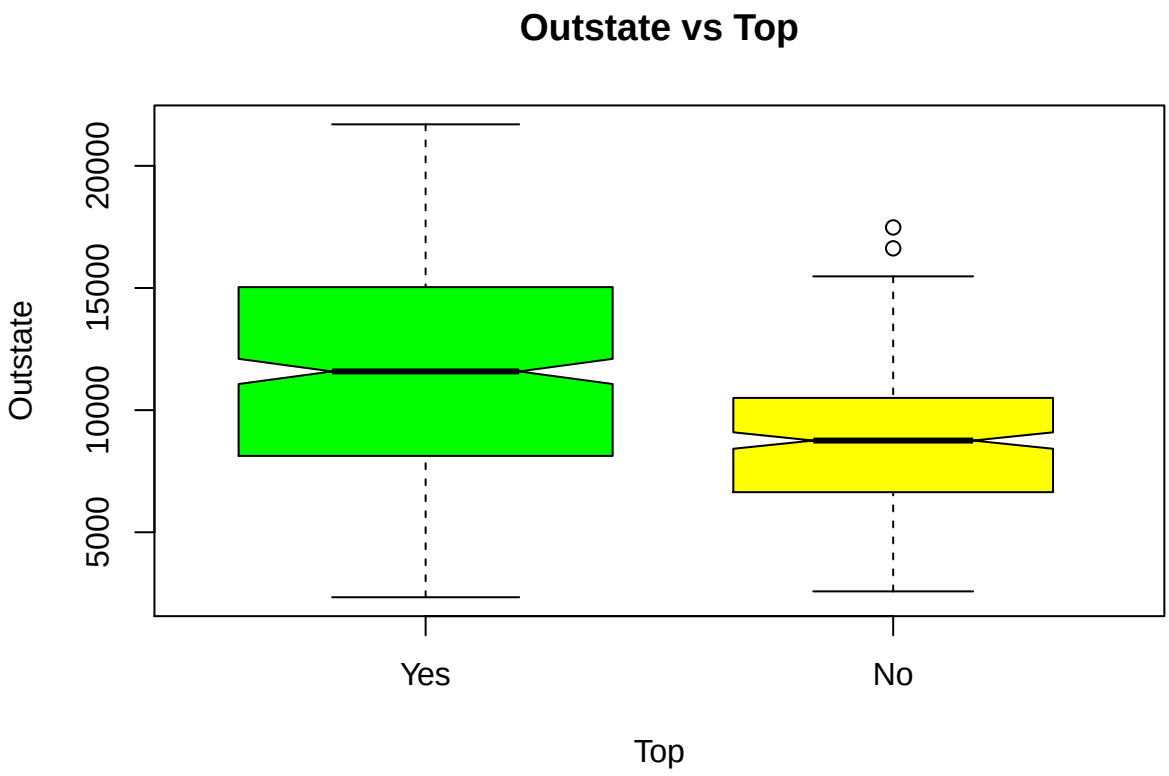
Use the summary() function to see how many top universities there are.

```
summary(Top)
```

```
##  Yes   No
##  449  328
```

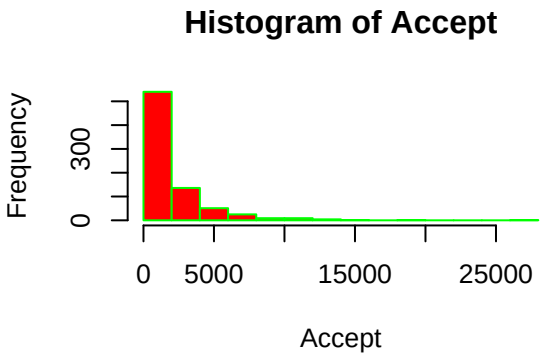
Now use the plot() or boxplot() function to produce side-by-side boxplots of Outstate with respect to the two Top categories (Yes and No). Ensure that this figure has an appropriate title and axis labels.

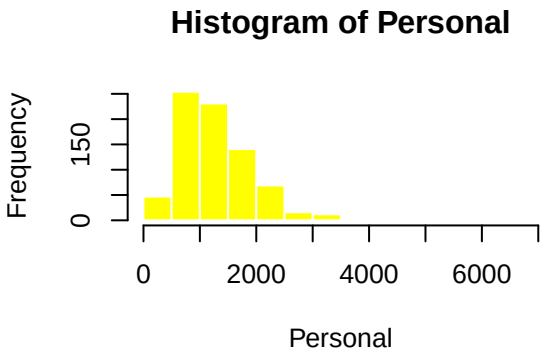
```
boxplot(Outstate ~ Top, data = college, xlab = "Top",
    ylab = "Outstate",
     main = "Outstate vs Top",
    notch = TRUE,
    varwidth = TRUE,
    col = c("green","yellow"),
    names = c("Yes","No"))
```

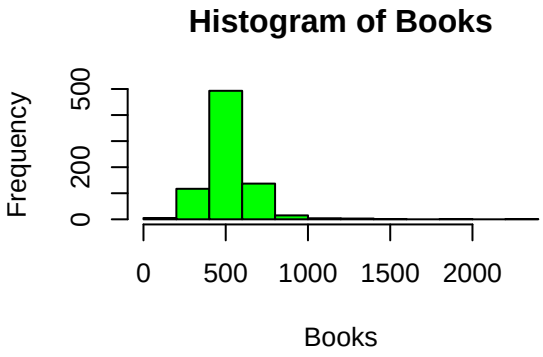


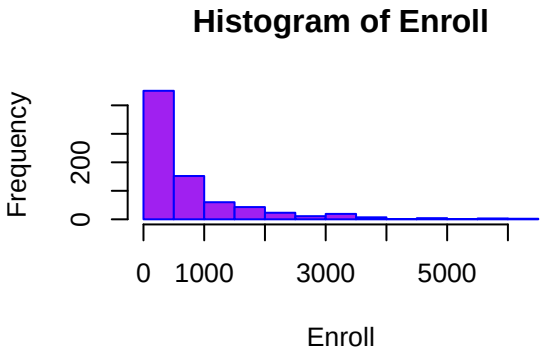
v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to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

```
par(mfrow=c(2,2))
Enroll<- college$Enroll
Books<- college$Books
Personal<- college$Personal
Accept<- college$Accept
hist(Accept, xlab = "Accept",col = "red",border = "green")
hist(Personal, xlab = "Personal",col = "yellow",border = "White")
hist(Books, xlab = "Books",col = "green",border = "black")
hist(Enroll, xlab = "Enroll",col = "Purple",border = "blue")
```


Histogram of Accept
Frequency
0
300
0
5000
15000
25000
Accept


Histogram of Personal
Frequency
0
150
0
2000
4000
6000
Personal


Histogram of Books
Frequency
0
200
500
0
500
1000
1500
2000
Books


Histogram of Enroll
Frequency
0
200
0
1000
3000
5000
Enroll