Homework 1- STAT 602

Snigdha Peddi

Question 2.4.2. 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.

Answer 2.a: Here we are dealing with a **Regression problem** as the dependent variable "CEO salary" is quantitative response variable. We are interested in **inference** as we are studying the relation between the independent variables and the CEO salary(dependent variable). Here, n=500 and p=3.

(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.

Answer 2.b: Here the dependent variable is a qualitative response of Success or Failure, hence a **Classification Problem**. It is a **Prediction** problem as we would like to know if the new product launch will be a success or failure. Here, n=20 and p=13.

(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.

Answer 2.c: The dependent variable here is %change in the USD/Euro exchange rate and is a quantitative response variable. Hence it is a **Regression Problem**.It is a **prediction** problem.Here, n=52(52 weeks in year 2012) and p=3.

Question 2.4.4. 4. 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.

Example 1: Classification is useful in diagnosis of a disease. By analyzing the clinical parameters one can provide a prognosis if the patient has a disease or not, or can provide information on if the disease is in early stages or advance state etc.

Response variable: Disease present or Absent, Mild state/ Advance state etc.

Predictors: Clinical parameters like urine analysis, blood work, number of hospital visits, number of inpatient admission, medication prescribed etc.

Goal: Can be prediction (if we want to use the historical data to predict the new patient condition) or Inference (if we would like to understand the factors effecting the presence or absence of a disease.)

Example 2: Classifying customers by banks before giving loans on their ability to pay the loans.

Response variable: Default or Non-Default

Predictors: Age, Income,monthly expense, education, salary/debt ratio etc.

Goal: Prediction(To predict if the customer will default on a loan or not)

Example 3: Classifying an Email as Spam or Not Spam

Response variable: Spam or Not Spam

Predictors: most common words used in emails

Goal: Prediction(to predict if the email is spam or not)

(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.

Example 1: Predicting revenue based on budget allocated on various marketing channels

Response variable: Revenue

Predictors: Budget allocated for each marketing channel like TV,Radio,Newspaper, social media etc

Goal: Inference (How advertising spending effects the revenue is what we would like to understand here)

Example 2: Agricultural scientists use regression techniques to measure effect of fertilizer and water on crop yields

Response variable: crop yields

Predictors: different fertilizers, amount of fertilizer, amount of water, number of times crops are watered, number of times fertilizer is used,active ingredient of fertilizer, PH level of water, PH lever of fertilizer etc

Goal: Inference (how the use of a particular fertilizer and water effect the crop yield)

Example 3: Effect of drug on blood pressure

Response variable: Blood pressure

Predictors: Drug name, dose of drug, time of administration, route of administration etc,.

Goal: Inference(understand the relation between the amount of drug and blood pressure of the subject)

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

Example 1: Cluster analysis is helpful in Identifying Fake news.The content of the news will be analyzed and words or corpus will be clustered. The words that are commonly used to sensationalize the news can be identified and predict if the news is fake or genuine.

Example 2: Clustering is also useful in market segmentation. Customer behavior and buying patterns can be clustered and each group of customers can be targeted with individual marketing strategy which would help increase the sales and build customer trust.

Example 3: It can be used to classify network traffic to the websites.Characteristics of Traffic sources are used to cluster them into groups of harmful traffic or areas driving growth.It is a faster process than auto-class method and help developer to grow the site and plan capacity effectively.

REFERENCES

- Statology Blog by Zach,*4 Examples of Using Linear Regression in Real Life*,May 19,2020.
- Datafloq Blog by Claire Whittaker,*7 Innovative Uses of Clustering Algorithms in the Real World*,April 4,2019.

Question 2.4.6. 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?

Answer:

Parametric methods are used to make inferences about population parameters.These methods assume a form(model) for the data.It learns from a predefined mapped function.They have fixed number of parameters.The method works well if the assumptions are correct.If the data does not fit the assumed model when the assumptions are incorrect.They are not flexible models.

If the data assumes to fit a linear model of form

$$\beta_0 + \beta_1 X_1 + \beta_2 X_2 = Y \tag{1}$$

All we need is to estimate the coefficients of the parameters to get Y value. The number of parameters remain same for any size of data. If the model assumed is right the predictions will be accurate. In case the model does not fit well the data can be transformed to its log form or squared to get better accuracy.

Advantages: They are simpler, easy to understand and interpretable.Computationally faster and requires less data.

Disadvantages: Constrained and stick to their assumptions, limited complexity, can result into poor fit.

Examples of Parametric methods include linear Regression,Logistic regression, Linear Discriminany analysis

Non-parametric methods do not assume a model. It learns from the training data.The number of parameters may increase as it learns from the data.These are flexible models.Compared to the parametric models these are not very interpretable.

Advantages: Higher performance and very flexible.

Disadvantages: Requires lot of data to learn and is slower. May lead to overfitting.

Examples of Non parametric methods include KNN, Decision Tree model.

REFERENCES

- Analytics Vidhya Blog ,*What is the difference between parametric and non-parametric regression?*.
- Blog by Lamiae Hana,*Parametric and Nonparametric Machine Learning Algorithms*,Aug 9,2020.

Question 2.4.8. 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.

```
##
##  Dimensions of dataset: 777 18
```

```
##
##  Number of missing values in dataset: 0
```

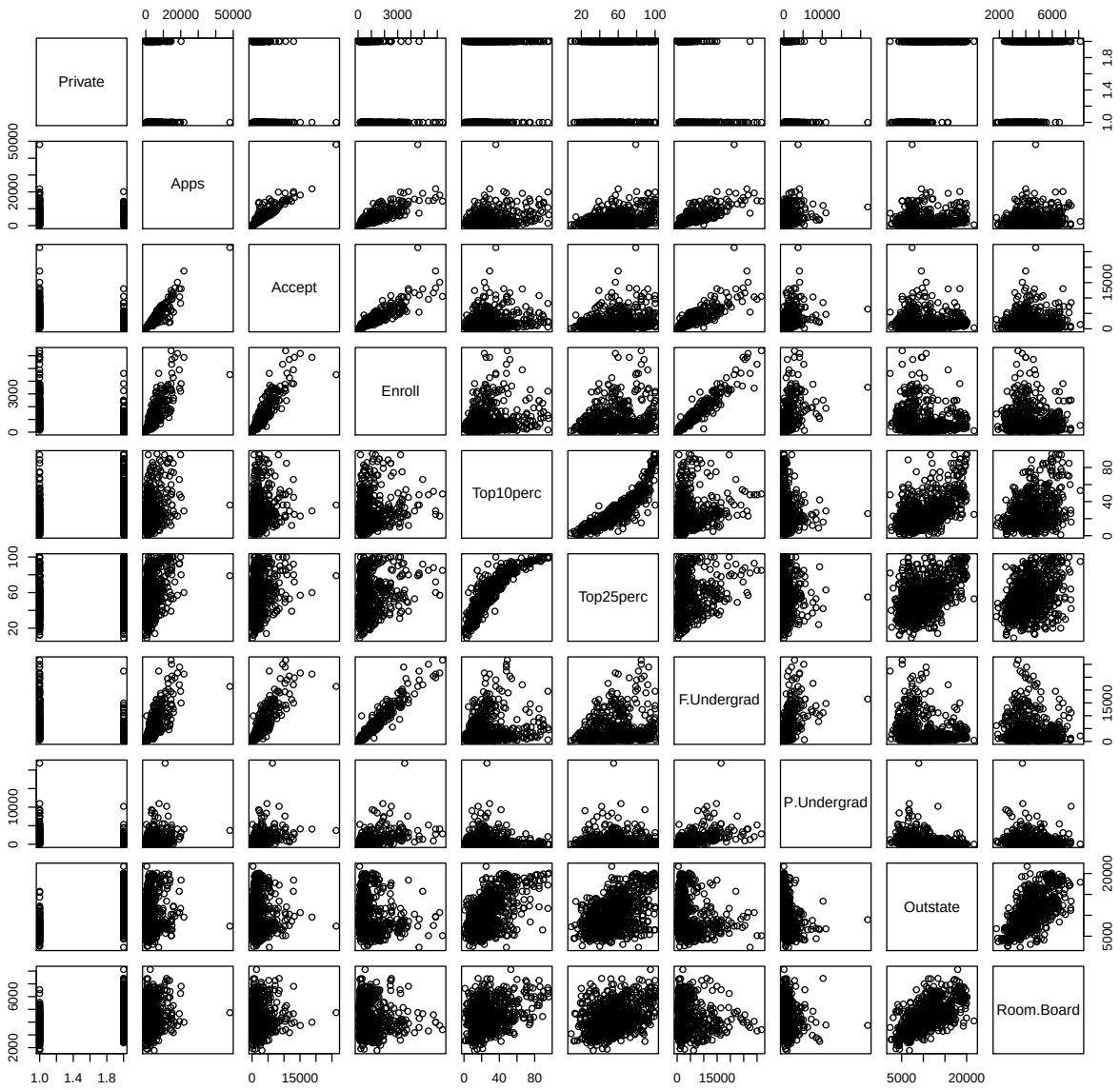
(c) i. Use the summary() function to produce a numerical summary of the variables in the data set.

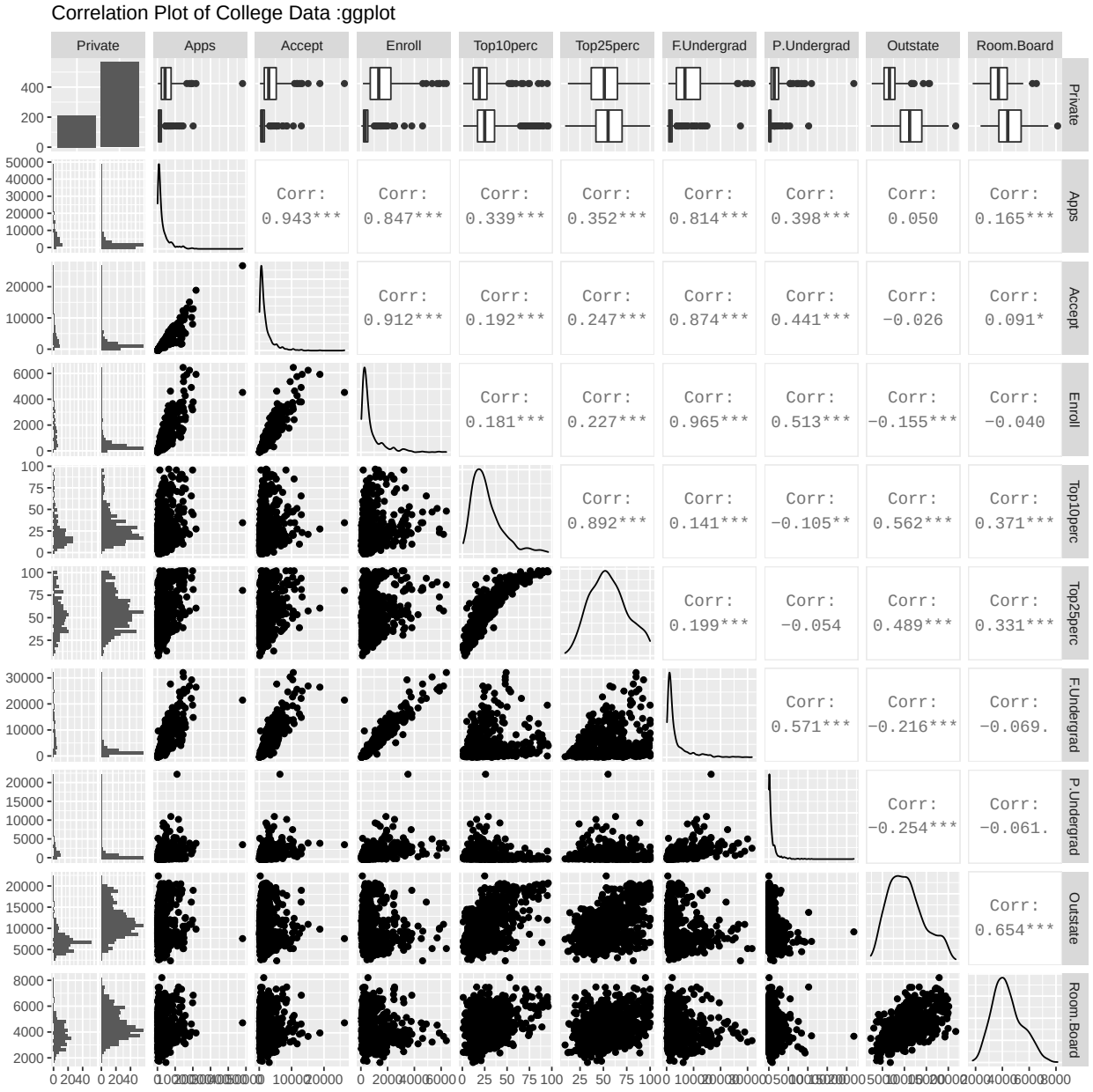
```
##
##  Summary of College Data:

##  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].

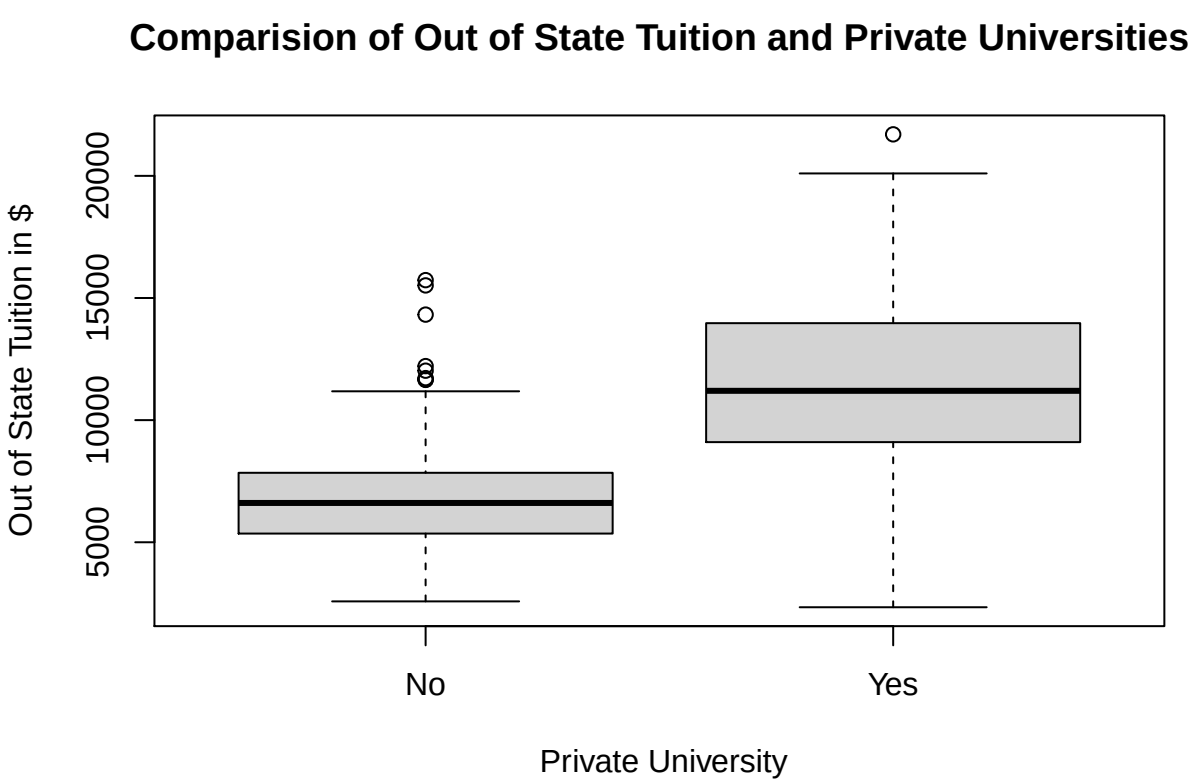
Correlation Plot of College Data



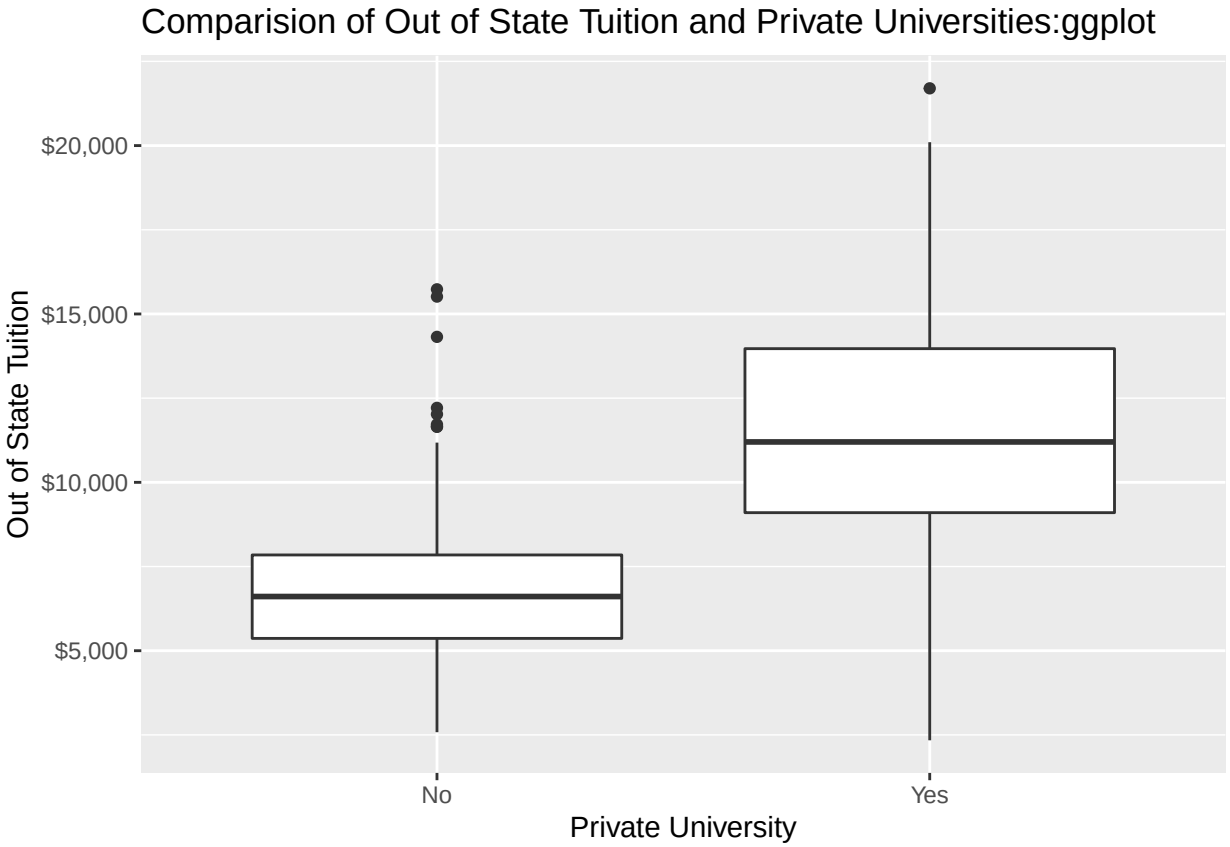


From the plot it is clear that 94.3% of applications were accepted.91.2% of accepted students enrolled into one of the Universities.96.5% of students enrolled are Full time students.

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



From the Plot it is clear that the median tuition of Private institutions is higher for the Out of state students than the Public Institutions.And over all fee for Out of state students is lower at Public institutions compared to private institutions.

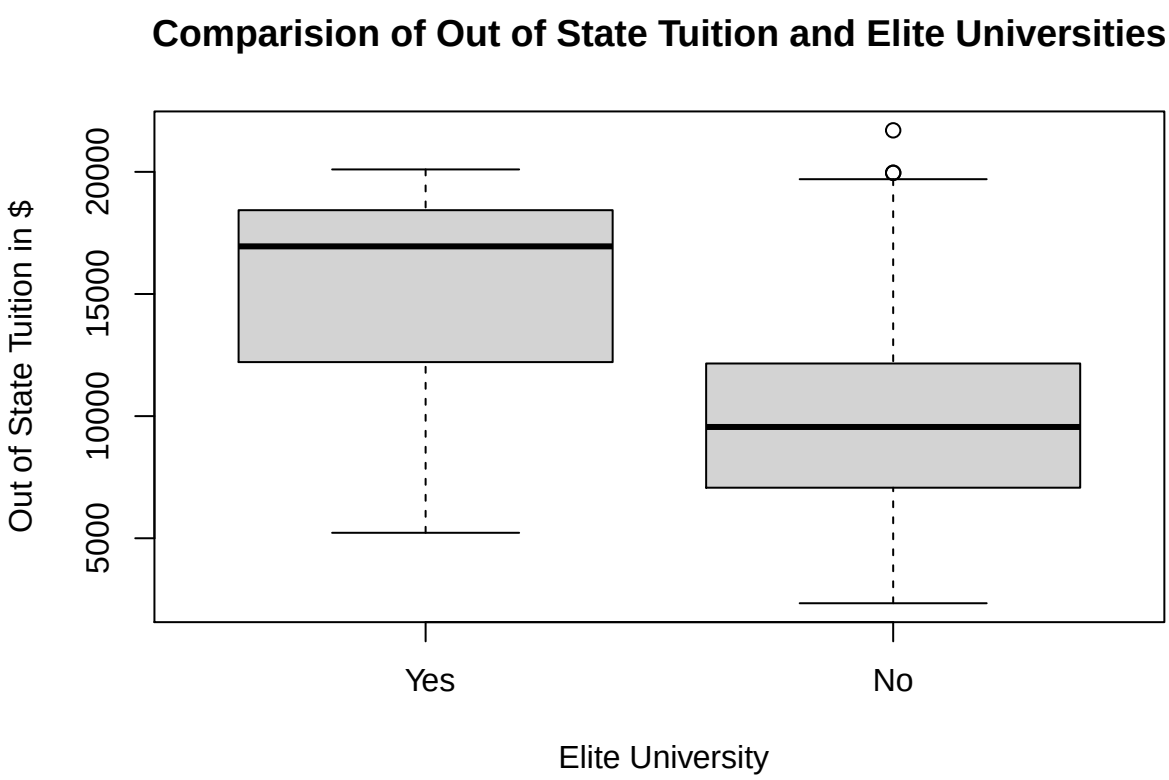


iv. Create a new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10 % of their high school classes exceeds 50 %.

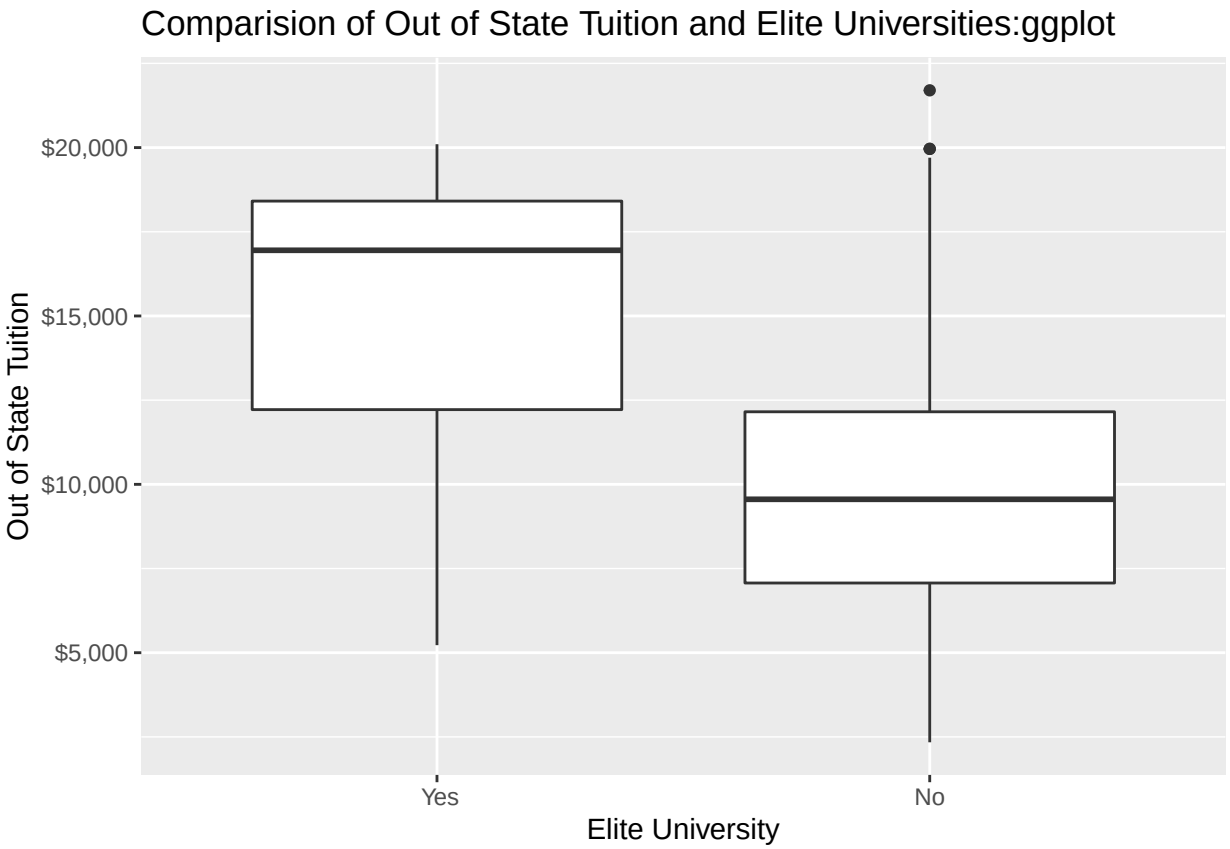
Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

```
##
##  Summary of Elite Variable:

##  Yes   No
##   78  699
```

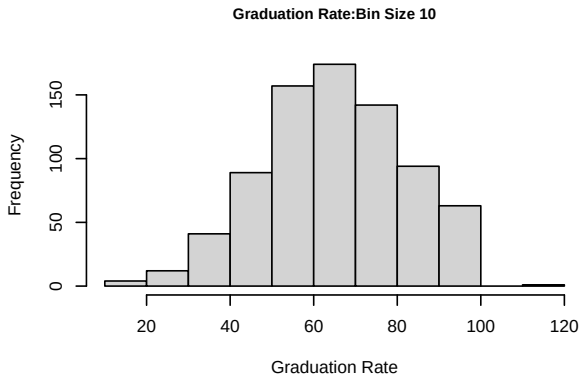


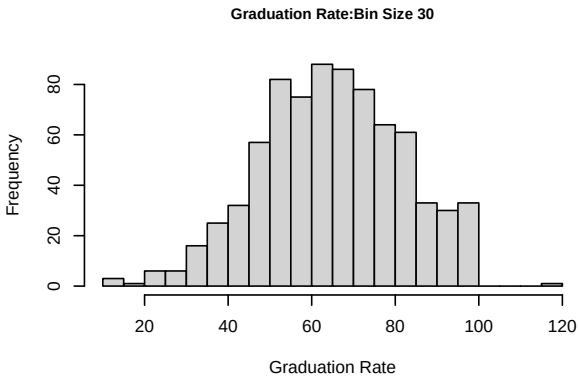
The median tuition fee for Elite Universities is higher for Out of State students compared to the other Universities.75th percentile of the tuition paid to Non-Elite colleges is lower than the 25th percentile of tuition paid to Elite universities by Out of state students.

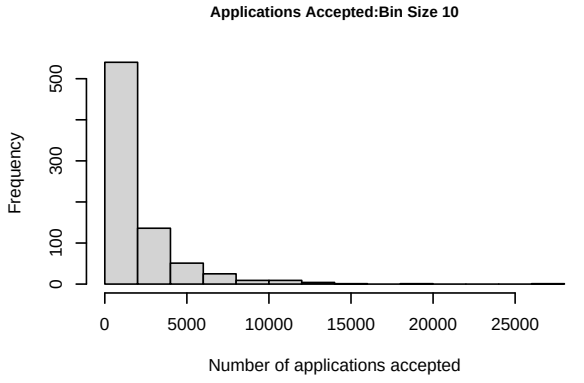


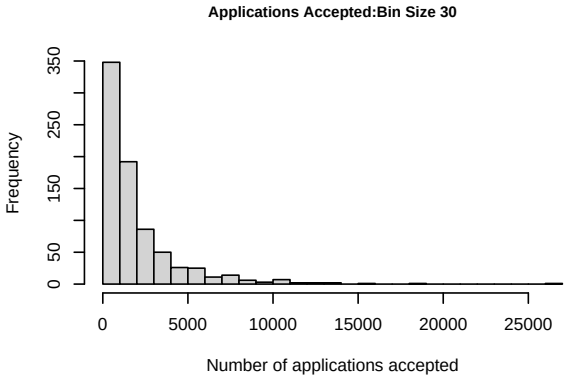
v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to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

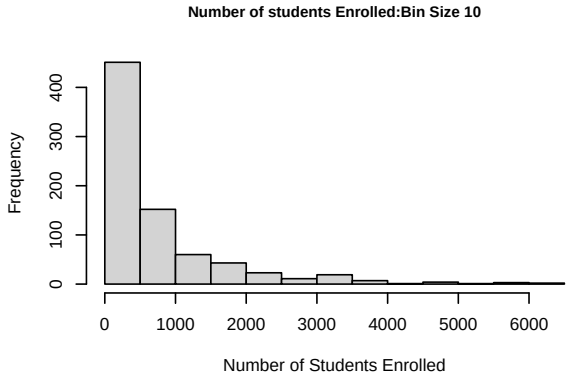
Histogram function is applied to few quantitative variables like Graduation rate, Number of Applications accepted, Number of students enrolled and Room/board costs. Applying hist() function from Graphics package on the Graduation Rate show that the data is kind of negatively skewed with 2 clusters.Highest number of Universities have a graduation rate between 60 -70%. Changing the Bin size to 30 did not change my opinion on the graduation rate of highest number of Universities.Histogram of Number of applications accepted is positively skewed with single cluster. The maximum number of applications accepted was about 525.Changing bin size to 30 did not change my interpretation.Histogram of Number of students enrolled is positively skewed with 3 clusters. The maximum number of applications accepted was about 430.Changing the bin size changed the maximum number of students enrolled to about 230 without change in the number of clusters.Histogram of Room/board costs is positively skewed with 2 clusters.Changing bin size to 30 didnot change much of the histogram.

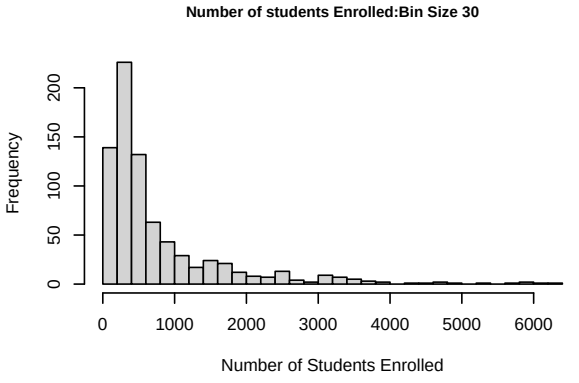

Graduation Rate:Bin Size 10
Frequency
0
50
100
150
20
40
60
80
100
120
Graduation Rate

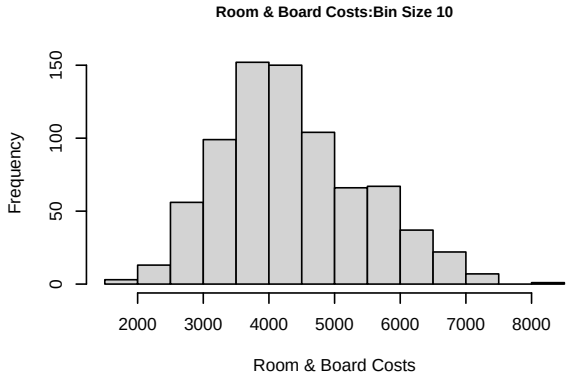

Graduation Rate:Bin Size 30
Frequency
0
20
40
60
80
20
40
60
80
100
120
Graduation Rate

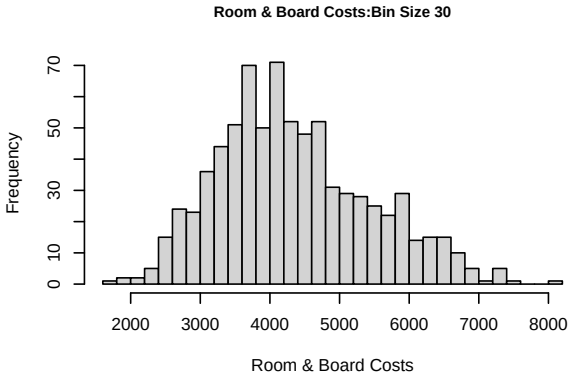

Applications Accepted:Bin Size 10
Frequency
0
100
300
500
0
5000
10000
15000
20000
25000
Number of applications accepted


Applications Accepted:Bin Size 30
Frequency
0
50
150
250
350
0
5000
10000
15000
20000
25000
Number of applications accepted

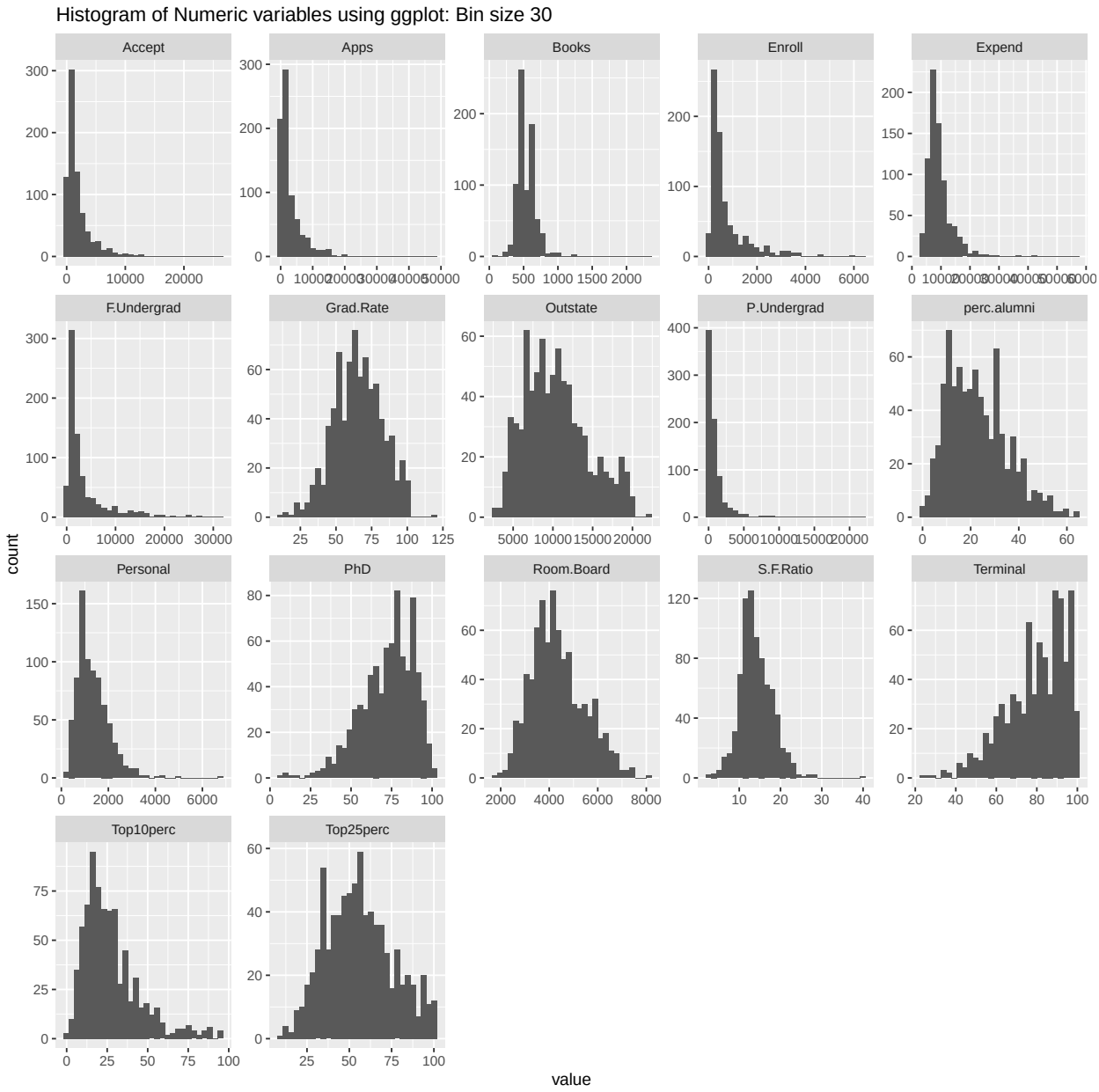

Number of students Enrolled:Bin Size 10
Frequency
0
100
200
300
400
0
1000
2000
3000
4000
5000
6000
Number of Students Enrolled

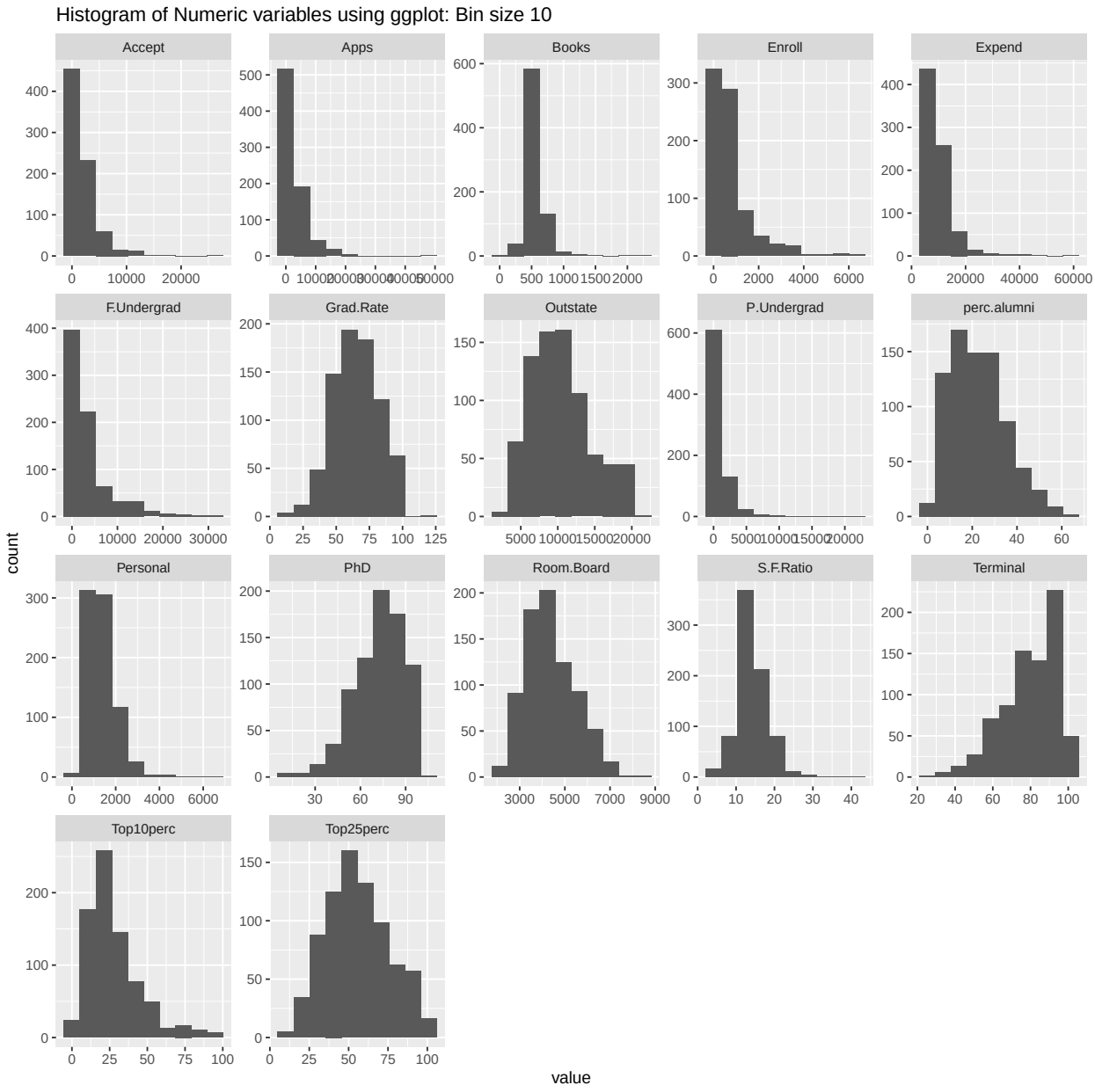

Number of students Enrolled:Bin Size 30
Frequency
0
50
100
150
200
0
1000
2000
3000
4000
5000
6000
Number of Students Enrolled


Room & Board Costs:Bin Size 10
Frequency
0
50
100
150
2000
3000
4000
5000
6000
7000
8000
Room & Board Costs


Room & Board Costs:Bin Size 30
Frequency
0
10
30
50
70
2000
3000
4000
5000
6000
7000
8000
Room & Board Costs

Similarly, Histograms of all the numerical variables are plotted with a bin size of 10 and 30 using ggplot.





Reference

- Blog BLOGR on *Quick plot of all variables*,July 13,2016

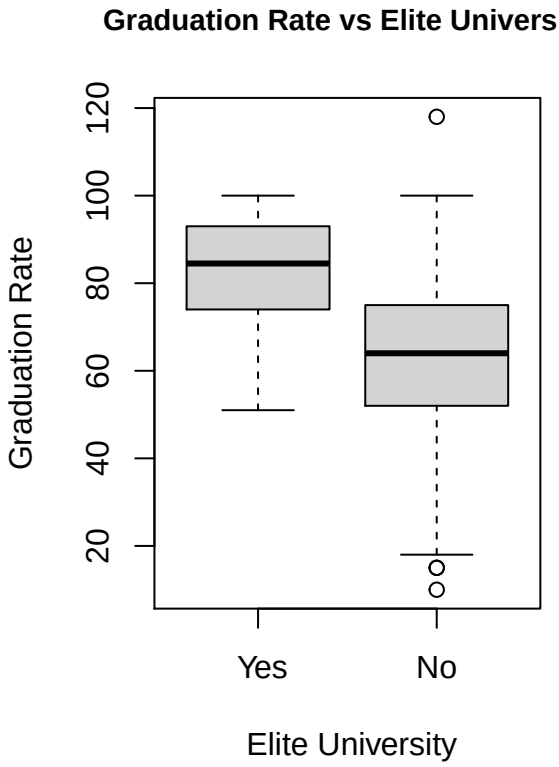
vi. Continue exploring the data, and provide a brief summary of what you discover.

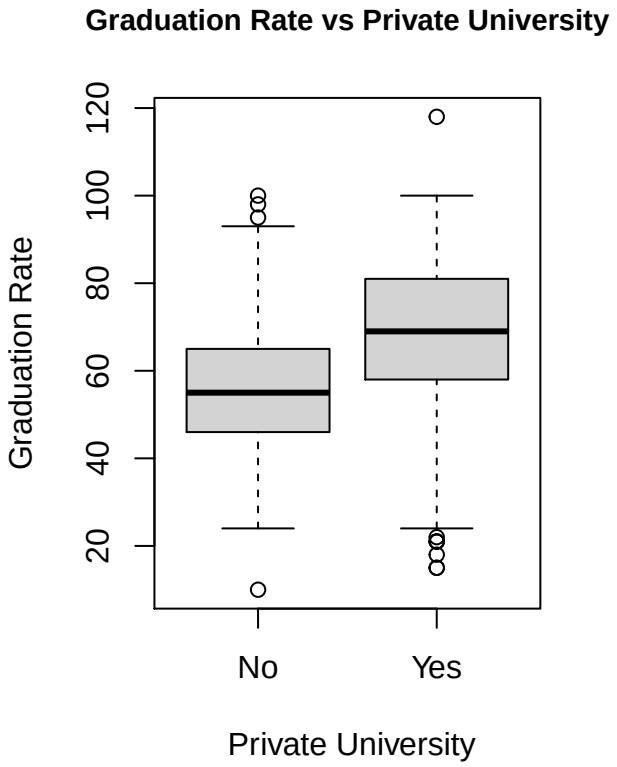
I have compared Graduation Rate, Room and board costs,Book costs,Personal costs and Instructional expenditure per student between students enrolled into Private Universities vs Public Universities and Elite Universities and Non-Elite Universities.

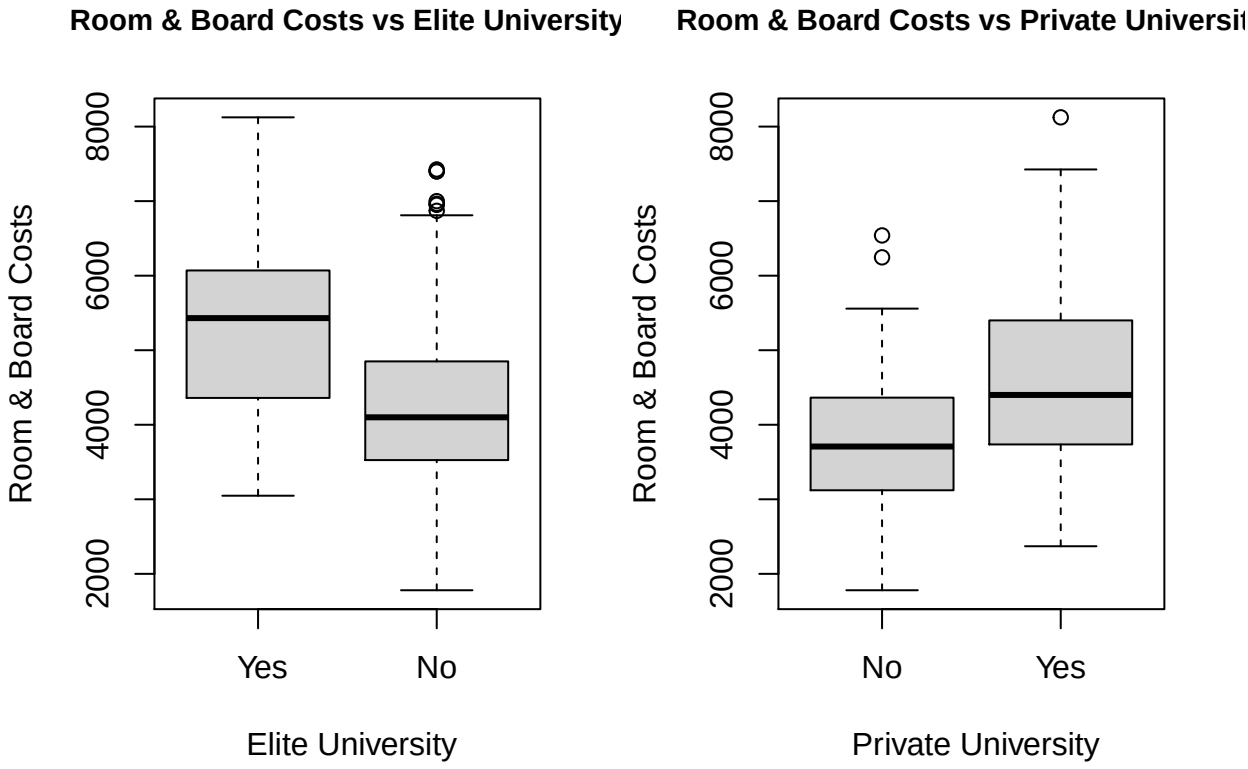
Observations:

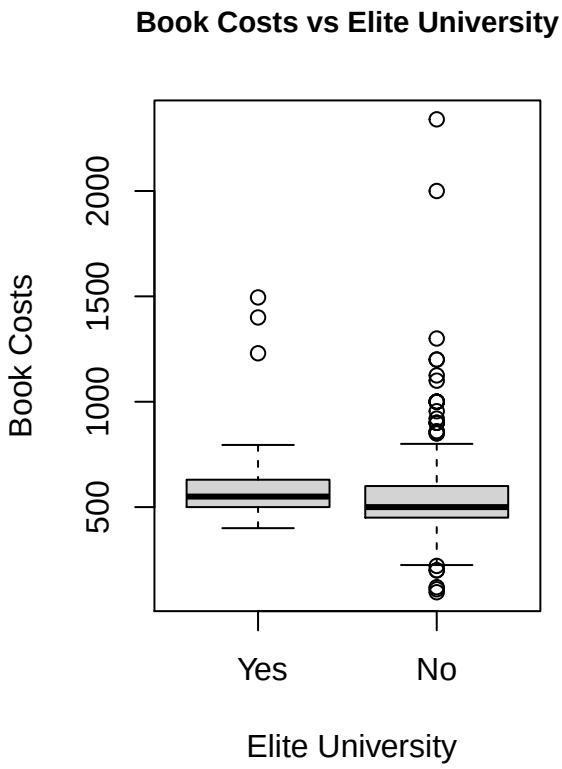
- Median Graduation rate of students enrolled in Elite and Private Universities is higher compared to students enrolled in Public and Non-Elite universities.

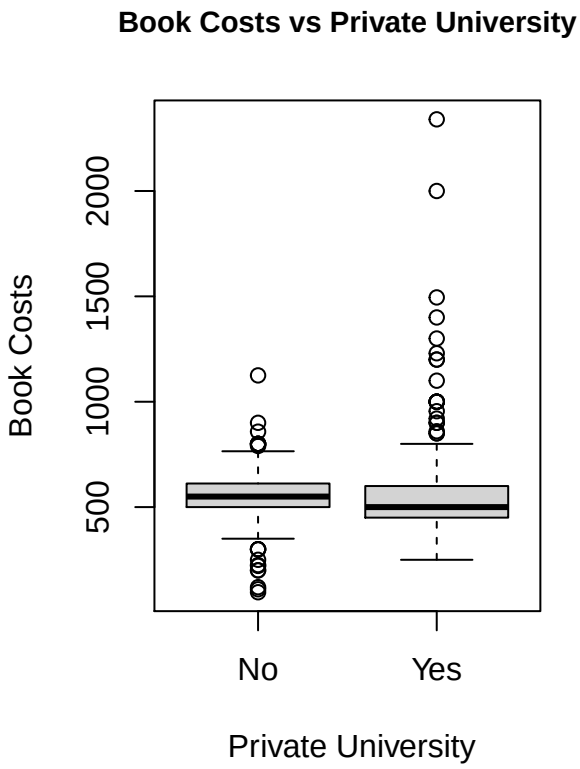
- Median Room & Board costs for students enrolled in Elite and Private Universities is higher compared to students enrolled in Public and Non-Elite universities.The 75th percentile costs incurred of students enrolled in Public and Non-Elite universities is close to the median value of costs incurred by students in Elite and Private Universities.
- Median Value of Book Costs for students enrolled in Elite and Public Universities is higher compared to students enrolled in Private and Non-Elite universities.However the costs incurred are are very close in both the cases.
- Median value of Personal Costs of students enrolled in Elite and Private Universities is lower compared to students enrolled in Public and Non-Elite universities.
- Median value of Instructional Expenditure per student of students enrolled in Elite and Private Universities is higher compared to students enrolled in Public and Non-Elite universities.

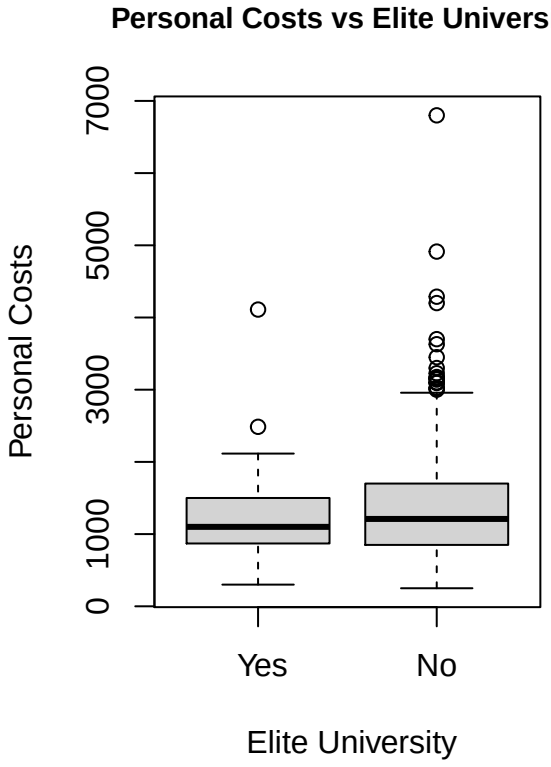


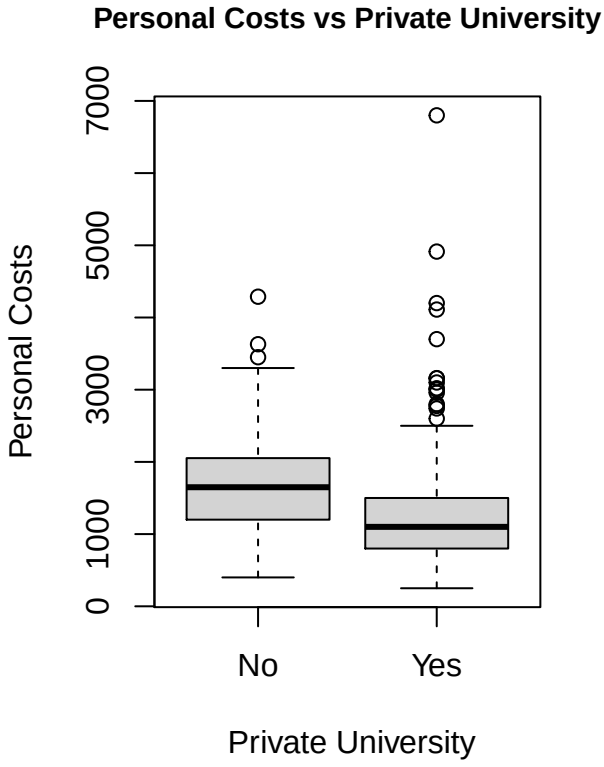


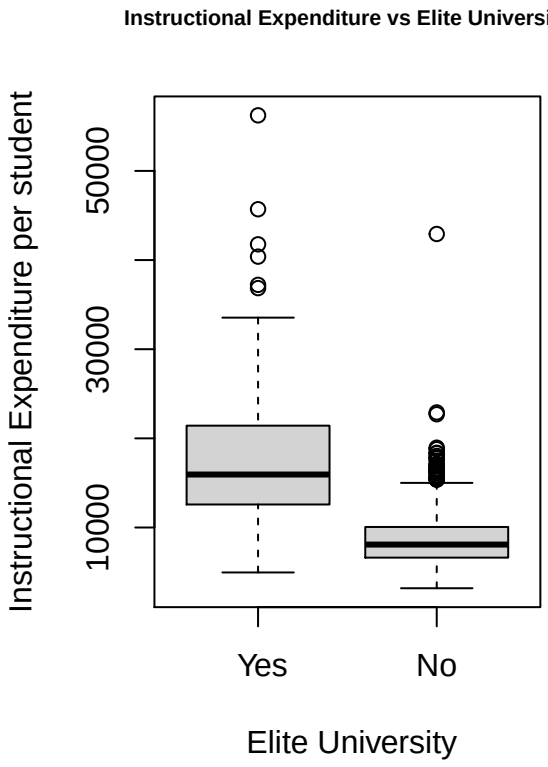

Room & Board Costs vs Elite University
Room & Board Costs
2000
4000
6000
8000
Yes
No
Elite University
Room & Board Costs vs Private Universit
Room & Board Costs
2000
4000
6000
8000
No
Yes
Private University

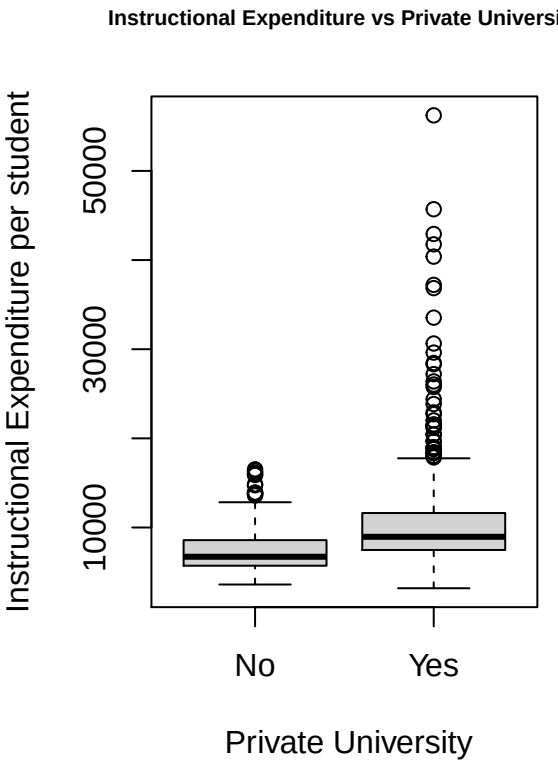

Book Costs vs Elite University
Book Costs
500
1000
1500
2000
Yes
No
Elite University


Book Costs vs Private University
Book Costs
500
1000
1500
2000
No
Yes
Private University


Personal Costs vs Elite University
Personal Costs
0
1000
3000
5000
7000
Yes
No
Elite University


Personal Costs vs Private University
Personal Costs
0
1000
3000
5000
7000
No
Yes
Private University


Instructional Expenditure vs Elite University
Instructional Expenditure per student
10000
30000
50000
Yes
No
Elite University


Instructional Expenditure vs Private University
Instructional Expenditure per student
10000
30000
50000
No
Yes
Private University