STAT 702
Shaheed Shihan
Assignment 1

Questions: 2.4.2,2.4.4,2.4.6,2.4.8

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.**

This is a supervised regression learning problem where we are interested in the inference. n=500, 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.**

This is a supervised classification learning problem where we are interested in the prediction. N = 20, p = 13

(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.**

This is a regression problem and we are interested in the prediction. N=52, p = 3

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.**

i. Predicting the likelihood of 10,000 patients either having disease A or B based on 10 symptoms. This is a classification problem with n=10,000 and p = 10. We are interested in prediction.
ii. Predicting if a particular political ad campaign is going to be successful or not. The response will be success or failure. The predictors may include run time, air time, cost, TV network, Polling data, etc. The goal is a prediction.
iii. Predicting whether customers will default on their credit card debt or not. The response will be default or not default. The predictors may include income, credit score, education level, bank balance, credit history, etc. The goal is prediction.

(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.

i. Prediction of ideal market price for houses. The response will be market price while the predictor variables will include location, proximity to school/grocery shops, house value, crime rate, etc. The goal is prediction.
ii. Understanding the association of heart attack to various factors including weight, family history, cholesterol level, body fat percentage, dietary habits, etc. The goal is inference.
iii. Prediction of a student GPA in college. The factors affecting it might include study habits, undergraduate GPA, work schedule, etc. Response is GPA obtained – this is a prediction problem.

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

i. Helping marketers discover distinct groups in their customer bases in order to make more informed business decisions. This knowledge can then be used to develop targeted marketing programs.
ii. Identifying groups of motor insurance policy holders with a high average claim cost.
iii. Understanding changing climate patterns by studying the patterns of the atmosphere and the sea.

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?**

In parametric methods one is often trying to infer a finite but unknown number of parameters, such as the mean and variance of a univariate distribution or covariance of a multivariate distribution. In practice, known distributions like the Gaussian or the Poisson are completely determined by well known parameters, hence the association between parametric methods and well defined distributions.

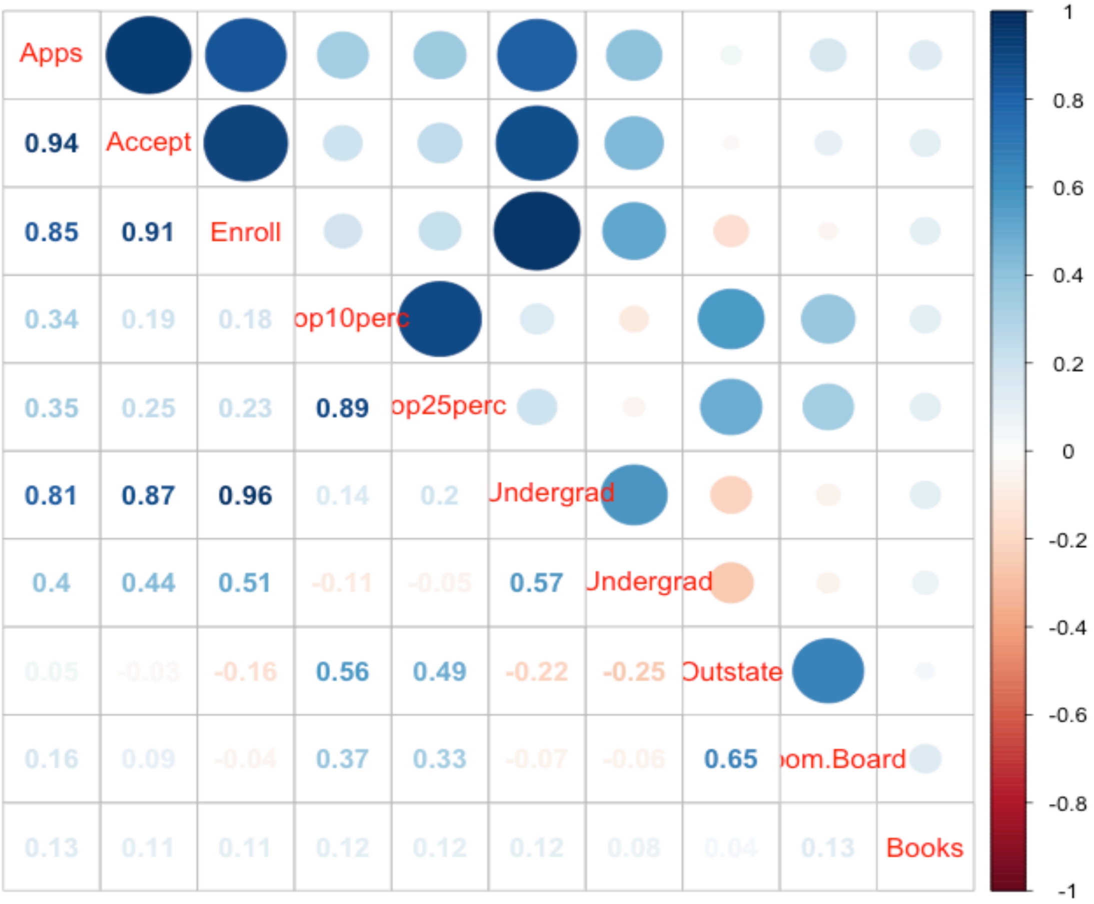
Nonparametric methods on the other hand try to precisely estimate an unknown distribution (nonparametric density estimation or regression or testing) where the number of parameters is at the very least high dimensional and is asymptotically infinite with the number of samples or observations.

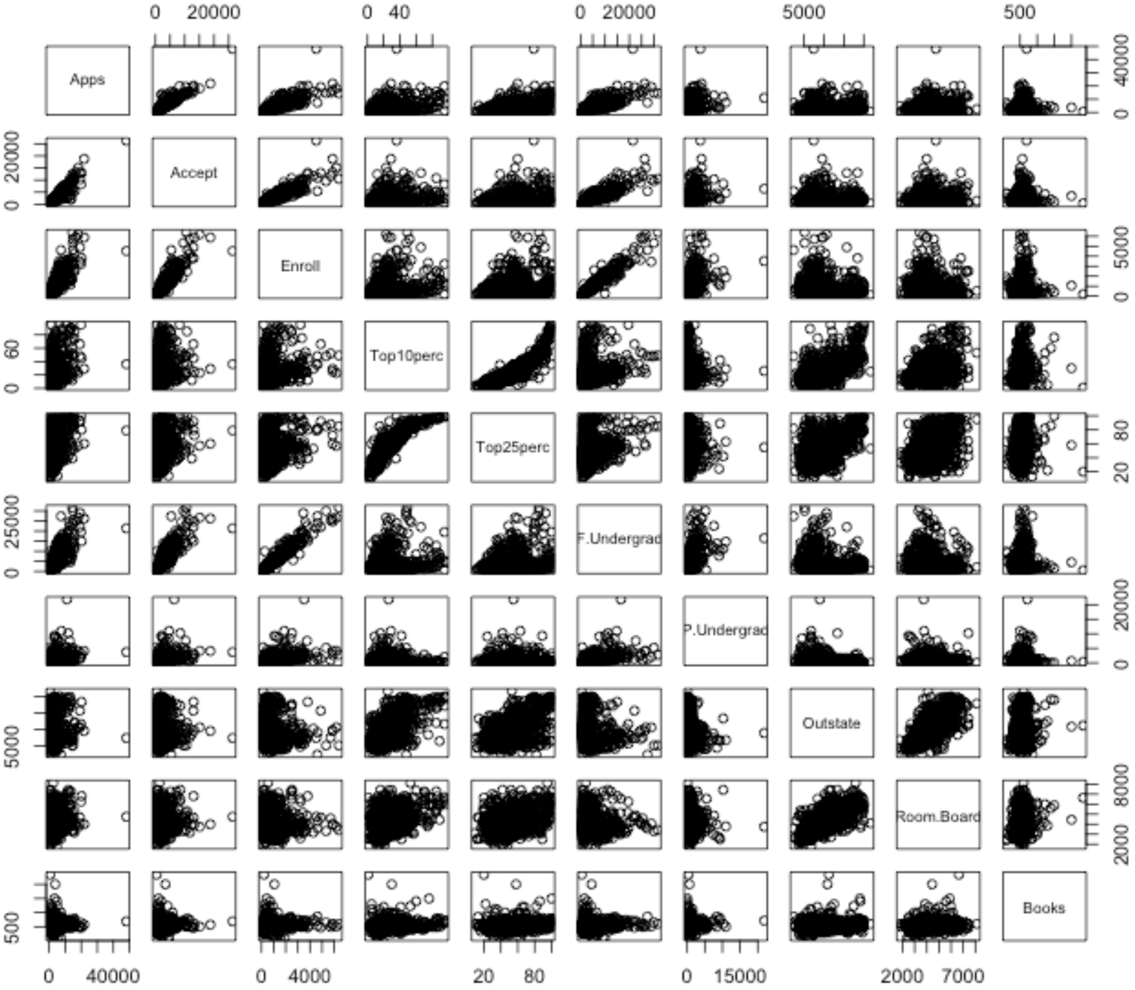
A parametric approach reduces estimating f to the problem by estimating parameters. However, overfitting may result as we try and reduce the error rate by making the model more flexible and including more parameters. Also in the case of a non-parametric approach a large number of estimates are required to obtain an accurate estimate.

8. c.

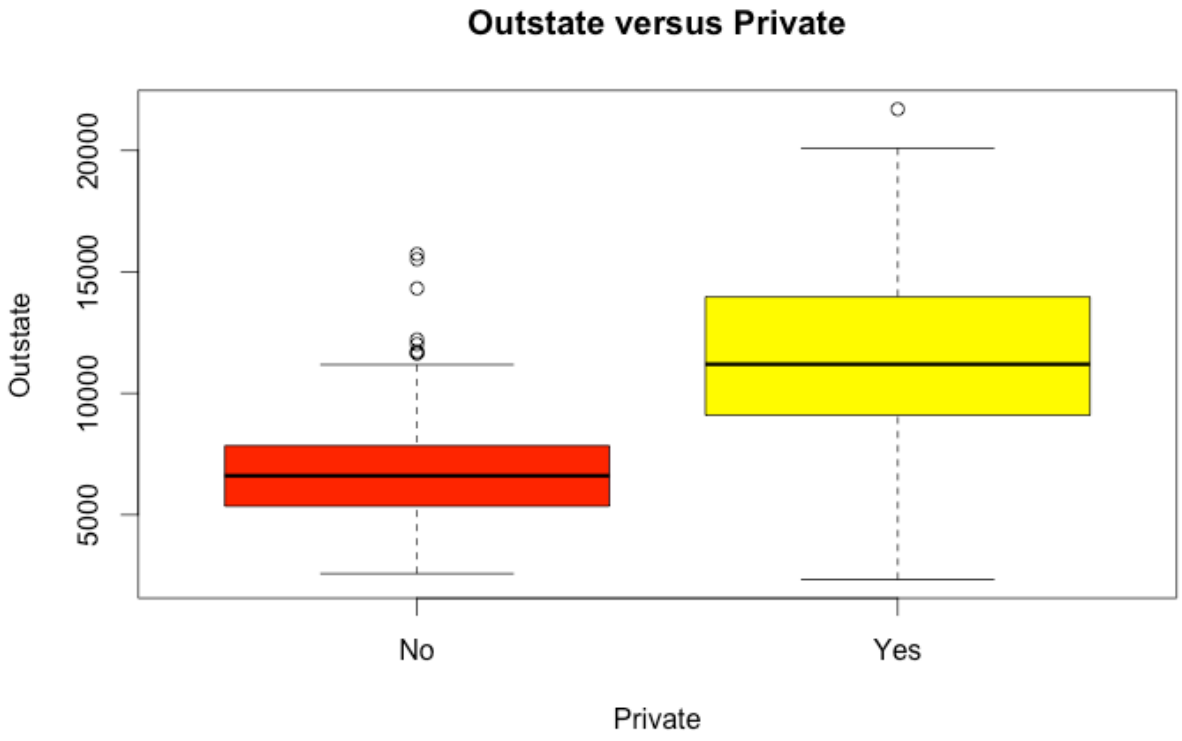
i. Check R-code.

ii. I used corrplot instead of pairs() just for better visual graphics.



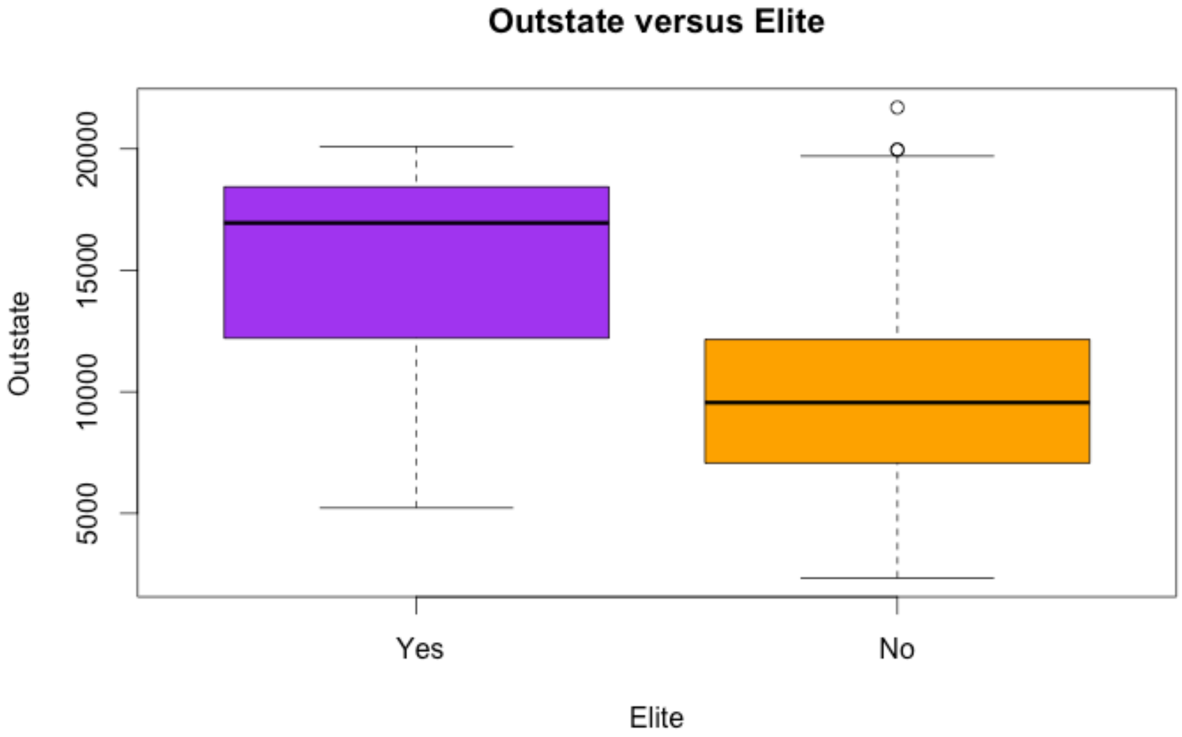

Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board
Books
0
20000
0
40
0
20000
5000
500
0
40000
0
5000
20
80
0
20000
2000
8000
0
20000
0
60
0
25000
5000
500
0
40000
0
4000
20
80
0
15000
2000
7000

iii.

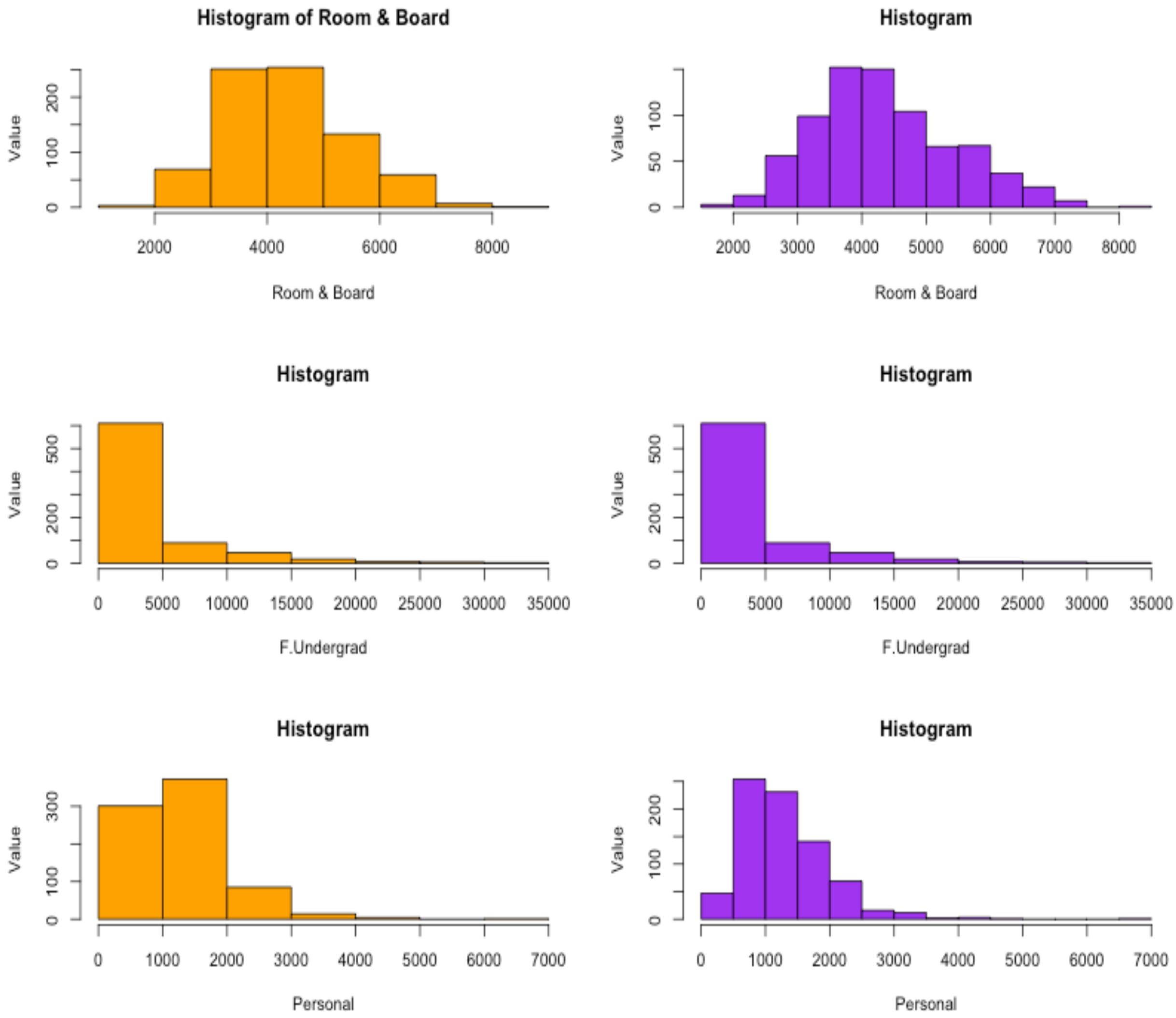


iv.

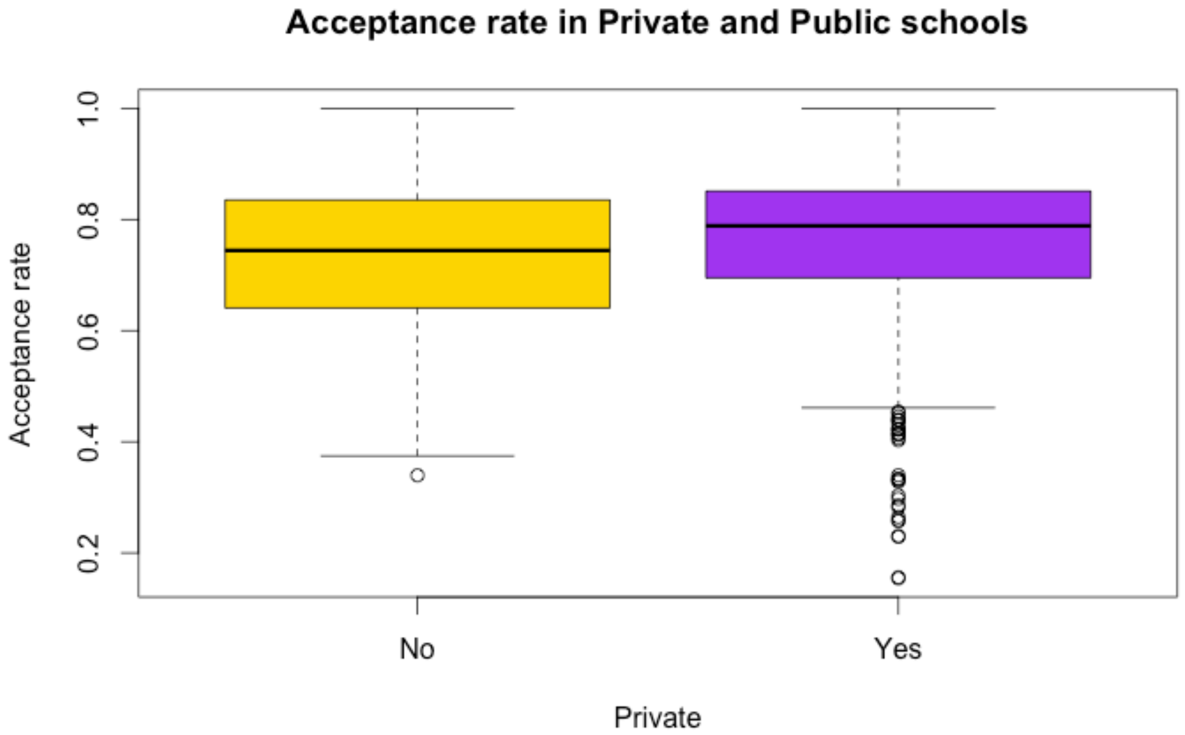
```
> summary(Elite)
 Yes   No
  78  699
```


Outstate versus Elite
Outstate
5000
10000
15000
20000
Yes
No
Elite

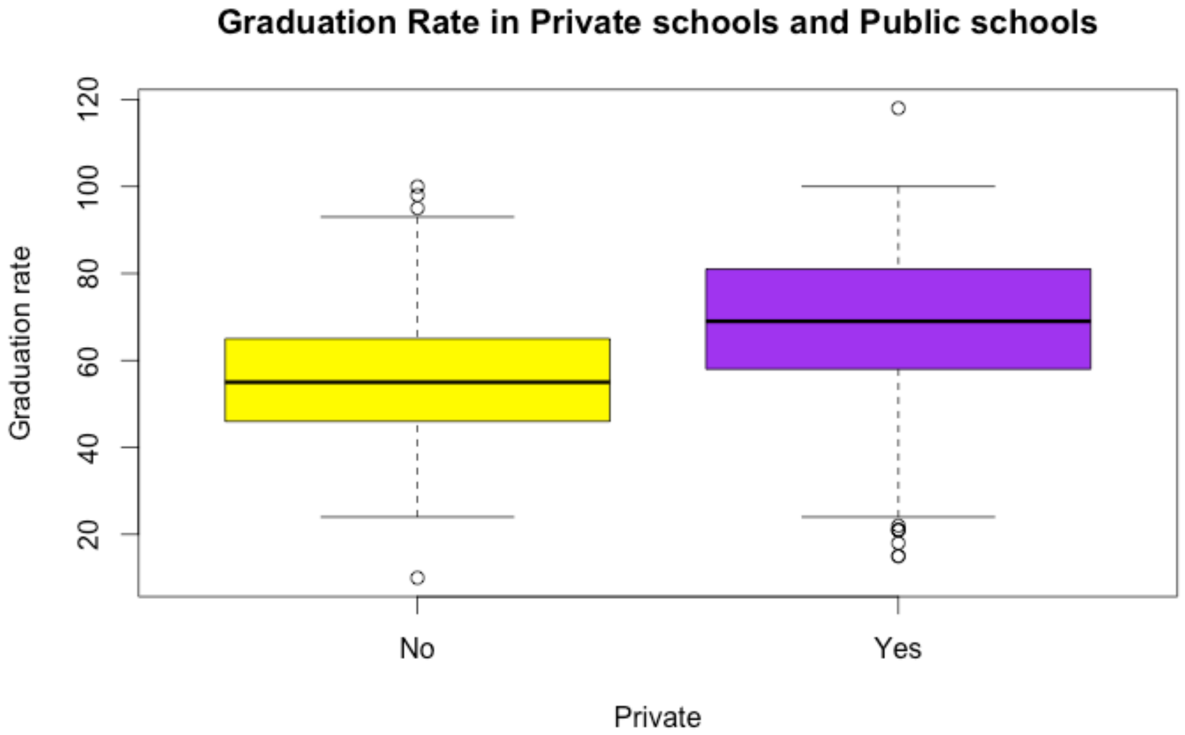
v.



vi.
I was interested in the distinction between public and private schools in terms of acceptance rate and graduation rate. The acceptance rate was computed by dividing the number of students accepted with the number who applied.
It was found using boxplots that private schools have slightly higher acceptance rate and significantly higher graduation rate than public schools.

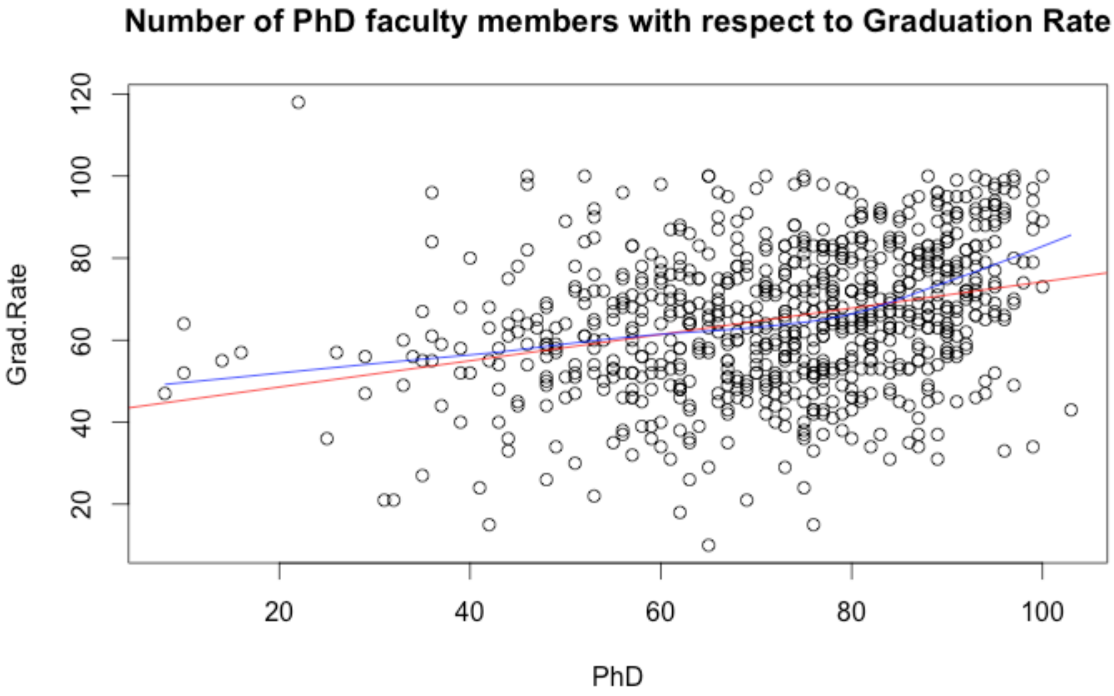


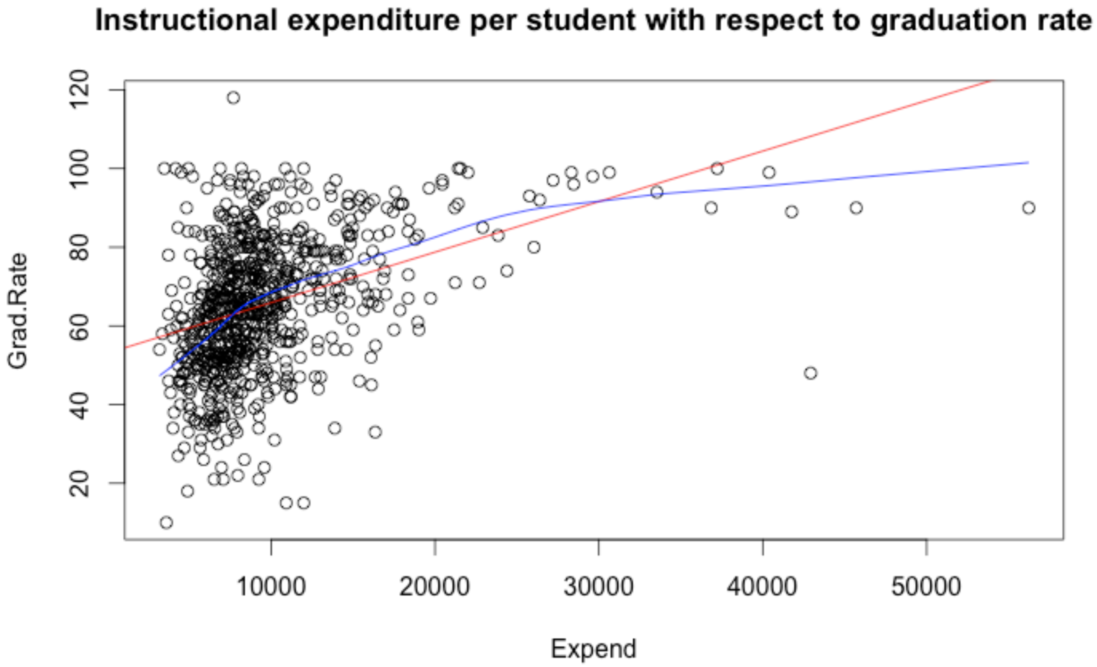
Acceptance rate in Private and Public schools



Graduation Rate in Private schools and Public schools

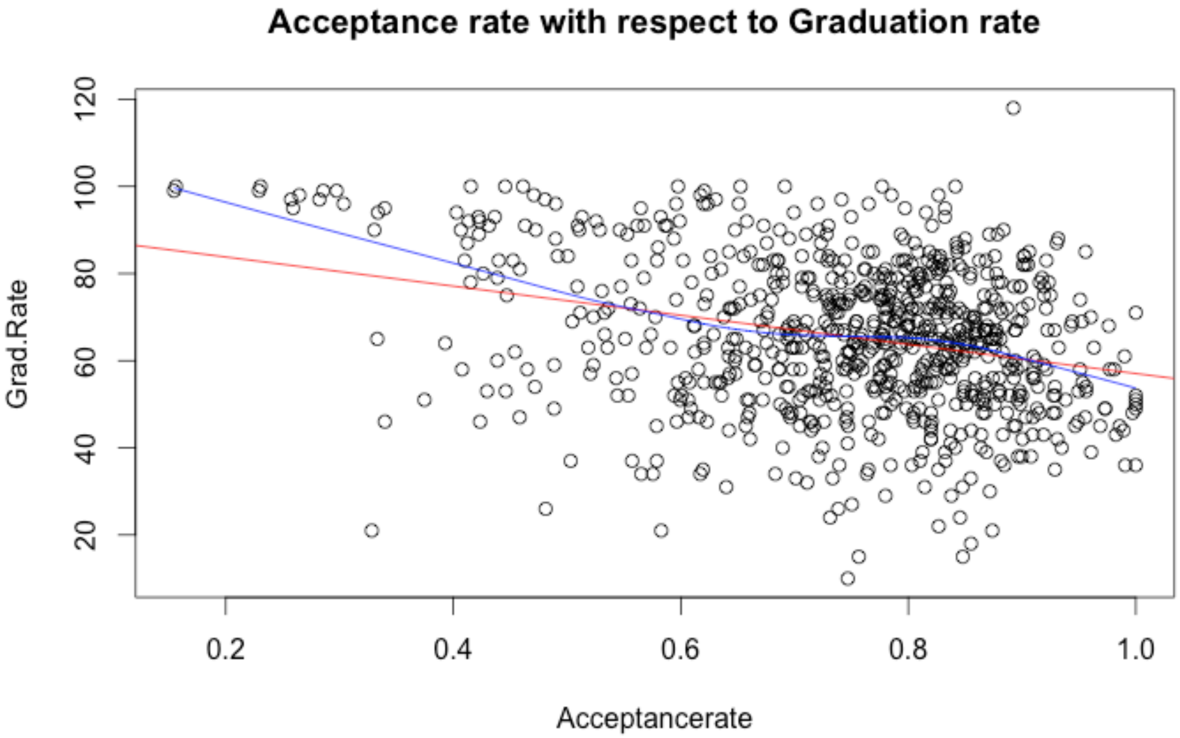
Next I took an interest to see if higher number of PhD faculty affected graduation rate. There does seem to be significance in this hypothesis, as demonstrated by the scatterplot. However, further analysis is required.





Instructional expenditure per student seemed to center around the $10000 mark. However, graduation rates did increase as this expenditure went up. It was surprising to see that schools spent above $20000 on each student had close to a 100% graduation rate.

Finally, I looked at if there was a correlation between acceptance rate and graduation rate. As expected, a negative correlation was found by looking at the plot.



More selective schools had higher graduation rates as compared to schools with higher acceptance rates.