Homework_1

Rylie Fleckenstein

1/16/2021

Please do the following problems from the textbook ISLR.

Question 2.4.2 pg 52

Question 2.4.4 pg 53

Question 2.4.6 pg 53

Question 2.4.8 pg 54-55

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.

We are trying to determine the salary which is a continuous variable (numeric) therefore, this scenario is a regression problem. Since we are not tasked with predicting what the CEO salary is and are instead tasked with determining what factors affect the CEO salary, this problem would be considered an inference regression problem. Finally, the variable n refers to the number of observations in the data set which would be 500 for this problem and the variable p refers to the number of parameters or predictors that are being taken into consideration which would be 3 (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.

We are trying to determine if a product will be a success or a failure making this problem a binary classification problem (response variable is categorical). Also, this is a prediction based scenario where we are not so concerned with how the variables are interacting and are more concerned with making accurate predictions. There are a total of 20 other products being observed making n = 20 and there are 13 predictors making p = 13.

(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.

We are looking to predict the % change in the USD/Euro market making this a regression problem (% change is a continuous variable). Also, we are not as concerned with how the predictions are made but are concerned with making accurate predictions therefore we are most interested in predictions not inference. The data consists of weekly changes for the year of 2012, since there are 52 weeks in a year n = 52. Lastly, there are 3 predictors for the 1 response variable making p = 3.

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.

One real life application in which classification would be useful is in determining if someone will default on their loans. The response variable would be default/not default and the predictors would be financial metrics gathered to describe the individual in question. Some examples of proper predictors would be credit score, income, amount of debt, age, status of employment and so on. The goal of this application would be prediction because we would like to predict if the potential client will default on the loan or not. However, we would also be interested in the application inference as well to determine how much of a risk they might pose on defaulting and we can then adjust our rates on a scale accordingly. A second real life example of classification might be determining if someone will die of heart failure or if they will not. The response would be death event/ lived. The predictors would be health metrics such as age, weight, height, blood pressure, diabetes, sex, level of platelets in the blood and so on. The goal of this problem would be prediction. The health organization would set a line for the data science and if the patients combined risk comes back passed that line, ie. 60% chance of death event from heart failure, then they would be considered at risk and brought in for preventative care. A third real life example of classification would be determining if an x-ray scan shows the patient having pneumonia or not having it. The response variable would be pneumonia/normal, and the predictors are slightly different then the previously described scenarios. For this scenario, the data scientist would use computer vision techniques which involve the use of convolutional neural networks. CNN's use predictors that are found during the feature extraction part of the network. These features or patterns are then used to make the classification calculations. These types of models are used solely for predictions as they are "black-box" approaches and give no insight into how the different features interact.

(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.

One real life application where regression applications might be useful would be predicting the percentage of change in the stock market by day. This scenario would be prediction based since we would only care about accurately determining the percentage of change, we are going to see in the stock market and do not necessarily need to know how our predictors interact with each other. The response variable would be the percentage of change which would be continuous, and the predictors would be market metrics which consist mostly of other continuous variables. A second real life application where regression applications would be useful would be in determining how much of an effect the number of bedrooms in a home has on its overall price. This goal of this application would be inference since we are wanting to know how the predictor bedroom size effects the response variable which is the price of the house. The predictors would be other metrics such as square footage, location, neighborhood crime rate, and so on. A third real world application of regression would be determining the rating a customer might give an application downloaded on their phone. The response variable would be the rating score out of 5 which is a continuous variable. The predictors would be attributes such as number of downloads, number of reviews, storage space required, price, and so on. This would be a prediction-based scenario since we are not looking to understand how the variables interact and are more concerned with accurate predictions.

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

According to the article titled "ingle-link cluster analysis of earthquake aftershocks: Decay laws and regional variations" written by Davis, S one real life application where cluster analysis might be useful would be finding earthquake epicenters. This was done and scientists were able to figure out where the dangerous zones were based upon clustering earthquake occurrences. Clustering analysis is used for unsupervised learning situations or situations where there is no known response variable and someone is trying to clump together like groups. A second real life application of cluster analysis would be detecting like groups of consumers. For example, a company tracking their sales data might run a cluster analysis on said data to find if there are different groups of spenders ie. groups of people who spend different amounts. Once they are able to determine the demographics of the different spending

groups, the company can then come up with marketing schemes to increase their sales in the lower spending groups or determine what factors influence how much a customer spends at their store. A third real life application of cluster analysis might be the detection of fraud or spam. According to the article titled "An incremental cluster-based approach to spam filtering" written by Hsiao, W. email spam is at least annoying and at most dangerous for consumers and therefore something that must be dealt with. Companies have developed clustering algorithms which take into consideration features from an email (the header, the sender, the subject) and then separate these emails into groups for later classification. The next step is to use some type of classification algorithm but the article talks about how the addition of the clustering analysis as the initial filter in the pipeline increases the accuracy of their models.

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?

   The parametric statistical learning method "involves a two-step model-based approach"(James, 2013). The first step of this method is to assume of the shape of the function f that describes the relationship between the predictors and response variable. A common parametric method is linear regression where we use some predetermined linear model as our base model to explain the function. The next step in the method is to use the selected model and fit or train the model using some training data. The goal is to calculate or estimate the coefficients of the model based upon the given data. A non-parametric approach is more flexible since no assumptions about the shape of the model are made. This type of approach can lead to a more complex model that is able to learn the training data quite well. One of the advantages to the parametric approach is that its complexity is lower and therefore explain ability is higher. Also, by assuming the shape of our function it simplifies the regression task because we are only required to estimate the parameters of f and not f itself. One of the draw backs of a parametric approach is that we are assuming the shape of the f and therefore may be completely wrong. Even if we are not completely wrong and get it somewhat close, we are still never going to be able to find the true form of f. Choosing improperly may lead to under fitting the training data which results in poor model performance. On the other hand, there are disadvantaged to choosing a non-parametric approach as well. Complex models where we do not make assumptions about their shape require estimating a lot more parameters to accurately represent the shape of the function. This can lead to overfitting of the training data which leads to poor model performance as well. Lastly, non-parametric models tend to coincide with the "black-box" terminology. We have little idea of how the different parameters interact within complex non-parametric models and therefore are not able to gain any insight into how they work or how they are making their predictions (James, 2013).

4. Question 2.4.8 pg 54-55

a.) Below I read in the "College" data and show the first 5 rows.

```
##                              Private Apps Accept Enroll Top10perc Top25perc
## Abilene Christian University     Yes 1660   1232    721        23        52
## Adelphi University               Yes 2186   1924    512        16        29
## Adrian College                   Yes 1428   1097    336        22        50
## Agnes Scott College              Yes  417    349    137        60        89
## Alaska Pacific University        Yes  193    146     55        16        44
## Albertson College                Yes  587    479    158        38        62
##                              F.Undergrad P.Undergrad Outstate Room.Board Books
## Abilene Christian University        2885         537     7440       3300   450
## Adelphi University                  2683        1227    12280       6450   750
## Adrian College                      1036          99    11250       3750   400
## Agnes Scott College                  510          63    12960       5450   450
```

```
## Alaska Pacific University            249         869     7560       4120   800
## Albertson College                    678          41    13500       3335   500
##                               Personal PhD Terminal S.F.Ratio perc.alumni Expend
## Abilene Christian University     2200  70       78      18.1          12   7041
## Adelphi University               1500  29       30      12.2          16  10527
## Adrian College                   1165  53       66      12.9          30   8735
## Agnes Scott College               875  92       97       7.7          37  19016
## Alaska Pacific University        1500  76       72      11.9           2  10922
## Albertson College                 675  67       73       9.4          11   9727
##                               Grad.Rate
## Abilene Christian University         60
## Adelphi University                   56
## Adrian College                       54
## Agnes Scott College                  59
## Alaska Pacific University            15
## Albertson College                    55
```

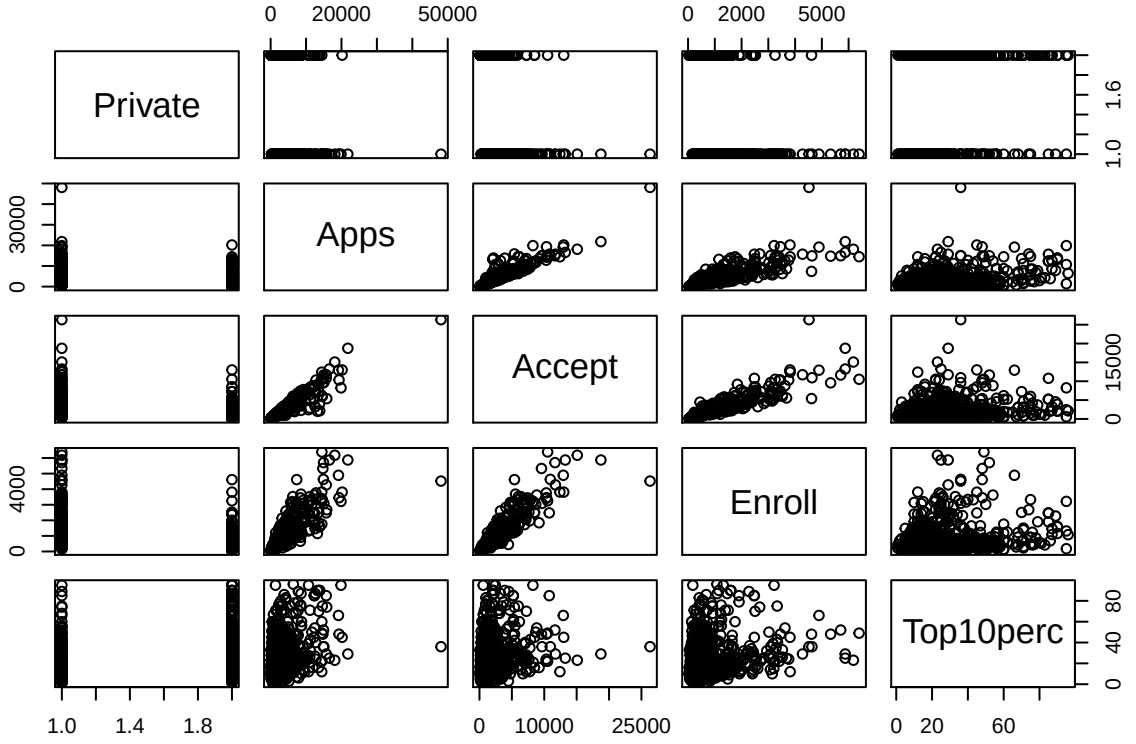
b.) The commands given in the text for part B are to use the code provided in order to replace the row.names of the college data frame with R generated values. The code provided, however, does not accomplish the intended task. Therefore, I have provided the code but commented it out so we can move on to part c.

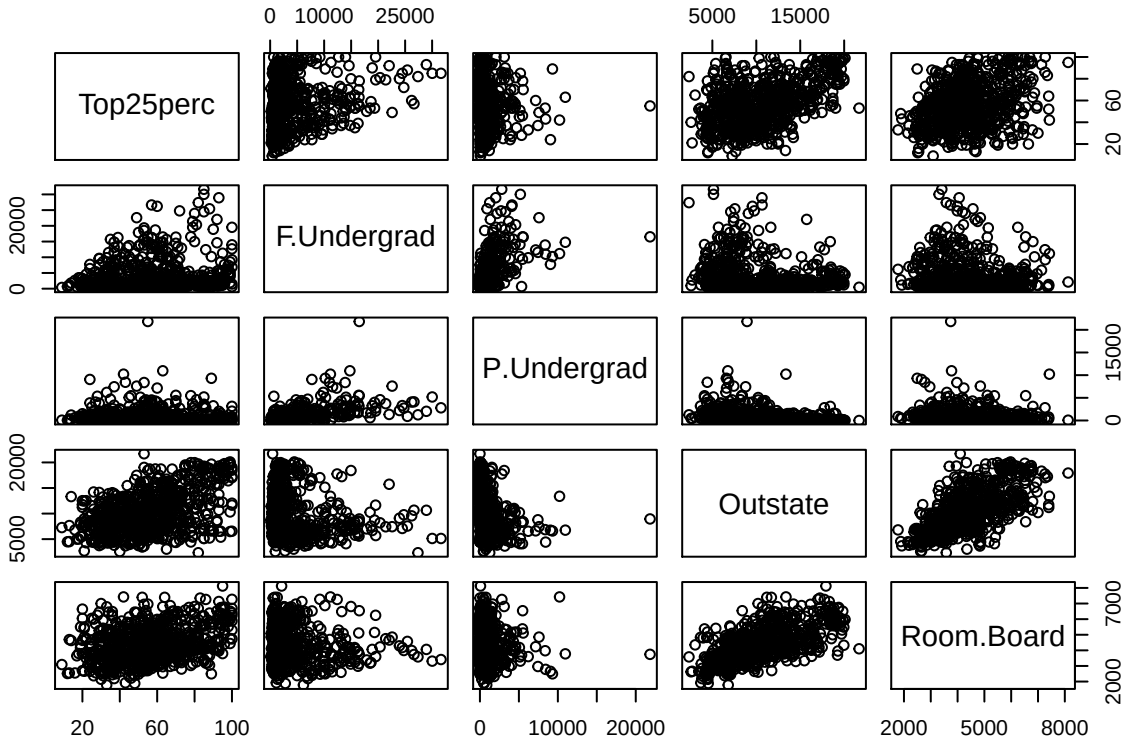
c.) For part c section i. I have printed a summary of the College data.frame.

```
##  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
```

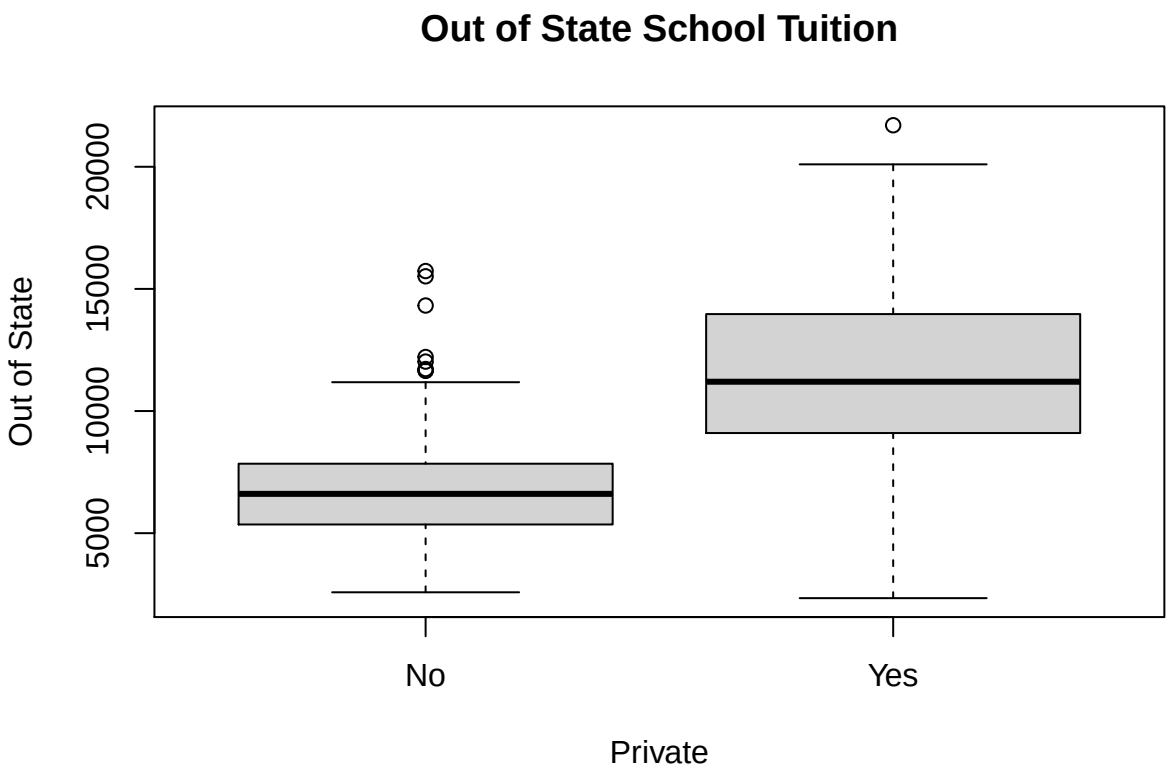
```
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

For part c section ii. I have used the pairs function to print a pair plot of the first 10 variables of the College data set. Following what the text says, however, leads to overplotting so I modified the task slightly and printed 2 different plots of 5 variables. This helped some of the overplotting issue but did not solve all of it.




Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
0 10000 25000
5000 15000
20 60
0 20000
0 15000
5000 20000
2000 7000
20 60 100
0 10000 20000
2000 5000 8000

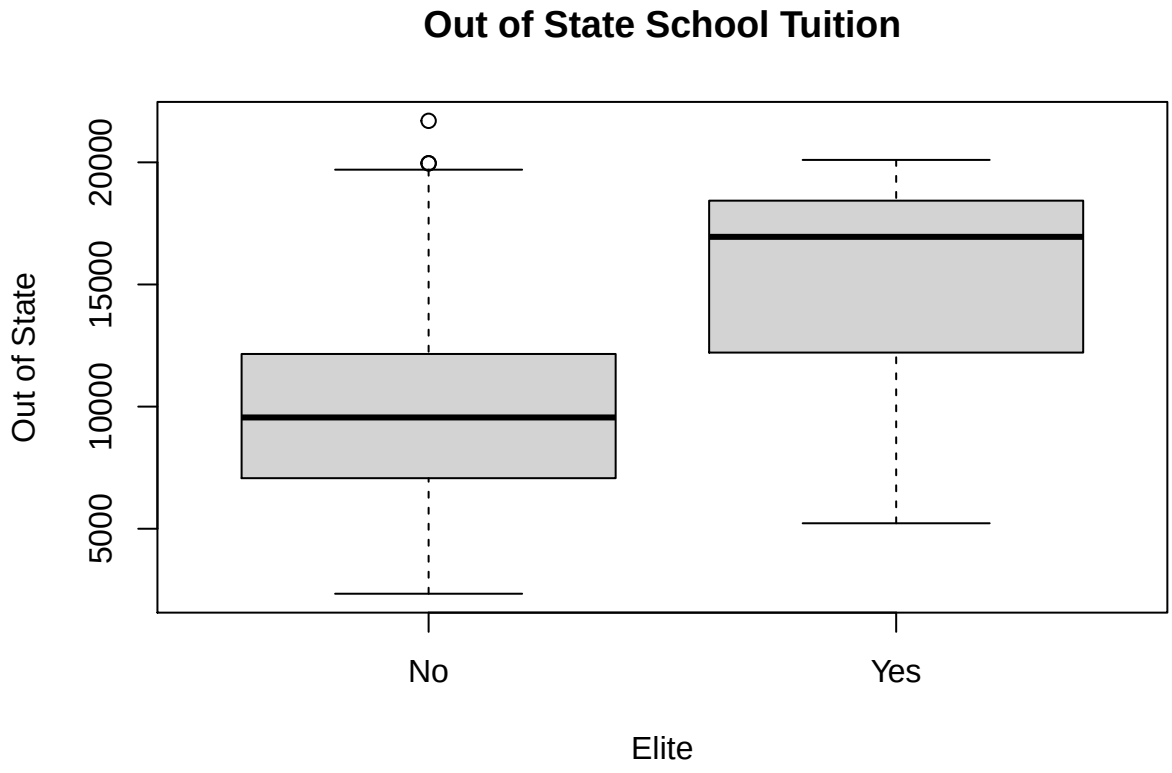
For part c section iii. I produced a side by side boxplot showing the distribution of out of state schools' tuition based upon if they are a private school or not. Referencing the plot we can see private schools have a wider range of tuition costs but their average is significantly higher then non-private schools. Also, the maximum tuition cost for private out of state schools is significantly higher then the maximum tuition cost for out of state non-private schools while the lowest cost of each is relatively the same.



Below I followed the code outlined in the text and created an "Elite" variable which tells us which schools are considered "Elite" and which are not. The criteria to be considered an "Elite" school us that at least 50% of the school population must be made up of individuals who were in the top 10% of their highschool graduating class. After running a summary on the College data containing the new "Elite" variable, we can see that there are 78 schools considered to be "Elite" and 699 that are not. Also, we can see in the side by side box plot the distribution of tuition prices among the out of state schools based upon their "Elite" status. We can conclude that on average tuition is more at the Elite schools, however, some of the most expensive non-elite schools are the same price as some of the most expensive elite schools.

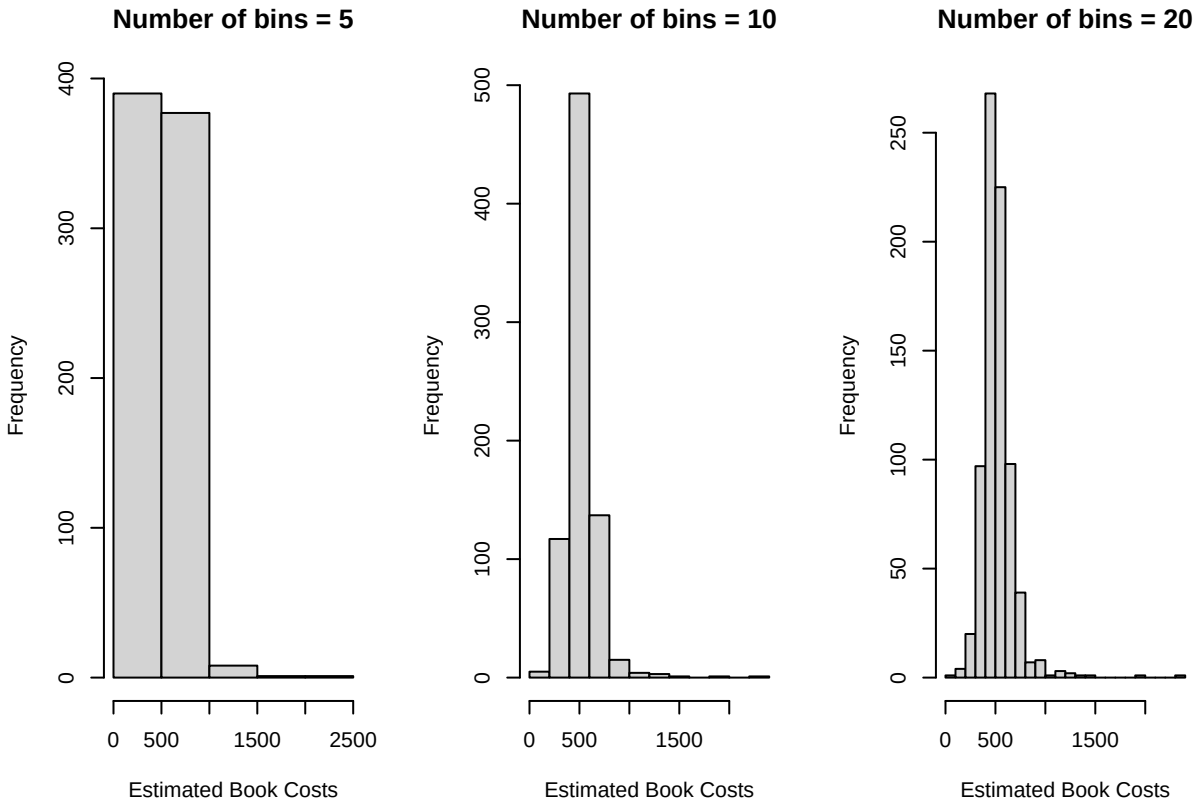
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00
##            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
```

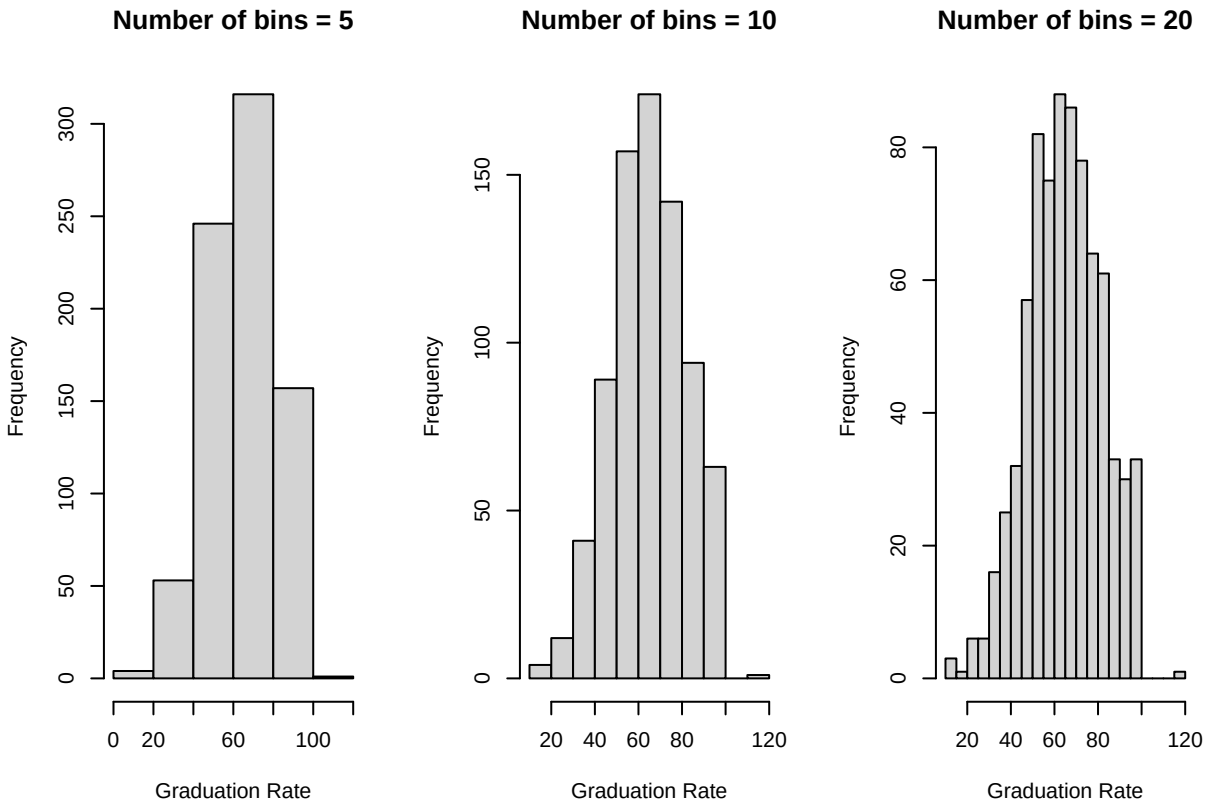
```
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           Personal          PhD
##  Min.   :1780   Min.   :  96.0   Min.   : 250   Min.   :  8.00
##  1st Qu.:3597   1st Qu.: 470.0   1st Qu.: 850   1st Qu.: 62.00
##  Median :4200   Median : 500.0   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      Elite
##  Min.   : 10.00   No :699
##  1st Qu.: 53.00   Yes: 78
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

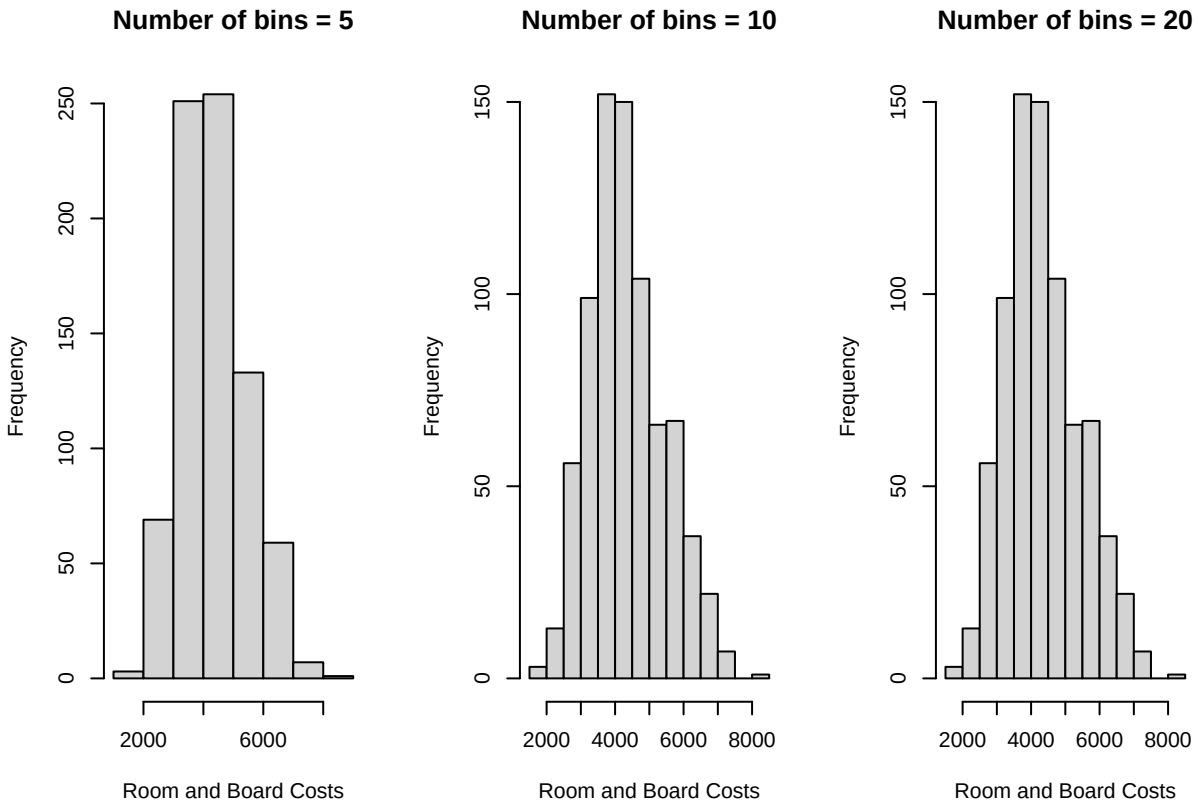


As instructed by part c section v. I chose three variables from the data set and plotted histograms of their distributions. I then plotted three versions of each variable using various bin sizes. The first variable I plotted was the "Books" variable to determine the distribution of Book costs across all colleges. The second variable is "Grad.Rate" which shows the distribution of graduation rates across the colleges and lastly I

plotted “Room.Board”. The histogram of “Room.Board” shows us the distribution of room and board costs across the colleges.




Number of bins = 5
Frequency
0
50
100
150
200
250
300
0
20
60
100
Graduation Rate
Number of bins = 10
Frequency
0
50
100
150
20
40
60
80
120
Graduation Rate
Number of bins = 20
Frequency
0
20
40
60
80
20
40
60
80
120
Graduation Rate



For part c section vi. the textbook calls for further data exploration and a brief summary on the findings. I was curious to know, on average, what the ratio was of money spent by the student to how much was invested back into said student through educational costs based upon if the school was an elite school or not. In order to complete this task I calculated the average Room and board, Personal expenditure, book cost, and Instructional expenditure for each subset of schools (Elite, non-Elite). I then divided the instructional expenditure average by the sum of the student expenditure averages to find that there is a 2.5:1 return in Elite schools and a 1.4:1 in non-elite. As in, the school spends 2.5 dollars on instructional expenditures per student for every dollar that student spends for their education at elite schools and 1.4 for every 1 at non-elite schools. I would say the difference is significant considering students at elite schools only spend roughly 1.15 times the amount students at non-elite schools spend on average for their education. One potential cause for this increase in spending is that the average percentage of professors with their Phd at elite schools is 89 while it is only 70 for non-elite schools. The increase in faculty costs may be one of the factors contributing to the discrepancy. Further analysis could lead to a more detailed understanding of the realtionship between Elite schools and non-Elite schools.

Table 1: School Averages

Variable	Elite	Non.Elite
Room and Board	5336.7949	4248.2518
Personal Expenditure	1188.1795	1357.6552
Books Expenses	594.9103	544.3004
Instructional Expenditure per Student	18404.8718	8684.3677

```
## Average percentage of Professors with Phd per School (Elite) 89.32051
```

```
## Average percentage of Professors with Phd per School (non-Elite) 70.80114
```

Sources: Davis, S., & Frohlich, C. (2012, September 20). Single-link cluster analysis of earthquake aftershocks: Decay laws and regional variations. Retrieved January 17, 2021, from https://agupubs.onlinelibrary.wiley.com/doi/abs/10.1029/90JB02634

Hsiao, W., & Chang, T. (2007, January 28). An incremental cluster-based approach to spam filtering. Retrieved January 17, 2021, from https://www.sciencedirect.com/science/article/abs/pii/S0957417407000279

Gareth James, Daniela Witten, Trevor Hastie, Robert Tibshirani. (2013). An introduction to statistical learning : with applications in R. New York :Springer,