Homework 1

Andrew Boschee

No collaborators

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.

Answer - The CEO's salary is a continuous variable and will be considered a regression and inference problem. n = 500, p = 3 (profit, # employees, 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.

Answer - The outcome of success or failure is a classification and prediction. n = 20 (products) and p = 13 (price, budget, competitor price, and other ten variables).

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 % chang in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Answer - The rate change will be a regression problem and we are making a prediction of those rates. n = 52 (weeks in year) and p = 3 (% change in each market besides USD/Euro)

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.

Fraud detection (Inference, Classification) - Response: account is classified as fraudulent/non-fraudulent. Predictors: Transaction frequency(dates), transaction amounts, recipient account activity.

Election Outcome (Prediction, Classification) - Response: Elected or not elected. Predictors: approval ratings, survey results, view of political subject, experience.

Health Risk (Prediction, Classification) - Response: Whether at risk for event. Predictors: Family history, genetic links, personal habits(eating, drug use, exercise frequency),pre-existing conditions

These are three popular applications that come to my mind quite often. Each of these events can have classification in multiple ways. I left the health risk vague since there are so many possible applications to see whether someone is likely to have a condition. Politics are constantly making predictions about who is most likely to win an election or measuring their approval. Finally, working in finance industry, I know that fraud detection is a common use of machine learning in this area to identify unusual use of money and transactions.

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.

Housing Prices (Regression, Prediction) - Response: Price of house. Predictors: Location, square footage, number of bedrooms, year built, number of bathrooms, lot size.

Fantasy Football (Regression, Prediction) - Response: Projected points. Predictors: Catches, Yards, Touchdowns

Company Turnover (Regression, Inference) - Response: Retention rate of customers. Predictors: Time with company, amounts paid, subscription type, number of orders.

Again, from common interactions in life, housing prices, fantasy football, and retention rates seem pretty easy for anyone to relate to. Was a little harder to view the inference aspect but a simple one that came to mind was to find what factors are most likely to contribute to the retention rate of customers and can be applied to basically any company. The other two topics are similar in determining how much each predictor factors into the response variable and giving expected output from that model.

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

Networking - Having common interests and similarities can be seen in clusters. Social media platforms cluster users based on preferences and activity to form networks. Examples: Linkedin, Facebook, Dating Apps

Marketing - Similar to social media, similarities can be found by segments of people on purchasing habits or personal interests. Ads can be targeted to specific demographics depending on websites visited, channels watched, and many other habits. Examples: Amazon, Ebay, Almost any website...

Epidemiology - Spread of illnesses/diseases can be sorted geographically or by genetic characteristics. Analyzing the spreading of the flu this time of year. This is just one small example of health trends that can go all the way to a global scale to fight epidemics over time.

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 non-parametric approach)? What are its disadvantages?

A parametric model has finite number of parameters in comparison to a non-parametric model. This has pros and cons. While a non- parametric approach can be very accurate and precise to the training set, this is overfitting and will not work well when applied to non-training data. The non-parametric model will pick up much more 'noise' than the parametric model. Parametric models are much simpler in most cases than non-parametric models and will most likely be faster and easier to explain. Non-parametric models may require more data and take much longer to train and explain clearly to those unfamiliar with concepts.

4. Question 2.4.8 pg 54-55

This exercise realtes to the College data set, which can be found in the College.csv. It contains a number of variables for 77 different universities and colleges in the US.

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.

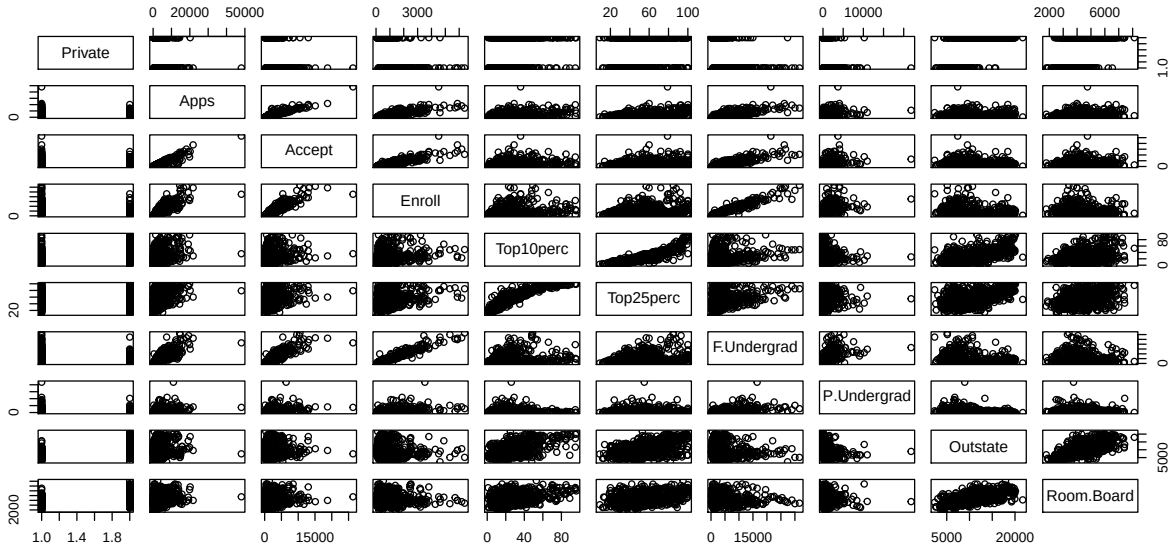
b. Look at the data using the fix() function. You should notice that the first column is just the name of each university. We don't really want R to treat this as data. However, it may be handy to have these names for later.

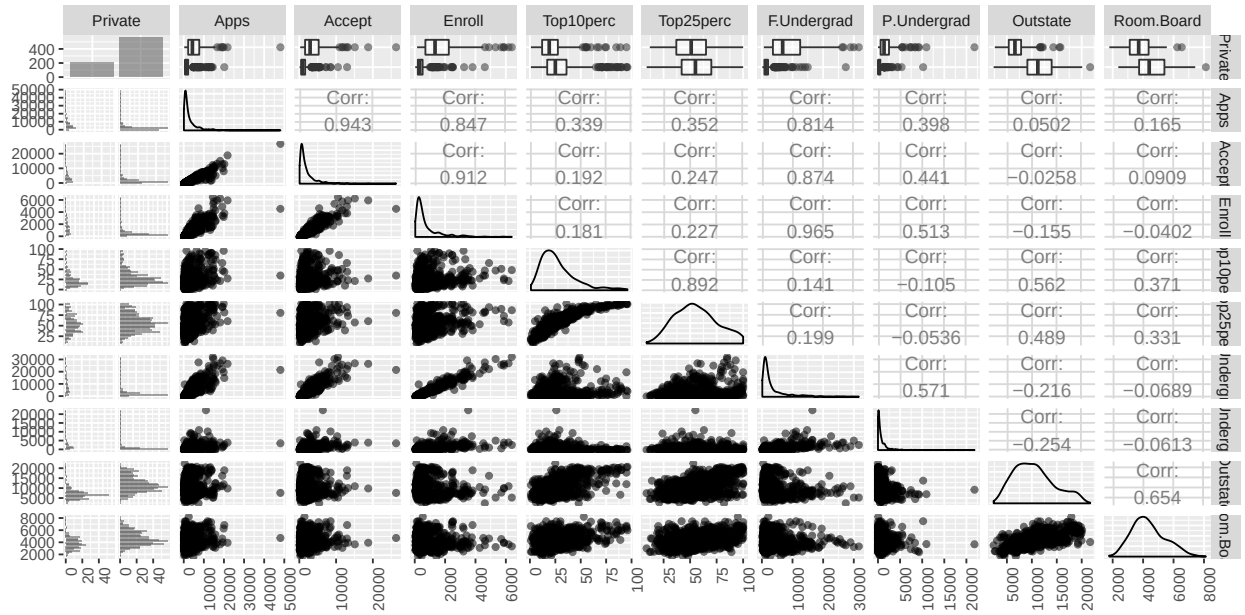
c(i) Use the summary function to produce a numerical summary of the variables in the data set.

```
##  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
```

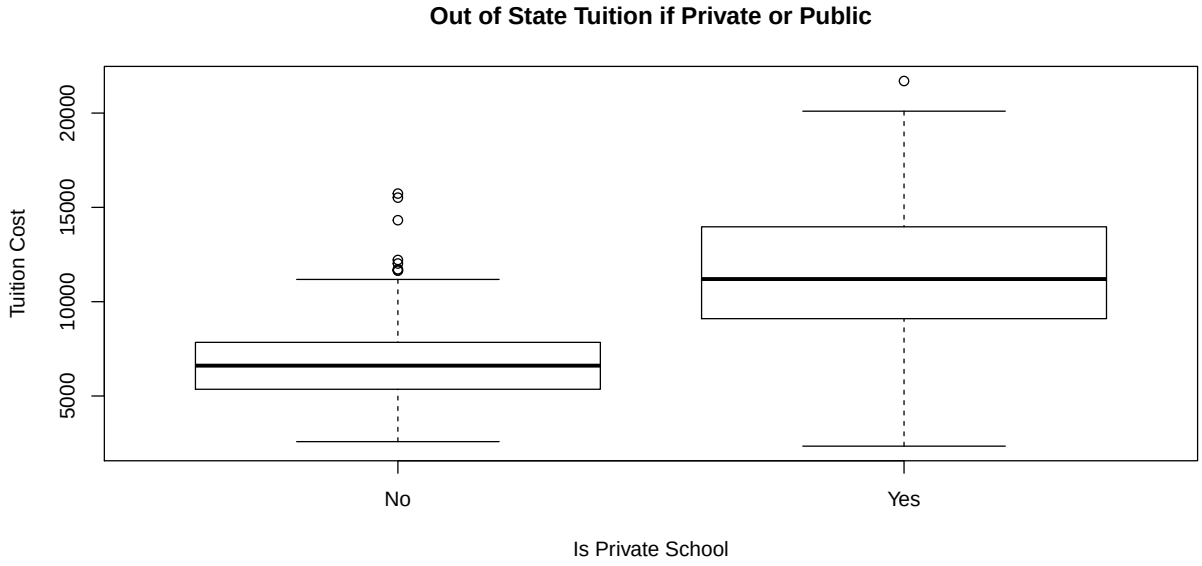
```
##  Median : 54.0   Median : 1707   Median :  353.0   Median : 9990
##  Mean   : 55.8   Mean   : 3700   Mean   :  855.3   Mean   :10441
##  3rd Qu.: 69.0   3rd Qu.: 4005   3rd Qu.:  967.0   3rd Qu.:12925
##  Max.   :100.0   Max.   :31643   Max.   :21836.0   Max.   :21700
##    Room.Board       Books           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
```

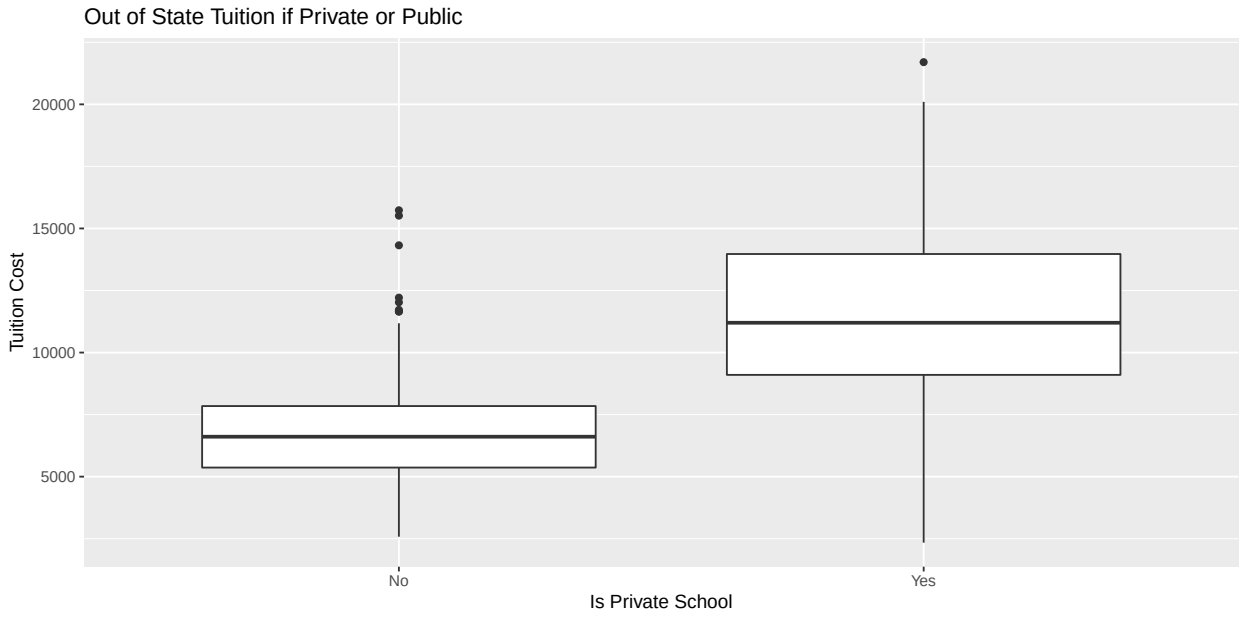
c(ii) use the pairs() function to produce a scatterplot matrix fo the first ten columns or variables.





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





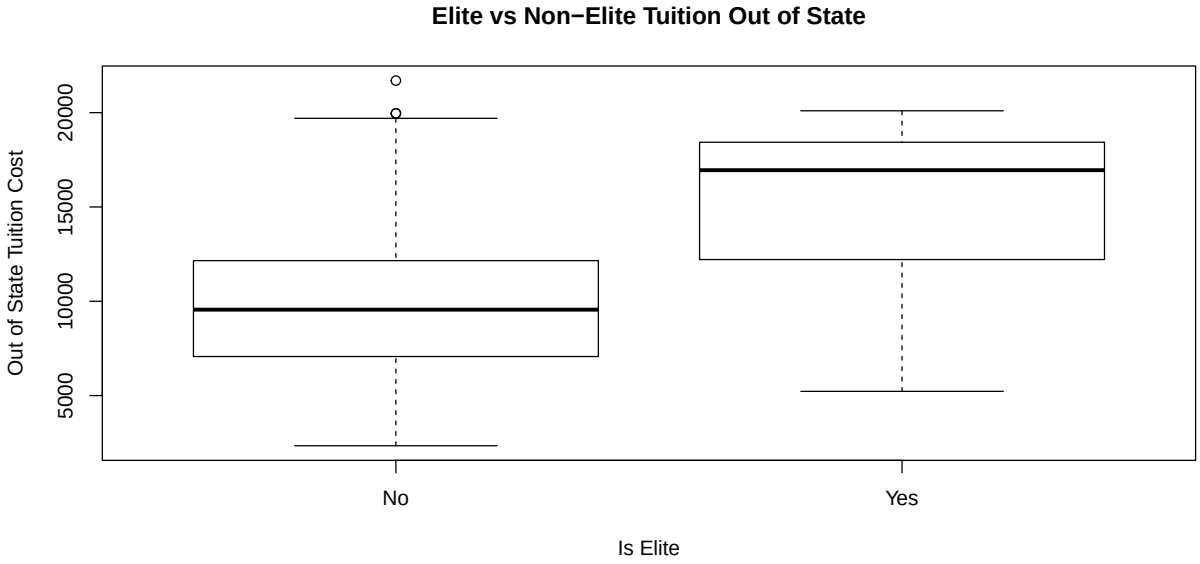
We can see that out of state tuition can be comparable for non-private schools in some instances but the average cost is much higher and the IQR is much broader for private schools.

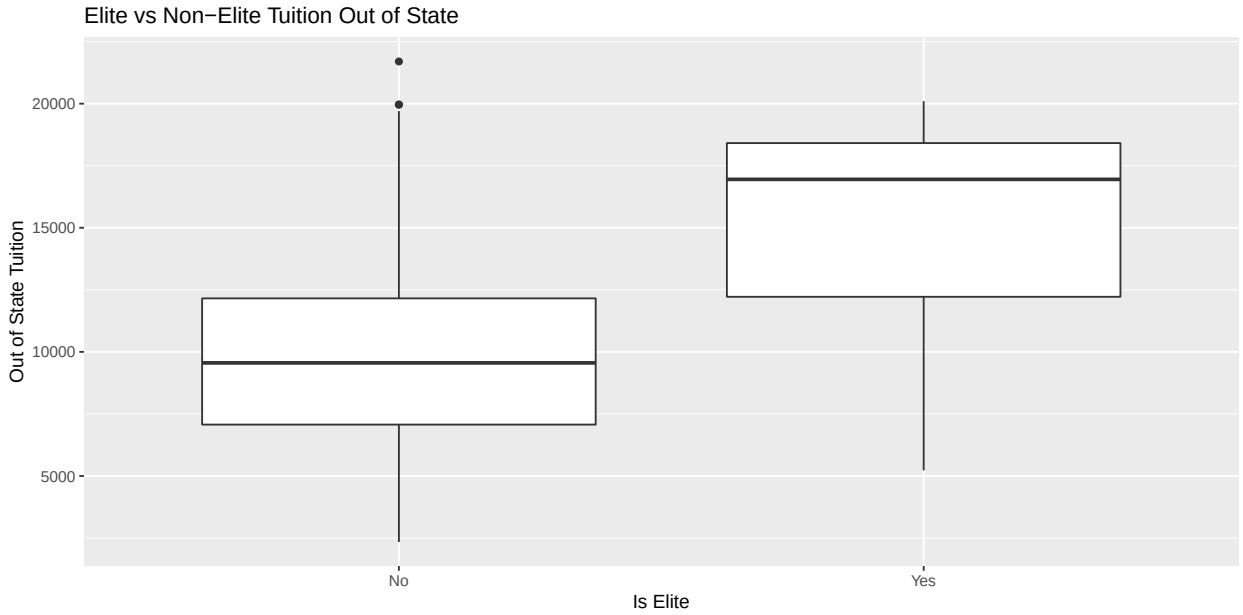
c(iv) Create new qualitative variable, called Elite, by binning the Top10perc variable. We are going to divide universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school classes exceeds 50%.

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.

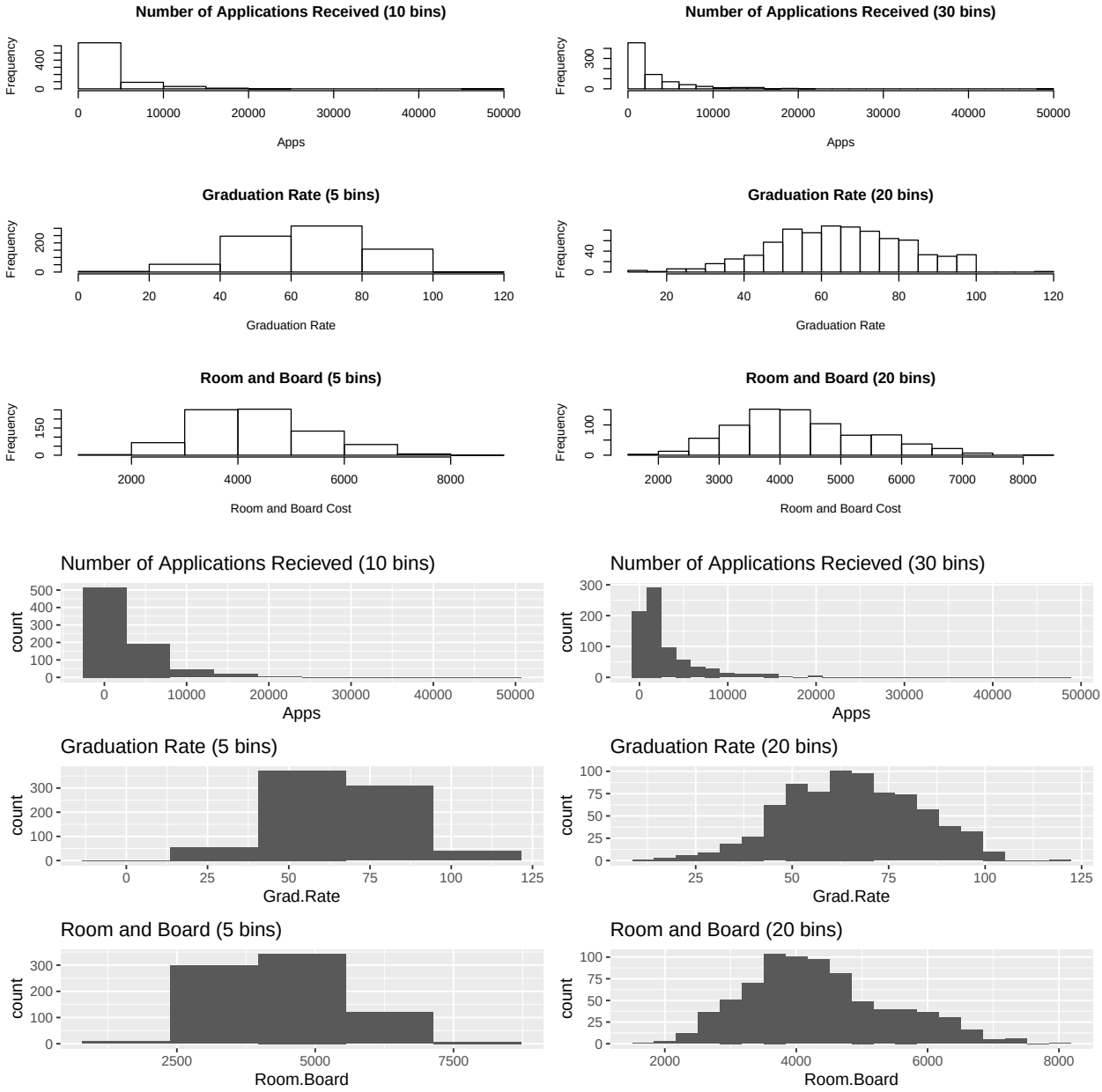
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
##            Median : 1558   Median : 1110   Median : 434   Median :23.00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





c(v) Use the hist() function to produce some histograms with differing numbers of bins for a few of the quantitative variables.



A couple quick take-aways - Graduation rate is above 100%. Second, distributions are righ-skewed in and is not hard to believe with the extreme costs of college and education. The above three variables have outliers and I'm sure that other variables do as well.

Colleges with 100% or Greater Graduation Rate Predictor

c(vi) Continue exploring the data, and provide a brief summary of what you discover

The first thing that stands out to me is the graduation rate plots. Clearly there is some inaccurate data or somebody misinterpreted the data being requested. Aside from exceeding 100 percent graduation rate, having 100 percent in general seems very unlikely. I believe that is worth exploring in this situation.

We can see that Cazenovia College is the school that had the excess of 100% from the summary of grad rates below.

Second, there are ten other schools with a 100 percent graduation rate. It brings to question if there was a formula that was just miscalcuated or personal error in gathering this data. An easy to notice fact is that all but one of the schools (Missouri Southern State College) in the group of ten are classified as private.

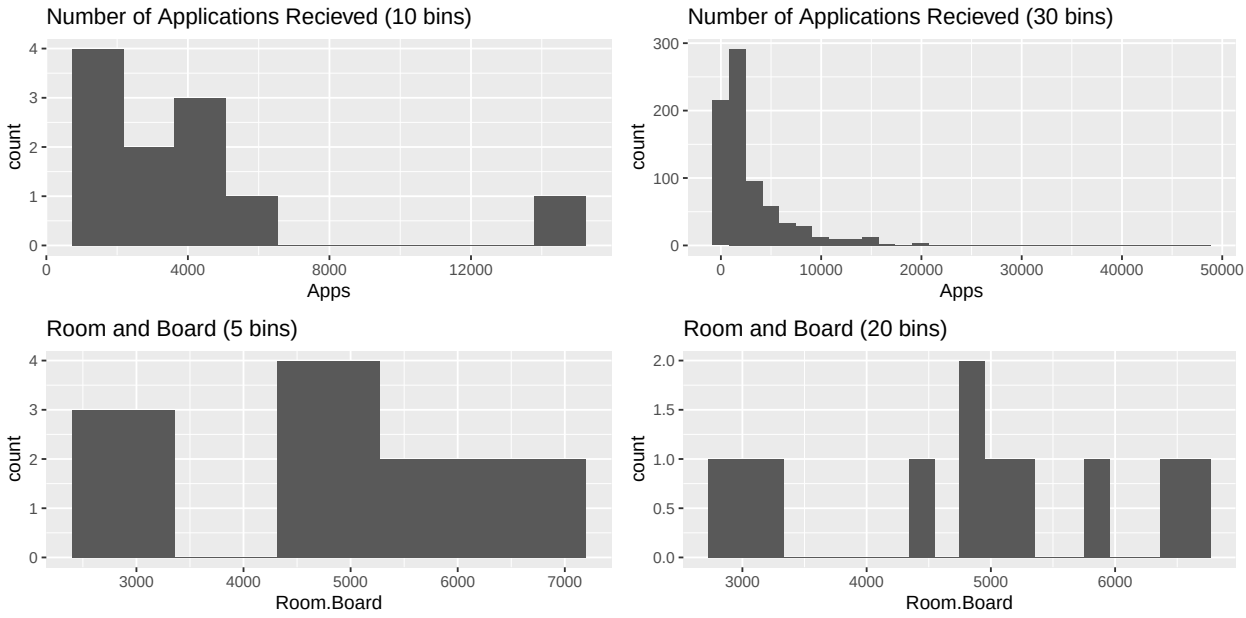
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   10.00   53.00   65.00   65.46   78.00  118.00
```

Table 1: Colleges with Graduation Rate = 100%

	Private	Apps	Accept	Enroll
Amherst College	Yes	4302	992	418
Cazenovia College	Yes	3847	3433	527
College of Mount St. Joseph	Yes	798	620	238
Grove City College	Yes	2491	1110	573
Harvard University	Yes	13865	2165	1606
Harvey Mudd College	Yes	1377	572	178
Lindenwood College	Yes	810	484	356
Missouri Southern State College	No	1576	1326	913
Santa Clara University	Yes	4019	2779	888
Siena College	Yes	2961	1932	628
University of Richmond	Yes	5892	2718	756

Table 2: College with Graduation Rate = 118%

	Private	Apps	Accept	Enroll
Cazenovia College	Yes	3847	3433	527



Using Linear Regression, I wanted to see what variables stood out as signficant when it comes to graduation rate. Below we can see that there are many that come into play. The ones that I believe should be focused on the most are Private status, Apps, p.Undergrad, and Outstate. I am a little surprised that Enroll and

Accept predictors aren't significant regarding graduation rate.

Table 3: Linear Regression - Predictor Significance Comparison

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	33.8925541	4.8522723	6.9848830	0.0000000
PrivateYes	3.4262050	1.7151733	1.9975853	0.0461186
Apps	0.0012936	0.0004428	2.9214052	0.0035881
Accept	-0.0006909	0.0008638	-0.7998820	0.4240297
Enroll	0.0021440	0.0023111	0.9277351	0.3538403
Top10perc	0.0465274	0.0851268	0.5465656	0.5848381
Top25perc	0.1374511	0.0564462	2.4350804	0.0151181
F.Undergrad	-0.0004648	0.0004026	-1.1545587	0.2486350
P.Undergrad	-0.0014809	0.0003907	-3.7903178	0.0001624
Outstate	0.0010197	0.0002339	4.3596337	0.0000148
Room.Board	0.0019067	0.0005926	3.2174441	0.0013484
Books	-0.0022140	0.0029189	-0.7584987	0.4483883
Personal	-0.0016620	0.0007703	-2.1576338	0.0312698
PhD	0.0882924	0.0571134	1.5459136	0.1225428
Terminal	-0.0751566	0.0624063	-1.2043112	0.2288452
S.F.Ratio	0.0746163	0.1595478	0.4676735	0.6401525
perc.alumni	0.2796432	0.0492353	5.6797292	0.0000000
Expend	-0.0004596	0.0001552	-2.9613830	0.0031583
eliteYes	0.4618984	2.5235781	0.1830332	0.8548210