Homework 1

Alex Soupir

January 18, 2020

Packages: ISLR

Collaborators:

Answer all questions specified on the problem and include a discussion on how your results answered/addressed the question.

Submit your **.rmd** file with the knitted **PDF** (or knitted Word Document saved as a PDF). If you are having trouble with .rmd, let us know and we will help you, but both the .rmd and the PDF are required.

This file can be used as a skeleton document for your code/write up. Please follow the instructions found under Content for Formatting and Guidelines. No code should be in your PDF write-up unless stated otherwise.

For any question asking for plots/graphs, please do as the question asks as well as do the same but using the respective commands in the GGPLOT2 library. (So if the question asks for one plot, your results should have two plots. One produced using the given R-function and one produced from the GGPLOT2 equivalent). This doesn't apply to questions that don't specifically ask for a plot, however I still would encourage you to produce both.

You do not need to include the above statements.

Please do the following problems from the text book ISLR.

1. Question 2.4.2 pg 52

Explain whether each scenario is a classificaiton 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 recorded profit, number of employees, industry and the CEO salary. We are interested in understanding which factors affect CEO salary.

- Regression
- Inference
- n = 500
- p = profit, 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.

- Classification
- Prediction
- n = 20
- p = price charged for the product, marketing budget, competition price, and ten other variables

c) We are interesting in predicting the % change in the US dollar 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 dollar, the % change in the US market, the % change in the British market, and the % change in the German market.

- Regression

- Prediction
- n = weekly 2012 data, 52 time points
- p = % change in the US stock marker, % change in the British market, % 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.

- **Whether someone has lung cancer or not; response is yes or no for cancer, predictors could be something as simple as just a radiology scan or things like family history, smoking, how much if they do smoke, age, maybe even location they live because of pollution; this would be prediction application because we want to know whether someone either has cancer due to the radiology scan or if they are at risk for having lung cancer.**
- **whether a plant is colonized by fungi or not; response would be yes or no for colonization, predictors could be something like rna sequencing results, a detailed scan of a certain plant tissue such as root or leaf, if looking for a benefical or harmful root fungi a predictor could just be root length and compare it to the control, and the goal of this application would be prediction of colonization based on high throughput methods rather than having to look at each sample individually.**
- **Another classification problem might be whether a plant is benefitting or not; the response would be whether or not it grew or performed better than a control set of plants, predictors could be the the abundance of specific microbes witin the plant discovered by microbial profiling from sequencing; the goal of this could be either prediction or inference depending how one wants to capitalize on the data and move forward, prediction would be just looking to see if a plant or crop would perform better than typical plants or inference to find out which microbe is potentially contributing to the plant health the most.**

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.

- **the amount the stock market is going to move; the response would be the % change of the stock market index, predictors could be % change of other indexes that are related, maybe twitter mentions and other words that are also found in the same tweets, possibly news mentions of the company, mentions of other world news whether like good things or bad things, market volume, past earnings, time of year, maybe even things like the state of the moon (since apparently that was correlated with the number of people going to the ER?), lag period for the last few days or even weeks for the market movement; this could be both inference and prediction because maybe there is something that is contributing the most to the movement of the stock market (might even reveal some highly correlated predictors like in the video Camy had done on P-values) and of course the main goal would be determining whether or not a stock is good to buy or better to sell.**
- **probability of an accident?; the response variable collected would be yes or no if an accident happened, predictors would be things like images of the surrounding environment which could give information like number of vehicles around, vehicle speed, maybe vehicle direction and position of the sun, weather condition (raining, snowing, sunny, cloudy, etc), time of day but that might be too related to sun position but maybe better?, temperature, population density?, number of lanes on the road, business district or residential, intersection or just 2 way road, maybe distance from home since a lot of accidents happen closer to home; this would be prediction because they would want to know the probability of a vehicle getting into an accident but you may be able to infer which predictor would contribute the most to the vehicle getting into an accident.**

- **the grade a student will get on a final exam; the response would be previous students final exam score for the same course, and predictors could be student credit load, previous assignment scores, time spent studying for the course, time of day the exam was at (if at different times each semster it could be an important predictor), amount of sleep the night before, possibly teacher, maybe the semester which the course was taken; this would be a prediction problem because the main goal is to just determine what the final exam score would be for a student.**

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

- **clustering people together based on shopping trends to see if people that shop online have similar purchasing habits to those that shop at physical locations, could also lead to target advertizing**
- **clustering single cell RNA-Seq results which could show how similar 2 different treatments are**
- **allowing unsupervised learning to try and determine whether there are distinct populations in microbiome samples**

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?

- **parametric learning makes assumptions about the data such as a linear relationship where non-parametric leanring doens't make doens't make direct assumptions about the function.**
- **parametric advantage is that it's generally easier to estimate the parameters but has the disadvantage that the model chosen, like in point one stating linear relationship, is potentially not a match for the actual data which then could make some predictions incredibly incorrect.**
- **nonparametric has the advantage that it is more flexible and can fit a wider range of predictions, but has the downside where more data is needed otherwise predictions could be way off.**

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. 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.

- **just using the *college* data located in the ISLR package.**

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:

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

You should see that there is now a row.names column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

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

Now you should see that the first data column is Private . Note that another column labeled row.names now appears before the Private column. However, this is not a data column but rather the name that R is giving to each row.

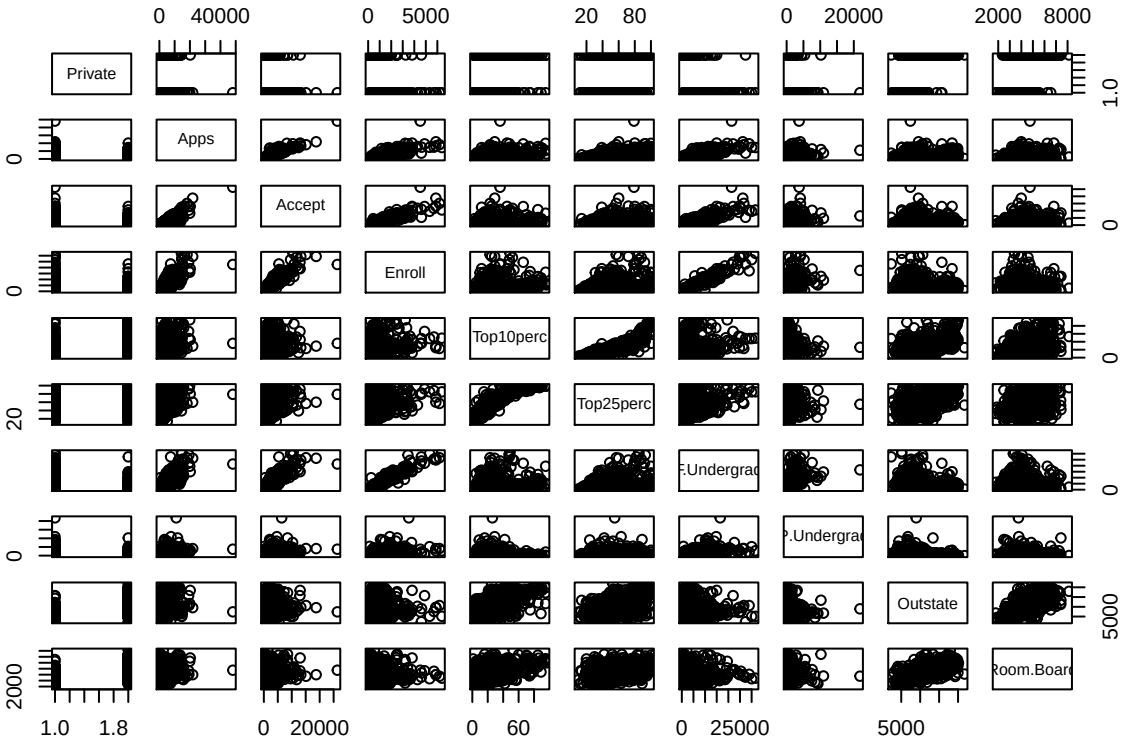
- **Don't have to do this with the data from the package because it is already read in with *Private* as the first column heading.**

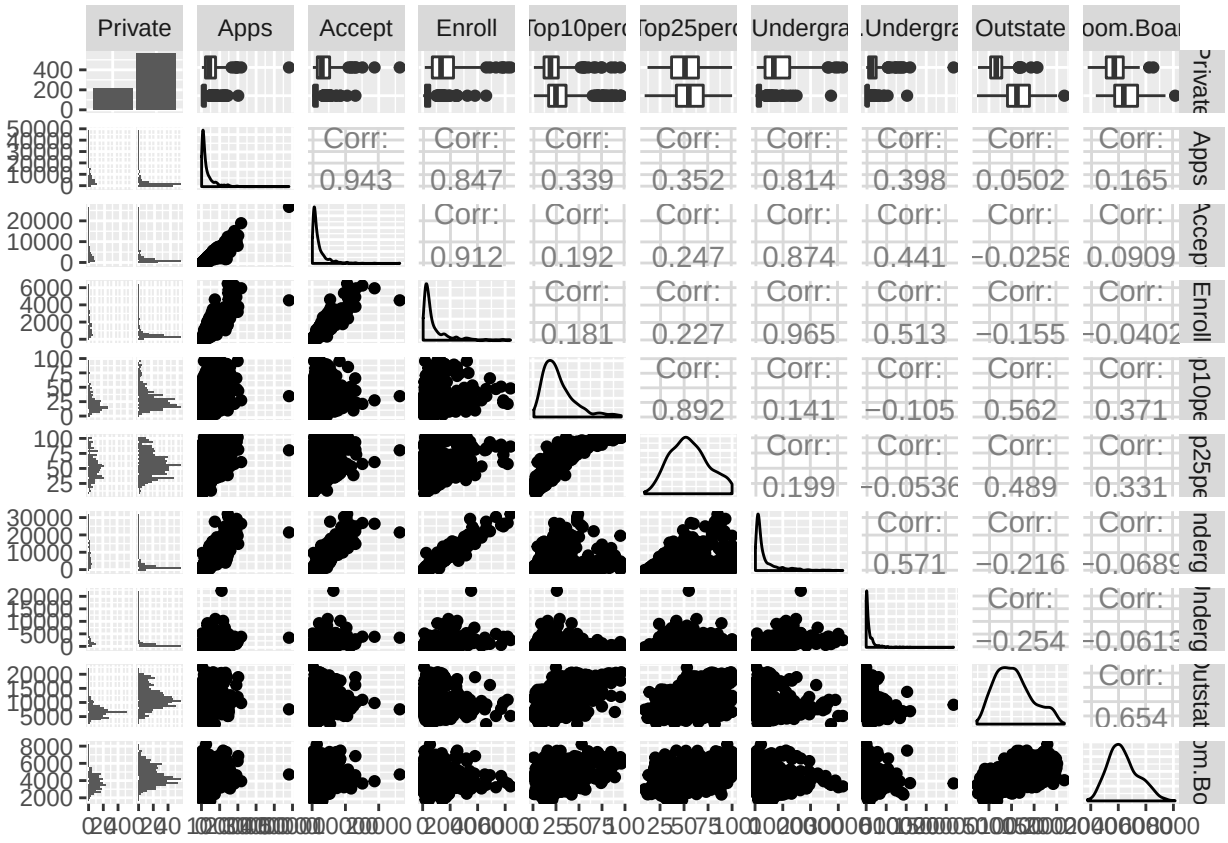
i. Use the summary() function to produce a numerical summary of the variables in the data set.

```
##  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
```

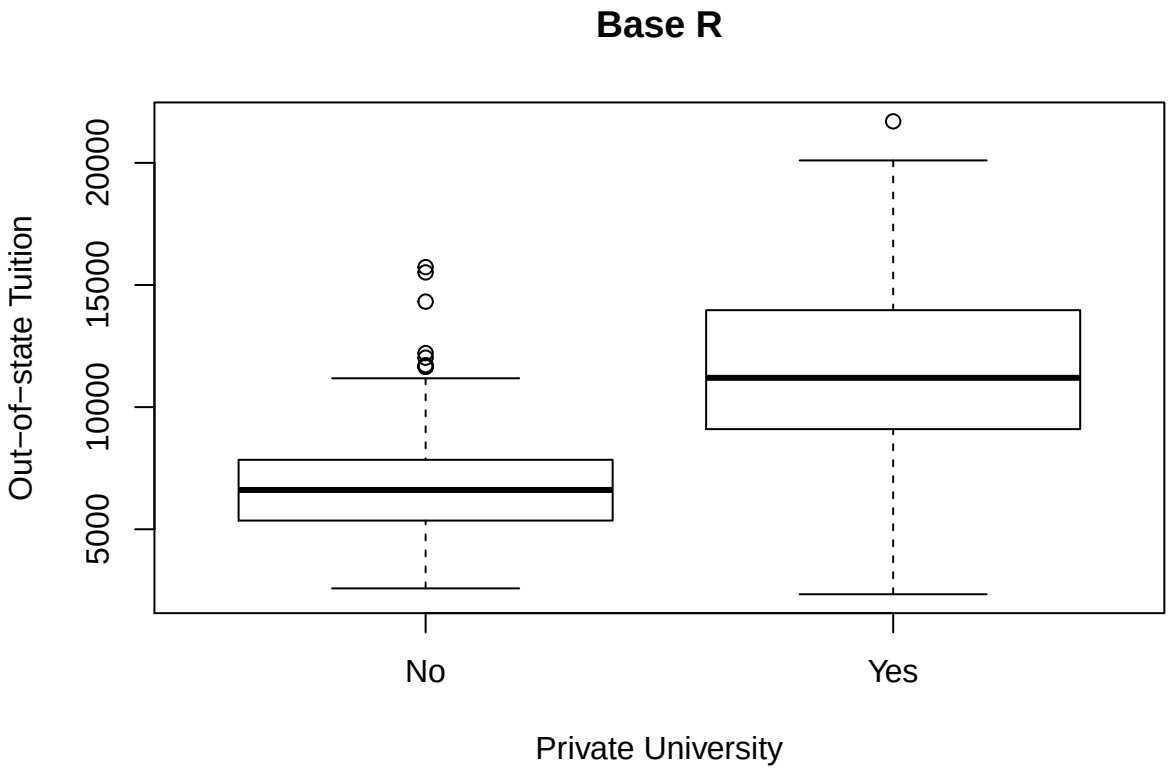
```
##  Mean   : 79.7   Mean   :14.09   Mean   :22.74   Mean   : 9660
##  3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00   3rd Qu.:10830
##  Max.   :100.0   Max.   :39.80   Max.   :64.00   Max.   :56233
##    Grad.Rate
##  Min.   : 10.00
##  1st Qu.: 53.00
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

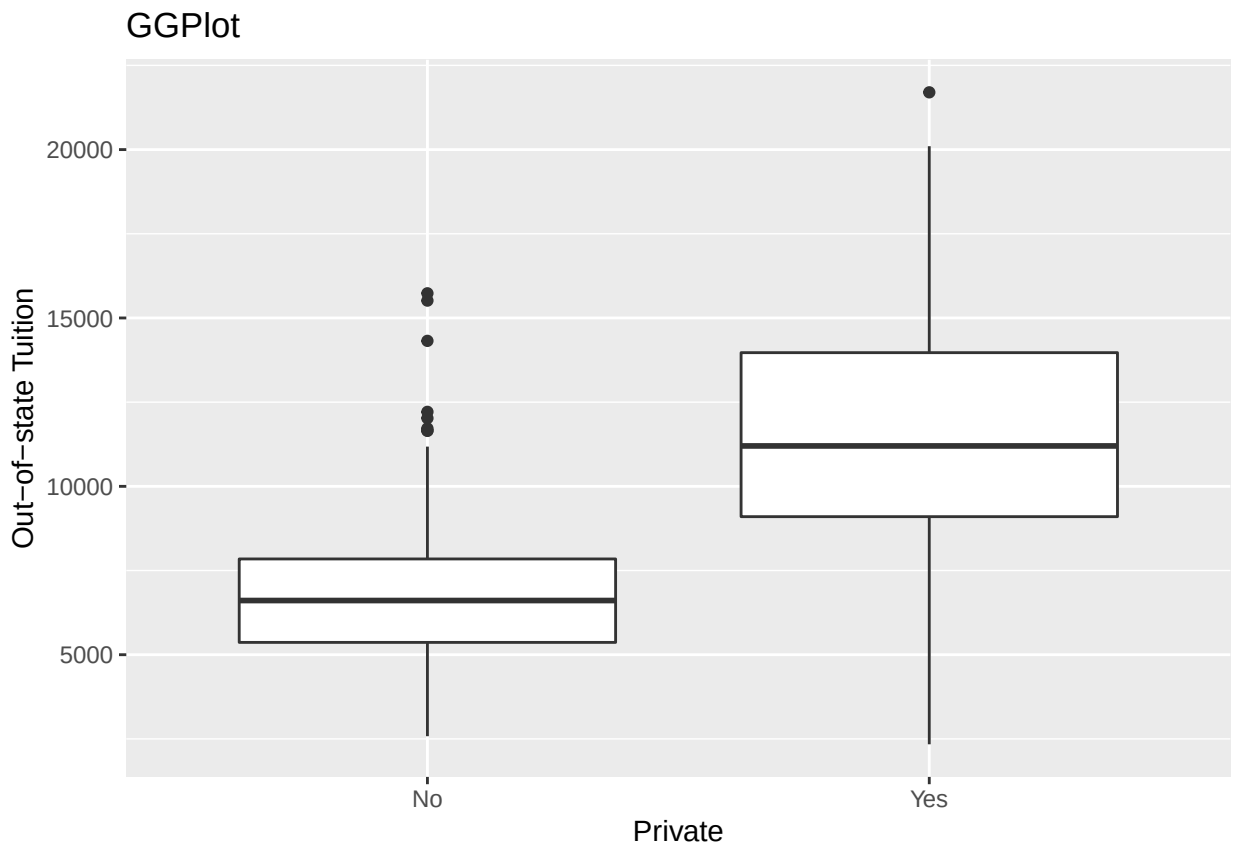
ii. Use the pairs() function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix A using A[,1:10].





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


Base R
Out−of−state Tuition
5000
10000
15000
20000
No
Yes
Private University

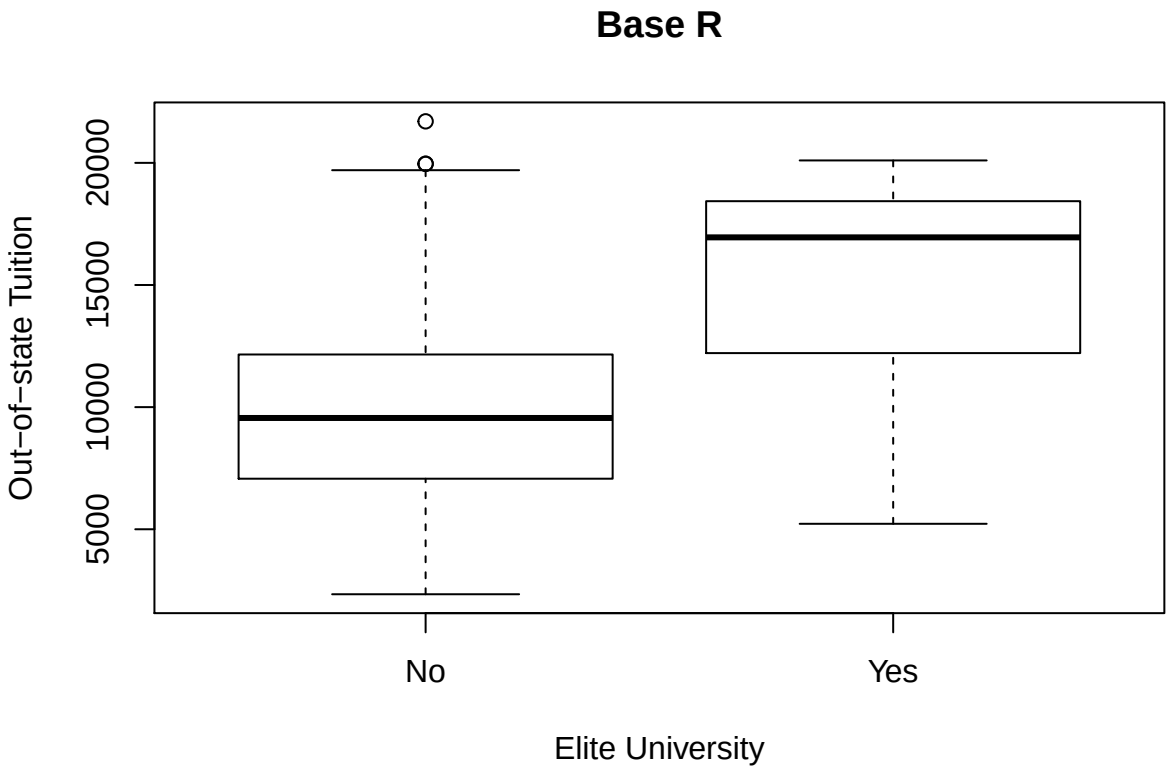


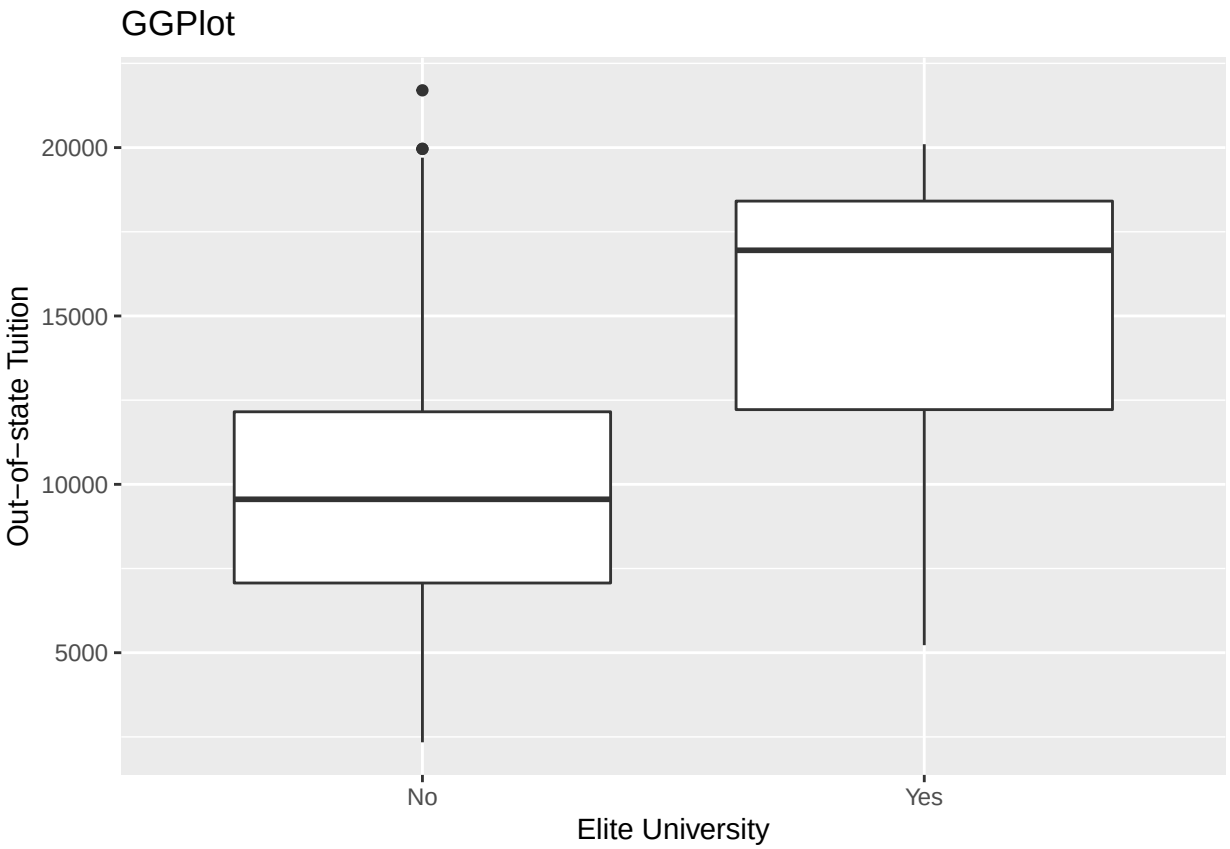
iv. Create a new qualitative variable, called Elite , by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

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

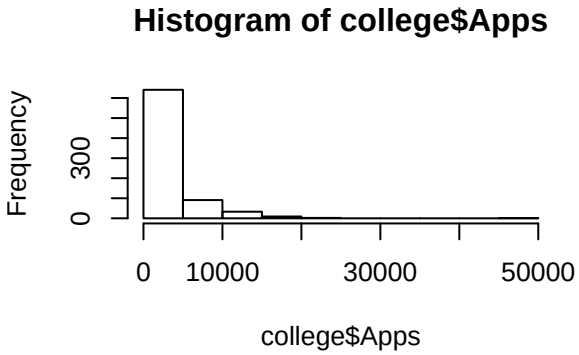
Use the summary() function to see how many elite universities there are. Now use the plot() function to produce side-by-side boxplots of Outstate versus Elite.

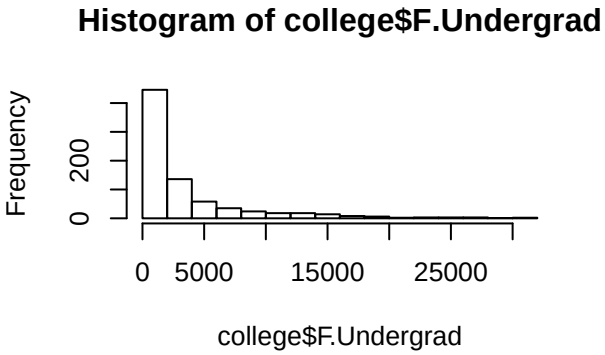
```
## Elite Universities

##  No Yes
## 699  78
```

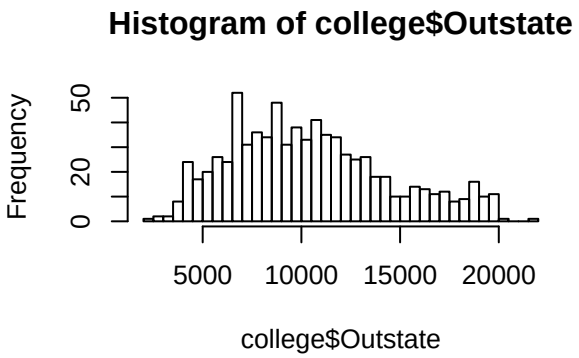

Base R
Out−of−state Tuition
5000
10000
15000
20000
No
Yes
Elite University

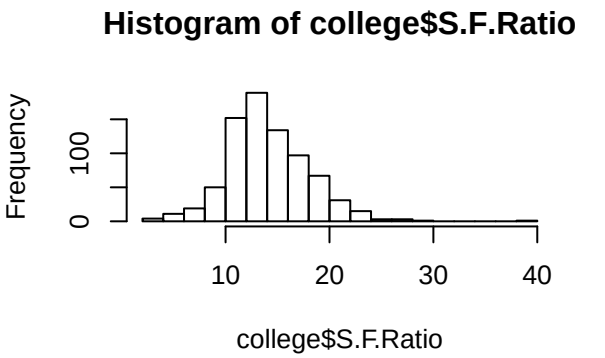


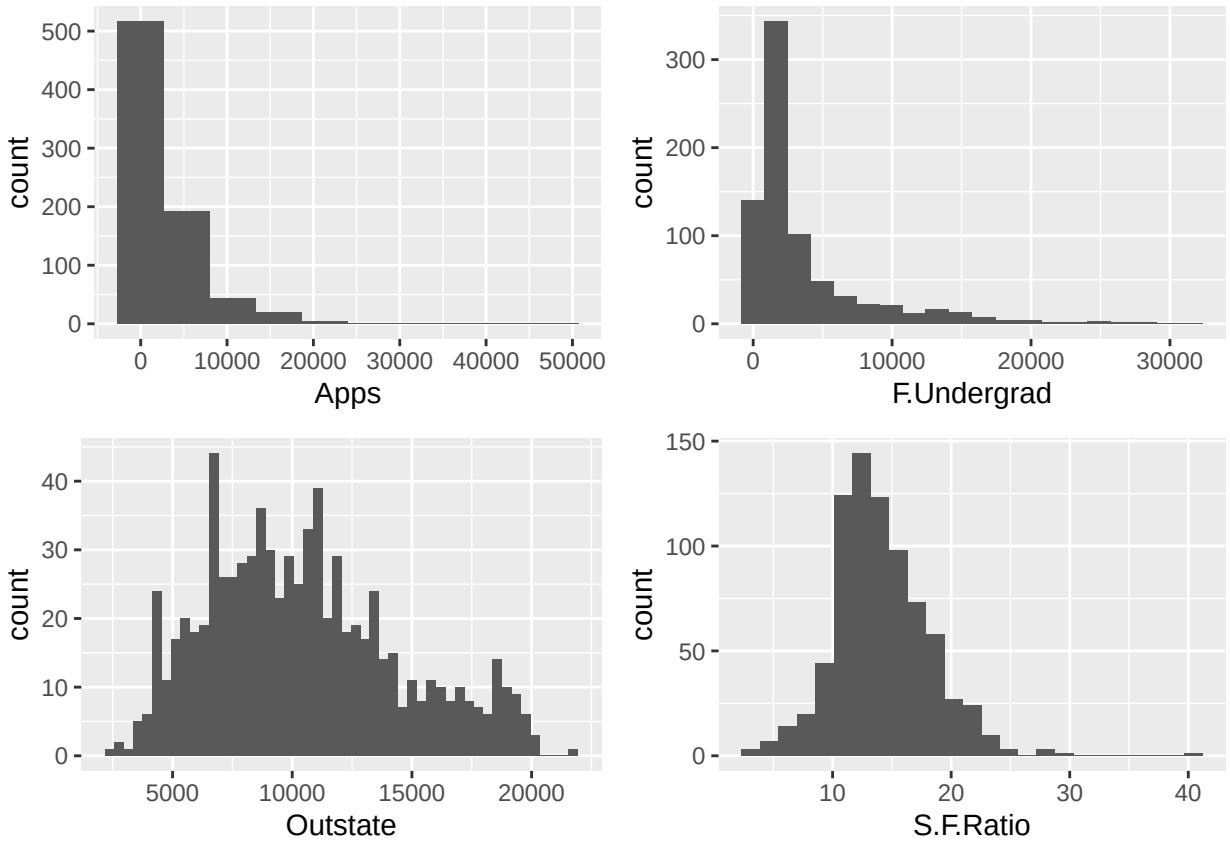
v. Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables. You may find the command par(mfrow=c(2,2)) useful: it will divide the print window into four regions so that four plots can be made simultaneously. Modifying the arguments to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


Histogram of college$Apps
Frequency
0
300
0
10000
30000
50000
college$Apps

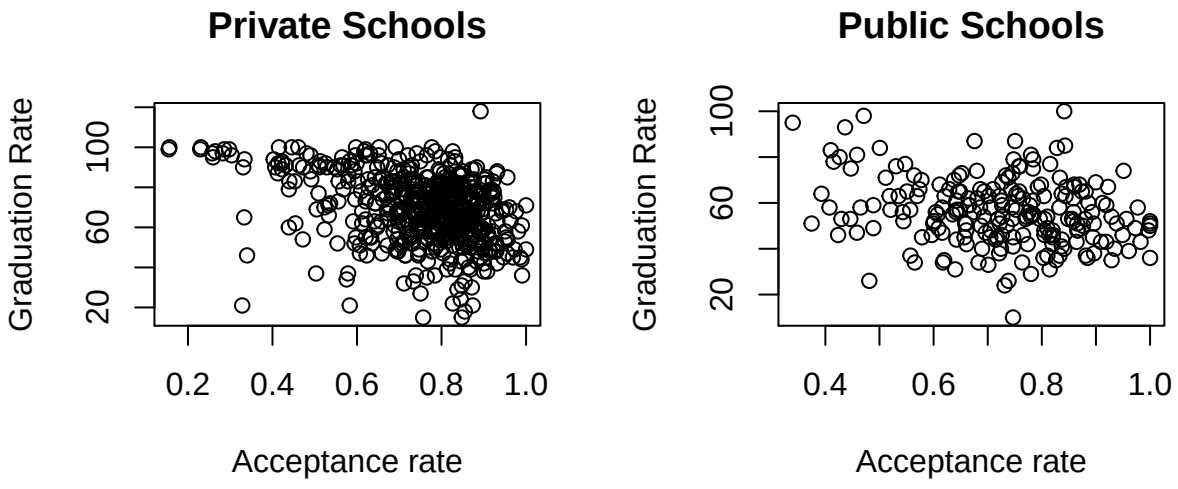

Histogram of college$F.Undergrad
Frequency
0
200
0
5000
15000
25000
college$F.Undergrad

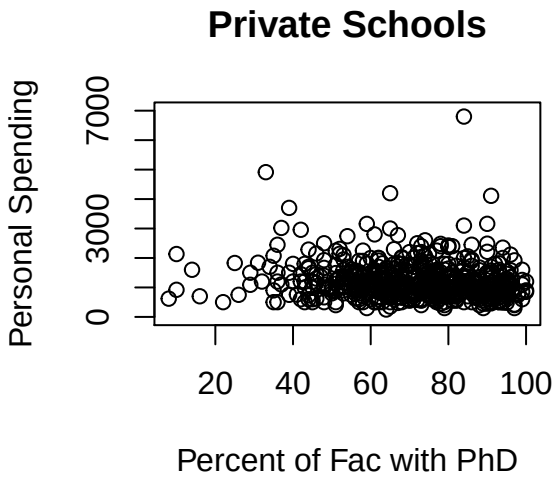

Histogram of college$Outstate
Frequency
0
20
50
5000
10000
15000
20000
college$Outstate

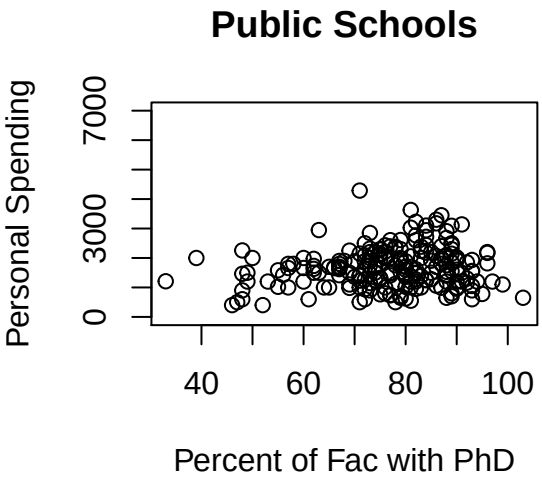

Histogram of college$S.F.Ratio
Frequency
0
100
10
20
30
40
college$S.F.Ratio

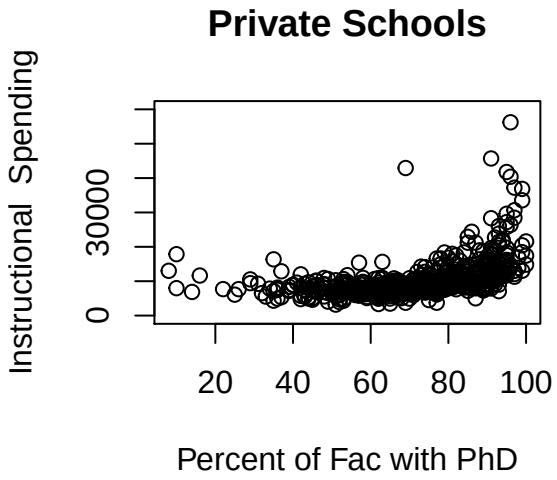


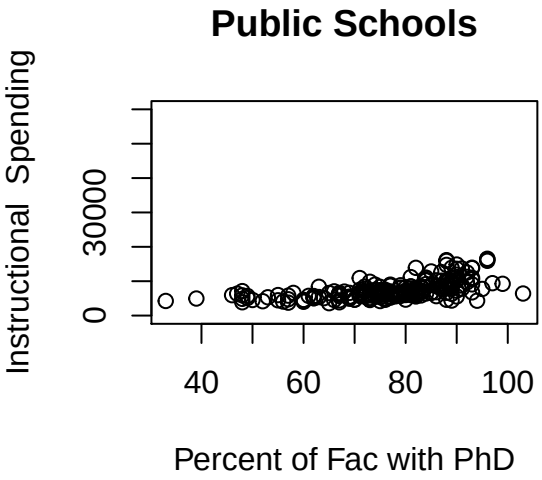
vi. Continue exploring the data, and provide a brief summary of what you discover.











I thought it would be interesting to explore how the differences of private and public schools have on other metrics. First I wanted to see how the acceptance rate and graduation rates different between the two. Private schools had shown a trend where, to an extent, lower acceptance rates had higher graduation rates. There was a point that did show a 118% graduation rate which was interesting to see; this school is Cazenovia College. Public schools show a slightly similar trend however its less visible and the points seem more 'random'.

Next I was interested in looking at the percent of faculty that had PhDs and the number of personal spending and also instructional expenditure per student. For personal expenditures, private schools seemed to be more focused between $500 and $1500 and public schools weren't as focused and seemed to expand as the percent of faculty with PhDs increased. The amount of personal spending had a larger range though for private schools.

Instructional spending was intersting for private schools where pretty evenly up to about 80% of faculty with PhDs instructional spending hovered around $10,000 but after that spending rockets up to at most $56,000 per student. Public institutes, however, only grow slightly as the percent of faculty with PhDs increase, maxing out at about $20,000 per student.

References:

- sthda.com
- stackoverflow.com
- rdrr.io/cran/ISLR