Homework 1

Prakash Paudyal

December 10, 2017

Collaborators: myself

Due January 16th

Answer all questions specified on the problem but if you see something interesting and want to do more analysis please report it as well. Don't forget to include discussion.

Submit your file with the knitted (or knitted Word Document saved as a PDF). If you are still 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 titled Format+STAT-702+HW.

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)

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
2. Question 2.4.4 pg 53
3. Question 2.4.6 pg 53

1. (Q 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.
   Ans: In this case we have quantitative variables and are predicting the salary of CEO. Hence, this is a regression problem. Since the goal of this analysis is to find which predictor variables are associated with the response variable, hence it is an inference problem.
   Number of observation (n):500
   Number of predictors (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.
Ans: Since this problem has qualitative response, it is a classification problem. We need to find whether it will be success or not, so we are interested in prediction.
n=20
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.
Ans: Since the response variable is quantitative, this is a regression problem. We are predicting the % change, so we are interested in prediction.
n=52 (total numbers of weeks in 2012)
p=3

2. (Q 2.4.4) 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.
Ans: 1. Determine whether a patient might suffer from heart attack using predictors age, gender, education, income, and blood pressure. The goal of this application is prediction and the response is a yes or no i.e. whether a patient might or might not have a heart attack.
2. Determine whether a customer will buy an item using predictors like age, income, ethnicity and location. The goal of this application is prediction and the response is a yes/no i.e. a customer will buy or will not buy an item.
3. Predict whether an advertisement is fake or not based on predictors like the number of positive and negative words used, length of the advertisement and author.

(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.
Ans: 1. Predict the number of car sales using predictors cost of car, location, fuel economy and demography
2. Predict house price using predictors neighborhood, size of house, Number of bedrooms, bathrooms and view.
3. Infer the relation between the number of bikes that are rented and the predictors like temperature, weather, day of the week, and population of the area.

(c) Describe three real-life applications in which *cluster analysis* might be useful.
Ans: 1. A pizza shop wants to find groups of customers with similar purchase history so that they can introduce different offers aimed towards the different types of customers.

2. The government wants to organize free health clinics, so they can group cities that have reported similar disease outbreak and send specialized medical professionals and medicines accordingly.
3. Students can be divided into different clusters based on their performance and different teaching techniques can be used to ensure that the groups that have poor performance participate more in the class.

3.(Q 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?

Differences:
Parametric approach makes an assumption about the functional form of the unknown function (**f**) and uses the training data to fit or train the model. On the other hand, non-parametric approach does not make any assumptions about the form of **f** and instead seeks an estimate of **f** that is as close as possible to the training data points.

Advantages:
Using a parametric approach to classification or regression simplifies the problem of estimating **f** because it is much easier to estimate a set of parameters that it is to fit an entirely arbitrary function **f**. Due to this, lesser amount of training data is required to obtain an accurate estimate of **f** when compared to non-parametric approach.

Disadvantages:
The disadvantage of parametric approach is that if the model that is chosen does not match the true form of **f**, our estimates will be poor.

4. Question 2.4.8 pg 54-55

a

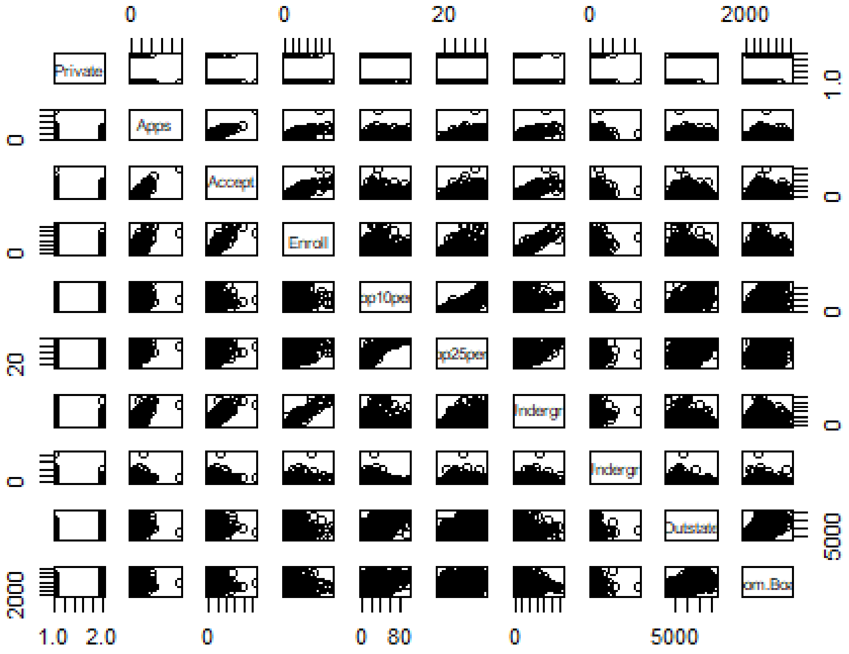
b

```
##   Private Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad
## 1     Yes 1660   1232    721        23        52        2885         537
## 2     Yes 2186   1924    512        16        29        2683        1227
## 3     Yes 1428   1097    336        22        50        1036          99
## 4     Yes  417    349    137        60        89         510          63
## 5     Yes  193    146     55        16        44         249         869
## 6     Yes  587    479    158        38        62         678          41
##   Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni
## 1     7440       3300   450     2200  70       78      18.1          12
## 2    12280       6450   750     1500  29       30      12.2          16
## 3    11250       3750   400     1165  53       66      12.9          30
## 4    12960       5450   450      875  92       97       7.7          37
## 5     7560       4120   800     1500  76       72      11.9           2
## 6    13500       3335   500      675  67       73       9.4          11
##   Expend Grad.Rate
## 1   7041        60
## 2  10527        56
## 3   8735        54
## 4  19016        59
## 5  10922        15
## 6   9727        55
```

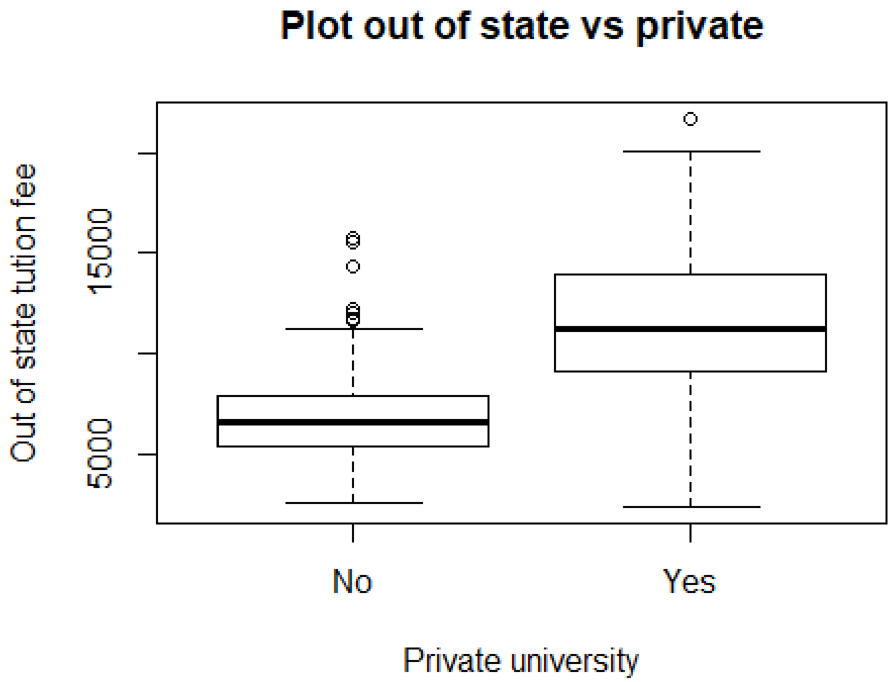
c(i)

```
##  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
```

c(ii)

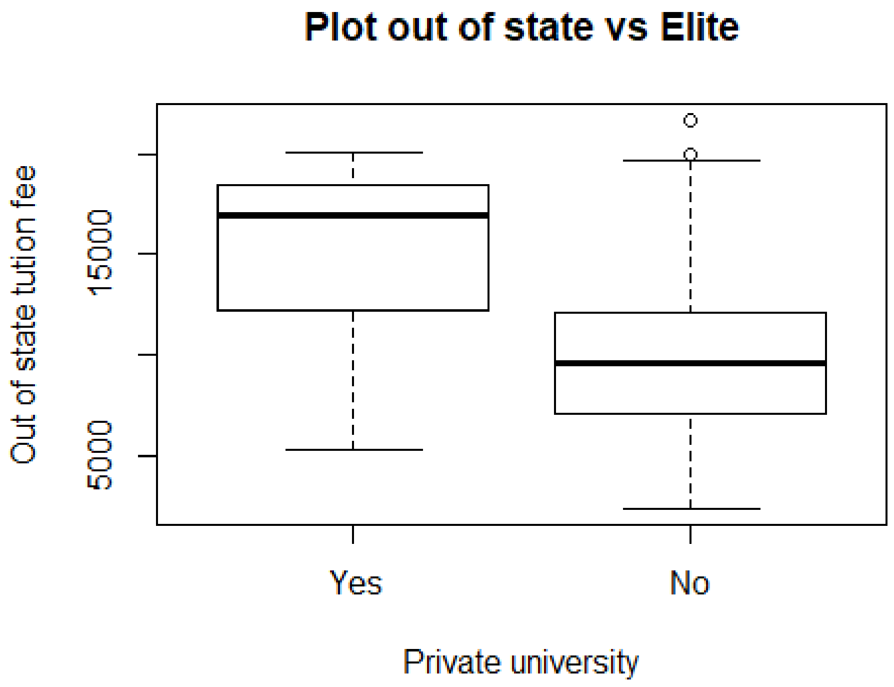


c(iii)

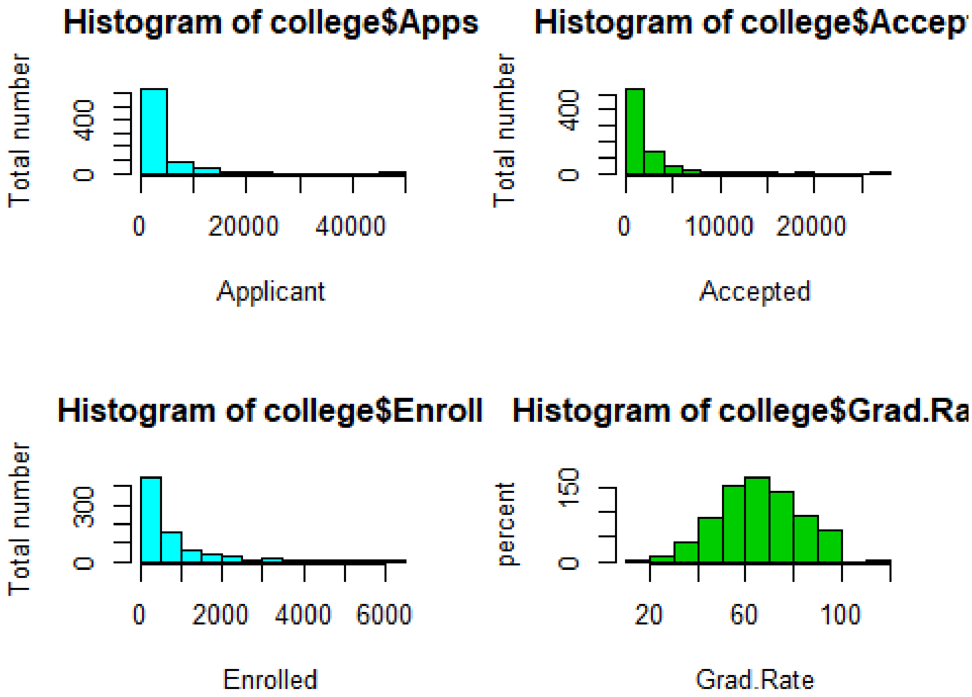


c(iv)

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


Plot out of state vs Elite
Out of state tution fee
5000
15000
Yes
No
Private university

c(v)

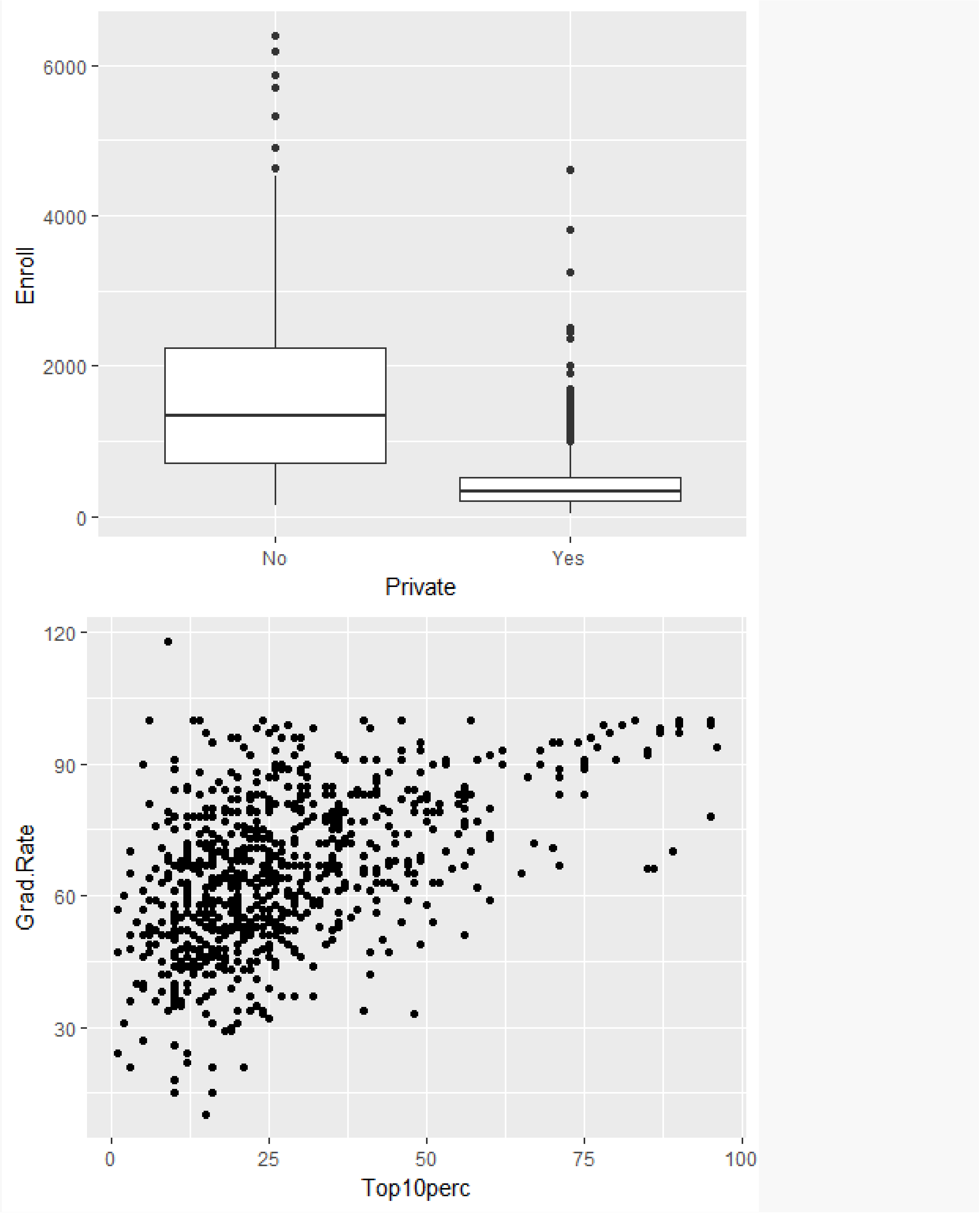

Histogram of college$Apps
Total number
0
400
0
20000
40000
Applicant
Histogram of college$Accep
Total number
0
400
0
10000
20000
Accepted
Histogram of college$Enroll
Total number
0
300
0
2000
4000
6000
Enrolled
Histogram of college$Grad.Ra
percent
0
150
20
60
100
Grad.Rate

c(Vi)

```
## Warning: package 'ggplot2' was built under R version 3.4.1

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

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


Enroll
6000
4000
2000
0
No
Yes
Private
Grad.Rate
120
90
60
30
0
25
50
75
100
Top10perc

Bar plot of private vs out of state shows that private universities have higher out of state tuition fee than public universities. Elite school has higher out of state tuition than non-elite school. Maximum enrollment of one of the college is 6392 and minimum enrollment is 35. Graduation rate is range from 10 to 118.

Private universities have less enrollment than public school. One of the university has 103% faculty with PhD. Graduation rate of the universities with higher top 10% student is higher than others.