Homework2

Student Name: Madhu Shri Rajagopalan

Student ID: 227260

ANLY 530 Assignment 2

Introduction to statistical learning, chapter 2, Exercise - 8

```
getwd()
```

```
## [1] "C:/Users/Madhu/Dohaa Tahha/ML1"
```

```
setwd("C:/Users/Madhu/Dohaa Tahha/ML1")
getwd()
```

```
## [1] "C:/Users/Madhu/Dohaa Tahha/ML1"
```

Check all the files in the working directory

```
dir()
```

```
## [1] "College.csv"      "homework2Rmd.log" "homework2Rmd.Rmd"
```

Read in dataset

```
college =read.csv("College.csv")
```

Structure of the data

```
str(college)
```

```
## 'data.frame':    777 obs. of  19 variables:
##  $ X          : Factor w/ 777 levels "Abilene Christian University",..: 1
2 3 4 5 6 7 8 9 10 ...
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : int  1660 2186 1428 417 193 587 353 1899 1038 582 ...
##  $ Accept     : int  1232 1924 1097 349 146 479 340 1720 839 498 ...
##  $ Enroll     : int  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : int  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : int  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: int  2885 2683 1036 510 249 678 416 1594 973 799 ...
##  $ P.Undergrad: int  537 1227 99 63 869 41 230 32 306 78 ...
##  $ Outstate   : int  7440 12280 11250 12960 7560 13500 13290 13868 15595
10468 ...
##  $ Room.Board : int  3300 6450 3750 5450 4120 3335 5720 4826 4400 3380 ...
##  $ Books      : int  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : int  2200 1500 1165 875 1500 675 1500 850 500 1800 ...
##  $ PhD        : int  70 29 53 92 76 67 90 89 79 40 ...
```

```
##  $ Terminal   : int  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: int  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : int  7041 10527 8735 19016 10922 9727 8861 11487 11644
8991 ...
##  $ Grad.Rate  : int  60 56 54 59 15 55 63 73 80 52 ...
```

```

Separating the first row that lists the college names from the rest of the data

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

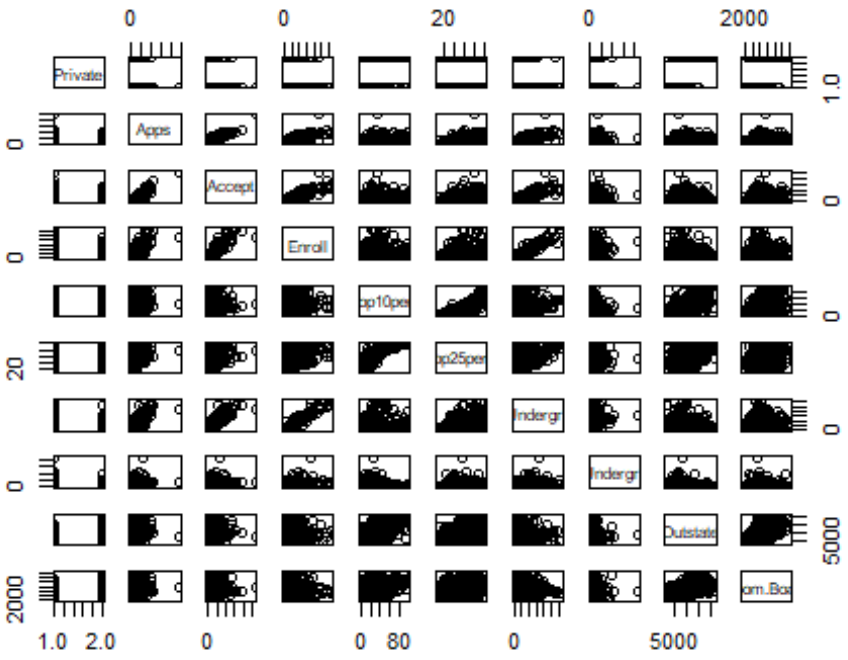
summary(college)

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
```

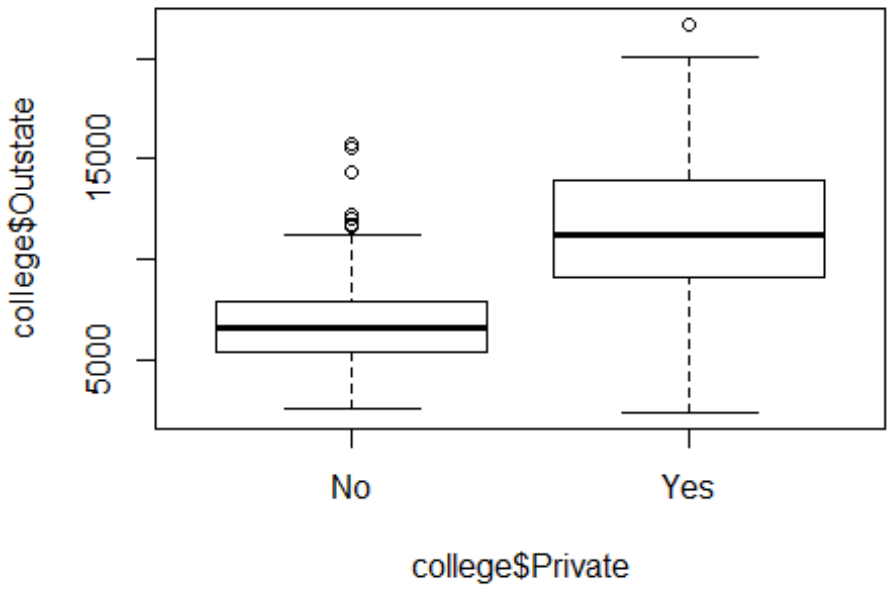
Creating plots for the first 10 variables

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



Creating box plot for the variables “Outstate” versus “Private”

```
plot(college$Outstate~college$Private)
```



Creating new variable to find out how many colleges are in the elite group

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

Using Summary to see if Elite colleges ahve been created in a separate column

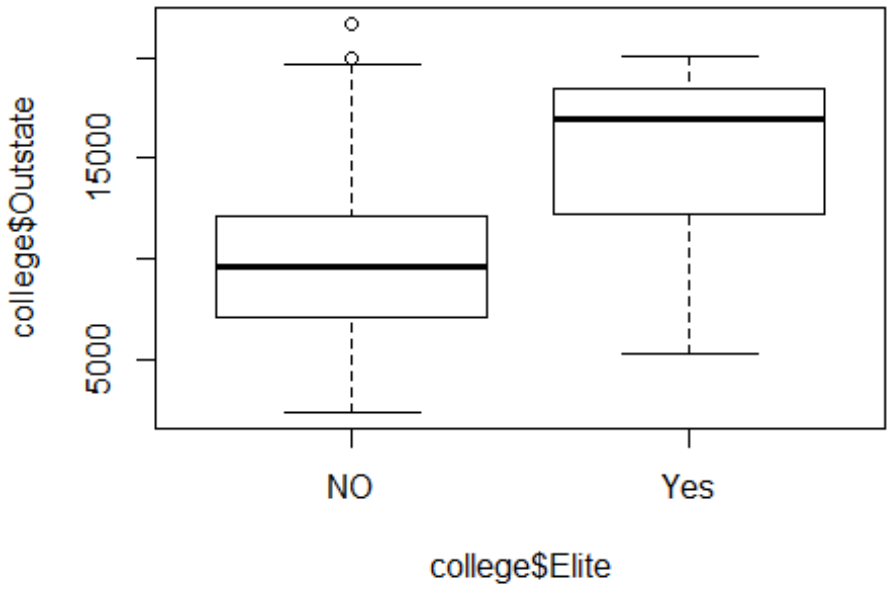
```
summary(college)
```

```
##  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
```

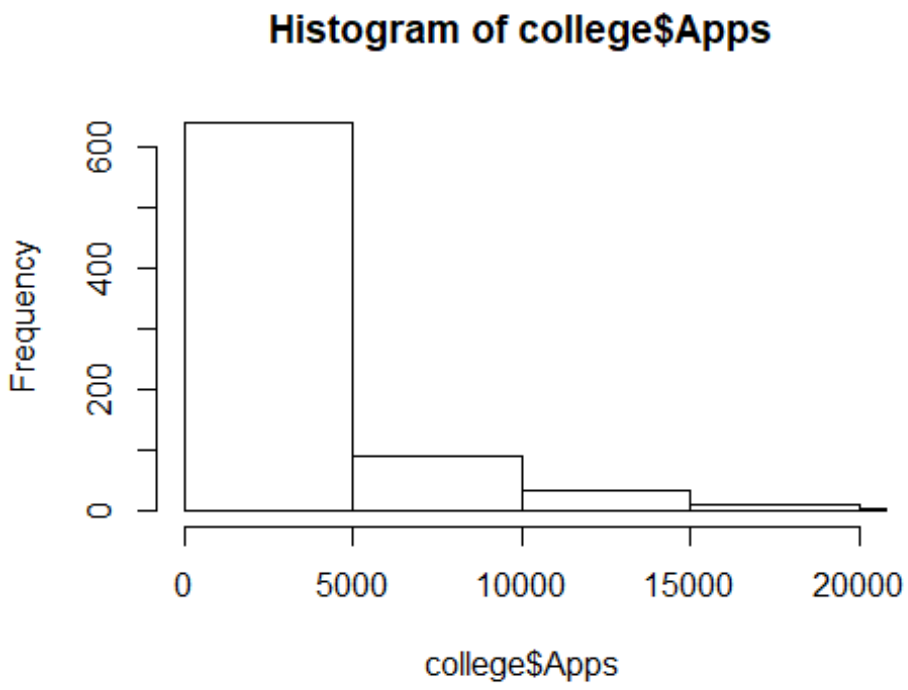
Creating box plot for “Outstate” vs “Elite”

```
plot(college$Outstate~college$Elite)
```

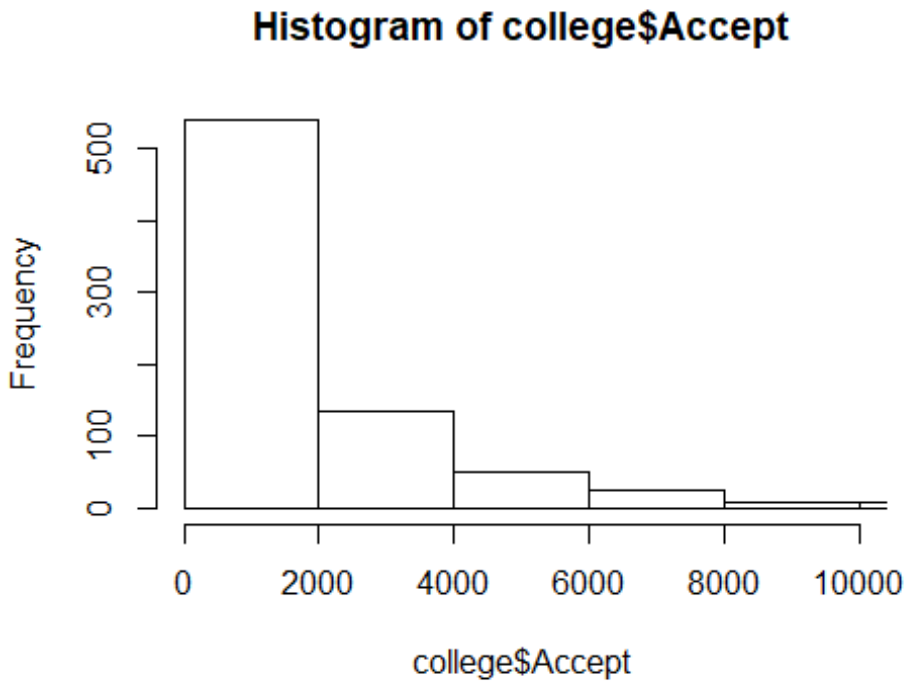


Plotting Histograms for variables namely “Apps”, “Enroll”, “Accept” , “Outstate”

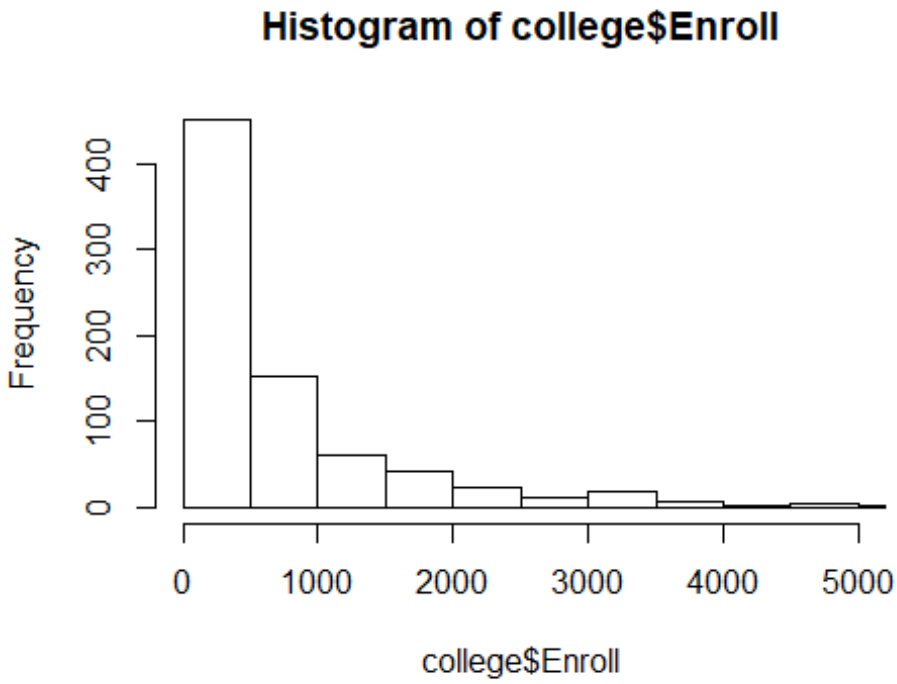
```
hist(college$Apps,xlim=c(0,20000))
```




```r
hist(college$Accept,xlim=c(0,10000))
```




```r
hist(college$Enroll,xlim=c(0,5000))
```




```r
hist(college$Outstate)
```