Homework 1 Solutions

Problem 1

a) When the sample size n is large and the number of predictors p is small, the samples will likely span the space of predictors and we can use a more flexible model to approximate the true function f.

b) When p is large and n is small, it would be best to use a simple model, because a model that is too flexible might overfit the data.

c) If the relationship between predictors and response is highly non-linear it would be good to use a flexible, nonparametric method. The roughness of $\hat{f}$ should be controlled to prevent overfitting.

d) If the variance of the error terms is large, it would be best to use an inflexible method, because a flexible method might fit the noise in the training sample.

Problem 2

a) This is a regression problem in which we are mostly interested in inference; $n = 500$, $p = 3$. Note: The variable industry is categorical; if there are many categories, this might be represented as several predictors.

b) This is a classification problem in which we are interested in prediction; $n = 20$, $p = 13$.

c) This is a regression problem in which we are interested in prediction; $n = 52$ (52 weeks in a year), $p = 3$.

Problem 3

Classification Example

- Predict whether it is going to rain or sunshine tomorrow given air pressure, temperature, humidity, last week's precipitation etc.
- Whether a new TV show will be a success or a failure. The response is success / failure. The predictors are Genre, air time, running time, producer, money spent. We want to predict the outcomes.
- Classify whether an email is spam/non-spam. We want to predict an email as spam or non spam. Response is either spam or non-spam. Predictor is the frequency of words in email, sender's email address, IP address of sender etc.

Regression Example

- Predict a movie's earnings (response) based on the following predictors: genre, budget, rating, pre-release analyst view etc. The goal may be inference, since production companies may want to understand how much the rating affects earnings for movies with the same genre and budget. They may also want to determine which factors are the most influential, in order to fund movies that are more likely to turn a profit, or so that they can influence the content of movies before they're released.
- Predict an individual's lifespan (response) based on the following predictors: gender, exercise, height, weight, and smoking status. If an insurance company wants to have the best possible estimate of how long someone is going to live, their goal is prediction not inference. In this scenario, they are more concerned with accuracy in order to provide plans that will maximize their profits.
- Infer about the important factors affecting crime rate of a city, predictors can be average income, average education level, average rent, unemployment rate etc.

Clustering Example

- Cluster analysis can be used to group parts of social networks, like connections on LinkedIn or characters that interact in a TV show. This process allows us to determine the social groups each person falls into. For example, clustering of LinkedIn connections may create groupings for schoolmates, friends, and people met through work.
- Cluster analysis can be used to group consumers into certain clusters for market segmentation. These groupings can be based on the frequency and types of purchases they've made (e.g. 10 household items, 5 pieces of sporting equipment, 3 electronic devices, etc. This process can help companies promote products to a select number of groups based on their shared interests.
- Cluster analysis can be used for image classification, for example to group a set of pictures based on who is in the image. These clusters can be formed based on pixel data, so that similar images, which we assume feature the same person, are grouped together.

Problem 4:

Advantages of a more flexible method: Less bias, may fit non-linear relationship better
Disadvantages of a more flexible method: High variance, may overfit to the noise in the training data

A flexible method is preferred when we have large sample size and fewer predictors, or when we expect the underlying relationship between the predictors and the response is highly non-linear. A less flexible method is preferred where we have small sample size and when we expect the true underlying relationship between the predictors and the response to be close to linear, Also, a less flexible method is more interpretable.

Problem 5)

Here is the code that was run, followed by the output:

```
# Question 5

# Formatting the data
college=read.csv("college.csv", header=T, na.strings = "?")
rownames(college)=college[,1]
college=college[,-1]
#fix(college)
print(summary(college))

# Looking at pairwise plots of the first 10 variables
jpeg("pairs.jpeg")
pairs(college[,1:10],main="Scatter Plot Matrix of 10 College Variables")
dev.off()

# Looking at the outstate tuition of private vs public schools
jpeg("outstatevsprivate.jpeg")
plot(college$Private,college$Outstate,xlab="Private",ylab="Outstate Tuition ($)",
     main="Outstate Tuition at Private and Public Colleges")
dev.off()

# Defining elite universities as having more than 50%
# of incoming students from the top 10% of the high school class
Elite=rep("No",nrow(college))
Elite[college$Top10perc>50]="Yes"
Elite=as.factor(Elite)
college=data.frame(college,Elite) # Add column to college matrix
print(summary(Elite))
jpeg("outstatevselite.jpeg")
plot(college$Elite,college$Outstate,xlab="Elite",ylab="Outstate Tuition ($)",
     main="Outstate Tuition at Elite and Not Elite Colleges")
dev.off()

# Creating some histograms
jpeg("histograms.jpeg")
par(mfrow=c(2,2))
hist(college$Apps, xlab="Number of Applications", ylab="Frequency",
     main="Number of Applications Histogram", breaks=20)
hist(college$Accept, xlab="Number Accepted", ylab="Frequency",
     main="Student Number Accepted Histogram", breaks=20)
hist(college$Books, xlab="Estimated Book Costs ($)", ylab="Frequency",
     main="Estimated Book Costs Histogram", breaks=40)
hist(college$Personal, xlab="Estimated Personal Spending ($)", ylab="Frequency",
     main="Estimated Personal Costs Histogram", breaks=40)
dev.off()

# Some more exploration: looking at acceptance rate and whether it is
# affected by private/public or elite/non-elite schools
jpeg("furtherdata1.jpeg")
par(mfrow=c(1,2))
acceptrate=100*college$Accept/college$Apps
plot(college$Private,acceptrate,xlab="Private",ylab="Acceptance Rate (%)",
     main="Acceptance Rate")
plot(college$Elite,acceptrate,xlab="Elite",ylab="Acceptance Rate (%)",
     main="Acceptance Rate")
dev.off()
```

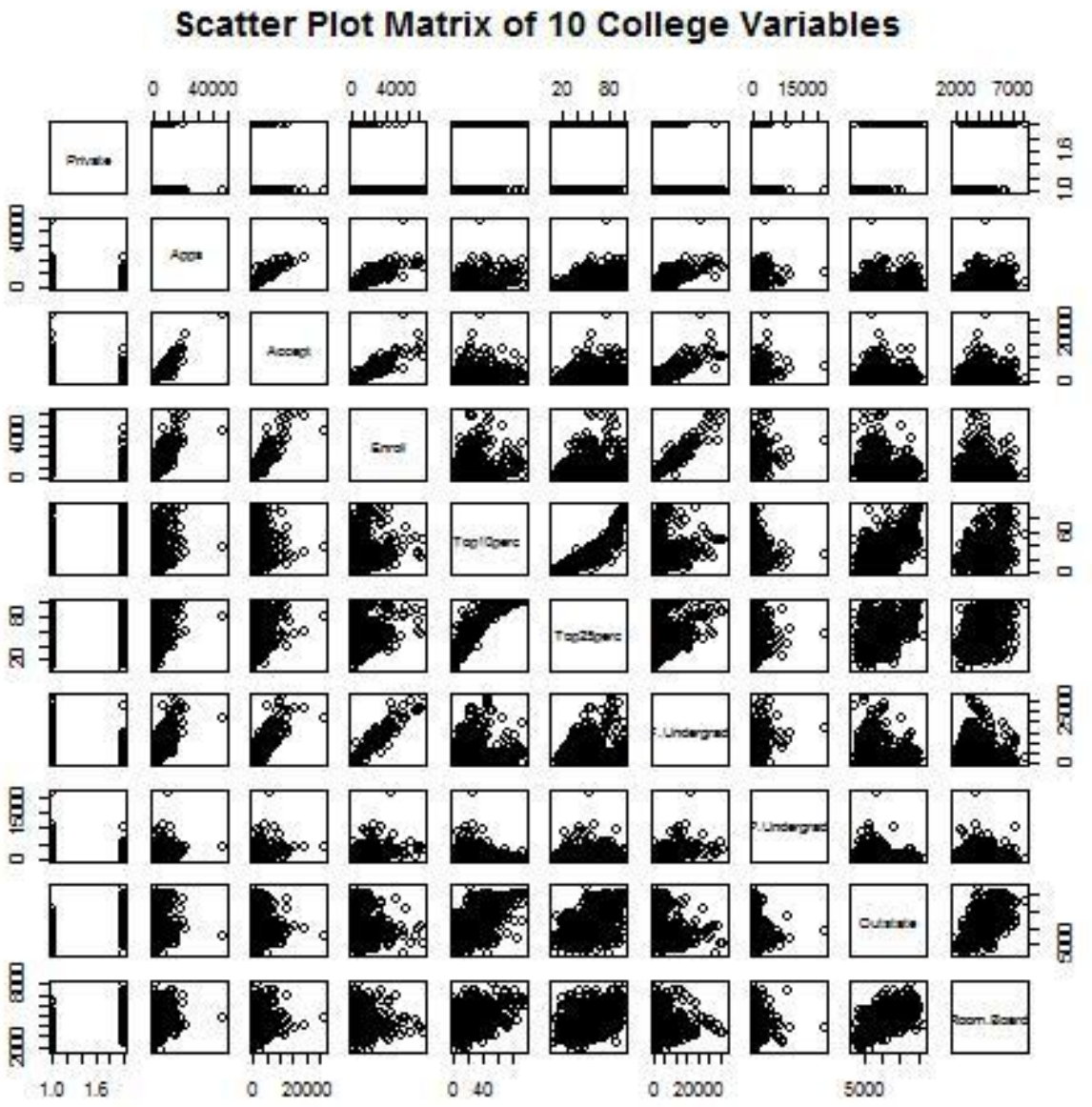
```r

# Now let us look at yield rate (enrollment/acceptance) and whether it is
# affected by private/public or elite/non-elite schools
jpeg("furtherdata2.jpeg")
par(mfrow=c(1,2))
yield=100*college$Enroll/college$Accept
plot(college$Private,yield,xlab="Private",ylab="Yield Rate (%)",
     main="Yield Rate")
plot(college$Elite,yield,xlab="Elite",ylab="Yield Rate (%)",
     main="Yield Rate")
dev.off()
```

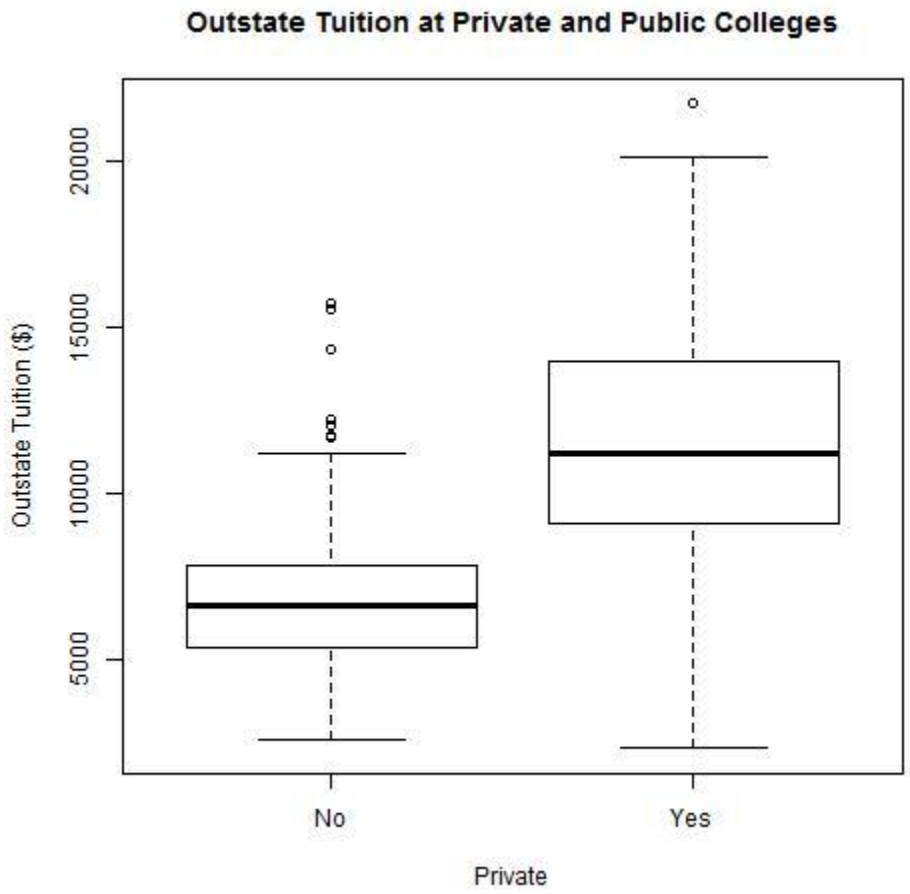
Summary of data printed into console:

```
Private        Apps           Accept          Enroll       Top10perc       Top25perc
No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00   Min.   :  9.0
Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00   1st Qu.: 41.0
          Median : 1558   Median : 1110   Median : 434   Median :23.00   Median : 54.0
          Mean   : 3002   Mean   : 2019   Mean   : 780   Mean   :27.56   Mean   : 55.8
          3rd Qu.: 3624   3rd Qu.: 2424   3rd Qu.: 902   3rd Qu.:35.00   3rd Qu.: 69.0
          Max.   :48094   Max.   :26330   Max.   :6392   Max.   :96.00   Max.   :100.0
 F.Undergrad     P.Undergrad         Outstate       Room.Board       Books
Min.   :  139   Min.   :    1.0   Min.   : 2340   Min.   :1780   Min.   :  96.0
1st Qu.:  992   1st Qu.:   95.0   1st Qu.: 7320   1st Qu.:3597   1st Qu.: 470.0
Median : 1707   Median :  353.0   Median : 9990   Median :4200   Median : 500.0
Mean   : 3700   Mean   :  855.3   Mean   :10441   Mean   :4358   Mean   : 549.4
3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925   3rd Qu.:5050   3rd Qu.: 600.0
Max.   :31643   Max.   :21836.0   Max.   :21700   Max.   :8124   Max.   :2340.0
   Personal         PhD           Terminal       S.F.Ratio      perc.alumni
Min.   : 250   Min.   :  8.00   Min.   : 24.0   Min.   : 2.50   Min.   : 0.00
1st Qu.: 850   1st Qu.: 62.00   1st Qu.: 71.0   1st Qu.:11.50   1st Qu.:13.00
Median :1200   Median : 75.00   Median : 82.0   Median :13.60   Median :21.00
Mean   :1341   Mean   : 72.66   Mean   : 79.7   Mean   :14.09   Mean   :22.74
3rd Qu.:1700   3rd Qu.: 85.00   3rd Qu.: 92.0   3rd Qu.:16.50   3rd Qu.:31.00
Max.   :6800   Max.   :103.00   Max.   :100.0   Max.   :39.80   Max.   :64.00
    Expend        Grad.Rate
Min.   : 3186   Min.   : 10.00
1st Qu.: 6751   1st Qu.: 53.00
Median : 8377   Median : 65.00
Mean   : 9660   Mean   : 65.46
3rd Qu.:10830   3rd Qu.: 78.00
Max.   :56233   Max.   :118.00
```

Pair-wise scatter plots:

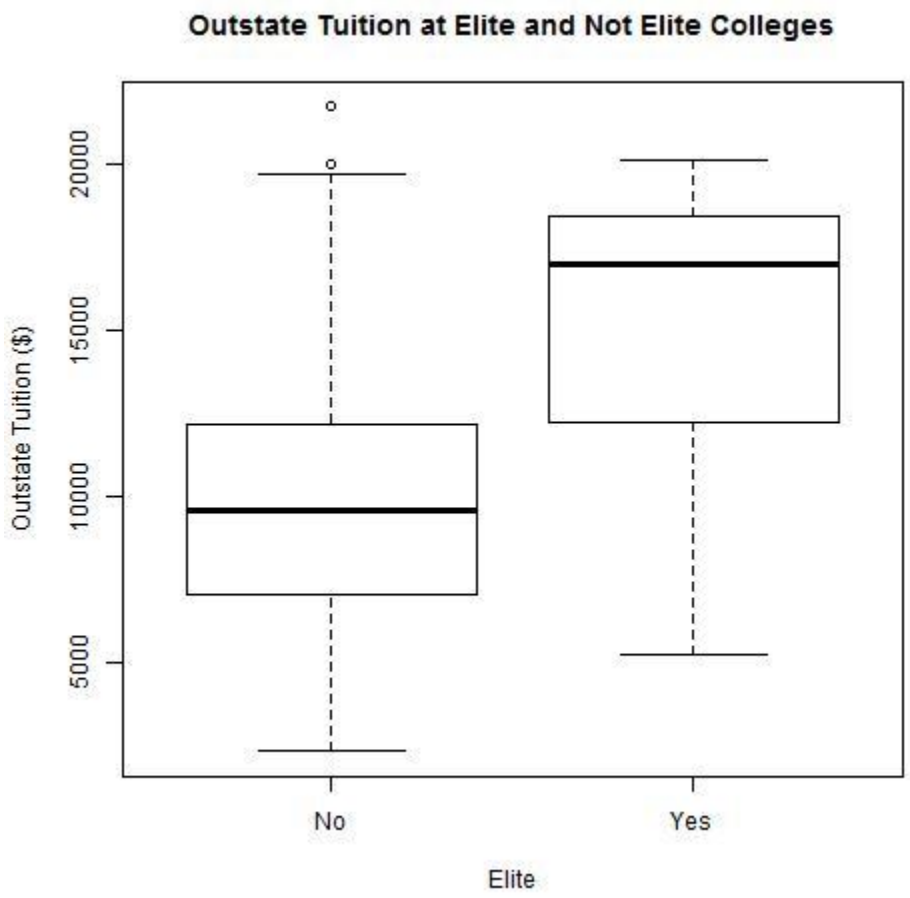


Side by side box plots of outstate tuition for private and public schools:



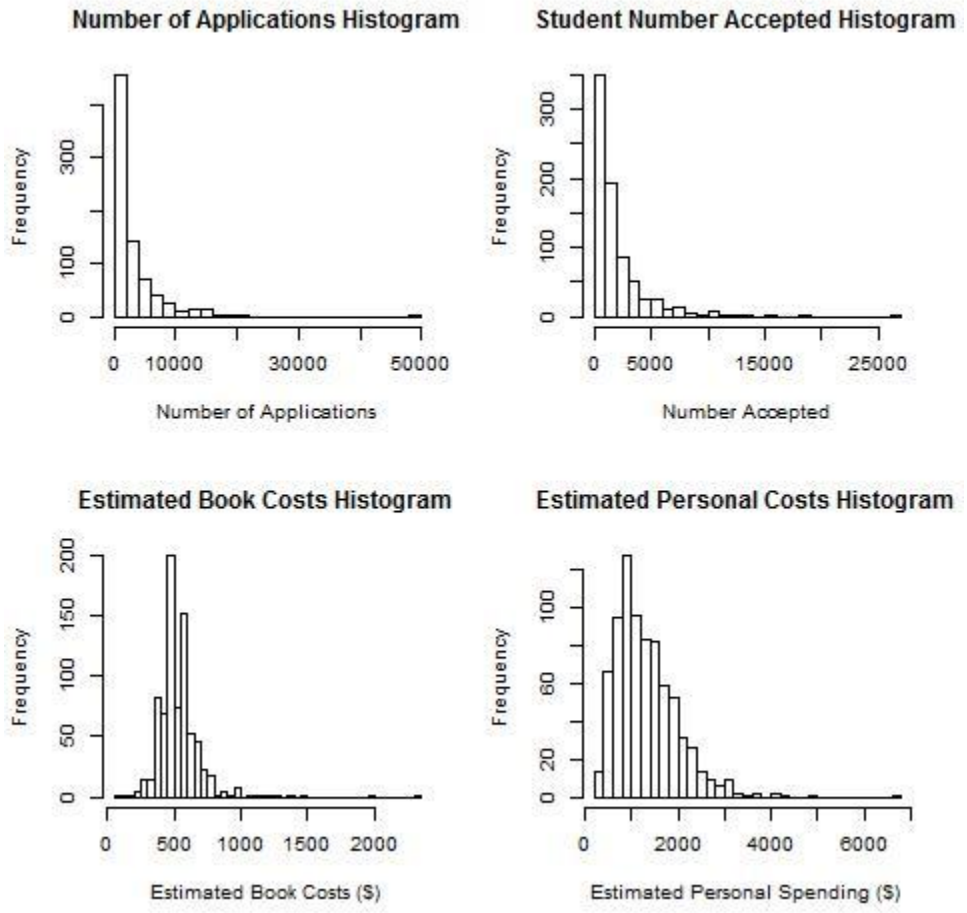
Short summary: as expected, private schools have a higher outstate tuition.

There are 78 elite universities. Side by side box plots of outstate tuition for elite and non-elite universities:



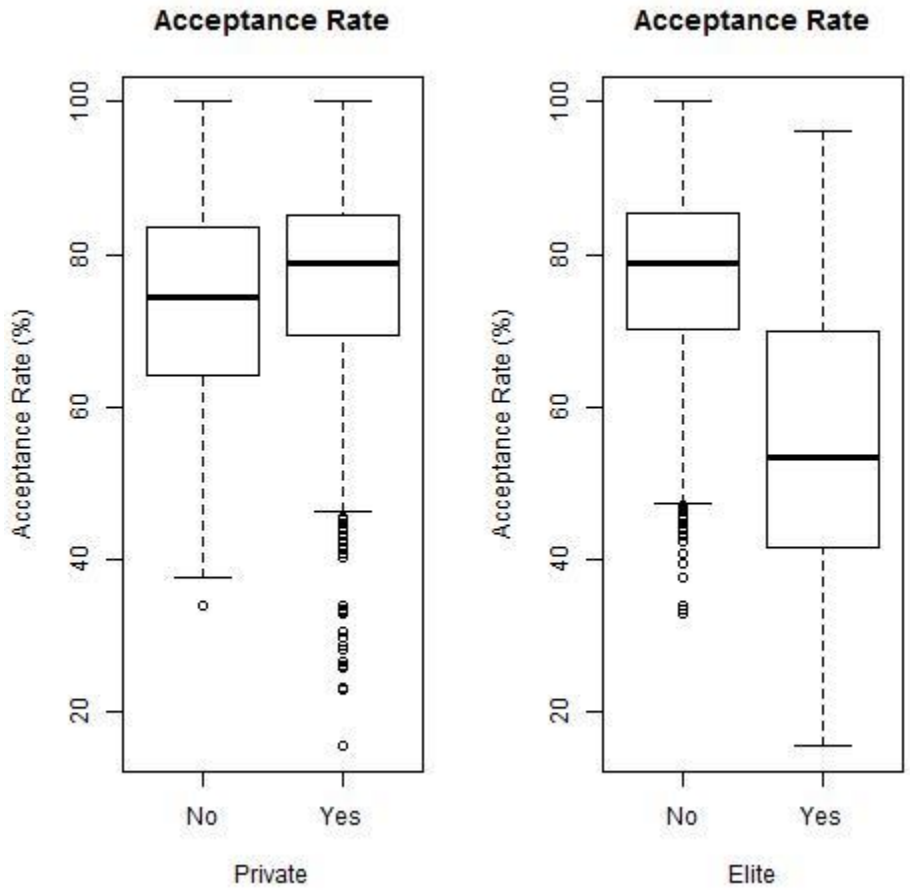
Short summary: elite schools clearly have a higher outstate tuition (since they are elite, they can probably afford to charge more, and attract students from out of state).

Some histograms exploring the data:



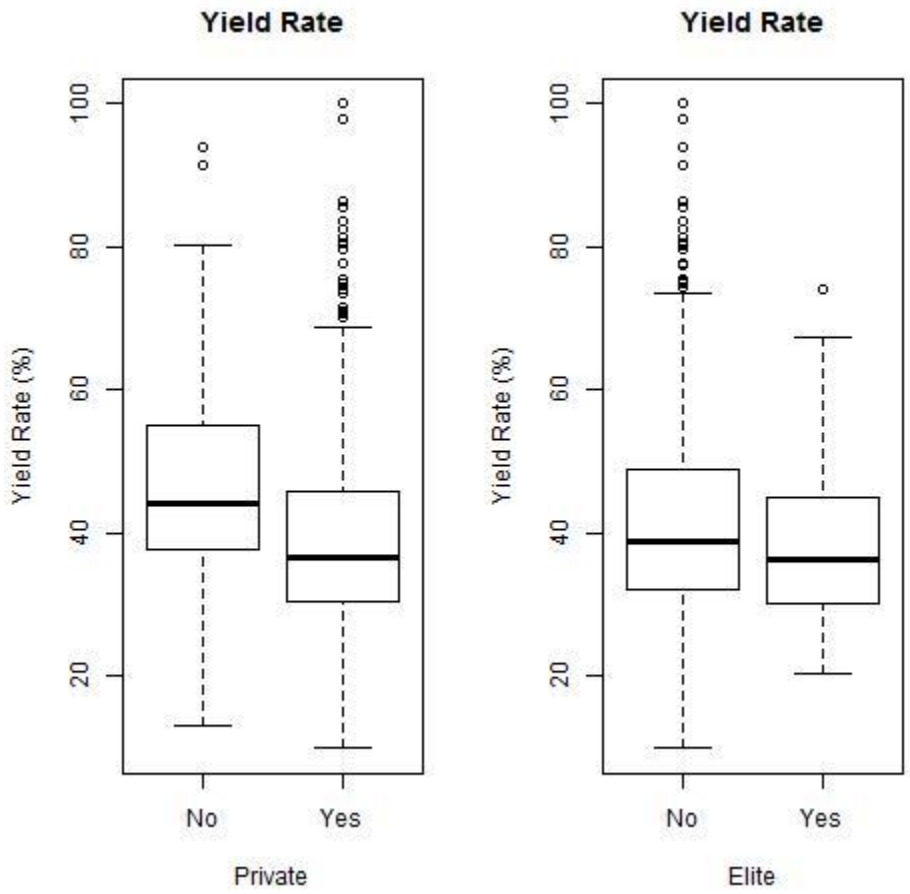
Short summary of finding: number of applications and number accepted drops of seemingly exponentially with a long tail, suggesting high outliers. Estimated book and personal costs are more or less normal, averaging around 500 and just above 1000 dollars respectively; again there are some high end outliers.

Finally, here are plots of acceptance rate at elite/not elite schools, as well as private/public schools:



Summary of findings: by looking at the initial scatter plot matrix we see a strong positive correlation between application number and acceptance number; but how do these numbers depend on whether the college is private or not? As we see here, private colleges on average actually have a higher acceptance rate (unexpected) – suggesting perhaps that students self-assort before applying. It is also notable that private colleges have more low outliers. When we look at acceptance rate and elite colleges, we see that it is actually much lower than for non-elite colleges, suggesting that private and elite school are not necessarily the same caliber!

Furthermore, here are plots of yield rate at elite/not elite schools, as well as private/public schools:



Summary of findings: after looking at acceptance rate, we can also see a strong positive correlation between number of accepted students and number of enrolled students in the initial scatter plot matrix. From this we can compute a yield. It this yield different for private vs public schools? Interestingly public schools have a higher yield (meaning more students accept their acceptance). Next it would be interesting to see how the yield differs from elite to non-elite schools. We would expect elite schools to have a higher yield rate as they are more prestigious – however we actually see a lightly lower yield rate!