Homework 1

Modern Applied Statistics II

James Young

Packages Used: ggplot2, GGally, gridExtra, ISLR

Resources: StackOverflow

Collaborators: None

NOTE : This report is longer due to incorporating the questions from the book into the report, that way answers can be viewed in the context of the actual question.

Please do the following problems from the text book ISLR.

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.

ANSWER

This is a regression problem as CEO salary is a continuous numerical variable. This could become a classification problem if the salaries were binned into groups of high, medium, and low, for example. The prompt states we are interested in ***understanding*** *what factors affect CEO salary, so we are more interested in inference. n = 500 firms and p = 3 variables (firm profit, number of employees, and 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.

ANSWER

This is a classification problem of a binary nature, where the results can be ***success*** *or* ***failure****. We are more interested in prediction than inference in this project based on the wording "we wish to* ***know*** *whether it will be a success or failure". n = 20 products and p = 13 variables.*

(c) We are interest 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

This is a regression problem interested in prediction moreso than inference. n =52 and p = 3 variables

2. Question 2.4.4 pg 53

(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.

ANSWER

Application 1: Predicting whether someone is positive or negative for a type of cancer. The response would be the diagnosis of being positive or negative for cancer and the predictors can be the quantified levels of biomarkers such as metabolites, proteins, or circulating RNA/DNA found in blood serum. The goal of this application is prediction.

Application 2: Predicting loan defaults. Predict the binary outcome of whether someone will default or make good on their loan based on input variables such as FICO score, longest line of credit, payment history, delinquency history, etc. The goal is prediction.

Application 3: Determining what behavioral variables are strongly predictive of heart attack. The goal of this application is inference, where behaviors that increase heart attack probability will be mitigated. While being strongly predictive does not necissarily make a variable "causitive", the inferences made can be joined with domain expertise to devise further studies. The benefit of this inferential focus rather than predictive, is the ability to try to intervene and prevent predictor variables associated with heart attacks from occuring, and in turn limit the amount of heart attacks.

(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.

ANSWER

Application 1: Predicting the Gestational Age of a baby. This was a project I worked on this summer from "DREAM challenges" (similar to kaggle, but strictly biological). Given transcriptome data, which is the expression levels of genes, and the respective gestational age of the patients, fit a model that predicts gestational age. The goal of this project was prediction. There were over 40,000 genes so this was a highly dimensional problem.

Application 2: Predicting the % change in the S&P 500 ETF "SPY". The response is the % change in price of "SPY". The predictor variables can be price of the previous day, 50 day moving average, 200 day moving average, MACD, Bollinger Band score, and RSI. All of the input predictors are derivatives of the trailing price data to predict the future price. The goal of this is prediction.

Application 3: Predicting the response of bioethanol yield based on predictors of sugar concentration, molecular weight of sugar, reactor temperature, microbe concentration, and magnesium concentration. The goal of this project is prediction to optimize bioethanol yield.

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

ANSWER

Application 1: Customer Segmentation. Variables associated with customers could be clustered to see if distinct sub-groups appear which could inform marketing efforts.

Application 2: Basketball player archetype determination based on playing time, scoring style, rebounds, steals, etc. This was actually done here https://github.com/klarsen1/NBA_RANKINGS and further used as input for a model that tried to predict basketball game outcomes.

Application 3: Clustering iris flowers based on their petal/sepal width/length to make groups.

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?

Answer

Parametric approaches make assumptions about the population that data comes from while non-parametric approaches do not make these assumptions. Some of the assumptions parametric tests make are that the data is normally distributed and independent, for example. The amount of data needed for a parametric model is smaller than non-parametric model. A non-parametric model may be better at fitting non-linear data structures than a parametric model.

4. Question 2.4.8 pg 54-55

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.

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.

I have called the college data from the ISLR package rather than using read.csv.

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

There is no need to turn column 1 into row.names becasue the data called from ISLR is already in this format. The factor "Private" is the first column.

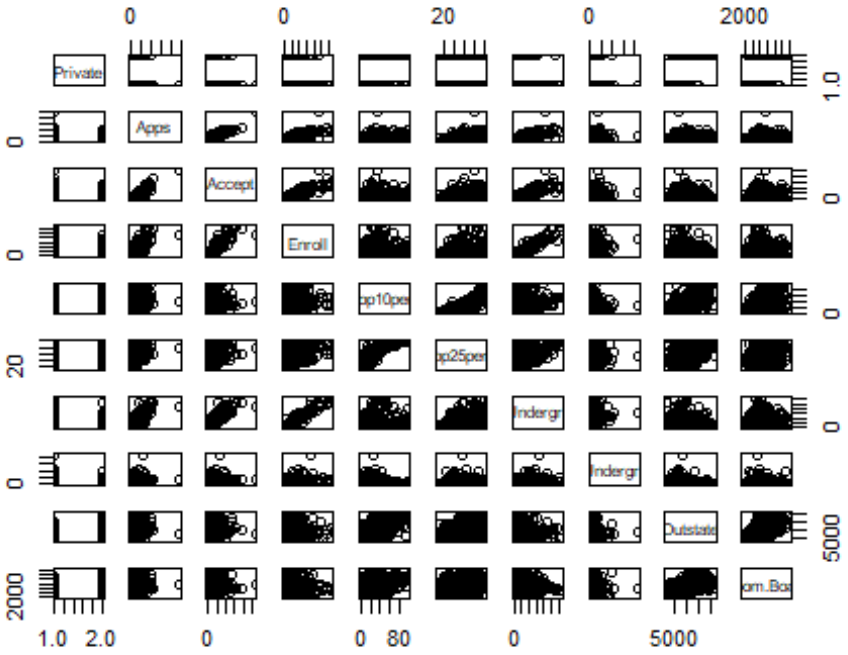
i. Use the summary() function to produce a numerical summary of the variables in the data set.

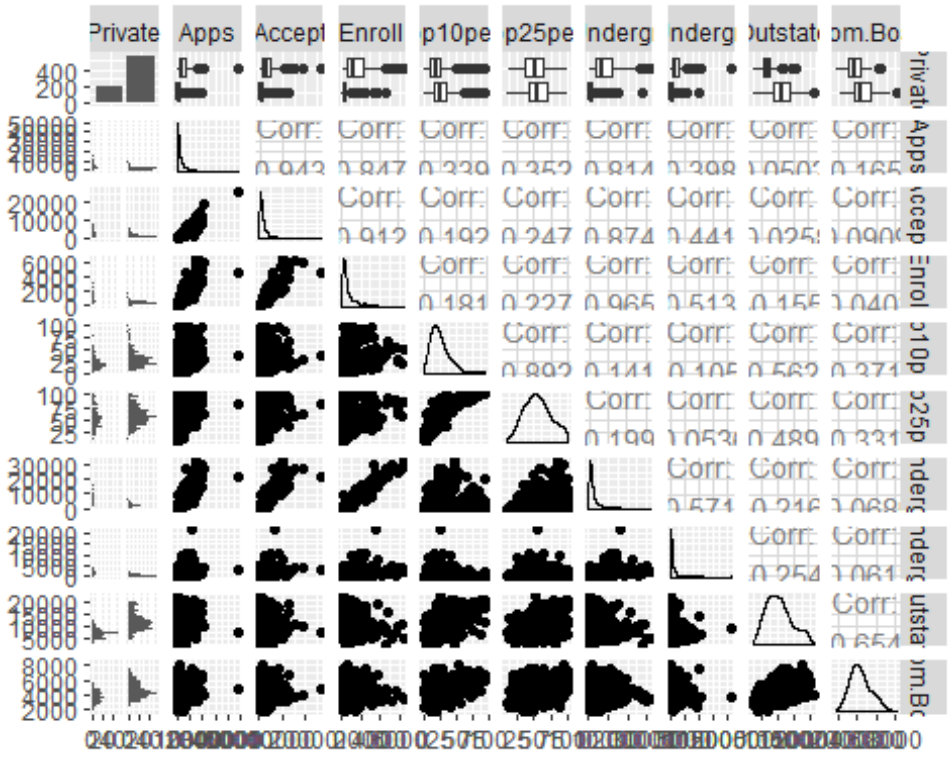
The summary is given below.

```
##  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].

Below I used the pairs() function as requested by the problem as well as the ggpairs() function to fulfill the ggplot requirement. Both plots are cumbersome due to having 10 variables included. However, the book did say to include the first 10 variables. We can see "Enroll" and "Accept" both have some correlation with "F.undergrad". There is also a prominent correlation between "Apps" and "Accept".

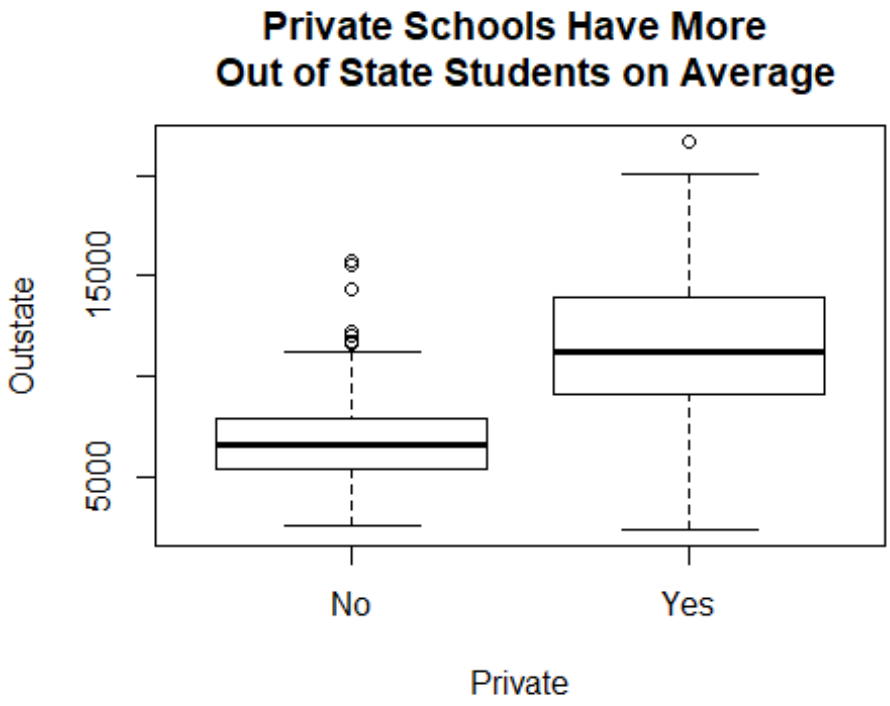


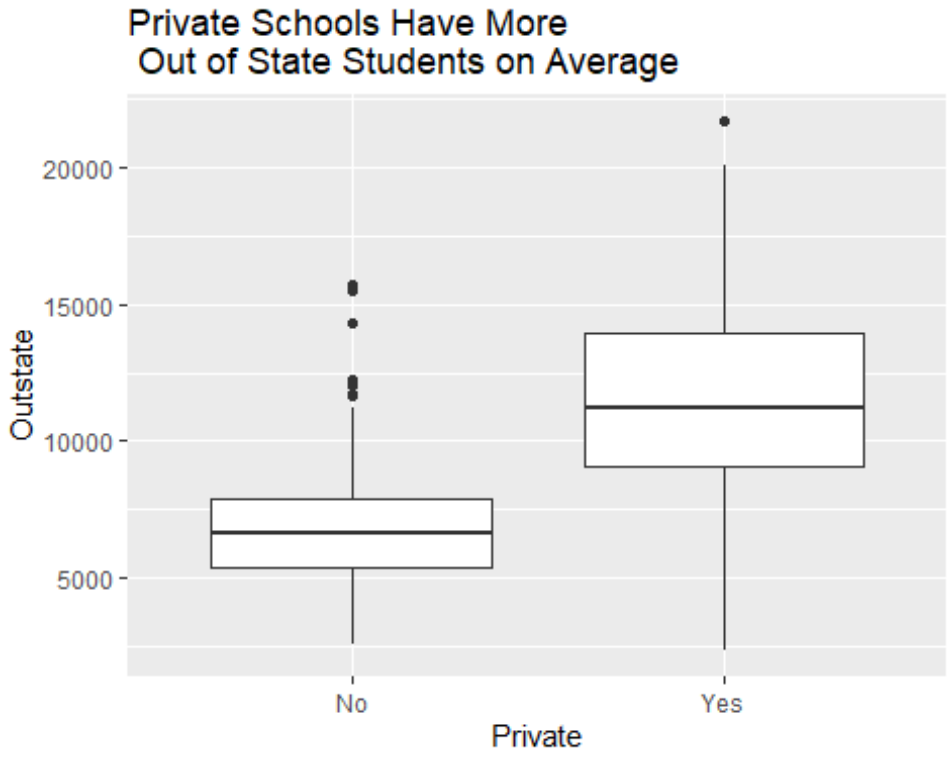


iii. Use the plot() function to produce side-by-side boxplots of Outstate versus Private. iv. Create a new qualitative variable, called Elite, bybinning the Top10perc variable. We are going to divide universities into two groups based on whether or not the

proportion of students coming from the top 10% of their high school classes exceeds 50%.

Below, the plots demonstrate the relationship between how many out of state students a school has and it's status as private or not. Private schools on average have more out of state students than non-private schools.





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

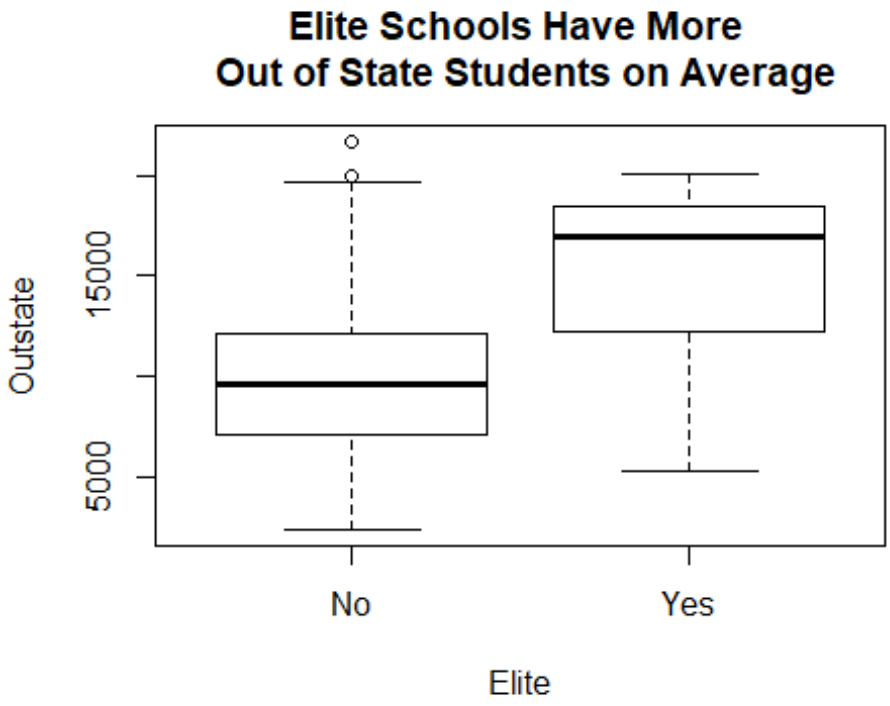
50%.Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

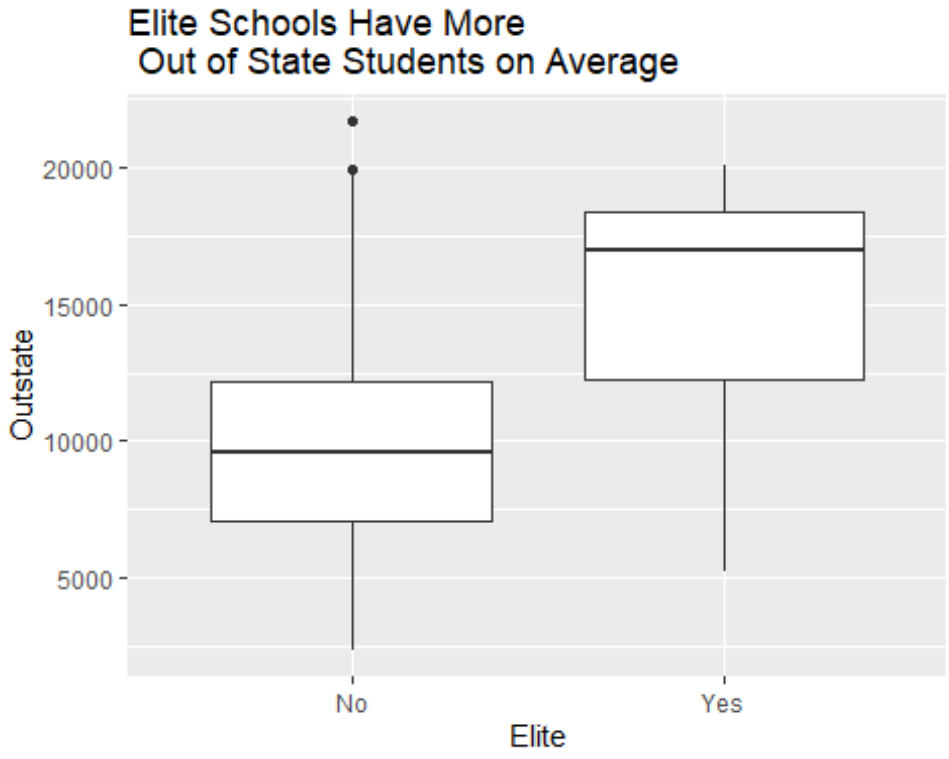
Below, the summary shows that there are approximately 10 times as many schools that are not considered elite than there are schools that fit the elite criteria.

```
## Elite Colleges Summary

##  No Yes
## 699  78
```

Below, the plots demonstrate the relationship between how many out of state students a school has and it's status as elite or not. Elite schools on average have more out of state students than non-elite schools.

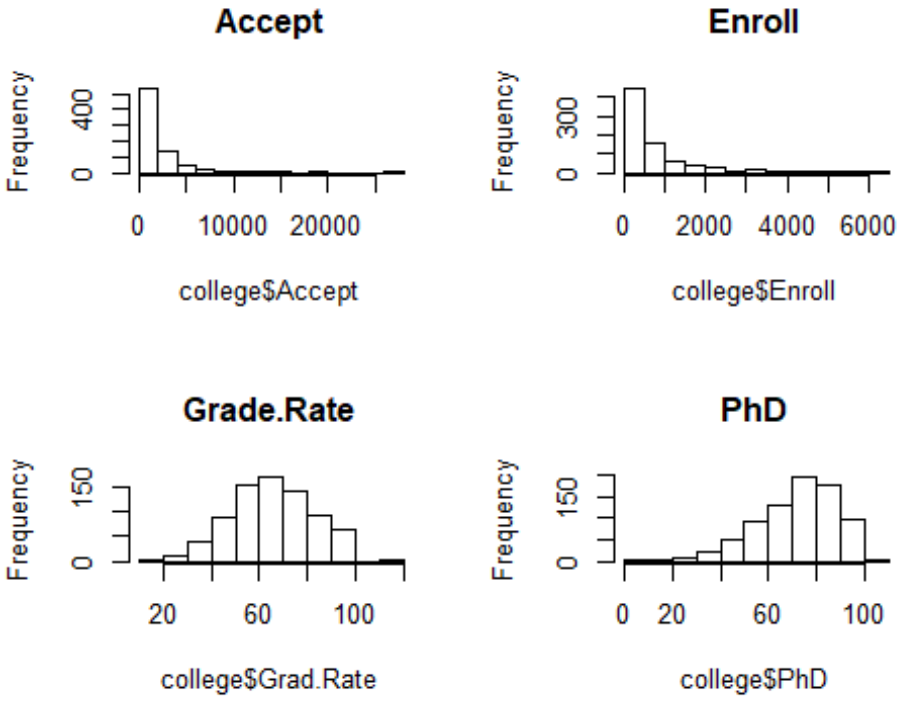


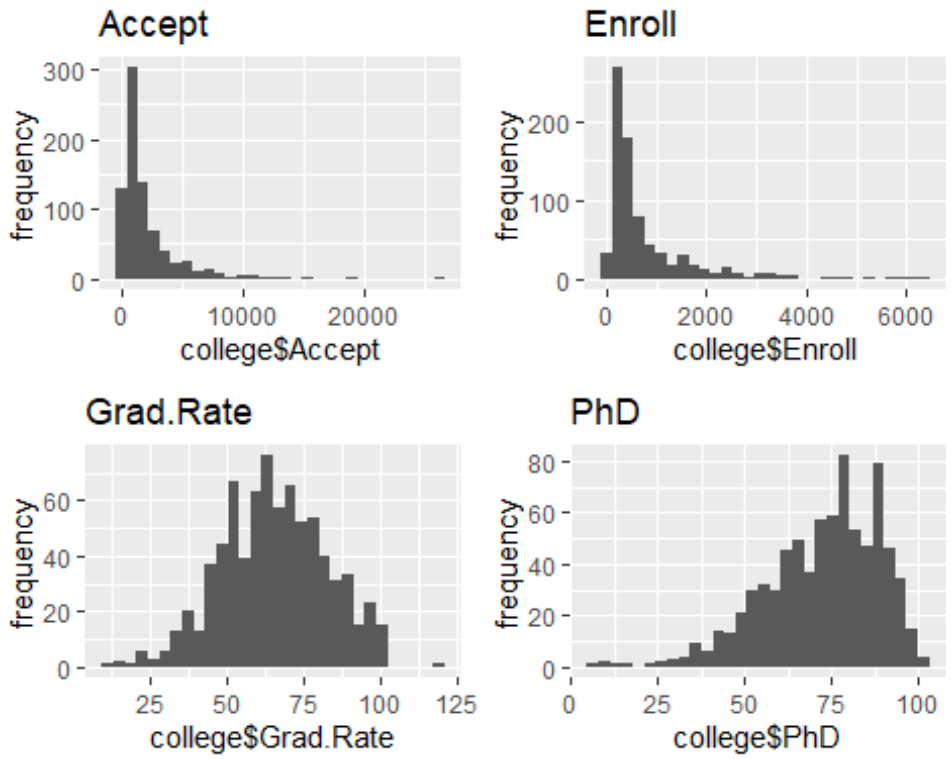


v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made

simultaneously. Modifying the arguments to this function will divide the screen in other ways.

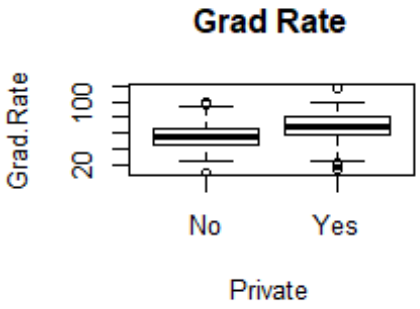
Below, the requested histograms were made in base r and ggplot. We can see acceptance and enrollment are not normal distributions, with a small portion of schools getting a much higher amount acceptance and enrollment. Grad rate and the PhD status of teachers are closer to a normal distribution.

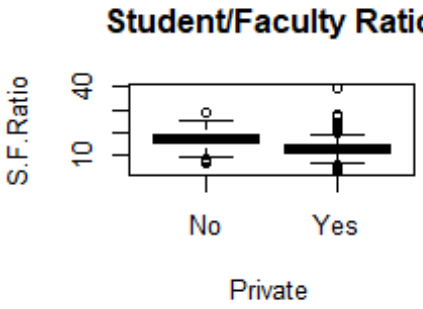


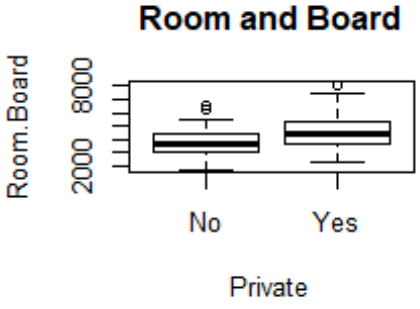


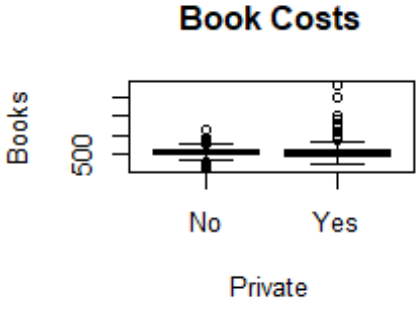
vi. Continue exploring the data, and provide a brief summary of what you discover.

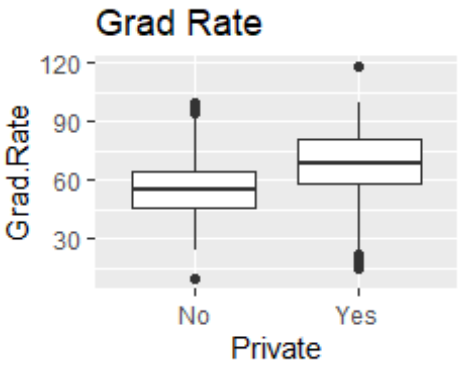
Below, we can see that private universities have (on average) a higher reported grad rate (one school even reported 118%, good for them!), lower student/faculty ratios, higher room and board costs, and approximately equivalent book costs.

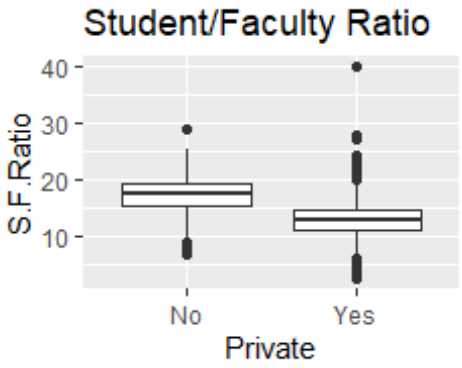

Grad Rate
Grad.Rate
20
100
No
Yes
Private

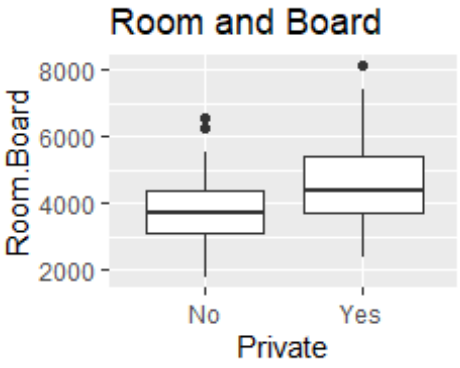

Student/Faculty Ratio
S.F.Ratio
10
40
No
Yes
Private

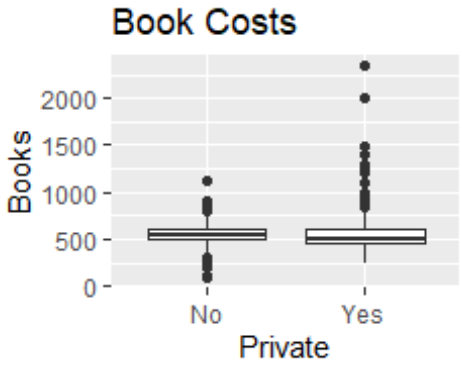

Room and Board
Room.Board
2000
8000
No
Yes
Private


Book Costs
Books
500
No
Yes
Private


Grad Rate
Grad.Rate
120
90
60
30
No
Yes
Private


Student/Faculty Ratio
S.F.Ratio
40
30
20
10
No
Yes
Private


Room and Board
Room.Board
8000
6000
4000
2000
No
Yes
Private


Book Costs
Books
2000
1500
1000
500
0
No
Yes
Private