Chapter 2: Statistical Learning-Problems: 8

Kartheek Raj

12/20/2019

Problem 8

packages required to excute code :ggplot

(a) Use the read.csv() function to read the data set set into R. Call the loaded data set set College. Make sure that you have the directory set to the correct location for the data set.

Answer

Reading the csv file using **read.csv** command

```
require(ggplot2)
```

```
## Loading required package: ggplot2
```

```
library(ggplot2)
College<-read.csv("College.csv")
```

First 10 records in College data set

```
head(College,10)
```

```
##                              X Private Apps Accept Enroll Top10perc Top25perc
## 1  Abilene Christian University     Yes 1660   1232    721        23        52
## 2            Adelphi University     Yes 2186   1924    512        16        29
## 3                Adrian College     Yes 1428   1097    336        22        50
## 4           Agnes Scott College     Yes  417    349    137        60        89
## 5     Alaska Pacific University     Yes  193    146     55        16        44
## 6              Albertson College     Yes  587    479    158        38        62
## 7       Albertus Magnus College     Yes  353    340    103        17        45
## 8                Albion College     Yes 1899   1720    489        37        68
## 9              Albright College     Yes 1038    839    227        30        63
## 10    Alderson-Broaddus College     Yes  582    498    172        21        44
##    F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
## 1         2885         537     7440       3300   450     2200  70       78
## 2         2683        1227    12280       6450   750     1500  29       30
## 3         1036          99    11250       3750   400     1165  53       66
## 4          510          63    12960       5450   450      875  92       97
## 5          249         869     7560       4120   800     1500  76       72
## 6          678          41    13500       3335   500      675  67       73
## 7          416         230    13290       5720   500     1500  90       93
## 8         1594          32    13868       4826   450      850  89      100
## 9          973         306    15595       4400   300      500  79       84
## 10         799          78    10468       3380   660     1800  40       41
##    S.F.Ratio perc.alumni Expend Grad.Rate
## 1       18.1          12   7041        60
## 2       12.2          16  10527        56
## 3       12.9          30   8735        54
## 4        7.7          37  19016        59
## 5       11.9           2  10922        15
```

```
## 6        9.4          11   9727        55
## 7       11.5          26   8861        63
## 8       13.7          37  11487        73
## 9       11.3          23  11644        80
## 10      11.5          15   8991        52
```

(b)Look at the data set 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 set. However, it may be handy to have these names for later.

Answer

Checking the default row names of the data set

```
head(rownames(College),10)
```

```
##  [1] "1"  "2"  "3"  "4"  "5"  "6"  "7"  "8"  "9"  "10"
```

Changing the row names from default number to university names i,e 1st column in the data set

```
rownames(College)<-College[,1]
```

Updating the data set using **fix()** command

```
fix(College)
```

Checking again data set using **head** for row names

```
head(College)
```

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

Duplicate feature "X" which contains again universities names has to eliminate

```
College<-College[,-1]
fix(College)
```

(c)

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

Answer

Command **Summary** gives numerical snapshot of the data set against each feature like mean,mode,max,min,etc.,.

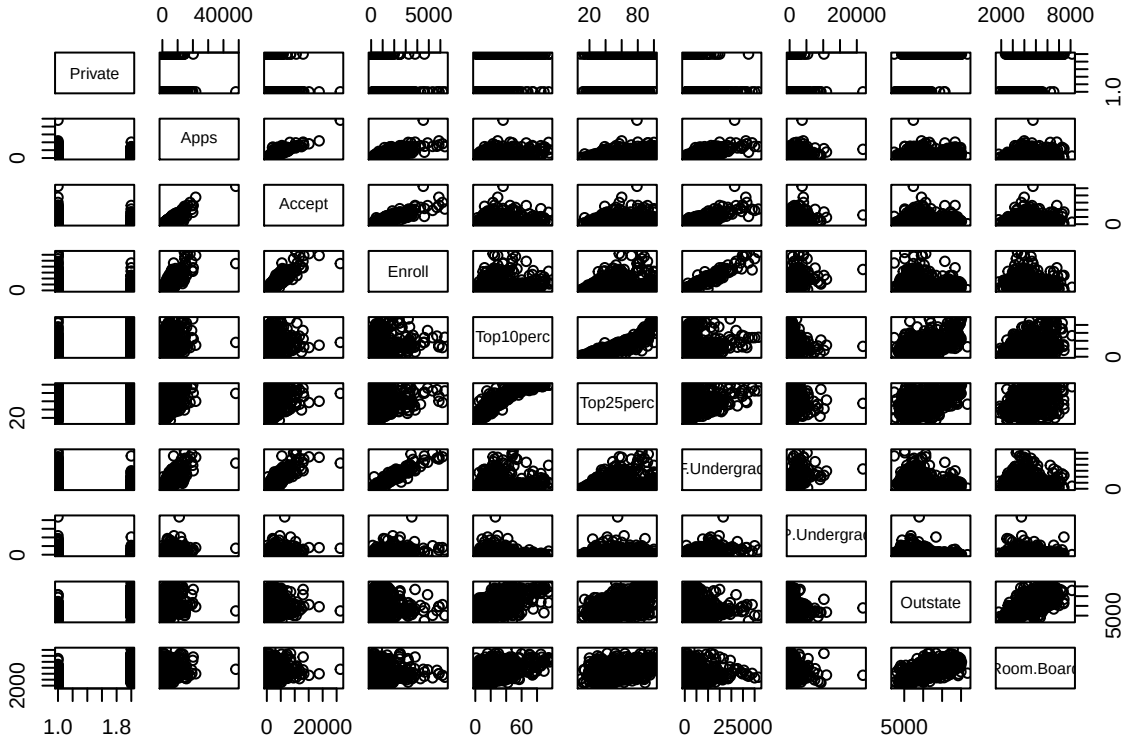
```
summary(College)
```

```
##  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].

Answer

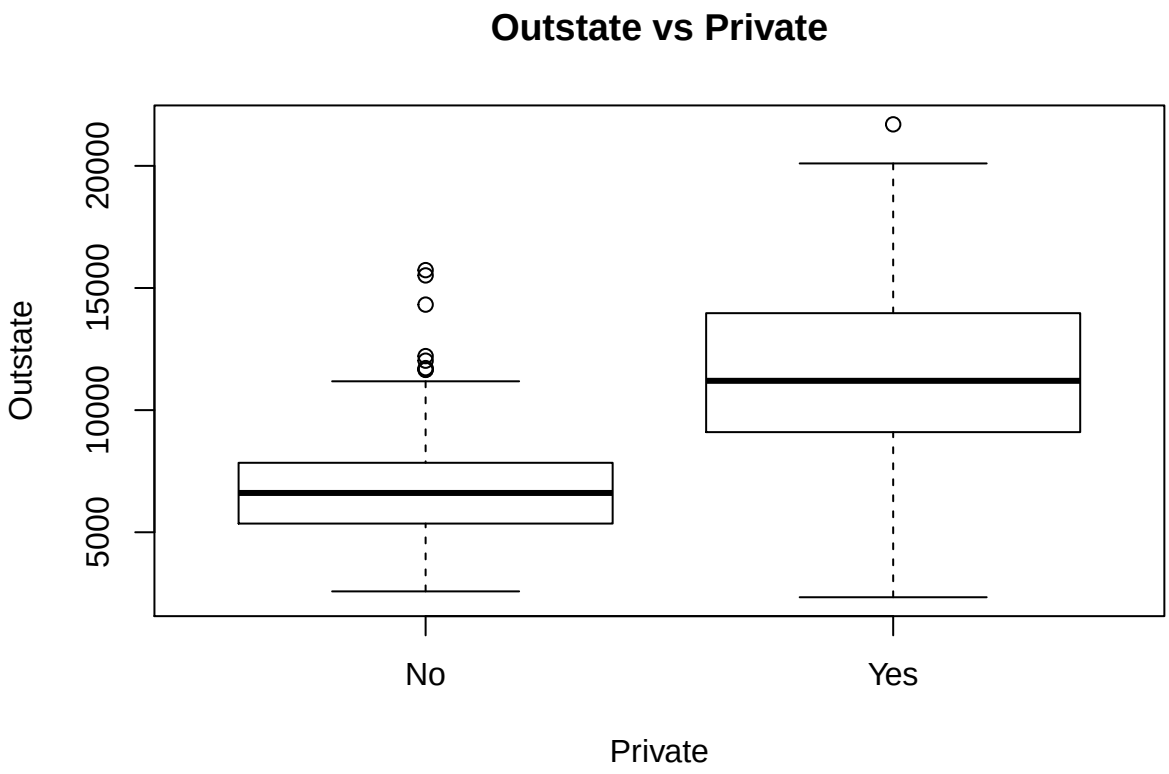
```
pairs(College[,1:10])
```



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

Answer

```
boxplot(College$Outstate~College$Private,ylab = "Outstate",xlab = "Private")
title("Outstate vs Private")
```



(iv)Create a new qualitative variable, called Elite, bybinning 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%.

Answer

creating the new vector with name **Elite** with same length as features in College data set.

```
Elite<-rep("No",length(College$Private))
```

Capturing proportion of students coming from the top 10% of their high school classes exceeds 50% into elite

```
Elite[College$Top10perc >50]=" Yes"
```

Checking the data type of the Elite vector

```
typeof(Elite)
```

```
## [1] "character"
```

Changing character data type to factor.

```
Elite<-as.factor(Elite)
```

Joining the ELite vector in College data set.

```
College<-data.frame(Elite,College)
```

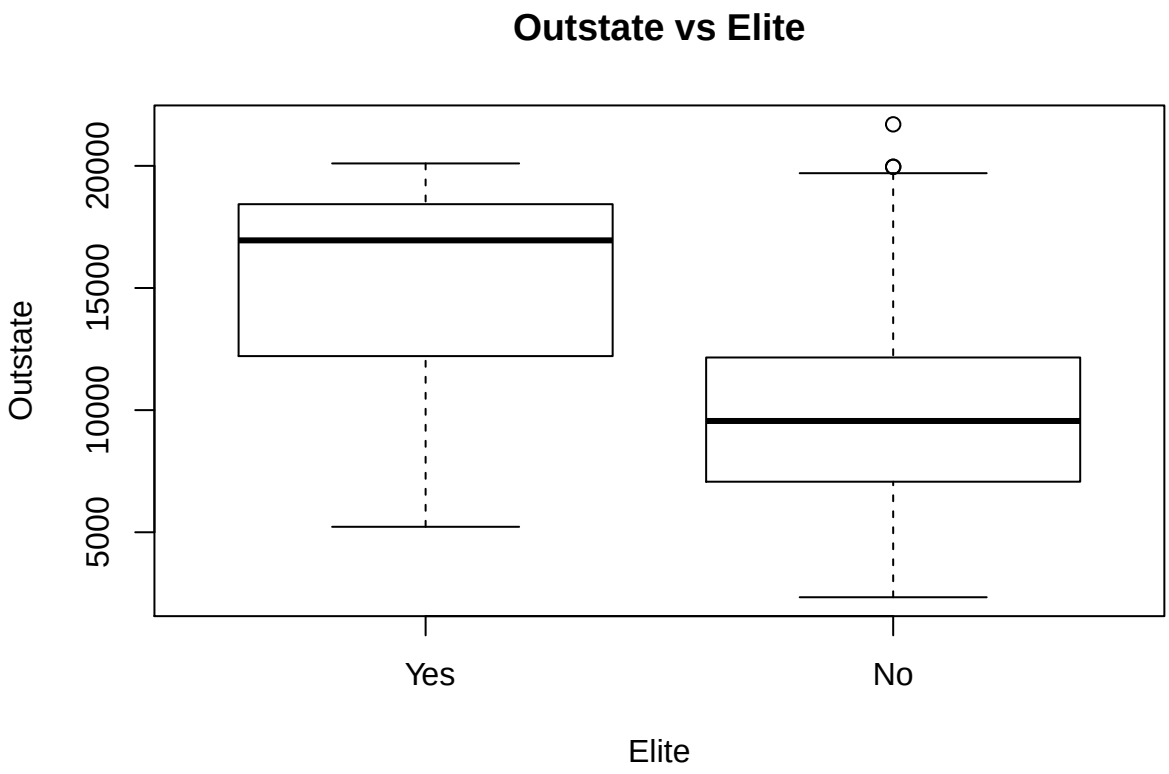
Using command **Summary** to see how many elite universities there are.

```
summary(College)
```

```
##   Elite     Private        Apps           Accept          Enroll
##   Yes: 78   No :212   Min.   :   81   Min.   :   72   Min.   :  35
##  No  :699   Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242
##                       Median : 1558   Median : 1110   Median : 434
##                       Mean   : 3002   Mean   : 2019   Mean   : 780
##                       3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902
##                       Max.   :48094   Max.   :26330   Max.   :6392
##    Top10perc       Top25perc      F.Undergrad     P.Undergrad
##  Min.   : 1.00   Min.   :  9.0   Min.   :  139   Min.   :    1.0
##  1st Qu.:15.00   1st Qu.: 41.0   1st Qu.:  992   1st Qu.:   95.0
##  Median :23.00   Median : 54.0   Median : 1707   Median :  353.0
##  Mean   :27.56   Mean   : 55.8   Mean   : 3700   Mean   :  855.3
##  3rd Qu.:35.00   3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0
##  Max.   :96.00   Max.   :100.0   Max.   :31643   Max.   :21836.0
##     Outstate       Room.Board       Books           Personal
##  Min.   : 2340   Min.   :1780   Min.   :  96.0   Min.   : 250
##  1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850
##  Median : 9990   Median :4200   Median : 500.0   Median :1200
##  Mean   :10441   Mean   :4358   Mean   : 549.4   Mean   :1341
##  3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0   3rd Qu.:1700
##  Max.   :21700   Max.   :8124   Max.   :2340.0   Max.   :6800
##       PhD            Terminal       S.F.Ratio      perc.alumni
##  Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
##  1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
##  Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
##  Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
##  3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
##  Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
##      Expend        Grad.Rate
##  Min.   : 3186   Min.   : 10.00
##  1st Qu.: 6751   1st Qu.: 53.00
##  Median : 8377   Median : 65.00
##  Mean   : 9660   Mean   : 65.46
##  3rd Qu.:10830   3rd Qu.: 78.00
##  Max.   :56233   Max.   :118.00
```

Generating boxplots of Outstate versus Elite

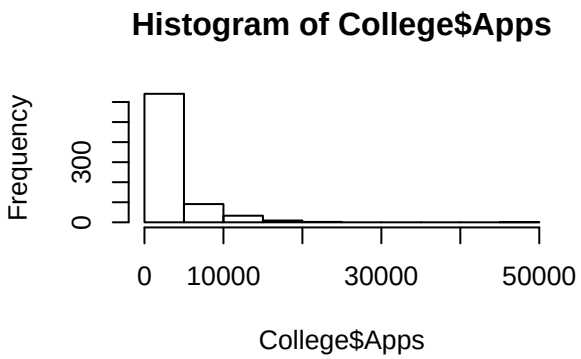
```
boxplot(College$Outstate~College$Elite,ylab = "Outstate",xlab = "Elite")
title("Outstate vs Elite")
```

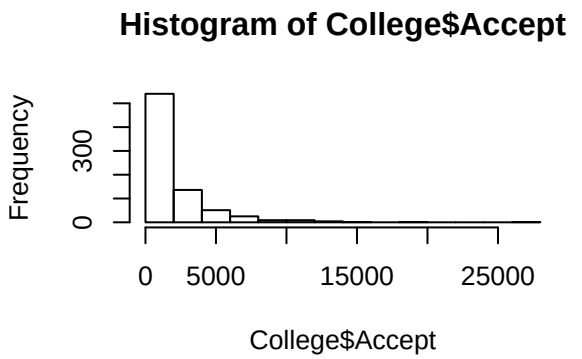


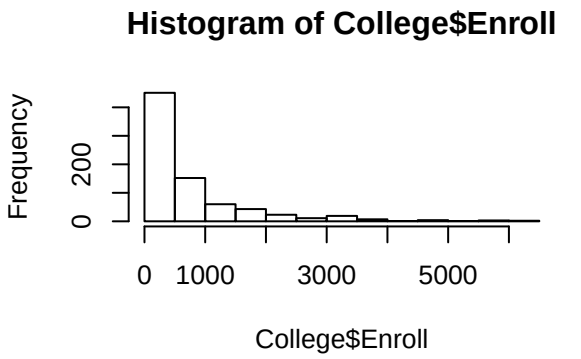
(v)Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.You may find the command par(mfrow=c(2,2)) useful:it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this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

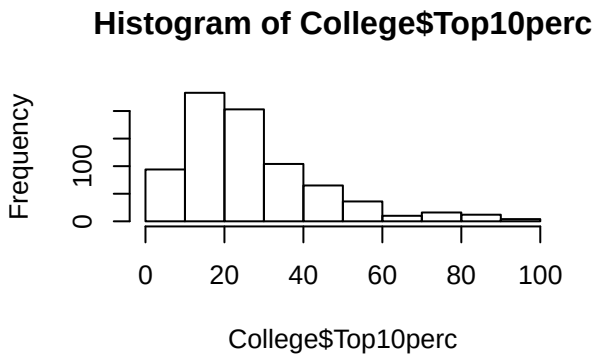
Answer

```
par(mfrow= c(2,2))
hist(College$Apps)
hist(College$Accept)
hist(College$Enroll)
hist(College$Top10perc)
```

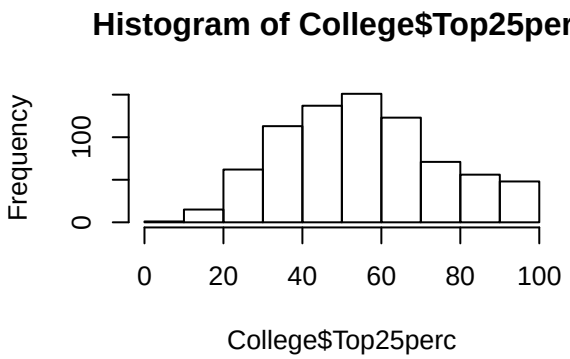


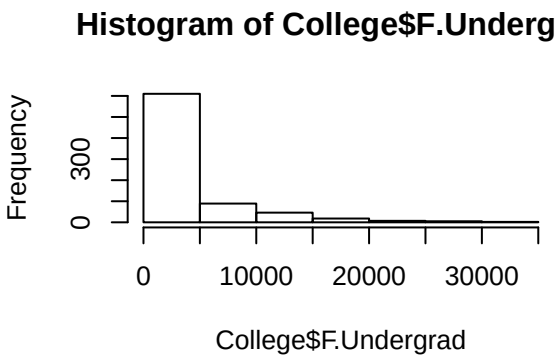


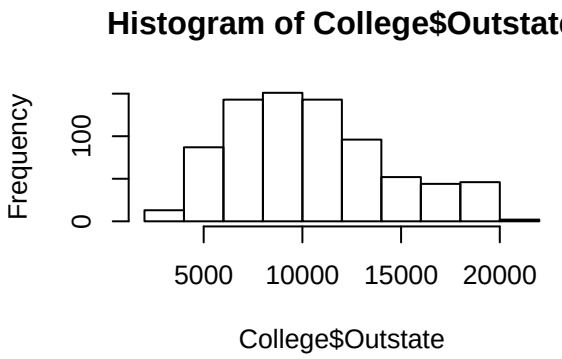


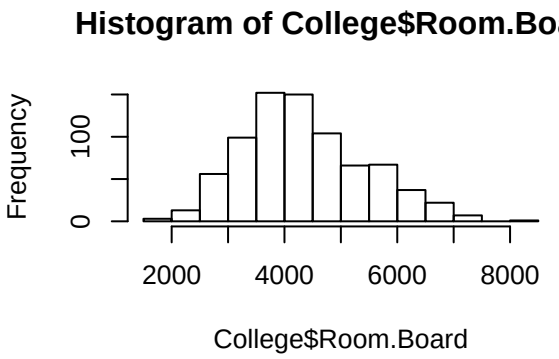



```
hist(College$Top25perc)
hist(College$F.Undergrad)
hist(College$Outstate)
hist(College$Room.Board)
```

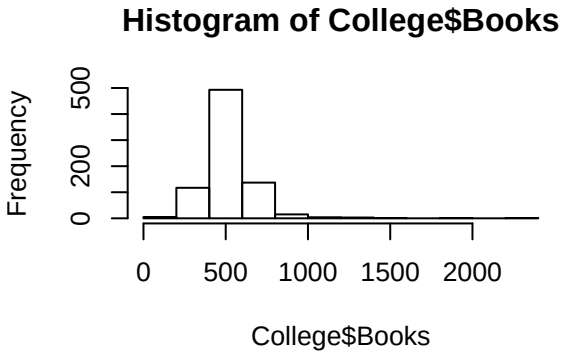








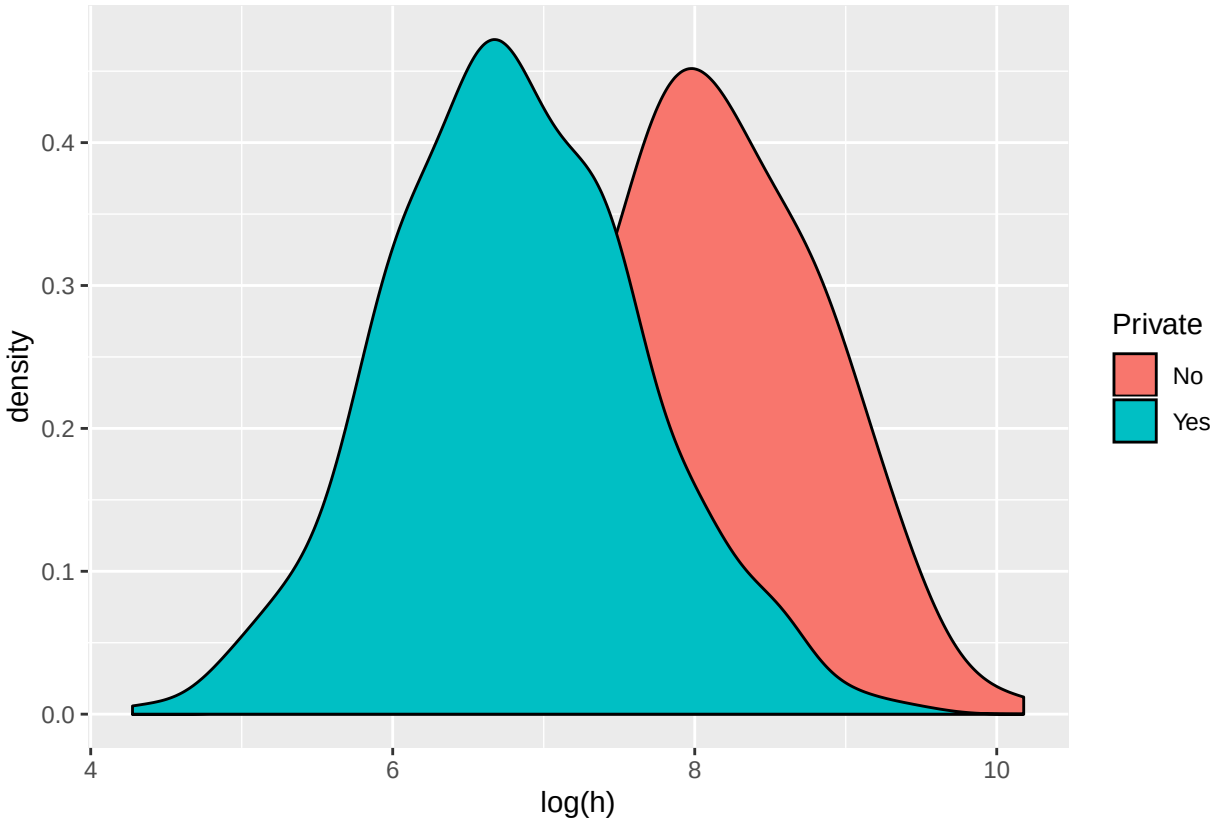

```r
hist(College$Books)
```



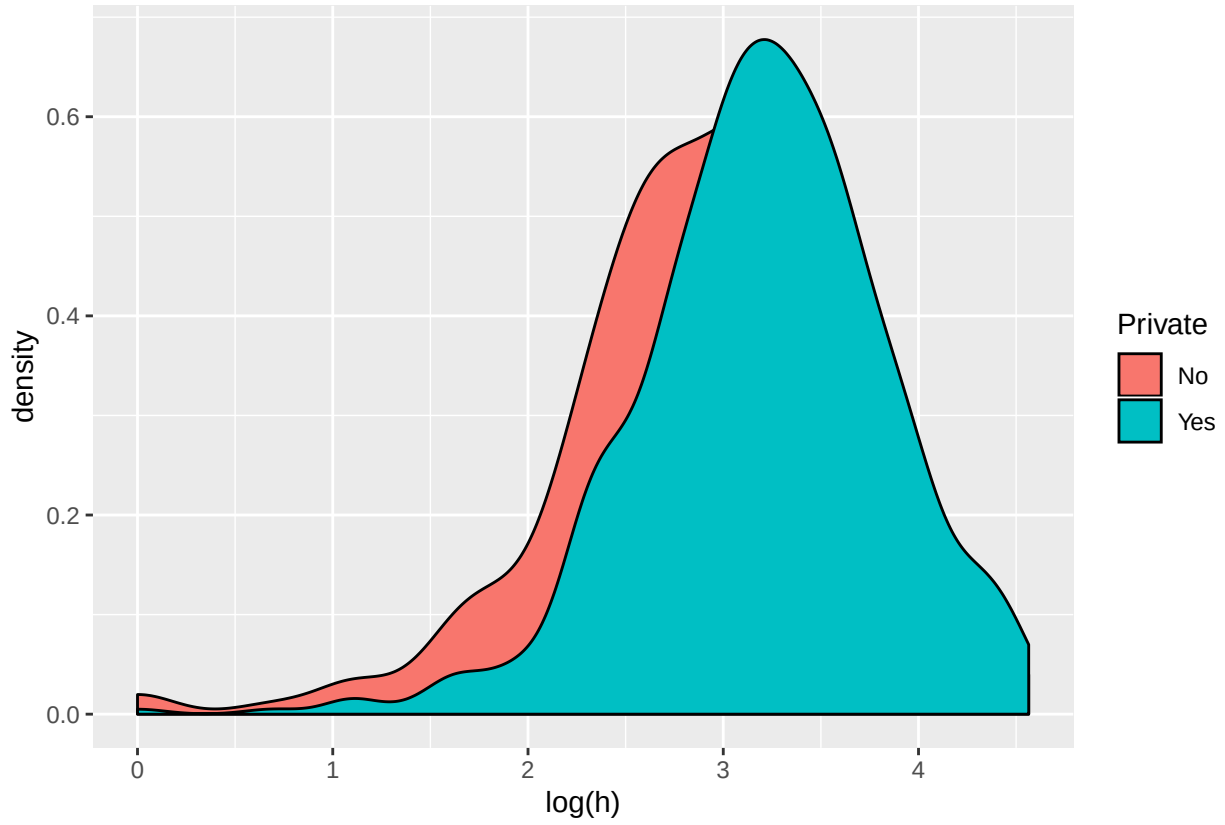
(vi)Continue exploring the data, and provide a brief summary of what you discover.

Answer

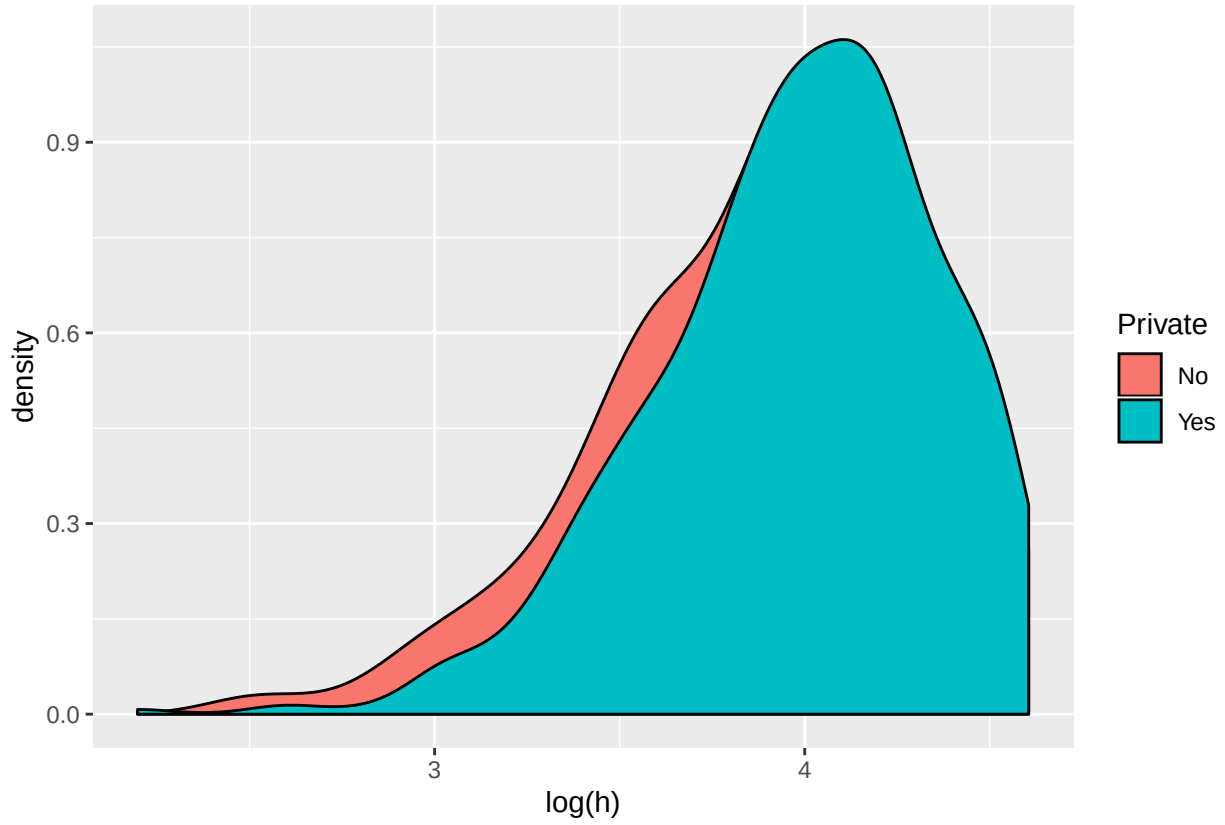
```
library(ggplot2)
collegeplot1=function(x){
  h=x
ggplot(College, aes(log(h), fill = Private))+geom_density()+ggtitle(paste(names(h)))

}
par(mfrow= c(2,2))
collegeplot1(College$Accept)
```



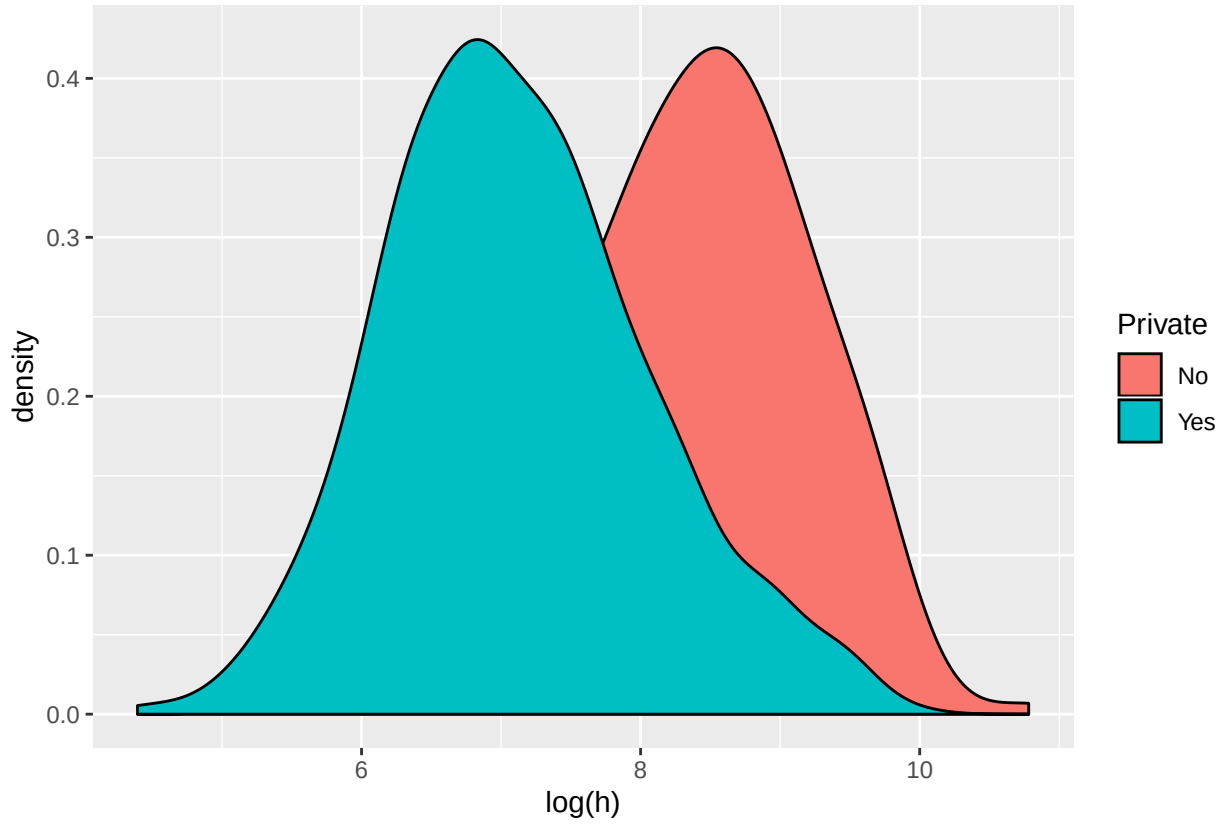

```
collegeplot1(College$Top10perc)
```




```
collegeplot1(College$Top25perc)
```



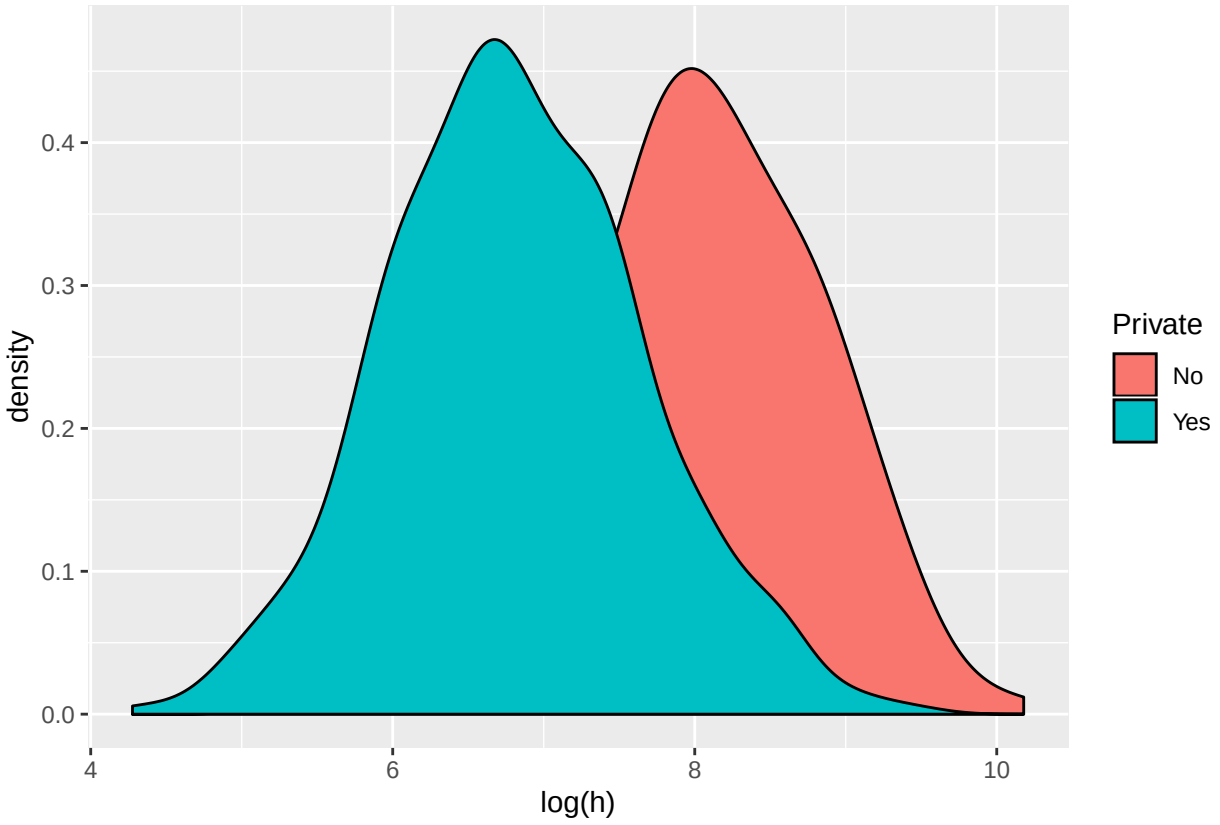

```
collegeplot1(College$Apps)
```



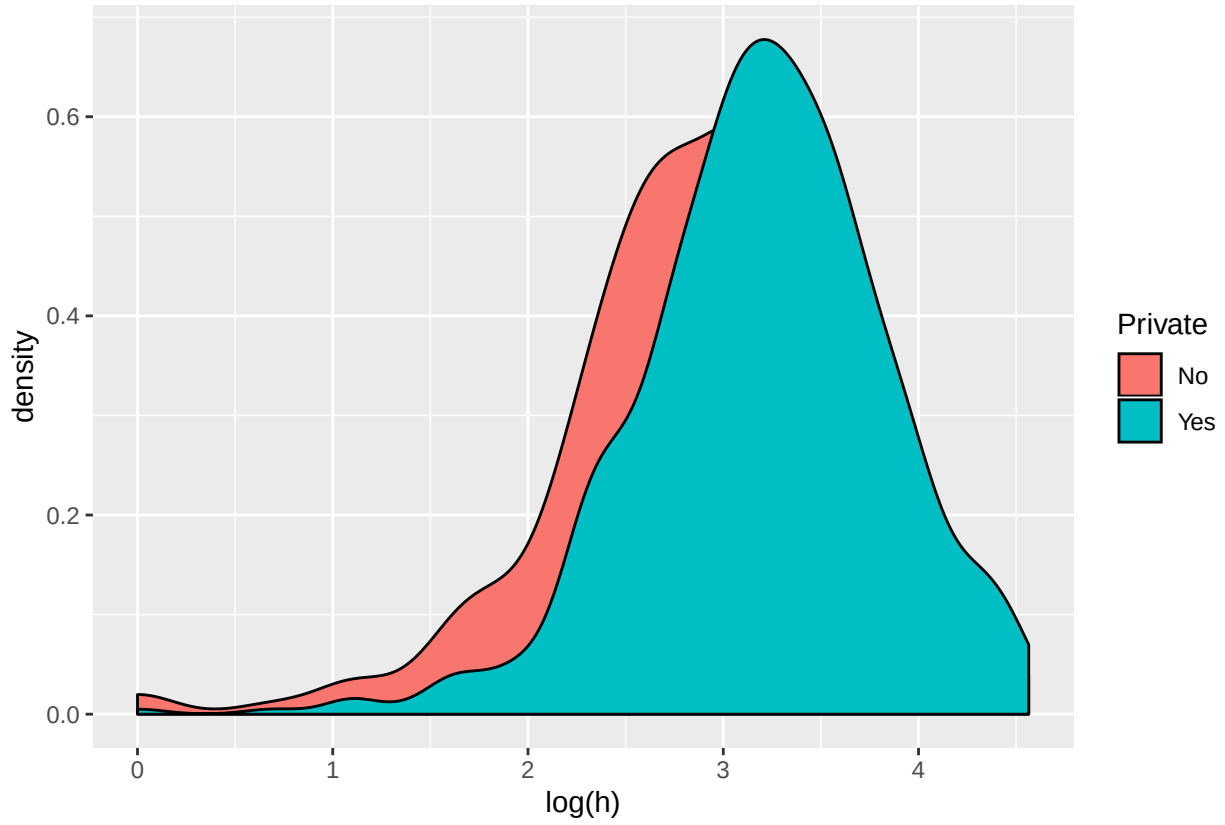
EDA Insight 1

Private college accept more students from top 10 percent in their high school acedemics.

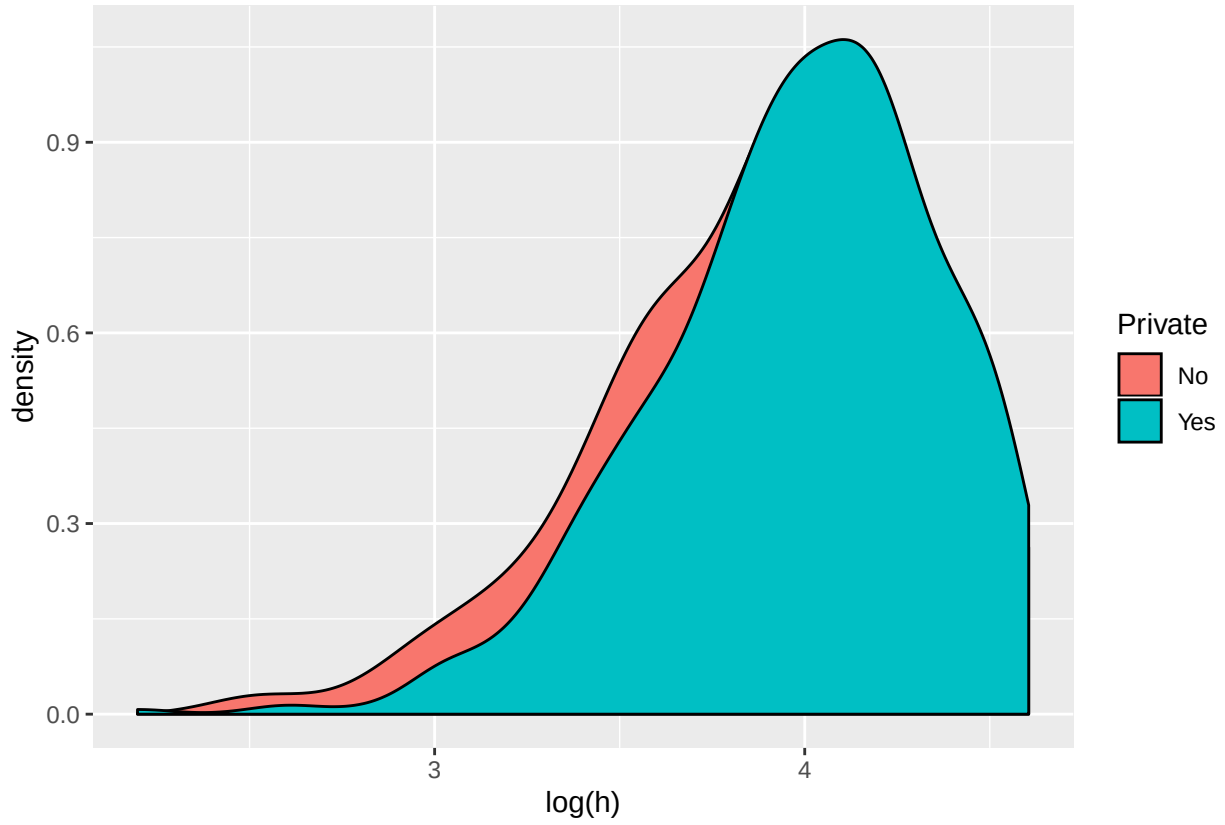
```
collegeplot1=function(x){
  h=x
ggplot(College, aes(log(h), fill = Private))+geom_density()+ggtitle(paste(names(h)))

}
par(mfrow= c(2,2))
collegeplot1(College$Accept)
```



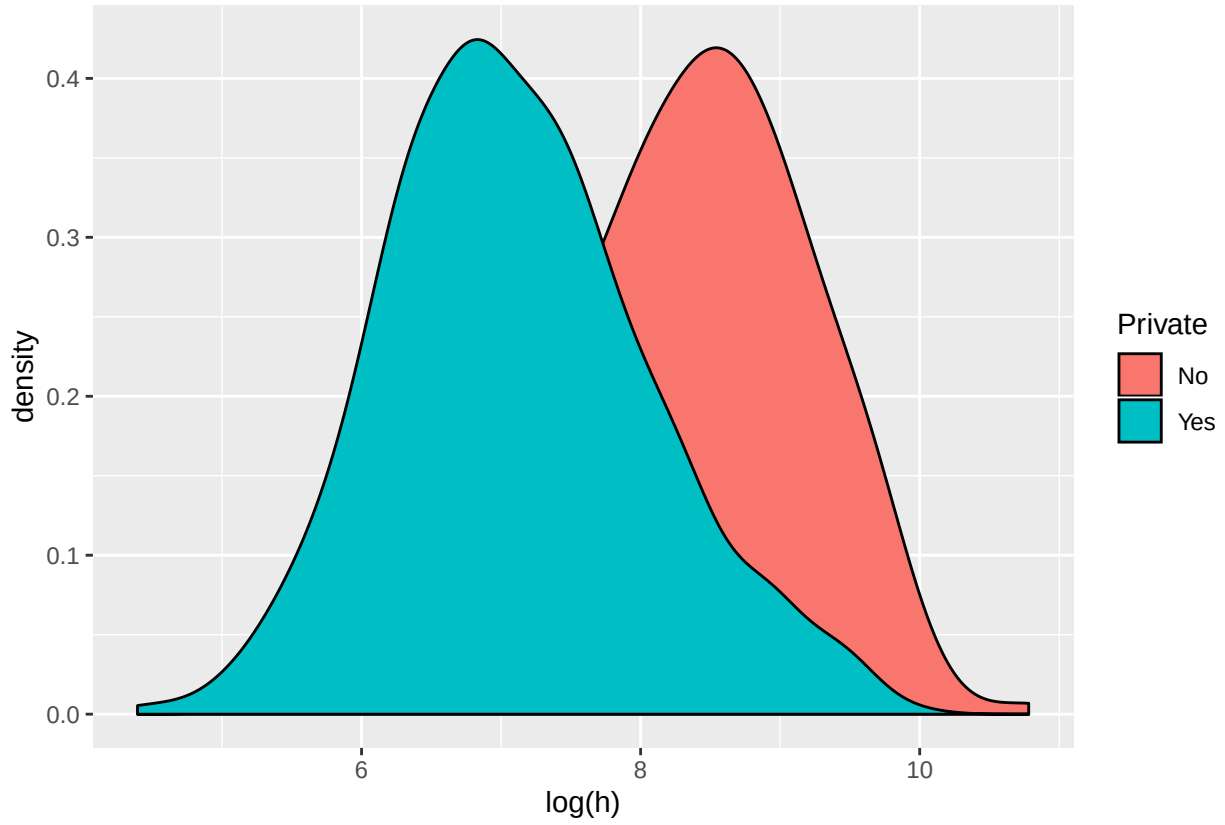

```
collegeplot1(College$Top10perc)
```




```
collegeplot1(College$Top25perc)
```



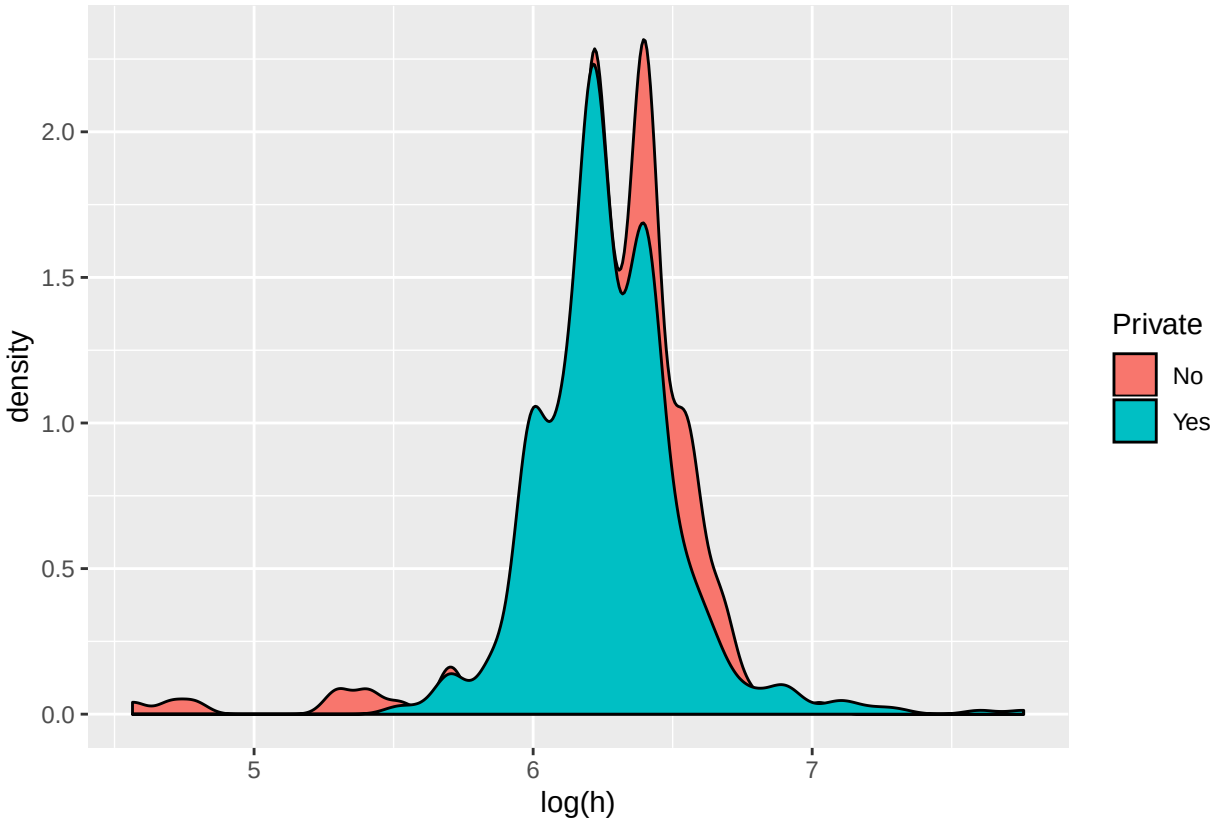

```
collegeplot1(College$Apps)
```



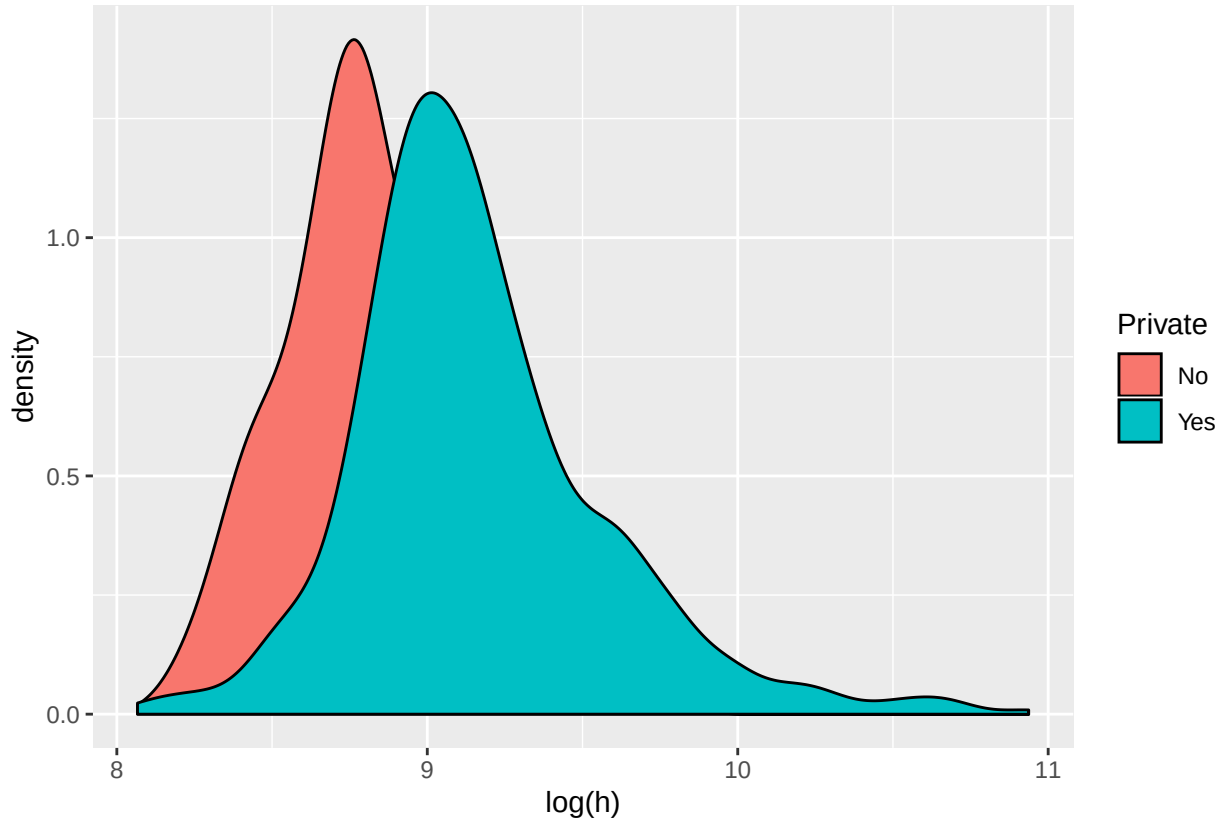
EDA Insight 2

Private college accept more students from top 10 percent in their high school acedemics. public school students spends more on Books,Expend,Personal but not on room&Boarding

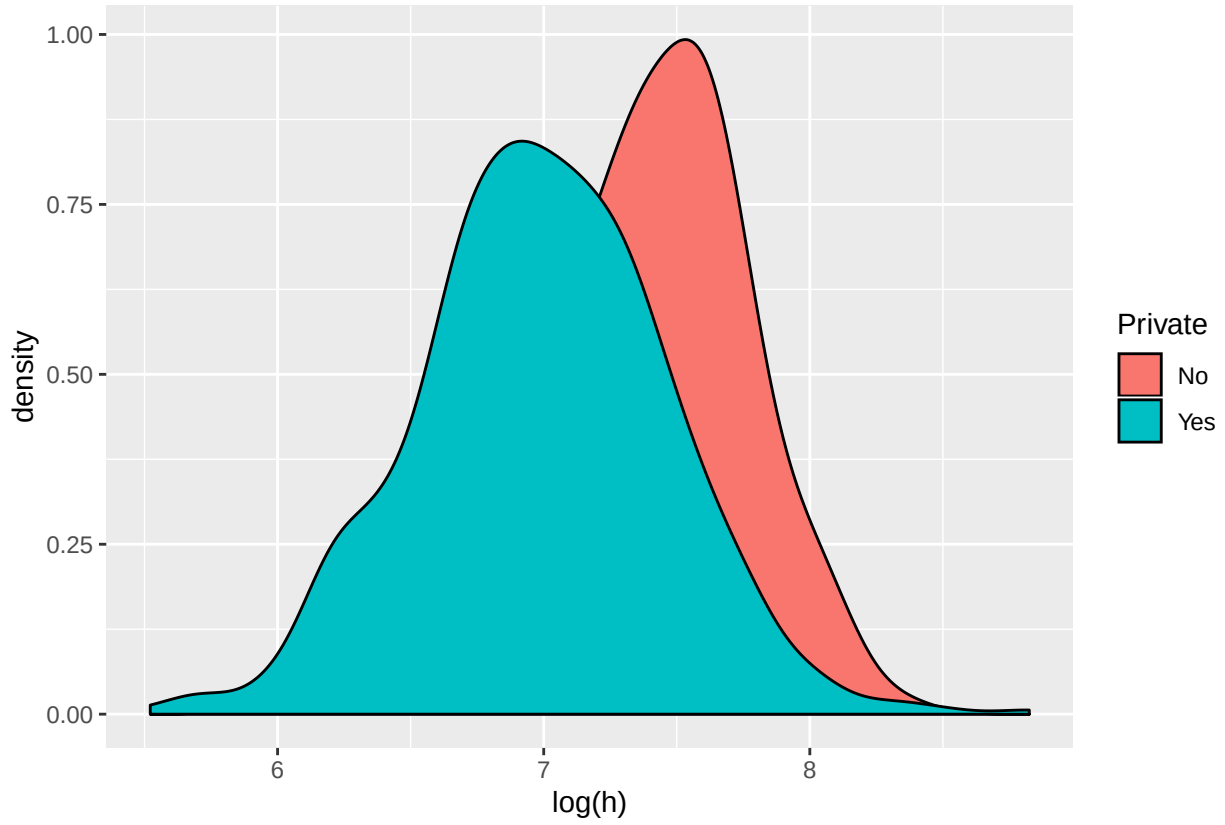
```
collegeplot2=function(x){
  h=x
ggplot(College, aes(log(h), fill = Private))+geom_density()+ggtitle(paste(names(h)))

}
par(mfrow= c(2,2))
collegeplot2(College$Books)
```



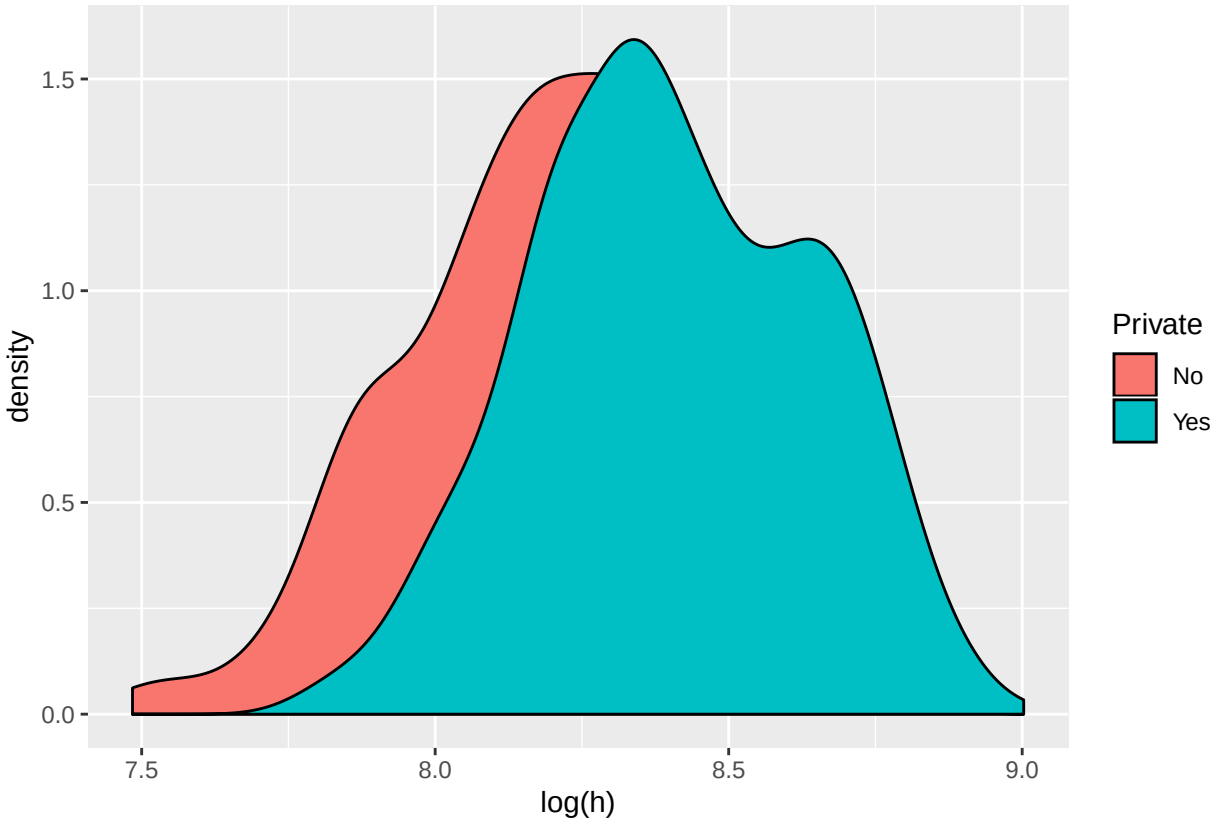

```
collegeplot2(College$Expend)
```




```
collegeplot2(College$Personal)
```




```
collegeplot2(College$Room.Board )
```


1.5
1.0
0.5
0.0
density
7.5
8.0
8.5
9.0
log(h)
Private
No
Yes