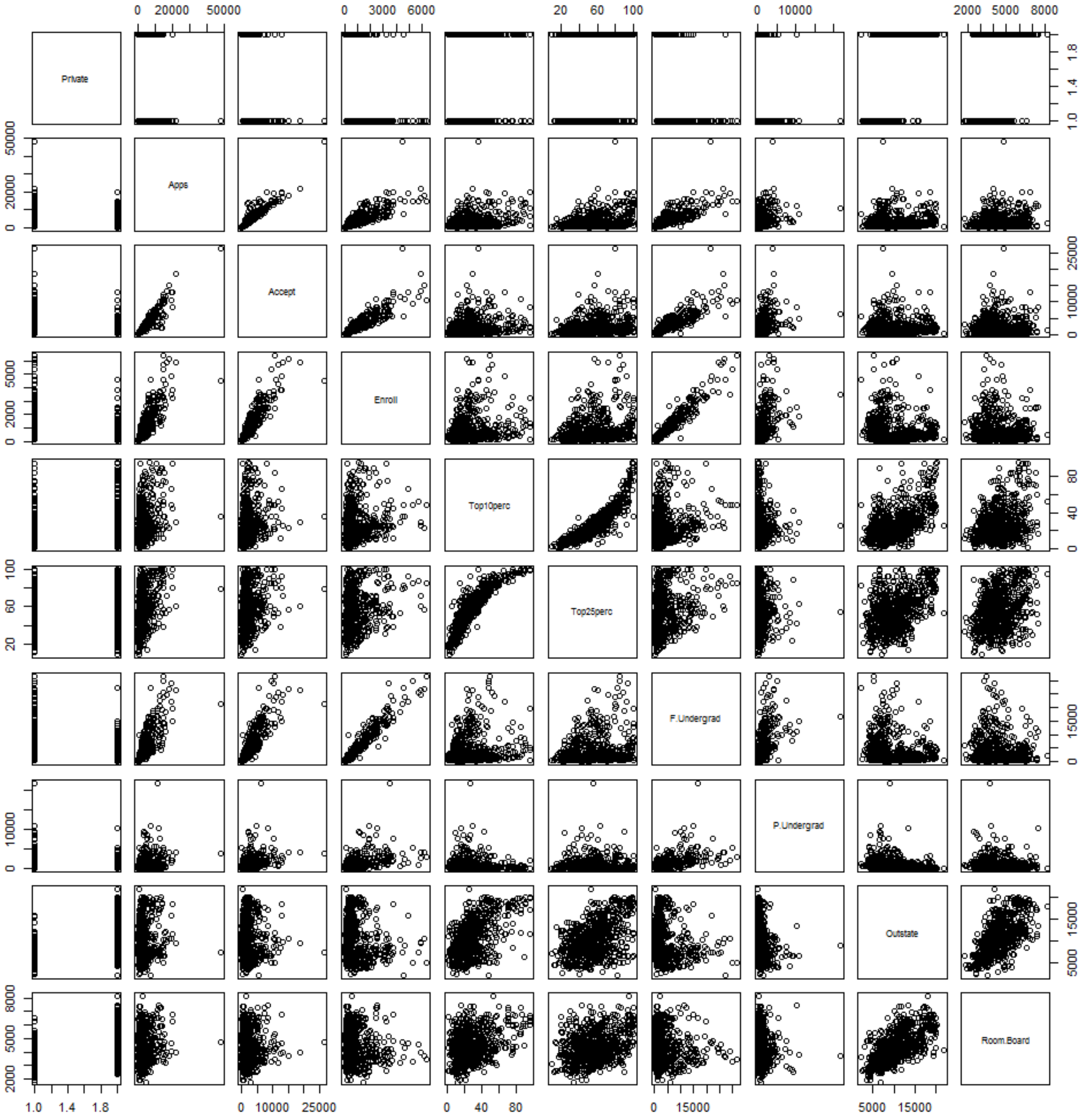
```r
#===================================================================================
#                        Using R: Introductory data exploration
#===================================================================================
#This exercise relates to the College data set, which can be found in the file
"College.csv" in D2L. The file 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
#Top25perc : New students from top 25
#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

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

#===================================================================================
#(a) Use the read.csv() function to read the data into a data frame in R. Call the
data frame college. Make sure that you have the directory set to the correct location
for the data (or that the data is in the same directory as the RStudio project).
college= read.csv(file="college.csv", head=TRUE, sep=",")

#===================================================================================
#(b) Look at the data using RStudio. 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]
View (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]

#and then view the data (either with the View command or clicking on the college data
frame in the RStudio workspace window) Now you should see that the first data column
is Private.

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

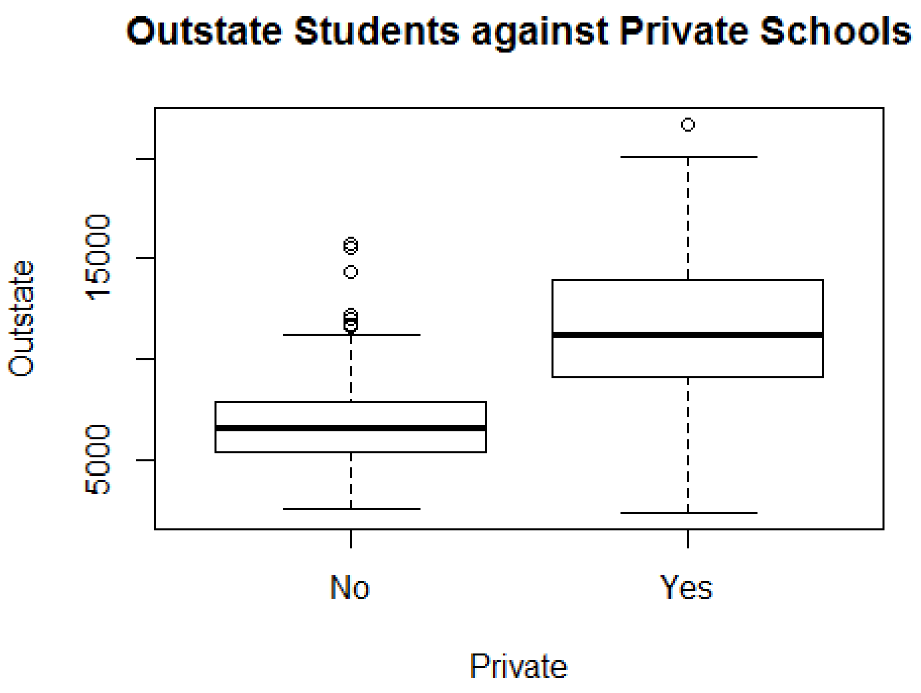
```r
data set.
summary(college)
```

```
##  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
```

```r
#ii. Access help for the pairs function and then use pairs to produce a scatterplot
matrix of the first ten columns. Recall that you can reference the first ten columns
of a matrix A using A[,1:10].
pairs(~Private+Apps+Accept+Enroll+Top10perc+Top25perc+F.Undergrad+P.Undergrad+Outstate
+Room.Board,data=college)
```



```
#iii. Use the plot() function to produce side-by-side boxplots of Outstate versus
Private. Label the axes and main title appropriately.
attach(college)
plot(Private,Outstate,main="Outstate Students against Private Schools",xlab="Private",
ylab="Outstate")
```

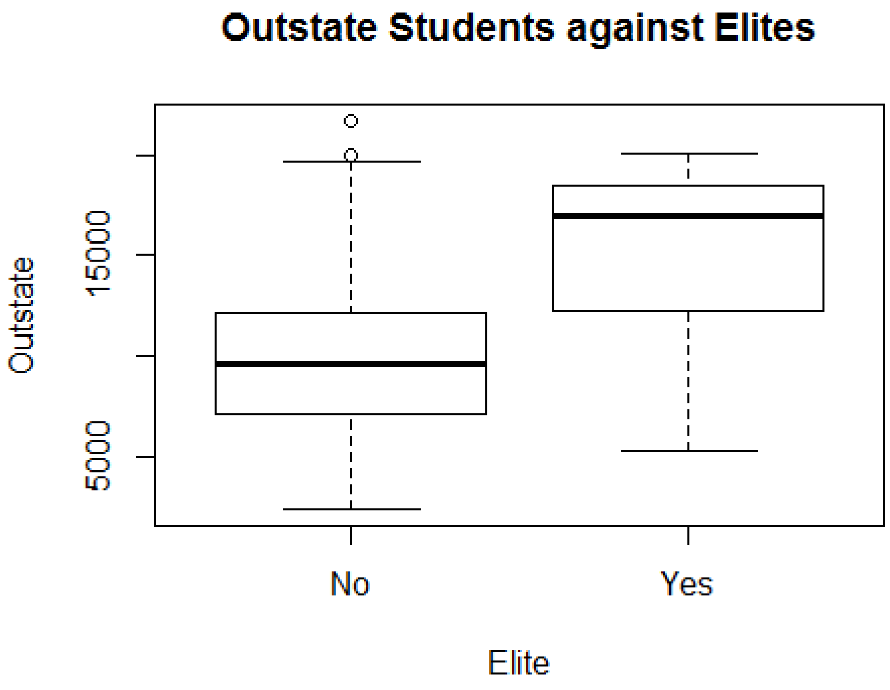


```
#iv. Using the following bit of code you will create a new qualitative variable,
called Elite by binning the Top10perc variable. That is, Elite will classify the
universities into two groups based on whether or not the proportion of students coming
from the top 10% of their high school classes exceeds 50%. Add comments to each line
below explaining what the corresponding code is doing and then run the code.
Elite <- rep("No", nrow(college))
#create a vector of No with size equal to number of rows in college
Elite[college$Top10perc >50] <- "Yes"
#in Elite it puts "Yes" in the cells that Top10pers is greater than 50
Elite <- as.factor(Elite)
#change the structure of Elite from character to factor
college <- data.frame(college ,Elite)
#Create a data table consisting of college and Elite and assign it to the college

#v. Use the summary() function to see how many elite universities there are.
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
```

```
#vi. Now use the plot() function to produce side-by-side boxplots of Outstate versus
Elite. Label the axes and main title appropriately.
attach(college)
```

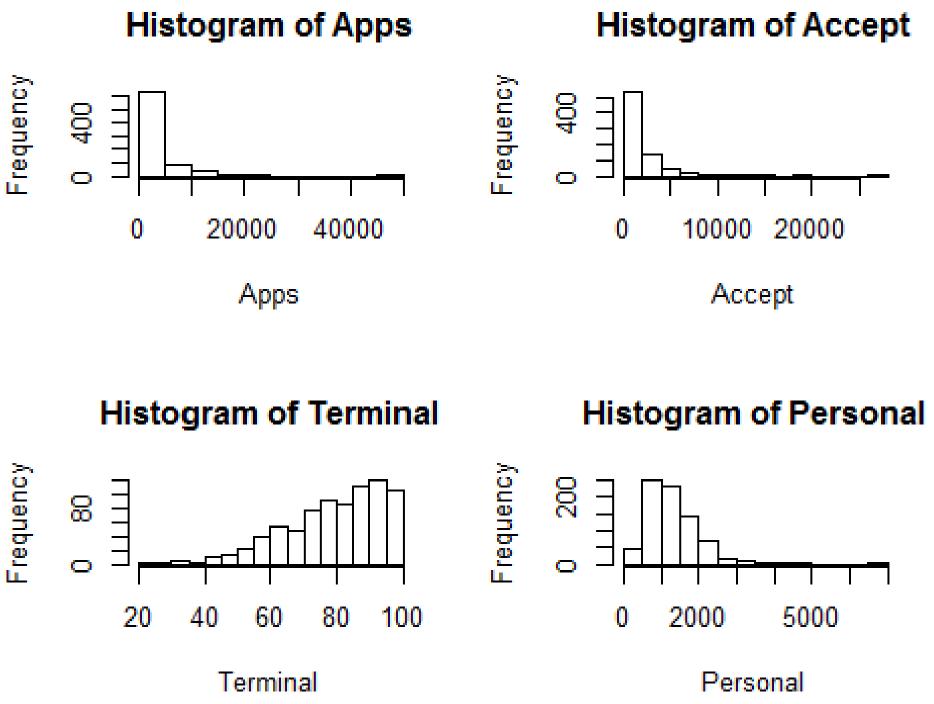
```
## The following object is masked _by_ .GlobalEnv:
##
##     Elite
##
## The following objects are masked from college (pos = 3):
##
##     Accept, Apps, Books, Enroll, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc
```

```
plot(Elite,Outstate,main="Outstate Students against Elites",xlab="Elite",
ylab="Outstate")
```



```
#vii. 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.
attach(college)

## The following object is masked _by_ .GlobalEnv:
##
##     Elite
##
## The following objects are masked from college (pos = 3):
##
##     Accept, Apps, Books, Elite, Enroll, Expend, F.Undergrad,
##     Grad.Rate, Outstate, P.Undergrad, perc.alumni, Personal, PhD,
##     Private, Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc
##
## The following objects are masked from college (pos = 4):
##
##     Accept, Apps, Books, Enroll, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc

par(mfrow=c(2,2))
hist(Apps)
hist(Accept)
hist(Terminal)
hist(Personal)
```


Histogram of Apps
Frequency
Apps
Histogram of Accept
Frequency
Accept
Histogram of Terminal
Frequency
Terminal
Histogram of Personal
Frequency
Personal