Homework #1

Justin Robinette

January 15, 2019

No collaborators for any problem

Problem #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 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.

Results: This would be a **regression** and **inference** problem. The **n = 500 observations**. The **p = 3 variables:** *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 *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 10 other variables.

Results: This would be a **classification** and **prediction** problem. The **n = 20 observations**. The **p = 13 variables:** *price, marketing budget, competitor price, 10 other 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 % change in the USD/Euro, the % change in the US market, the % change in the British market, and the % change in the German market.

Results: This would be a **regression** and **prediction** problem. The **n = 52 observations (weeks)**. The **p = 3 variables:** *US market % change, British market % change, and German market % change*.

Problem #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.

Results: 1) Using the data from STAT 600 and 601 to predict whether or not a student would get an *A* in STAT 602 or not. This would be a **prediction** and **classification** problem because we are working with a qualitative response variable of whether or not an *A* grade will be earned or not and we are trying to predict what the outcome will be.

2) Exploring whether or not a voter voted for Donald Trump for president or not. This would be a **inference** and **classification** problem because we are again working with a binary qualitative variable of *yes* or *no* regarding the vote cast. It is an inference problem because we are not trying to predict anything; we are just trying to understand factors that influenced the voters' choice.

3) Using data on a person's health (height, weight, smoker/non-smoker, drinker/non-drinker, etc.) to predict whether or not someone will have a heart attack. This is a **prediction** and **classification** problem because we are working with a discrete qualitative response variable and attempting to predict if someone will have a heart attack based on the predictor variables.

(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.

Results: 1) Trying to predict what your house is worth using data gathered on other houses sold in your area. This would be a **prediction** and **regression** problem because we are attempting to predict a continuous quantitative variable of house price.

2) Analyzing average number of wins, by each college football team, over the last 50 years based on various factors (school location, conference, stadium size, etc.). This would be an **inference** and **regression** problem because we are analyzing the factors that contribute to the average number of wins (quantitative) to see if there are any highly correlated predictors.

3) Studying the change in average temperature, world-wide, over the past 10,000 years. This would be an **inference** and **regression** problem because, in my example, we are not predicting future temperatures. We would just be analyzing the continuous quantitative response variable to see if we could deduce anything about climate change from various predictor variables.

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

Results: 1) Segmenting prospective customers for marketing purposes would be an example of useful cluster analysis. This would allow you to target your marketing to customers based on the cluster in which your analysis places them. You could cluster customers based on demographic information (age, marital status, if they have children, estimated annual income, home ownership status, etc.), by previous purchase history, or by some other method.

2) Clustering can be used in social networking to place people into clusters of similar users for recommending people/companies to follow or connections to make. This would allow you to make the social network application more useful for the user by providing recommendations to topics that they may be interested in based on their 'cluster'.

3) You can use clustering for color compression in photos. You can take a photo that may have millions of colors represented and compress it down to only 10-20 colors based on clustering similar colors in the original image. These 10-20 colors can then be used to replace each of the original colors based on the cluster to which the color belongs to produce a new image. The resultant image can be incredibly similar to the original.

Problem #2.4.6 pg 53: Describe the difference between 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?

Results: The difference in parametric and non-parametric approaches is that with parametric approaches make assumptions about the distribution of the data. Non-parametric approaches do not make these assumptions.

One advantage of parametric approaches is that they reduce the problem of estimating the data down to one of estimating a set of parameters. It is generally easier to estimate the parameters than it is to fit an entire function as you would in a non-parametric approach. A potential disadvantage of the parametric approach is that, because of our assumptions, the model we choose will generally not match the true form of the unknown *f*. Fitting a more flexible model can help mitigate this problem because we can estimate more parameters. This can lead to another potential disadvantage known as *overfitting* which means that the model follows the data *too closely*.

Non-parametric approaches, as mentioned about, do not make assumptions about the function. They try, instead, to estimate the function that best fits the data. This is an advantage because it allows these

methods to fit a wider range of possible *f* shapes more accurately. Non-parametric approaches suffer from a possible disadvantage in that they require a larger number of observations than parametric approaches. When using non-parametric approaches, *overfitting* is also a concern as it is with parametric approaches.

Problem #2.4.8 pg 54-55: This exercise relates to the college data set, which can be found in the file **College.csv**. It contains a number of variables for the 777 different universities and colleges in the US.

Before reading the data into R, it can be vieiwed in Excel or a text editor.

Part 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.

Results: Here I've read in the CSV from the ISLR website (http://www-bcf.usc.edu/~gareth/ISL/data.html) and printed the header to begin examining the data set.

```
##                                X Private Apps Accept Enroll Top10perc
## 1 Abilene Christian University     Yes 1660   1232    721        23
## 2           Adelphi University     Yes 2186   1924    512        16
## 3               Adrian College     Yes 1428   1097    336        22
##   Top25perc F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD
## 1        52        2885         537     7440       3300   450     2200  70
## 2        29        2683        1227    12280       6450   750     1500  29
## 3        50        1036          99    11250       3750   400     1165  53
##   Terminal S.F.Ratio perc.alumni Expend Grad.Rate
## 1       78      18.1          12   7041        60
## 2       30      12.2          16  10527        56
## 3       66      12.9          30   8735        54
```

Part 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. Try the following commands:

Results: Now I've added row names to the original data set as instructed and printed the header to confirm.

```
##                                                    X Private Apps
## Abilene Christian University Abilene Christian University     Yes 1660
## Adelphi University                     Adelphi University     Yes 2186
## Adrian College                             Adrian College     Yes 1428
##                                Accept Enroll Top10perc Top25perc F.Undergrad
## Abilene Christian University   1232    721        23        52        2885
## Adelphi University             1924    512        16        29        2683
## Adrian College                 1097    336        22        50        1036
##                                P.Undergrad Outstate Room.Board Books
## Abilene Christian University         537     7440       3300   450
## Adelphi University                  1227    12280       6450   750
## Adrian College                        99    11250       3750   400
##                                Personal PhD Terminal S.F.Ratio perc.alumni
## Abilene Christian University     2200  70       78      18.1          12
## Adelphi University               1500  29       30      12.2          16
## Adrian College                   1165  53       66      12.9          30
##                                Expend Grad.Rate
## Abilene Christian University   7041        60
## Adelphi University            10527        56
## Adrian College                 8735        54
```

You should see that there is now a **row.names** column with the name of each university recorded. This means that R has given each row a name corresponding to the appropriate university. R will not try to perform calculations on the row names. However, we still need to eliminate the first column in the data where the names are stored. Try

```
##                               Private Apps Accept Enroll Top10perc
## Abilene Christian University     Yes 1660   1232    721        23
## Adelphi University               Yes 2186   1924    512        16
## Adrian College                   Yes 1428   1097    336        22
##                               Top25perc F.Undergrad P.Undergrad Outstate
## Abilene Christian University         52        2885         537     7440
## Adelphi University                   29        2683        1227    12280
## Adrian College                       50        1036          99    11250
##                               Room.Board Books Personal PhD Terminal
## Abilene Christian University       3300   450     2200  70       78
## Adelphi University                 6450   750     1500  29       30
## Adrian College                     3750   400     1165  53       66
##                               S.F.Ratio perc.alumni Expend Grad.Rate
## Abilene Christian University      18.1          12   7041        60
## Adelphi University                12.2          16  10527        56
## Adrian College                    12.9          30   8735        54
```

Now you should see that the first data column is **Private**. Note that another column labeled **row.names** now appears before the **Private** column. However, this is not a data column but rather a name that R is giving to each row.

Part C(i): Use the **summary()** function to produce a numerical summary of the variables in the data set.

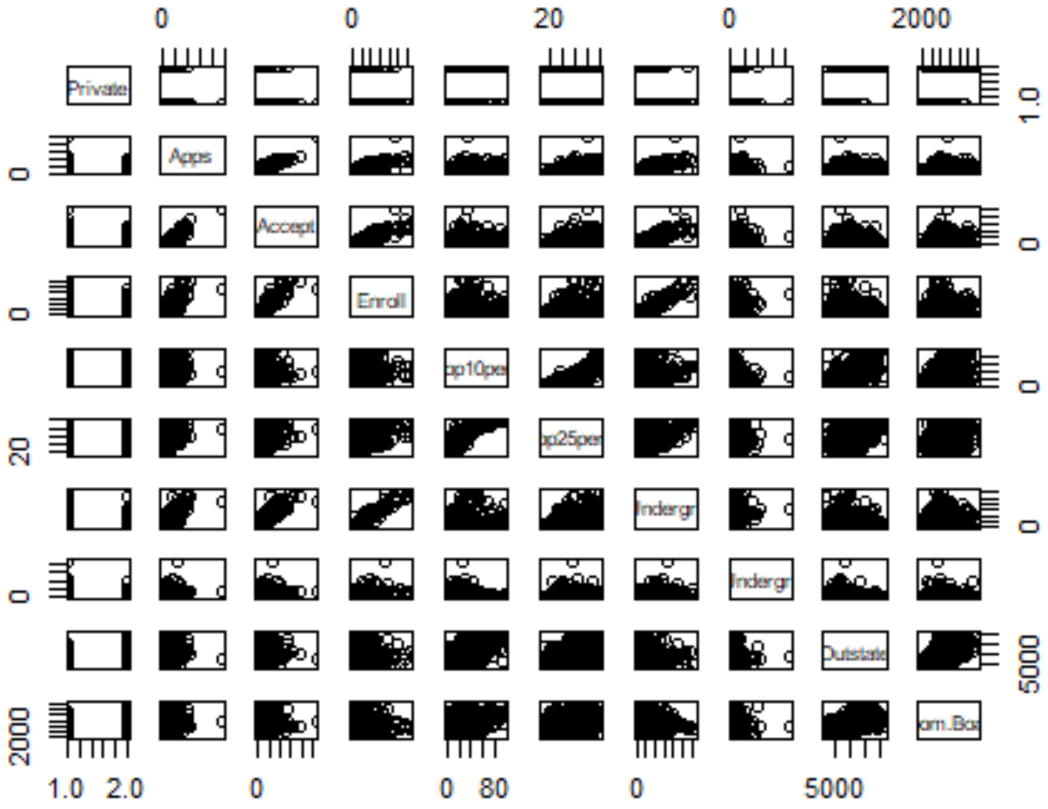
Results: Looking at a summary of the dataset, we can see that the 'x' variable from before (university name) is now gone. There are more private than public universities and no missing values.

```
##  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
```

Part C(ii): Use the **pairs()** function to produce a scatterplot matrix of the first ten columns or variables of the data. Recall that you can reference the first ten columns of a matrix *A* by using **A[,1:10]**.

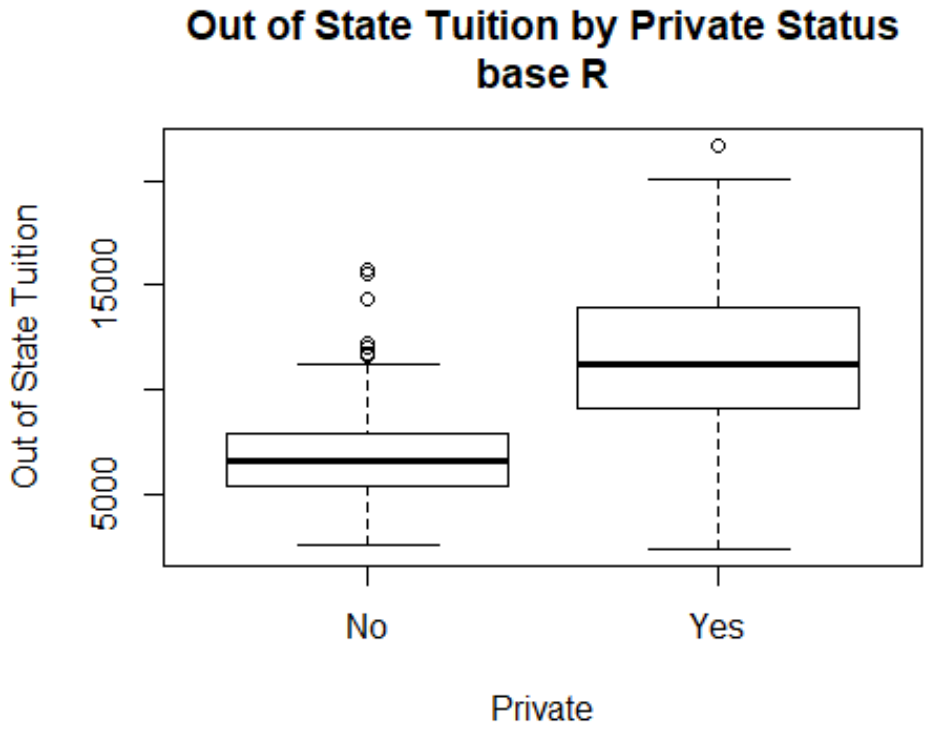
Results: Below we see the scatterplot matrix, as requested in the assignment, using the pairs() function. Per our assignment instructions, I used an extension of the ggplot2 package known as 'GGally'. Some parameters were changed to attempt to make the plot more readable, but this is hard to do as requested in the assignment (10 columns).

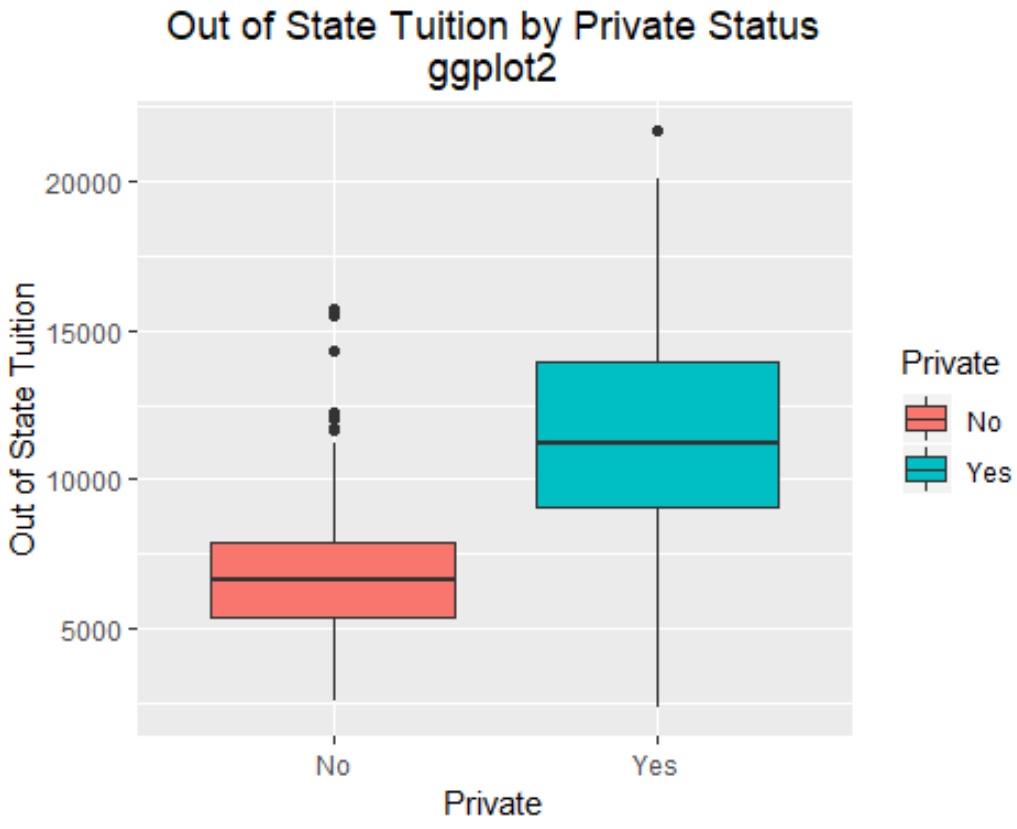
From the second plot, we can see that there are some highly correlated variables. For example (and as we would expect), the number of Accepted apps is highly correlated with the total number of apps (0.943) and the Top25perc with the Top10perc (0.892). Also we can see that the number of Top 25% students and Top 10% students is pretty highly correlated with the Outstate Tuition (0.562 & 0.489, respectively).


Private
Apps
Accept
Enroll
Outstate
Corr:
0.943
0.847
0.339
0.352
0.814
0.398
0.0502
0.165
0.912
0.192
0.247
0.874
0.441
-0.0258
0.0909
0.181
0.227
0.965
0.513
-0.155
-0.0402
0.892
0.141
-0.105
0.562
0.371
0.199
-0.0536
0.489
0.331
0.571
-0.216
-0.0689
-0.254
-0.0613
0.654

Part C(iii): Use the **plot()** function to produce side-by-side boxplots of **Outstate** versus **Private**.

Results: Below we have two plots, one in base R and one in ggplot. We can see that private universities have a higher range of tuitions, as well as typically a higher tuition. With that being said, we can see that the lowest tuition value appears to be a private university (also the highest tuition amount comes from a private university).



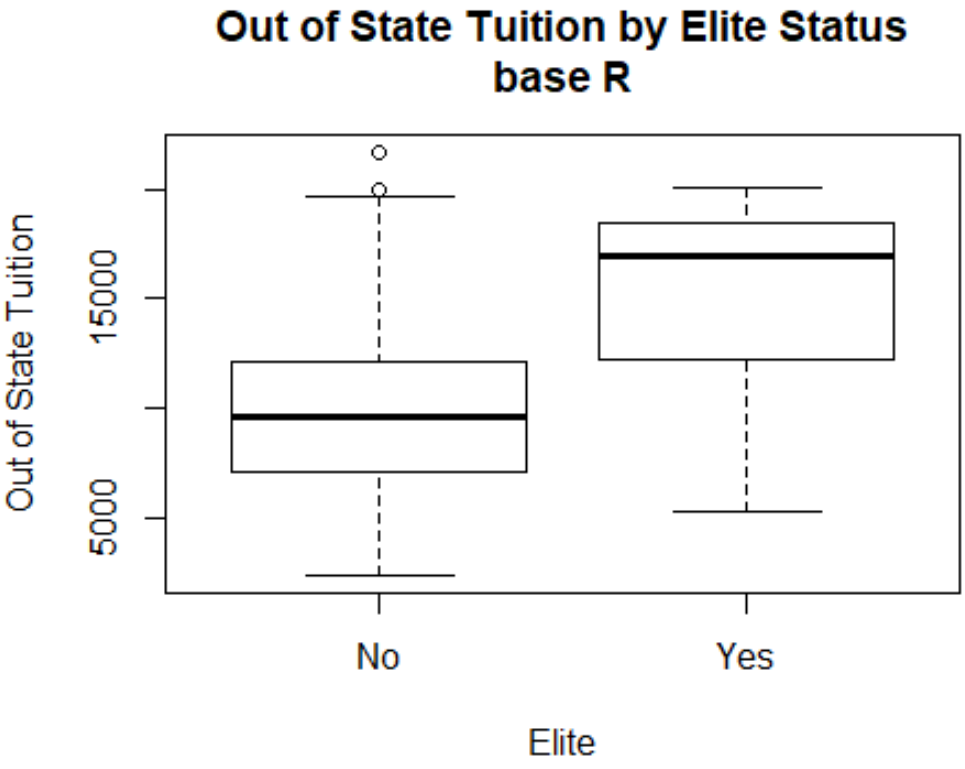


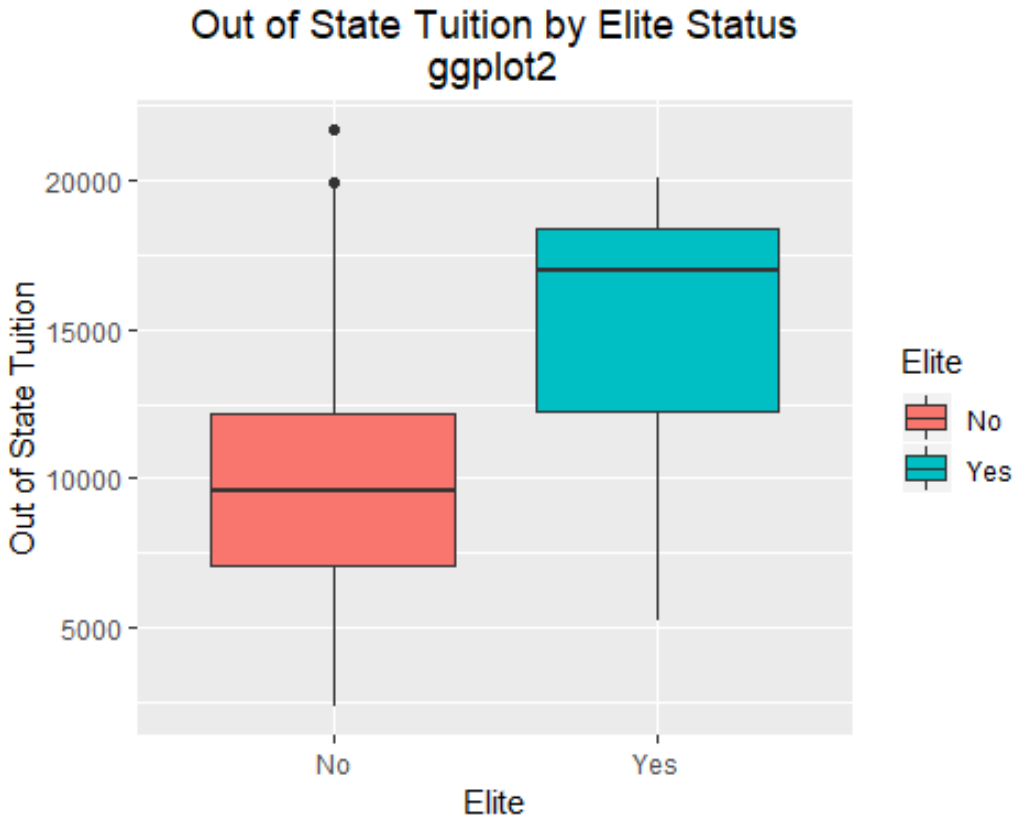
Part C(iv): Create a new qualitative variable, called **Elite**, by *binning* the **Top10perc** variable. We are going to divide the universities into two groups based on whether or not the proportion of students coming from the top 10% of their high school class exceeds 50%. Use the **summary()** function to see how many elite universities there are. Now use the **plot()** function to produce side-by-side boxplots of **Outstate** versus **Elite**.

Results: Here, I've added the 'Elite' variable to the college data set as instructed. We print a summary to confirm that Elite is now a qualitative variable in the data set.

Next, there are 2 plots, one in base R and one in ggplot. These boxplots show the Out of State Tuition by 'Elite' status. We see that the non-elite universities have a higher range of tuition amount and, generally, a lower out of state tuition amount. With that being said, we should also notice that the highest out of state tuition amount comes from a non-elite university. I've extracted this observation below.

```
##  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
```





Maximum Outstate Tuition School

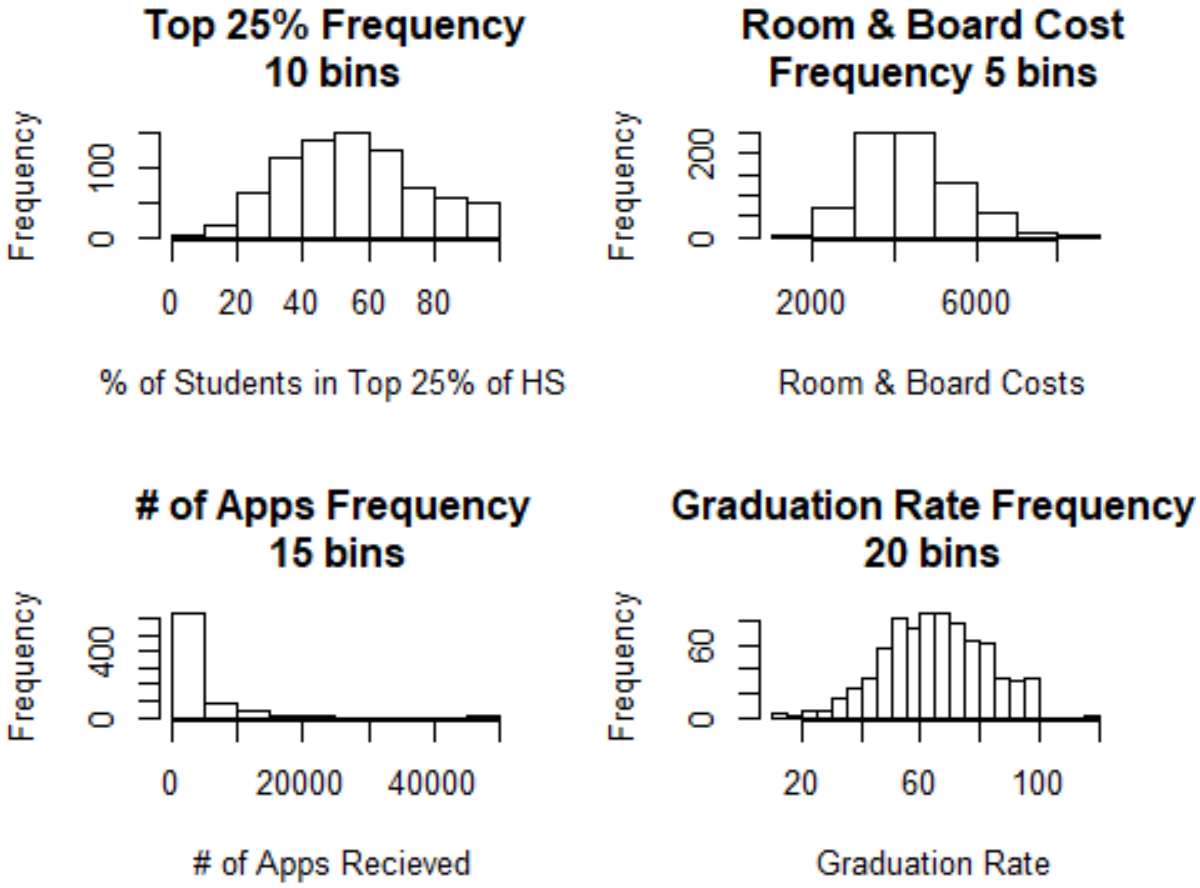
	Outstate	Elite
Bennington College	21700	No

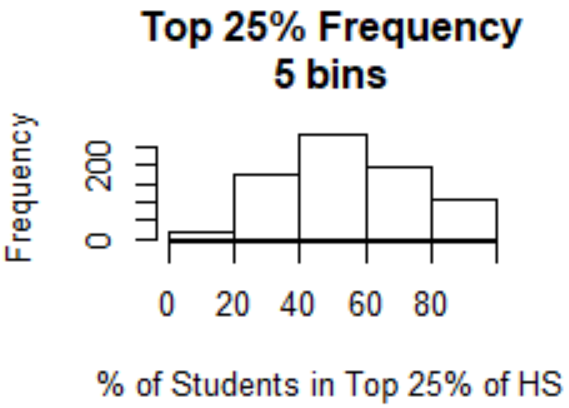
Part C(v): Use the **hist()** function to produce some histograms with differing numbers of bins for a few of the quantitative variables. Use **par(mfrow=c(2,2))**.

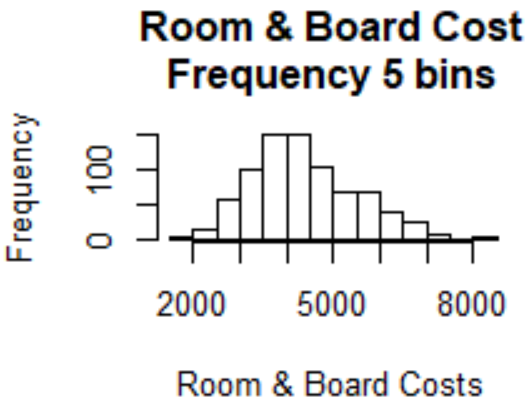
Results: Using the hist() function, and corresponding geom_histogram() functionality of ggplot, I examined 4 different variables using differing numbers of bins, per the instructions.

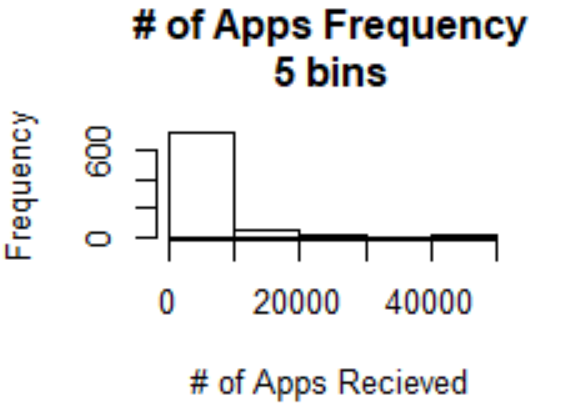
As we might expect, the distribution for the Top25perc variable is relatively normal. The # of Applications received variable is heavily skewed but it appears this may be due to a lack of reporting as there are many '0' values listed. Also interesting is that we see that some schools are reporting a graduation rate of higher than 100% which indicates that there are some errors in the data - or possibly students were given credit for graduating twice for advanced degrees or double majors. Either way, this is something we would want to examine deeper when analyzing this data set.

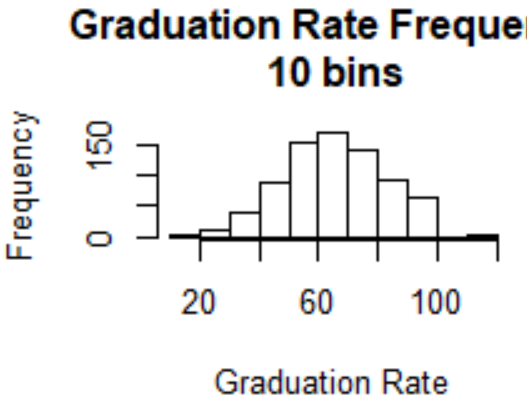
To compare the effects of changing the bins I printed the histograms for the 4 variables using differing numbers of bins. The plot most affected by changing the bin numbers was the # of Apps Received histogram.

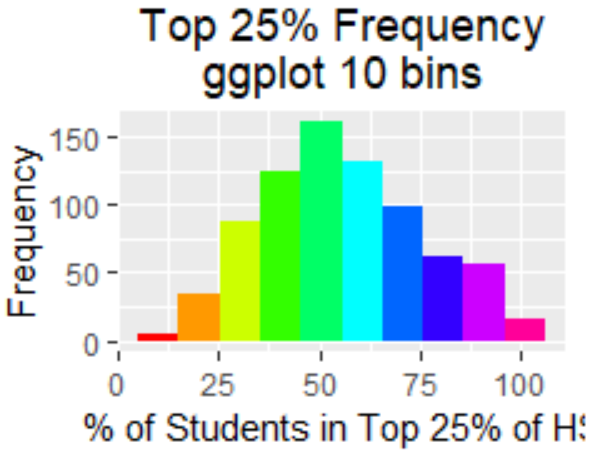


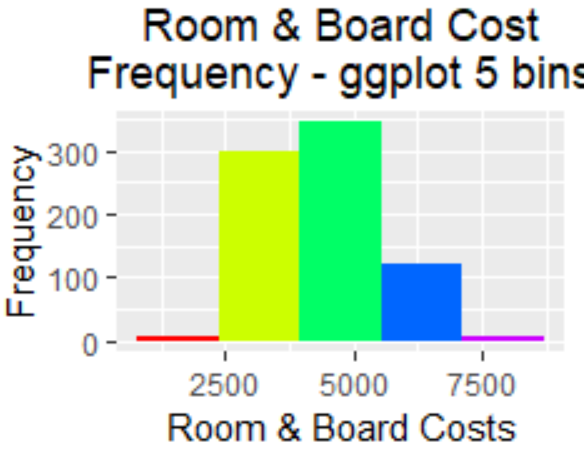

Top 25% Frequency
5 bins
Frequency
0
200
0
20
40
60
80
% of Students in Top 25% of HS

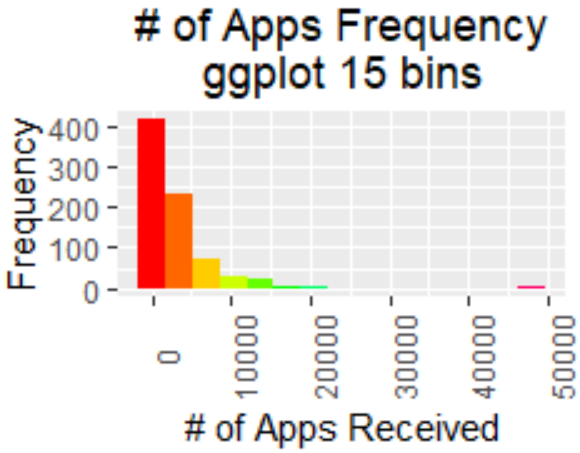

Room & Board Cost
Frequency 5 bins
Frequency
0
100
2000
5000
8000
Room & Board Costs

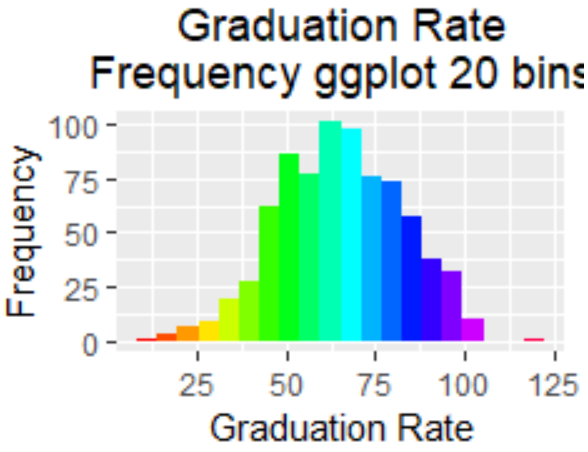

# of Apps Frequency
5 bins
Frequency
0
600
0
20000
40000
# of Apps Recieved

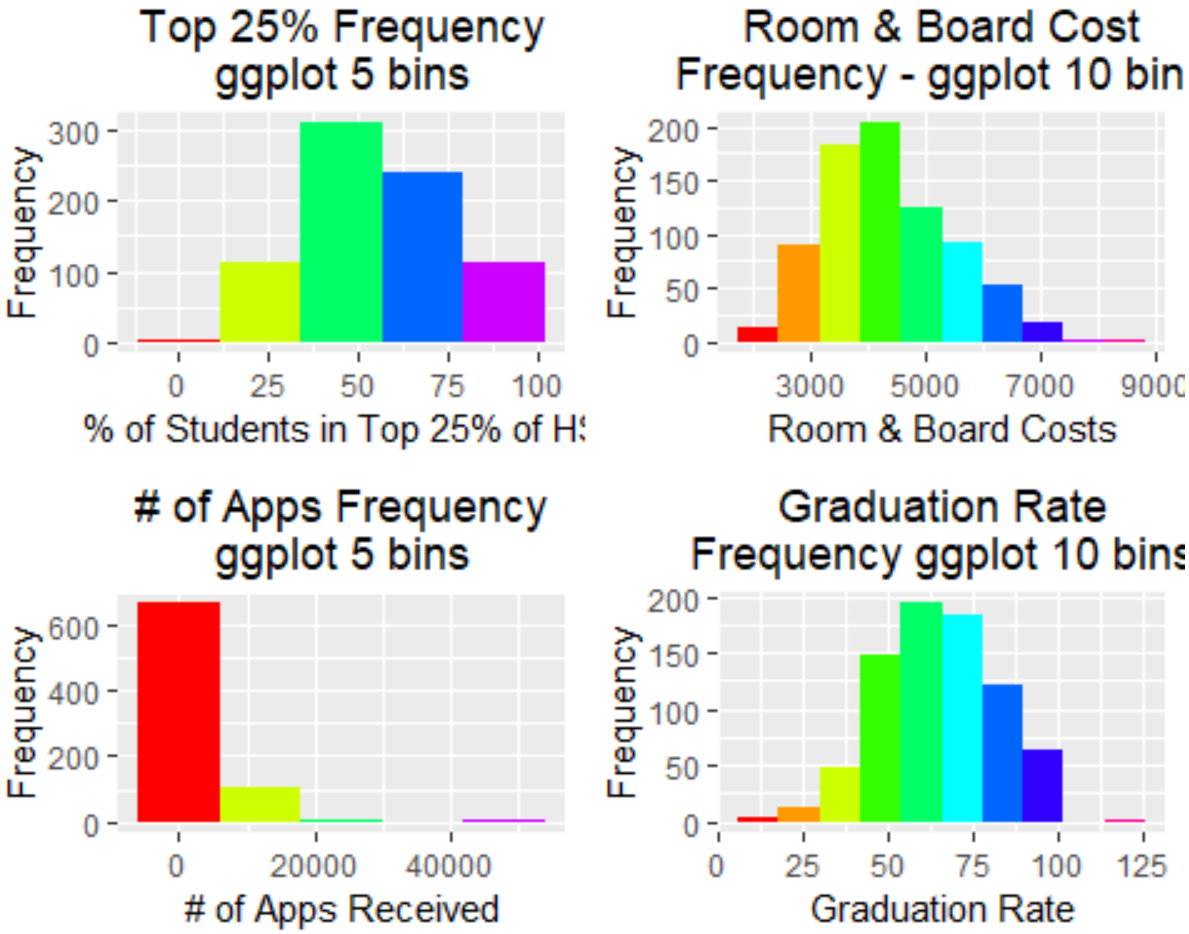

Graduation Rate Frequency
10 bins
Frequency
0
150
20
60
100
Graduation Rate


Top 25% Frequency
ggplot 10 bins
Frequency
0
50
100
150
0
25
50
75
100
% of Students in Top 25% of HS


Room & Board Cost
Frequency - ggplot 5 bins
Frequency
0
100
200
300
2500
5000
7500
Room & Board Costs


# of Apps Frequency
ggplot 15 bins
Frequency
0
100
200
300
400
0
10000
20000
30000
40000
50000
# of Apps Received


Graduation Rate
Frequency ggplot 20 bins
Frequency
0
25
50
75
100
25
50
75
100
125
Graduation Rate



Part C(vi): Continue exploring the data, and provide a brief summary of what you discover.

Results: My first curiousity, from the above summaries, is that there appears to be some schools with a graduation rate greater than or equal to 100%. Printing the list, we can see 11 schools report a graduation rate in this range. Notice, Casenovia College reports a graduation rate in excess of 100%. As I mentioned above, maybe they are reporting students who double major as graduating twice or maybe this is an error. Either way, we'd want to examine it further or we may want to exclude the school due to this anomaly.

I also notice that there were some high Faculty PhD rates. The next table shows the schools reporting that 100% or more of their faculty have PhD's. Again, this is something that I'd want to examine further to determine why Texas A&M is reporting a greater than 100% value.

Next, I had noticed that the Apps variable appears to have some large values, relative to the mean. We can see from the table the 10 schools who received the most Applications. Two takeaways, for me, are that Rutgers has more than double the 2nd highest school. Also, I see that 6 of the top 10 are members of the Big 10 Conference. From what I know about this conference's research prowess, it's not terribly surprising to me that they're well represented on here.

Next, we have a base R plot and ggplot exploring the Graduation Rate vs. the Out of State Tuition Rate. I notice that there does seem to be a reasonable amount of correlation between a school's tuition rate and their graduation rate.

Next, we have a base R plot and analogous ggplot looking at the % of students that were in the top 10% of their High School class versus the Graduation Rate. Again, I see a reasonable correlation between these two variables.

Schools with >= 100% Graduation Rate

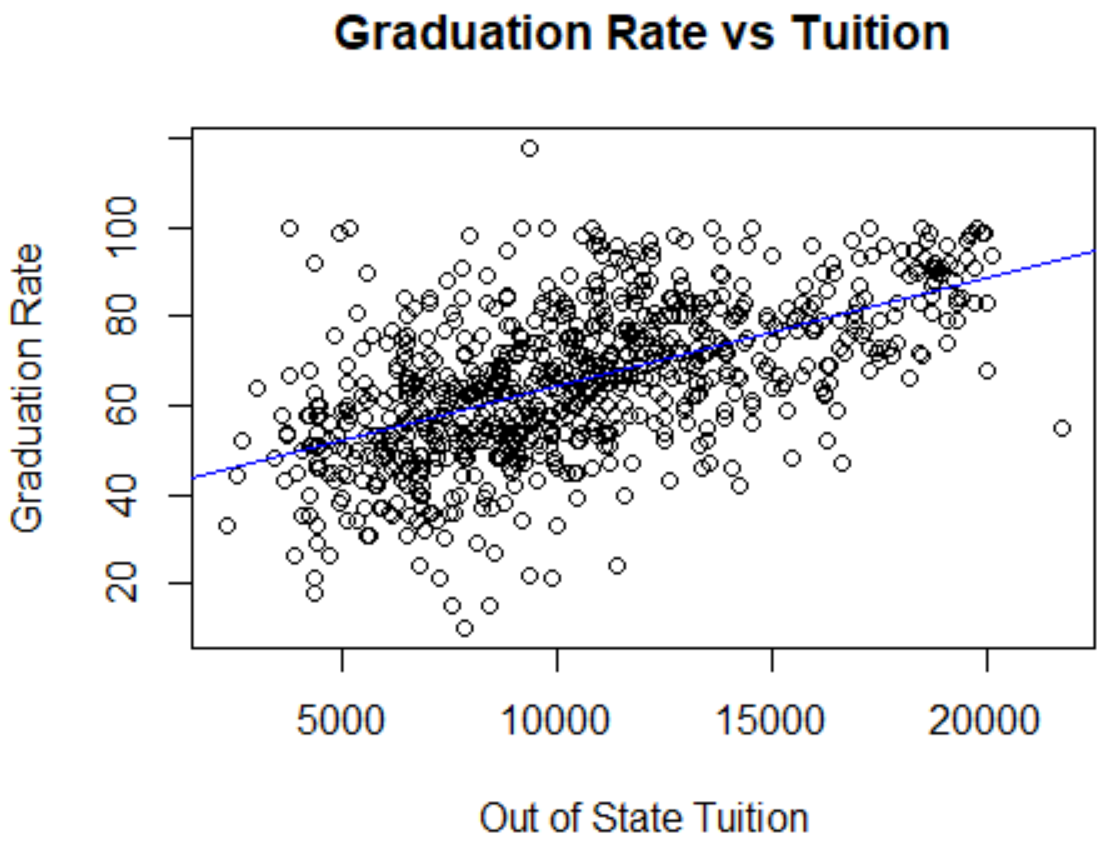
	Grad.Rate
Amherst College	100
Cazenovia College	118
College of Mount St. Joseph	100
Grove City College	100
Harvard University	100
Harvey Mudd College	100
Lindenwood College	100
Missouri Southern State College	100
Santa Clara University	100
Siena College	100
University of Richmond	100

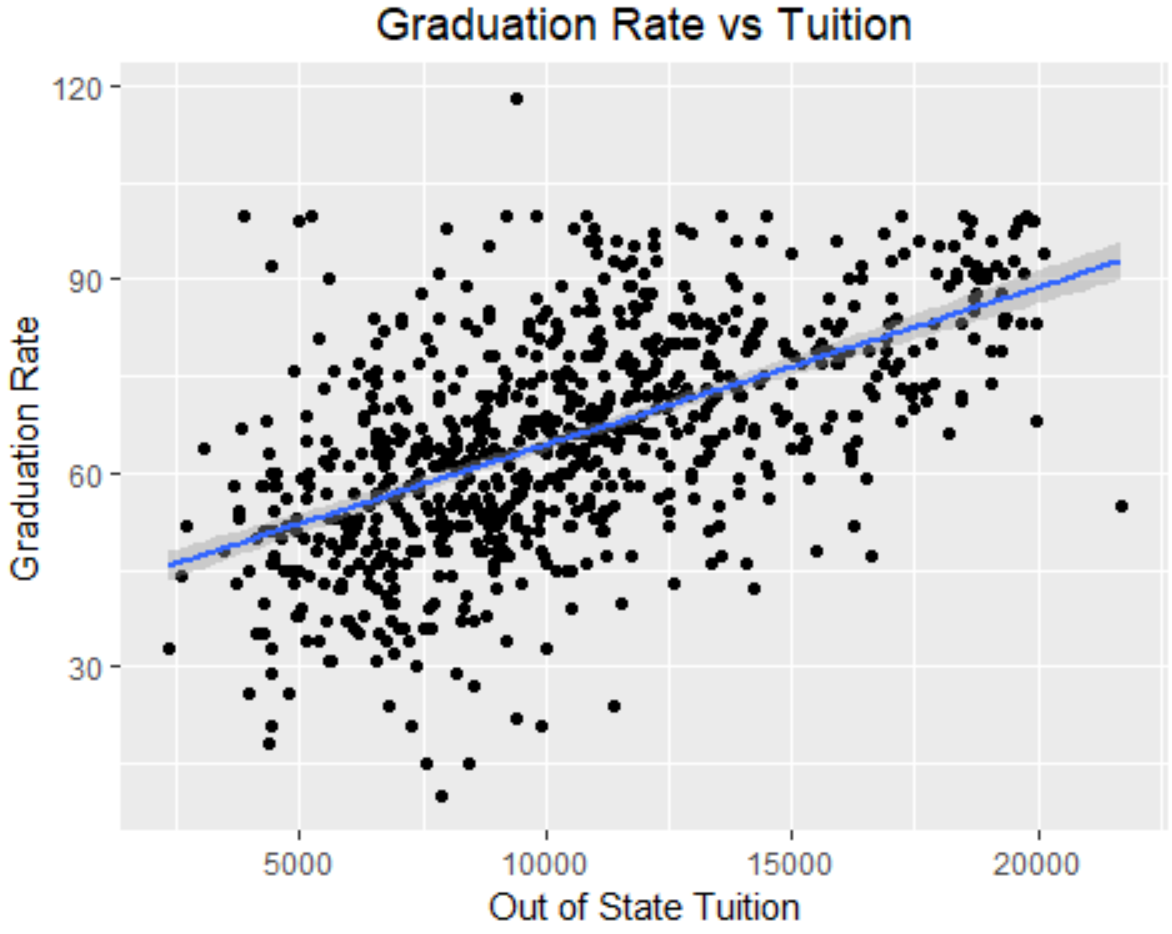
Schools with >= 100% PhD Faculty

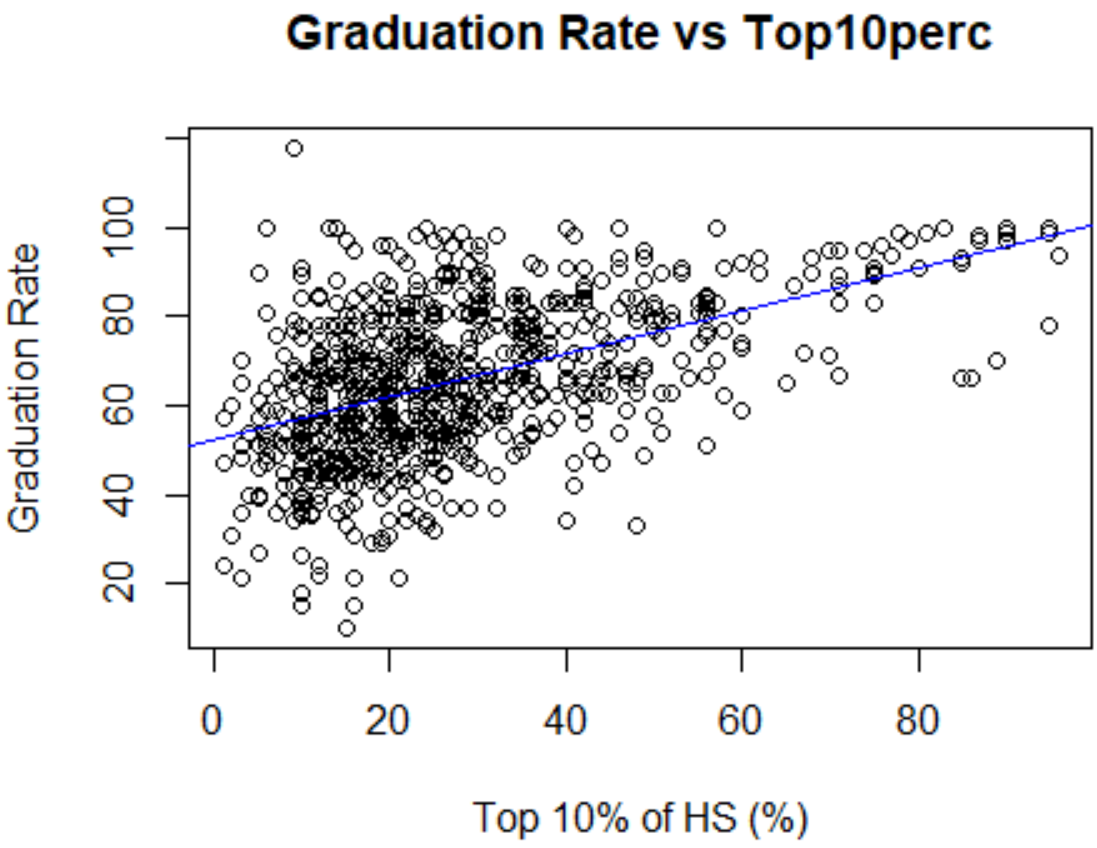
	PhD
Bryn Mawr College	100
Harvey Mudd College	100
Pitzer College	100
Texas A&M University at Galveston	103

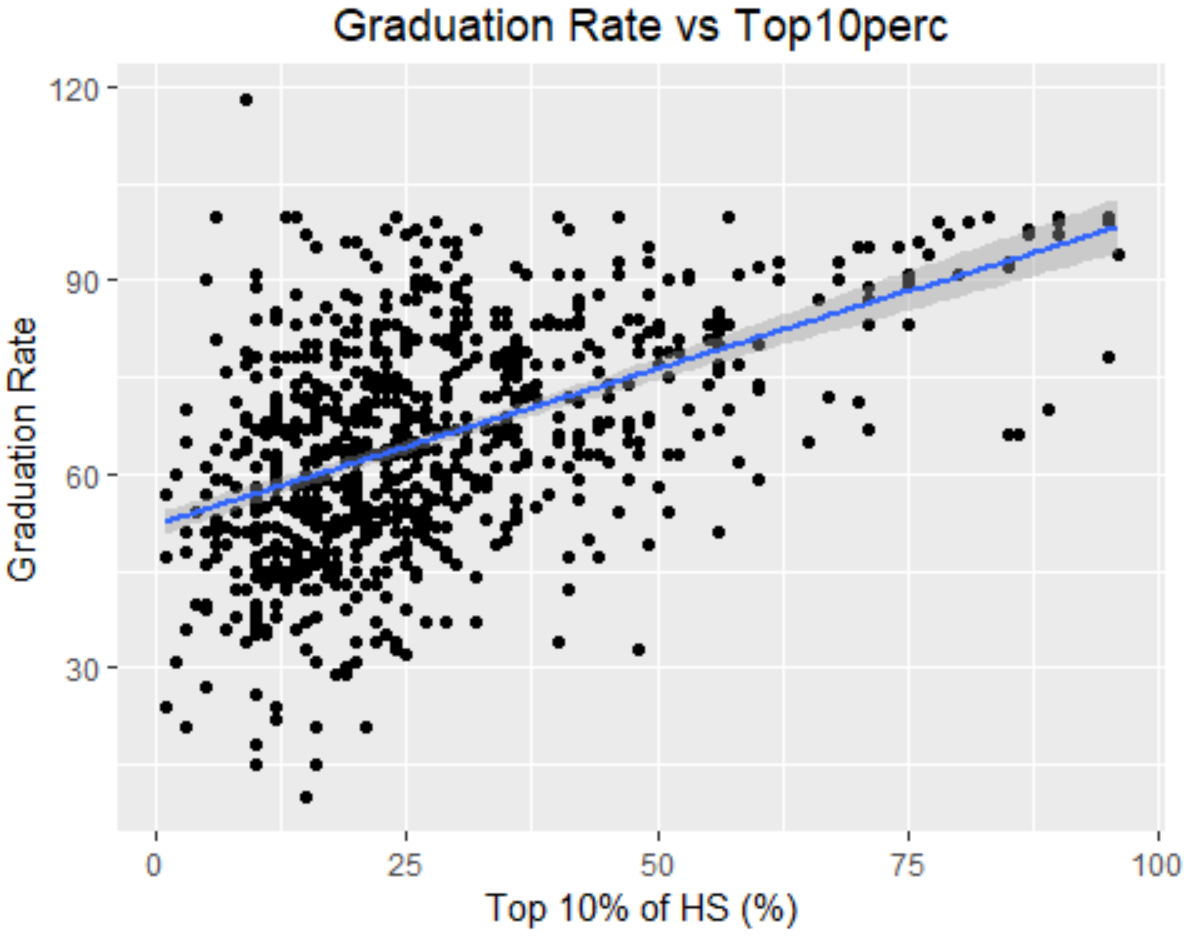
Schools with Most Applications

College	Apps
Rutgers at New Brunswick	48094
Purdue University at West Lafayette	21804
Boston University	20192
University of California at Berkeley	19873
Pennsylvania State Univ. Main Campus	19315
University of Michigan at Ann Arbor	19152
Michigan State University	18114
Indiana University at Bloomington	16587
University of Virginia	15849
Virginia Tech	15712


Graduation Rate vs Tuition
Graduation Rate
20
40
60
80
100
5000
10000
15000
20000
Out of State Tuition


Graduation Rate vs Tuition
Graduation Rate
120
90
60
30
5000
10000
15000
20000
Out of State Tuition


Graduation Rate vs Top10perc
Graduation Rate
20
40
60
80
100
0
20
40
60
80
Top 10% of HS (%)


Graduation Rate vs Top10perc
Graduation Rate
120
90
60
30
0
25
50
75
100
Top 10% of HS (%)