Conceptual question from An Introduction to stastical Learning

Yamuna Dhungana

Exercises

1. Question 2.4.2 pg 52

Explain whether each scenario is a classification or regression problem, and indicate whether we are most interested in inference or prediction. Finally, provide n and p. a.We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

The given scenario is the regression problem because the salary is obtained from the continuous measurement with the other variables. The question is the inference problem since we have to evaluate the effect of the predictor with the response variables.The total no of a sample (n) is 500, and the number of the predictor (p) is 3 (profit, number of employees, industry)

b.We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price,and ten other variables.

Since we have the two outcomes success or failure, therefore, it is a binary classification.It is the prediction problem. We will have the dependant and the independent variable that helps in predicting the outcomes. The no of samples (n) is 20, and the number of predictors (p) is 13 (price charged for the product, marketing budget, competition price, and ten other variables).

c.We are interested in predicting the % change in the USD/Euro exchange rate in relation to the weekly changes in the world stock markets. Hence we collect weekly data for all of 2012. For each week we record the % change in the USD/Euro, the % change in the US market, the % change in the British market,and the % change in the German market.

The problem is the regression problem because the percentage change is the continuous measurement. This is the prediction problem.The number of observations (n) is 52 (the number of weeks in 2012 is 52) The number of predictors (p) is 3 (the % change in the US market, the % change in the British market, and the % change in the German market).

2.Question 2.4.4 pg 53

You will now think of some real-life applications for statistical learning. a.Describe three real-life applications in which classification might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer.

Application 1:

Analyzing whether it will be a rainy day or a sunny day.The response of this analysis will be rainy or sunny, and the predictors will be temperature, humidity, and windy. The goal is to predict; since we have to predict whether it will be a rainy day or a sunny day.

Application 2:

Is the newly discovered coronavirus vaccine success or a failure? The response for this project is the vaccine able to develop antibodies or not. The predictor of the project will be age, geographical condition, ethnicity,

underlying conditions, control/test group, etc.The goal of the project is to predict.

Application 3:

Analyzing either the person running for the presidency will win or lose.The response for this project is win and loss. The predictor of the project is advertisement, communication, policies, popularity, funding, qualification of candidates, etc. The goal of the project is to predict the outcome of the election.

b.Describe three real-life applications in which regression might be useful. Describe the response, as well as the predictors. Is the goal of each application inference or prediction? Explain your answer

Application 1:

Suppose we want to find the consumption of water in the unites states.We have a certain parameter that we set, G being the water consumption, and the predictors will be no of family, properly installed regulators or drinking lines, environmental temperature, etc. This is the case of inference.

Application 2:

We want to find out the average sale price of the house in the specific area over the next ten years.The average house that will be sold in the for x next year, y the year after, z after that, etc., and the predictor are parking, schools, crime rate, the average income of the family, neighbors, etc. This is the condition of the inference.

Application 3:

Suppose we want to find out the growth in the number of students in the particular school by 2030. The response of the project is the no of students.The predictor is the performance of students, qualified teachers, extracurricular activities, tuition fees, etc.The above problem is the inference.

c.Describe three real-life applications in which cluster analysis might be useful.

Application 1:

Suppose we want to Identify the conspiracy theorist on Twitter. For this project, we will be using clustering. Based on the keywords that a person post on his posts, the theorist can be identified. Similar keywords from the posts can be identified from another user as well.The goal in this project will be prediction.

Application 2:

The clustering also can be used to find the product's popularity among the age group. The purchase behavior of the customer can be used for the promotion of this product to a specific age group, target advertisement, new product development.The goal for this is to first find the inferance then predict.

Application 3:

It can be used to cluster mutation of a new strain involved in coronavirus.Here, the response will be the abnormal RNA strain other than the previously found strain. Predictors will be single point mutations, deletion, insertion, duplication, etc.In this case, the Goal will be Prediction.

3.Question 2.4.6 pg 53

Describe the differences between a parametric and a non-parametric statistical learning approach. What are the advantages of a parametric approach to regression or classification (as opposed to a nonparametric approach)? What are its disadvantages?

In a non-parametric model, The complexity of the model increases with the increase of the training data. Whereas, in a parametric model, there are a fixed number of model. In a non-parametric model, the information about the population is unknown. Whereas, In the parametric model, information about the population is known completely.The null hypothesis is free from the parameters in the non-parametric model, and the null hypothesis is made on the parameters of the population distribution.Results for the non-parametric model cannot be significantly affected by the outliers. However, the results of the parametric model can be significantly affected by the outliers.

Advantage of parametric models:

1. It requires less data for training the data.
2. It is simpler model and easy to understand.

Disadvantage of parametric models:

1. The size of the sample is always big that makes difficult to carry the whole test.
2. The result get affected by the outliers.

Advantage of non-parametric models:

1. Assumption of the distribution is not required.
2. This model is applicable to all kind of data.

Disadvantage of non-parametric models:

1. This model is less effective than the parametric model.
2. It takes more data to train data and longer time to train data.

4.Question 2.4.8 pg 54-55 (skip Part b. For Part a -'College' can be found in the `ISLR` library).

This exercise relates to the College data set, which can be found in the file College.csv. 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 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 R. Call the loaded data college. Make sure that you have the directory set to the correct location for the data.

```
## [1] 777  19
```

I am currently working in the correct working directory. For setting the current working directory. We can use the a code

```
setwd(C:/Users/Yamuna/Documents/Acamedic_materials/STAT_602/Week_2/Homework)
```

(c)

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

I have used loop in this question to see the summaries of the variables because it saves the time. Since, I do not need to write the name of the variable.

```
## The summary of: X
```

```
##            Abilene Christian University                       Adelphi University
##                                       1                                        1
##                           Adrian College                     Agnes Scott College
##                                       1                                        1
##               Alaska Pacific University                       Albertson College
##                                       1                                        1
##                  Albertus Magnus College                          Albion College
##                                       1                                        1
##                          Albright College               Alderson-Broaddus College
```

```
##                                     1                                     1
##                        Alfred University                      Allegheny College
##                                     1                                     1
## Allentown Coll. of St. Francis de Sales                          Alma College
##                                     1                                     1
##                           Alverno College          American International College
##                                     1                                     1
##                           Amherst College                    Anderson University
##                                     1                                     1
##                         Andrews University                 Angelo State University
##                                     1                                     1
##                         Antioch University            Appalachian State University
##                                     1                                     1
##                           Aquinas College     Arizona State University Main campus
##                                     1                                     1
##          Arkansas College (Lyon College)                Arkansas Tech University
##                                     1                                     1
##                         Assumption College           Auburn University-Main Campus
##                                     1                                     1
##                          Augsburg College                      Augustana College
##                                     1                                     1
##                      Augustana College IL                        Austin College
##                                     1                                     1
##                          Averett College                       Baker University
##                                     1                                     1
##                  Baldwin-Wallace College                          Barat College
##                                     1                                     1
##                               Bard College                       Barnard College
##                                     1                                     1
##                          Barry University                      Baylor University
##                                     1                                     1
##                           Beaver College                     Bellarmine College
##                                     1                                     1
##                    Belmont Abbey College                     Belmont University
##                                     1                                     1
##                           Beloit College                 Bemidji State University
##                                     1                                     1
##                      Benedictine College                     Bennington College
##                                     1                                     1
##                          Bentley College                         Berry College
##                                     1                                     1
##                          Bethany College                       Bethel College
##                                     1                                     1
##                         Bethel College KS                 Bethune Cookman College
##                                     1                                     1
##              Birmingham-Southern College                      Blackburn College
##                                     1                                     1
##          Bloomsburg Univ. of Pennsylvania                     Bluefield College
##                                     1                                     1
##                          Bluffton College                      Boston University
##                                     1                                     1
##                           Bowdoin College          Bowling Green State University
##                                     1                                     1
##                          Bradford College                      Bradley University
```

```
##                                     1                                     1
##                    Brandeis University                       Brenau University
##                                     1                                     1
##                 Brewton-Parker College                    Briar Cliff College
##                                     1                                     1
##                    Bridgewater College      Brigham Young University at Provo
##                                     1                                     1
##                       Brown University                       Bryn Mawr College
##                                     1                                     1
##                    Bucknell University                    Buena Vista College
##                                     1                                     1
##                      Butler University                       Cabrini College
##                                     1                                     1
##                      Caldwell College         California Lutheran University
##                                     1                                     1
##        California Polytechnic-San Luis   California State University at Fresno
##                                     1                                     1
##                          Calvin College                    Campbell University
##                                     1                                     1
##                 Campbellsville College                      Canisius College
##                                     1                                     1
##                     Capital University                      Capitol College
##                                     1                                     1
##                      Carleton College             Carnegie Mellon University
##                                     1                                     1
##                       Carroll College                 Carson-Newman College
##                                     1                                     1
##                      Carthage College        Case Western Reserve University
##                                     1                                     1
##                Castleton State College                       Catawba College
##                                     1                                     1
##         Catholic University of America                     Cazenovia College
##                                     1                                     1
##                    Cedar Crest College                    Cedarville College
##                                     1                                     1
##                      Centenary College                              (Other)
##                                     1                                   678
```

```
## The summary of: Private
```

```
##  No Yes
## 212 565
```

```
## The summary of: Apps
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      81     776    1558    3002    3624   48094
```

```
## The summary of: Accept
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      72     604    1110    2019    2424   26330
```

```
## The summary of: Enroll
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##      35     242     434     780     902    6392
```

```
## The summary of: Top10perc
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00   15.00   23.00   27.56   35.00   96.00
```

```
## The summary of: Top25perc
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     9.0    41.0    54.0    55.8    69.0   100.0
```

```
## The summary of: F.Undergrad
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     139     992    1707    3700    4005   31643
```

```
## The summary of: P.Undergrad
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     1.0    95.0   353.0   855.3   967.0 21836.0
```

```
## The summary of: Outstate
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2340    7320    9990   10441   12925   21700
```

```
## The summary of: Room.Board
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1780    3597    4200    4358    5050    8124
```

```
## The summary of: Books
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    96.0   470.0   500.0   549.4   600.0  2340.0
```

```
## The summary of: Personal
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     250     850    1200    1341    1700    6800
```

```
## The summary of: PhD
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    8.00   62.00   75.00   72.66   85.00  103.00
```

```
## The summary of: Terminal
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    24.0    71.0    82.0    79.7    92.0   100.0
```

```
## The summary of: S.F.Ratio
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    2.50   11.50   13.60   14.09   16.50   39.80
```

```
## The summary of: perc.alumni
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   13.00   21.00   22.74   31.00   64.00
```

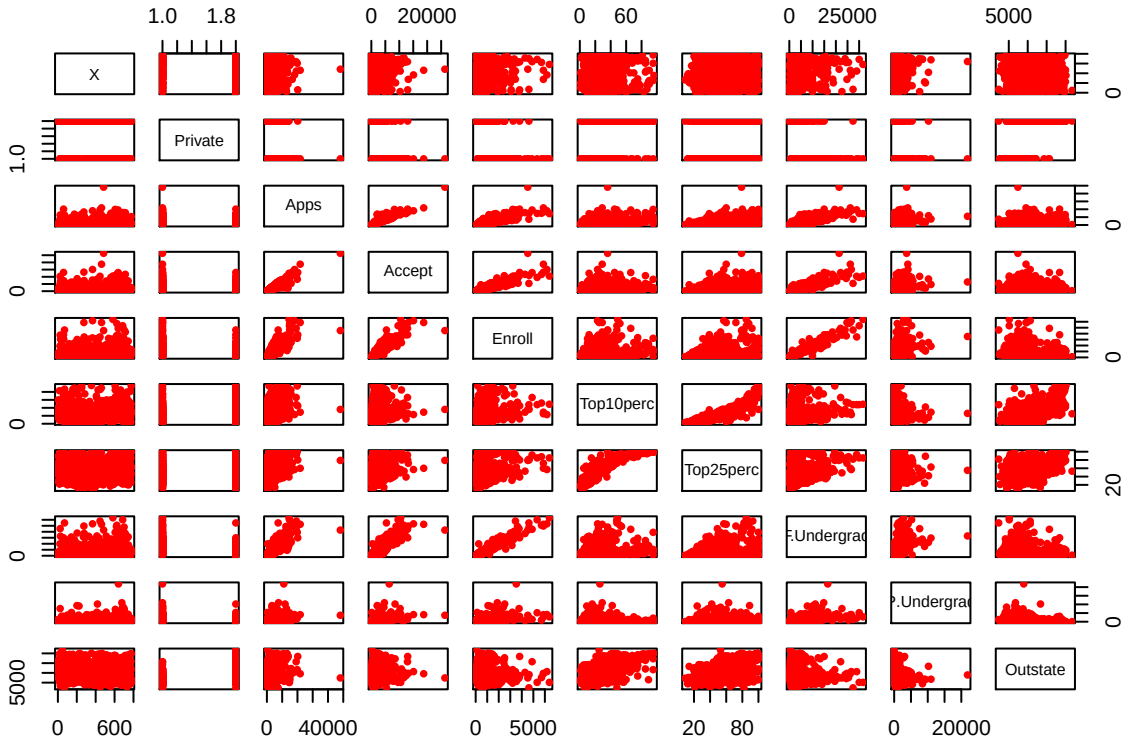
```
## The summary of: Expend
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    3186    6751    8377    9660   10830   56233
```

```
## The summary of: Grad.Rate
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   53.00   65.00   65.46   78.00  118.00
```

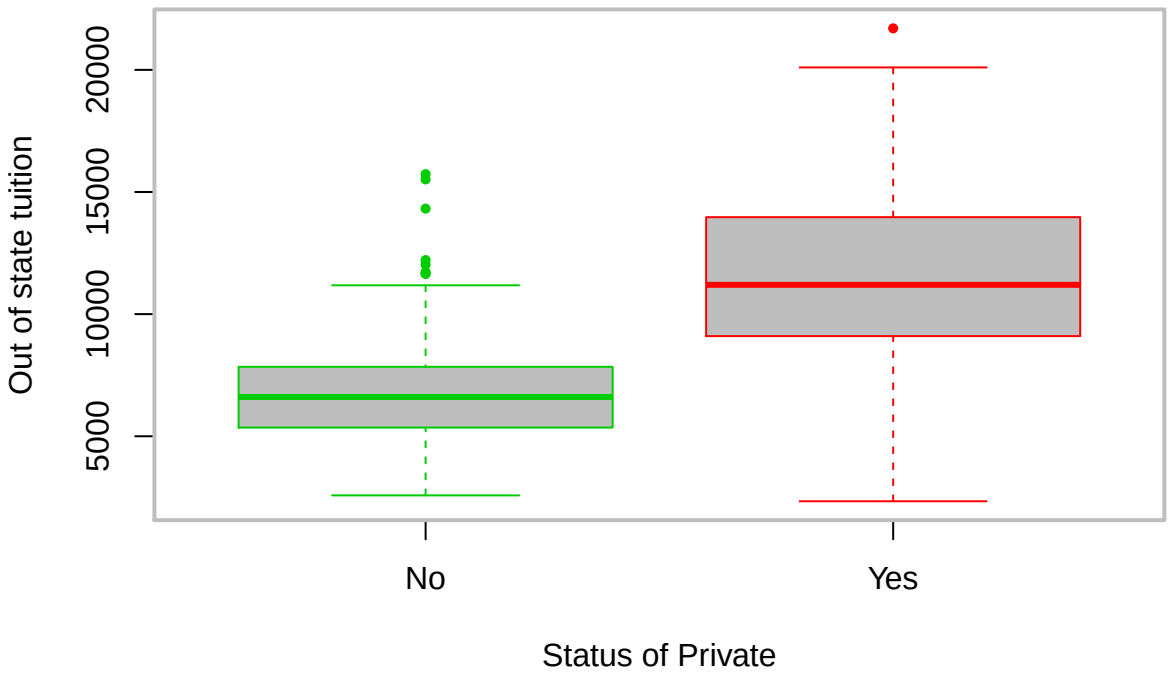
ii.Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall thatyou can reference the first ten columns of a matrix A using A[,1:10].

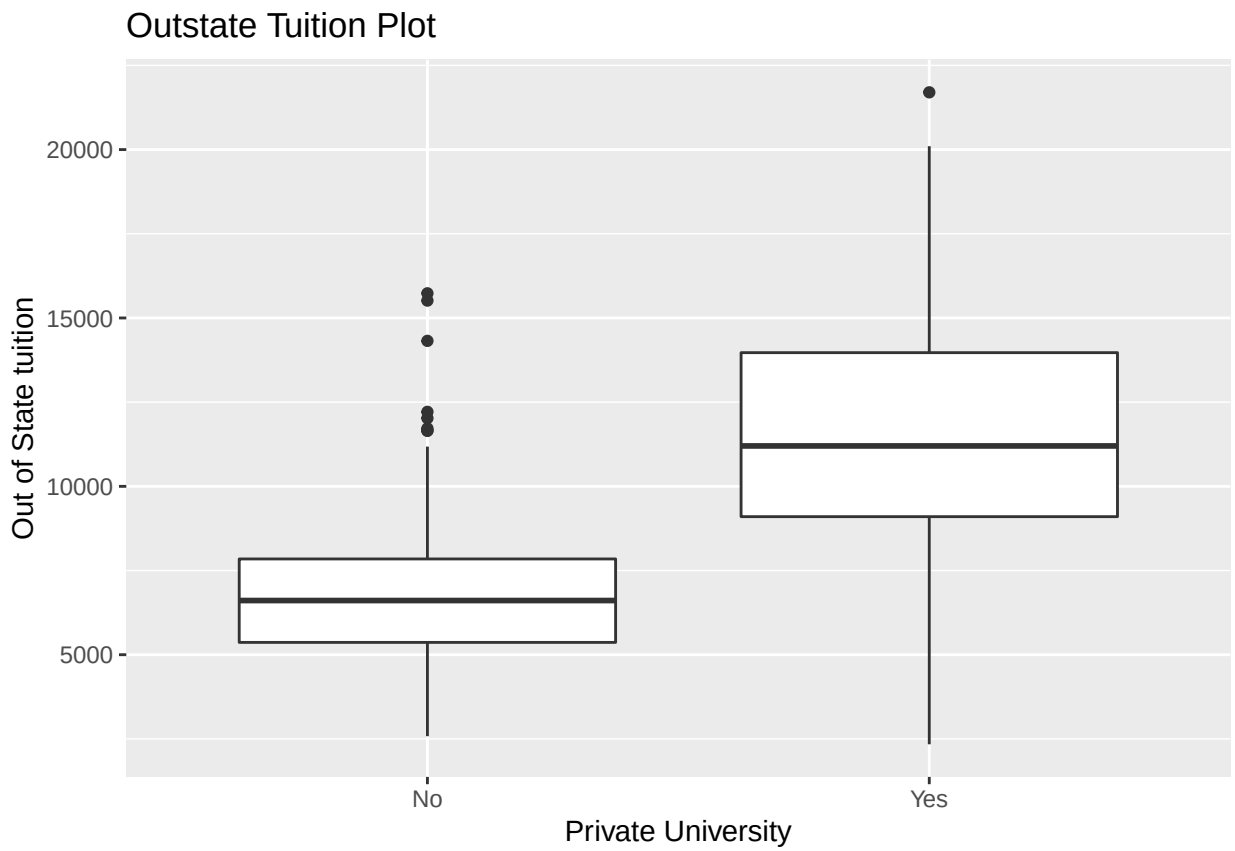


From the scatter plot we can see the linear and the quadratic relationship among the variables. Variables such as between F.Undergard vs Enroll, Accept vs Enroll, Apps vs Accept etc and the variables Outstate vs Room.Board have the positive relationship. The variables Top10Perc vs Top25Perc has quadratic relationship.

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

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```


Out of state tuition
5000
10000
15000
20000
No
Yes
Status of Private

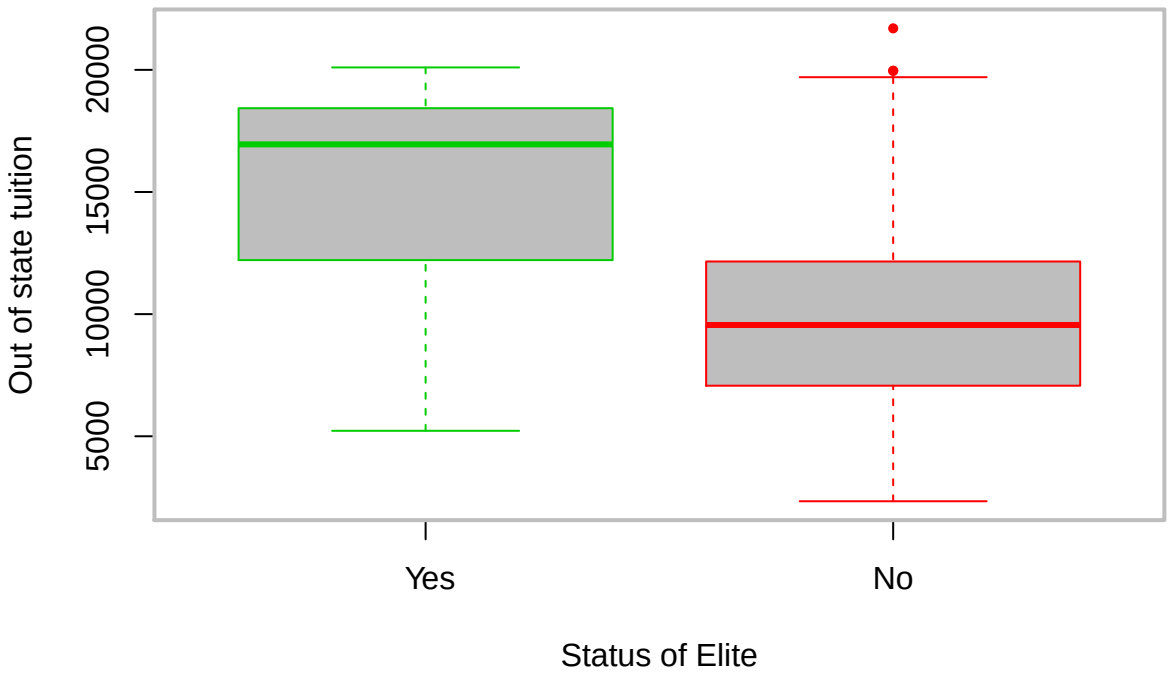


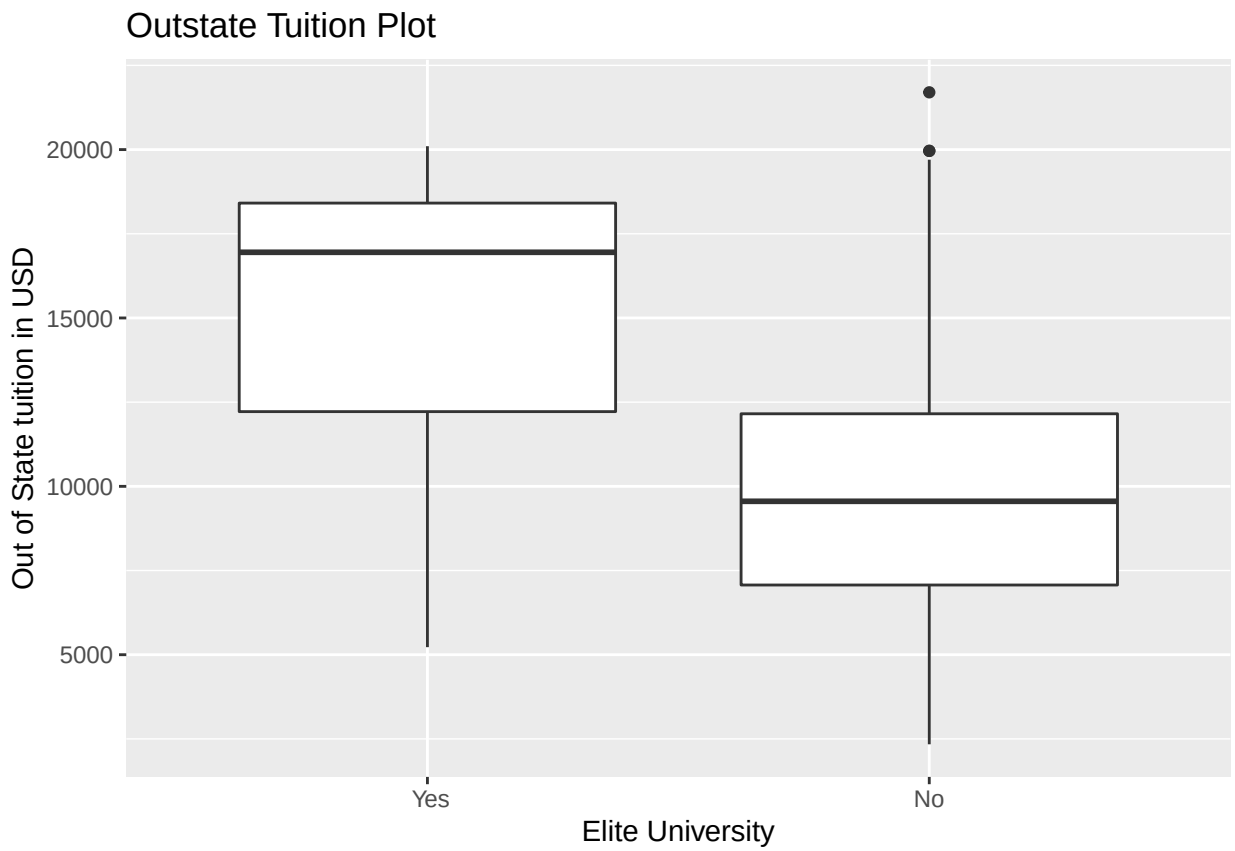
The plot shows that the the out of state tution is higher in the private colleges than the public colleges. Public college has lower tuition compared to the private college.

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

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.

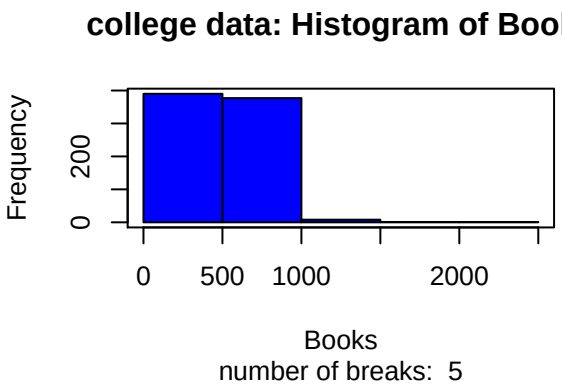
```
##  Yes   No
##   78  699
```

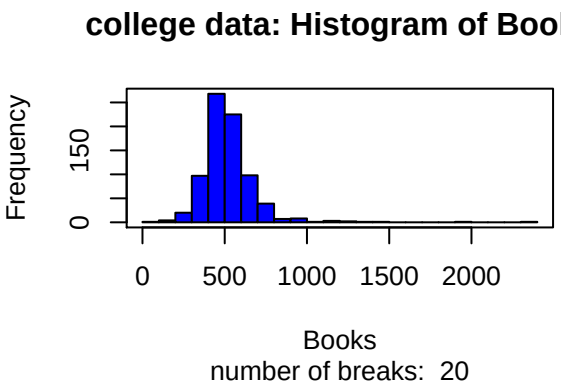

Out of state tuition
5000
10000
15000
20000
Yes
No
Status of Elite

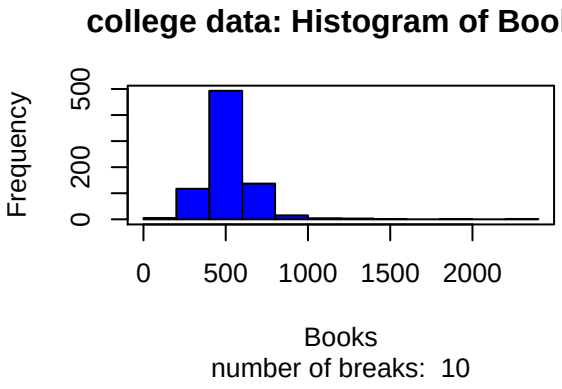


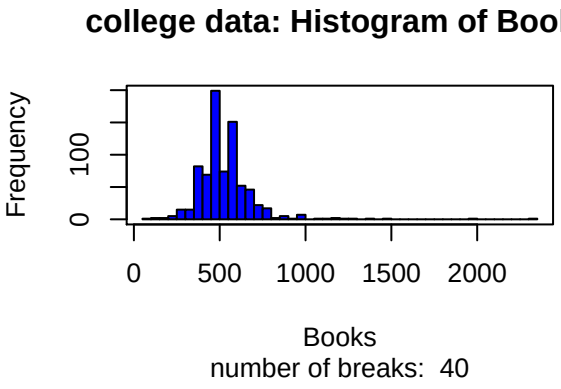
The summary of the data shows that 78 of the colleges out of 777 colleges falls under the elite group whereas, 699 do not fall under the elite group. From the figure, we can see that the out-of-state tuition is much higher for elite universities. Also, the first quartile of the elite university with the out-of-state tuition is higher than the third quartile.

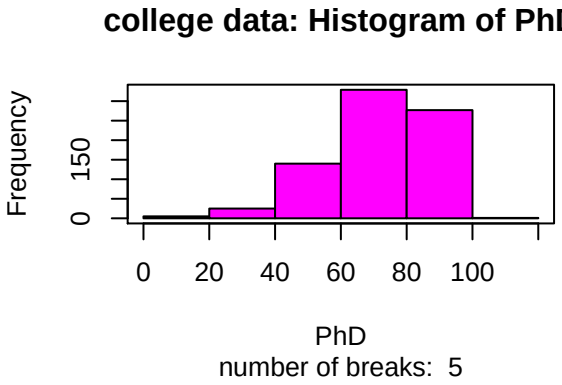
v.Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.You may find the command par(mfrow=c(2,2)) useful:it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to this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

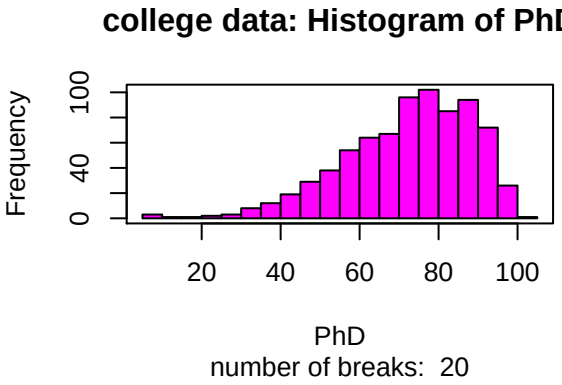

college data: Histogram of Books
Frequency
0
200
0
500
1000
2000
Books
number of breaks: 5

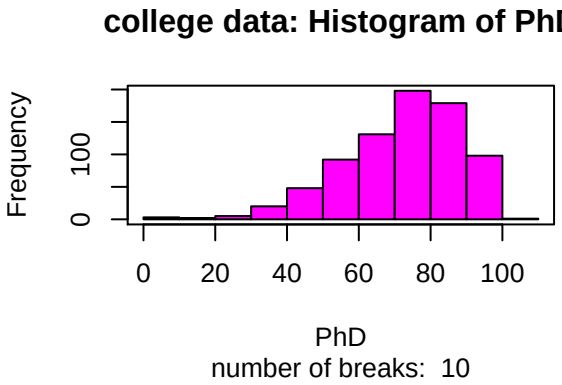

college data: Histogram of Books
Frequency
0
150
0
500
1000
1500
2000
Books
number of breaks: 20

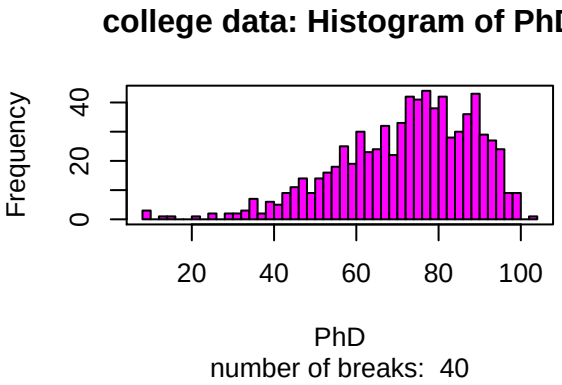

college data: Histogram of Books
Frequency
0
200
500
0
500
1000
1500
2000
Books
number of breaks: 10

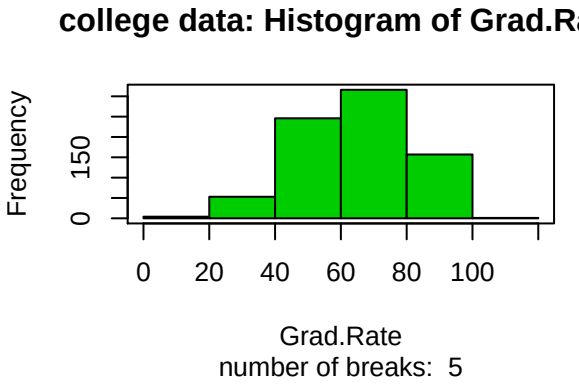

college data: Histogram of Books
Frequency
0
100
0
500
1000
1500
2000
Books
number of breaks: 40

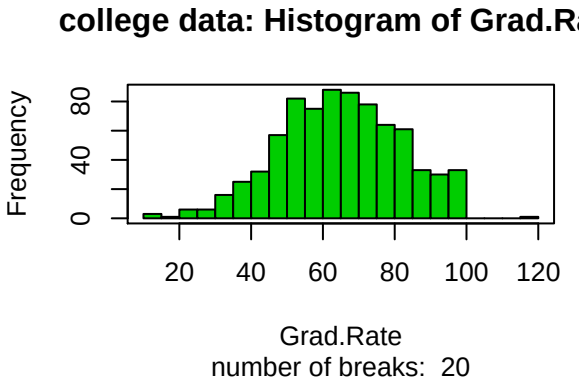

college data: Histogram of PhD
Frequency
0
150
0
20
40
60
80
100
PhD
number of breaks: 5

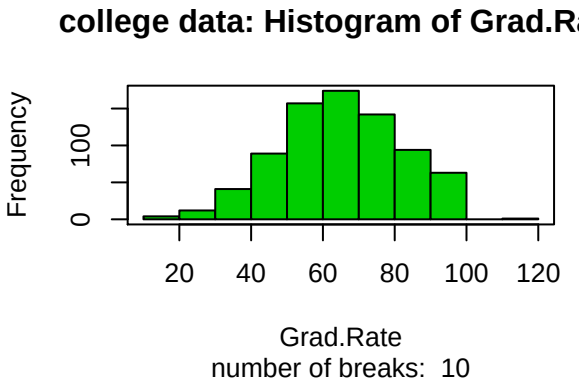

college data: Histogram of PhD
Frequency
0
40
100
20
40
60
80
100
PhD
number of breaks: 20

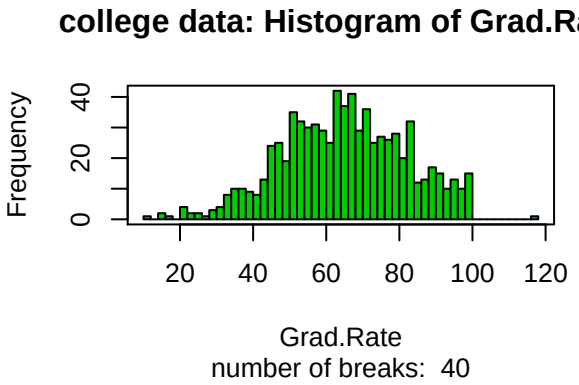

college data: Histogram of PhD
Frequency
0
100
0
20
40
60
80
100
PhD
number of breaks: 10

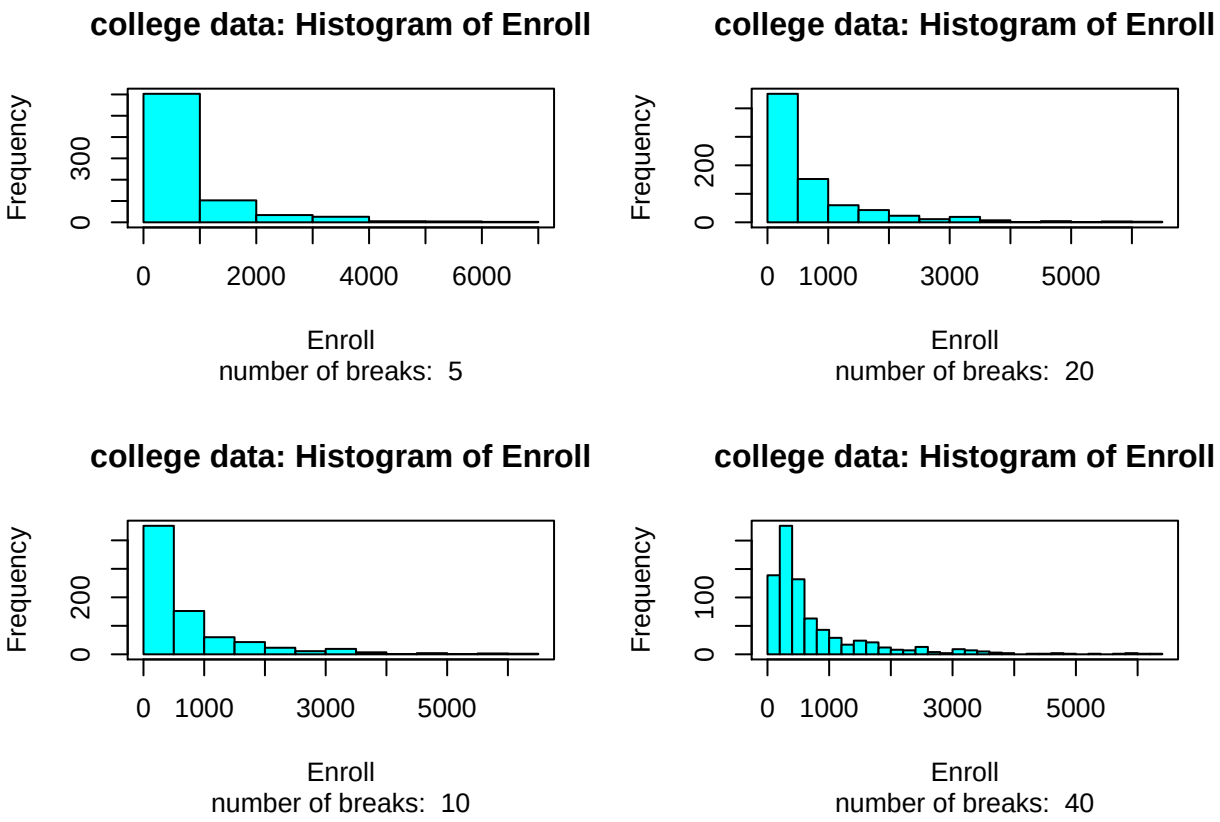

college data: Histogram of PhD
Frequency
0
20
40
20
40
60
80
100
PhD
number of breaks: 40


college data: Histogram of Grad.Rate
Frequency
0
150
0
20
40
60
80
100
Grad.Rate
number of breaks: 5


college data: Histogram of Grad.Rate
Frequency
0
40
80
20
40
60
80
100
120
Grad.Rate
number of breaks: 20

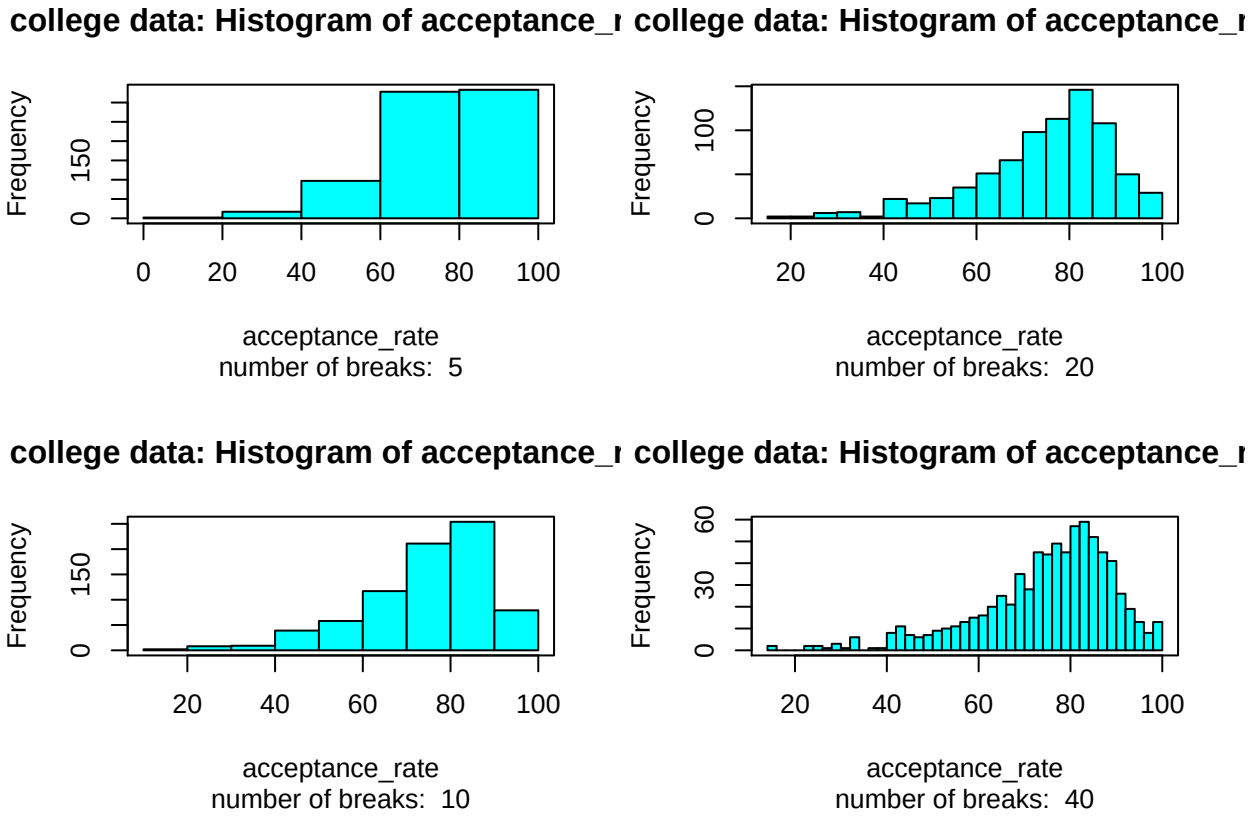

college data: Histogram of Grad.Rate
Frequency
0
100
20
40
60
80
100
120
Grad.Rate
number of breaks: 10

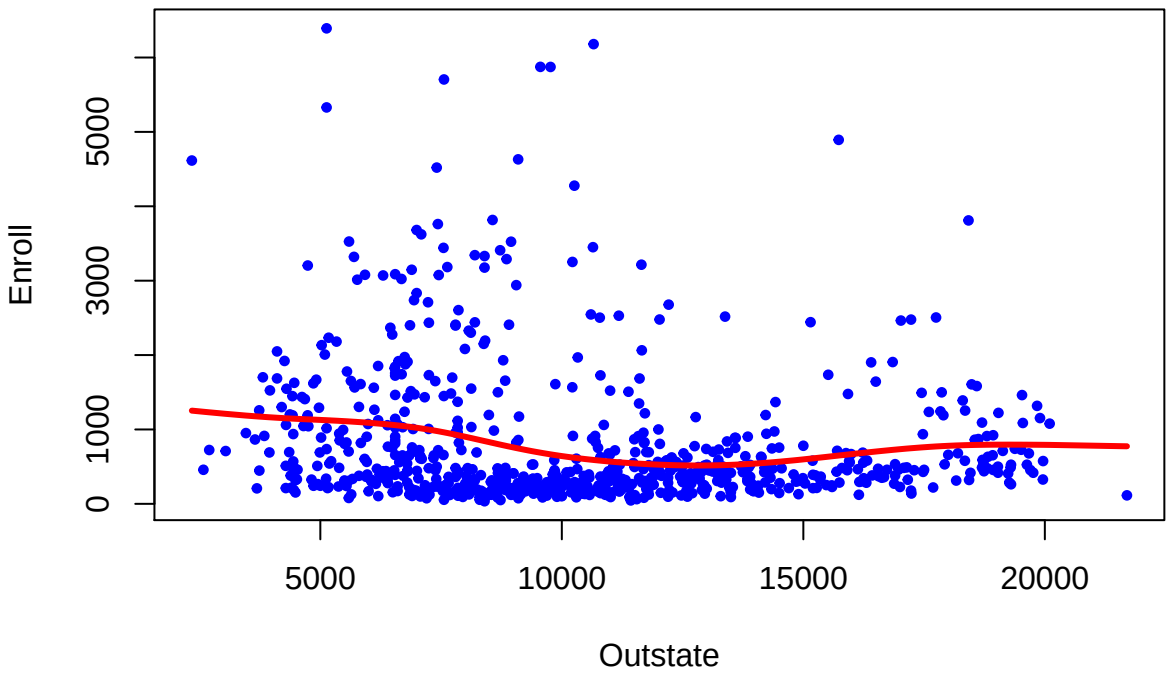

college data: Histogram of Grad.Rate
Frequency
0
20
40
20
40
60
80
100
120
Grad.Rate
number of breaks: 40

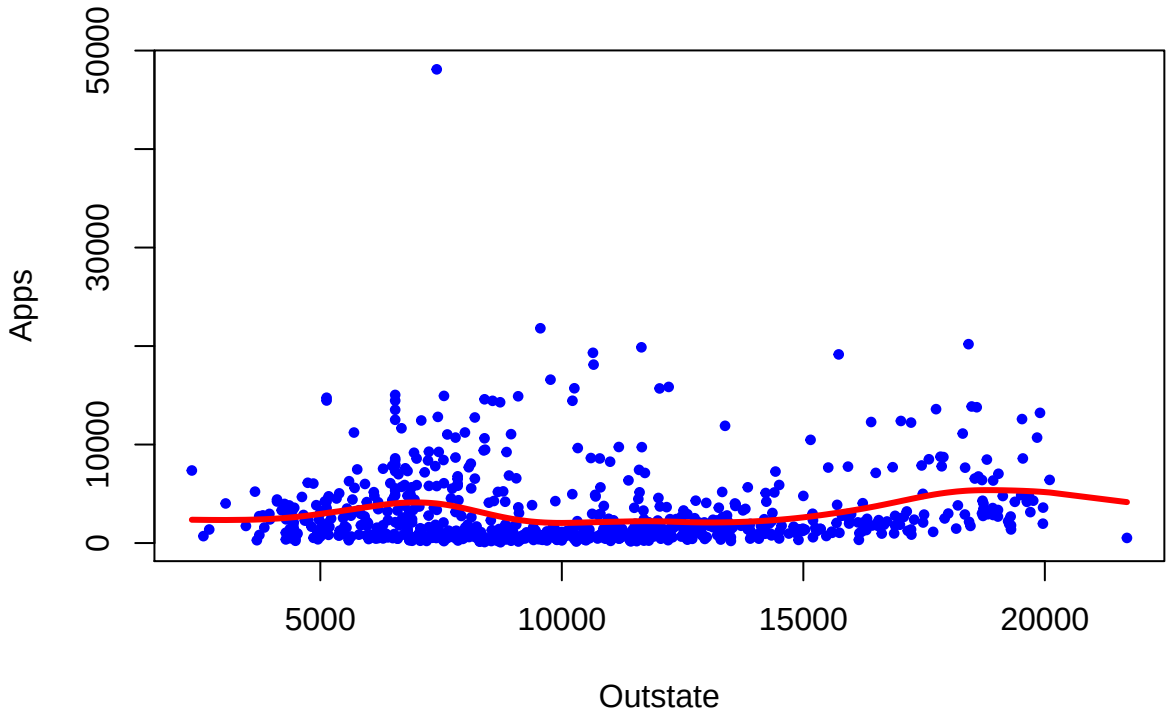


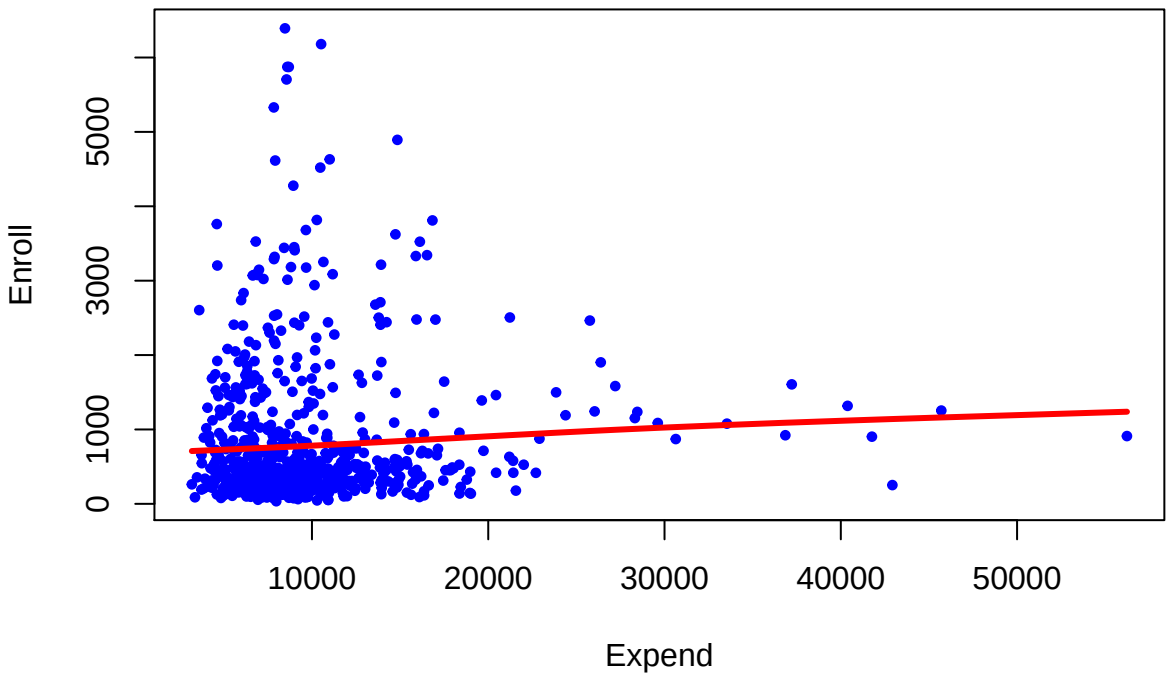
From the plot above, it shows that the cost of the book is normally distributed with the long right tail. The property of the distribution is not distinguishable with the break = 5. The distribution visible in break 20 and 10. From the plot named Histogram of Ph.D., it seems the data is mostly left-skewed. We also can know that most of the colleges do have the Ph.D. faculty in their college. The histogram is named grad rate shows that the graduation rate is normally distributed. Likewise, The histogram of enrollment shows that the enrollment seems exponentially distributed.

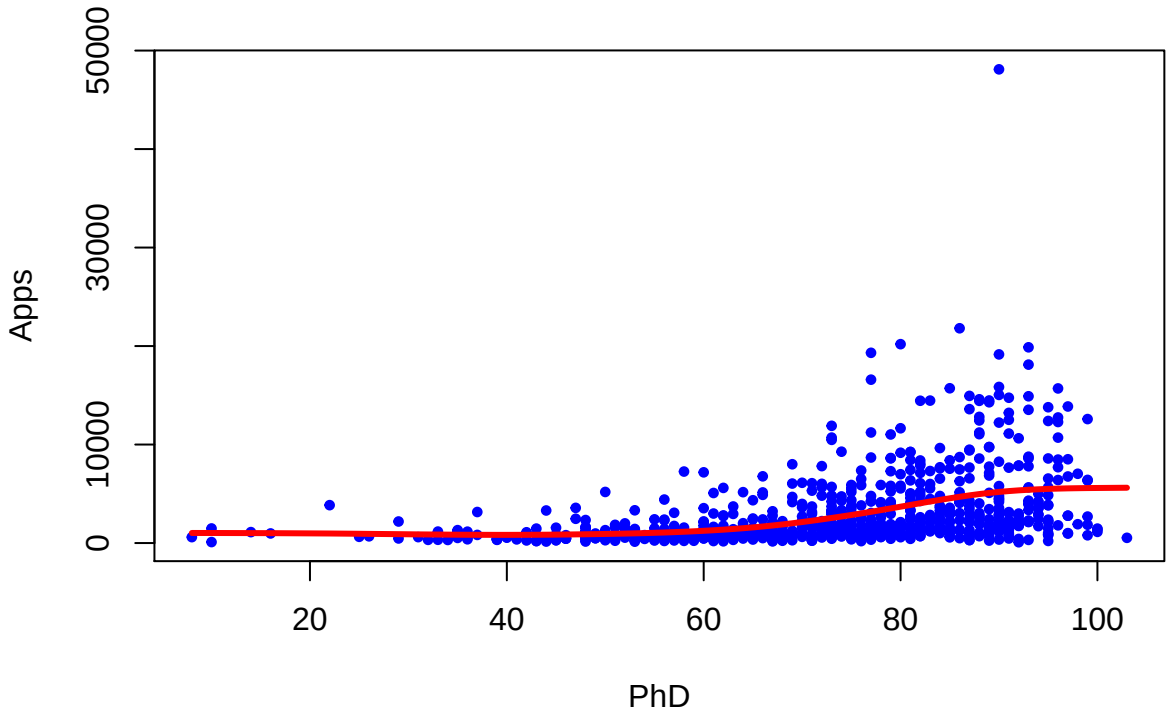
vi.Continue exploring the data, and provide a brief summary of what you discover.

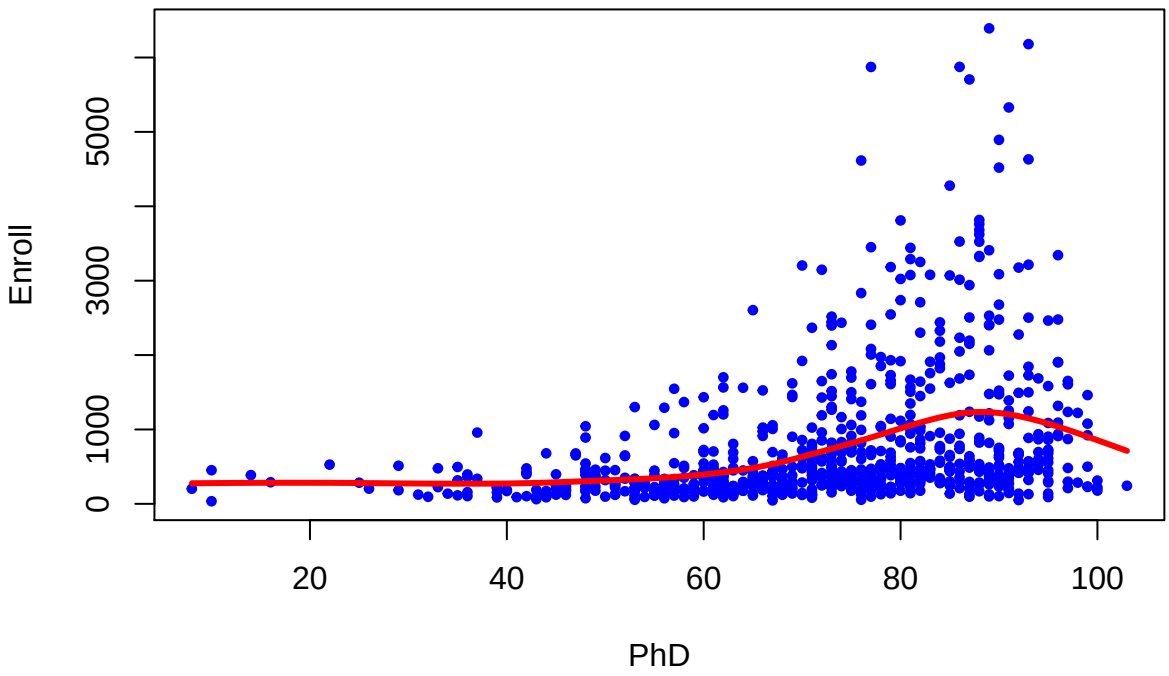

college data: Histogram of acceptance_r
Frequency
0
150
0
20
40
60
80
100
acceptance_rate
number of breaks: 5
college data: Histogram of acceptance_r
Frequency
0
100
20
40
60
80
100
acceptance_rate
number of breaks: 20
college data: Histogram of acceptance_r
Frequency
0
150
20
40
60
80
100
acceptance_rate
number of breaks: 10
college data: Histogram of acceptance_r
Frequency
0
30
60
20
40
60
80
100
acceptance_rate
number of breaks: 40

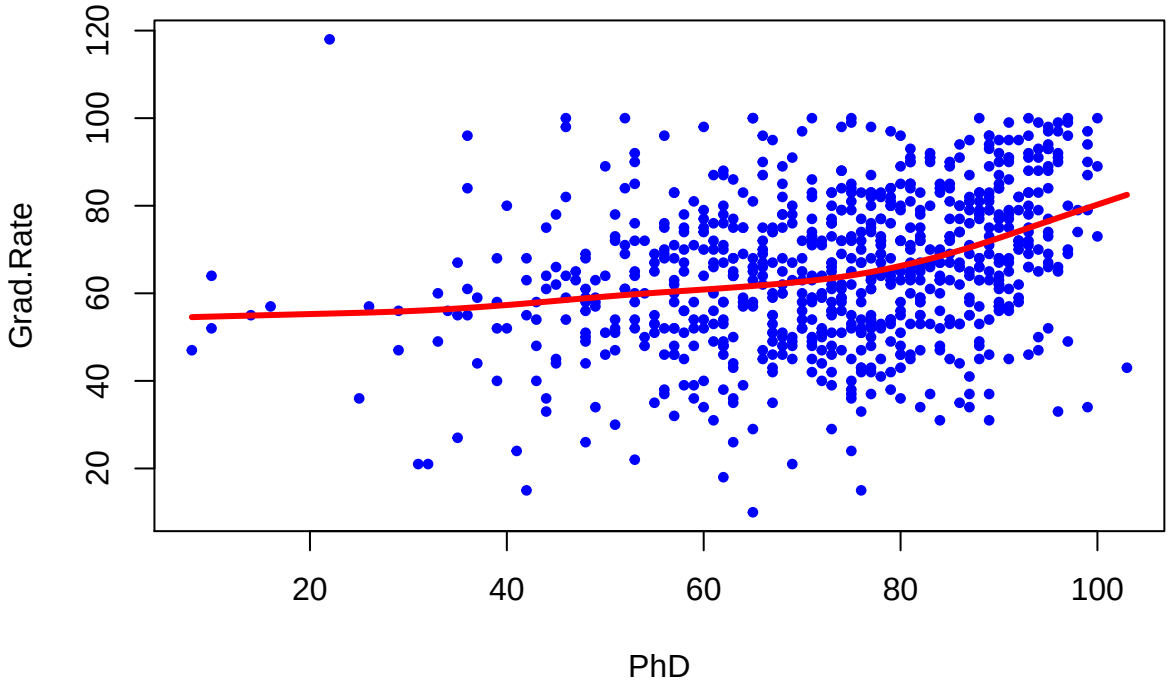

Enroll
0
1000
3000
5000
5000
10000
15000
20000
Outstate


Apps
0
10000
30000
50000
5000
10000
15000
20000
Outstate


Enroll
0
1000
3000
5000
10000
20000
30000
40000
50000
Expend


Apps
0
10000
30000
50000
20
40
60
80
100
PhD


Enroll
0
1000
3000
5000
20
40
60
80
100
PhD



I wanted to check the distribution of acceptance rate for the application, so I generated the variable `acceptance_rate` by dividing the acceptance of application by the no of application and multiplying the function with 100. The histogram of the acceptance rate shows that the data is skewed to the left with the long and heavy tail towards the right. I also wanted to know how the cost affects the other variables. So, the graph shows that the out of state tuition cost does not affect much the number of application.Additionally, I also wanted to see if the Ph.D. faculties help in the application rates. From the graphs, it is visible that there is no such relation between the application and the Ph.D. faculties. Further, I also wanted to see the enrollment and the graduation rate for the Ph.D. students. It has the positive relationship between the graduation rate and the enrollment with Ph.D.