Homework 1

Segbehoe, Lawrence Sethor

January 11, 2019

Collaboration: none

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.

(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.

(c) We are interest 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.

Solution to Question 1

(a) The scenario described is **regression** because the response variable, CEO salary, is continuous or a quantitative variable. **Inference** because the interest is in understanding which factors (variables) affect CEO salary as opposed to determining CEO salary given a set of factors. $n = 500$ and $p = 3$.

(b) The scenario described is **classification** because the response variable, *success or failure* of a new product, binary/qualitative. **Prediction** since the interest is to know whether the new product will be a *success or failure.* $n = 20$ and $p = 13$.

(c) The scenario described is **regression** since the response variable, % change in the USD/Euro exchange rate, is quantitative/continuous. **Prediction** because the interest is predicting the % change in the USD/Euro exchange rate. $n = 52$ and $p = 3$.

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.

(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.

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

Solution to Question 2

(a) **Classification**

i. Iris flower is an ornamental plant with sword-shaped leaves and showy flowers typically purple, yellow or white. 50 samples each of iris flower species we collected. The species are iris setosa, verisicolor and virginica. For each of these samples the following measurements were taken: sepal length, sepal width, petal length and petal width. The interest is to predict which species a sample belong using the measurements taken. $n = 150$ and $p = 4$. The response is the species and the predictors are sepal length, sepal width, petal length and petal width. The goal is prediction since the interest is to determine which species (class) each sample belong using the measurements taken.

ii. Vole is a small, typically burrowing, mouse-like rodent with a rounded muzzle, found in both Eurasia and North America. Two species of this vole are *Microtus multiplex* and *Microtus subterraneus.* Skulls of 288 specimen from voles found at various places in central Europe collected. For 89 of the skulls, the chromosomes were analyzed to identify their species; $N_1 = 43$ specimens were from *Microtus multiplex* and $N_2 = 46$ from *Microtus subterraneus.* Eight different kinds of measurements were taken from each skull to give 8 variables. Species was not determined for the remaining 199 specimens. The interest here is predict the species of each of the 89 skulls analyzed based on the 8 variables. The response is the species and the predictors are the eight variables. The goal of is prediction since the interest is to determine the kind of species of each sample.

iii. In Bioinformatics, 4189 RNA gene levels from 180 samples of cancer tumor were read. Each of the 80 samples comes from either a male or female. The interest is to predict whether a sample comes from a male or a female based on the 4189 RNA gene levels. The response gender (male or female) and predictors are 4189 RNA gene levels. The goal is prediction since the interest is to predict gender of a sample based on the genes.

Reference

Airoldi, J.-P., Flury, B., and Salvioni, M. 1996. Discrimination between two species of Microtus using both classified and unclassified observations. *Journal of Theoritical Biology* **177**, $247 - 262$.

(b) **Regression**

i. A director of admissions of a small college selected 120 students at random from the new freshmen class in a study to determine whether a student's grade point average (GPA) at the end of the freshman year can be predicted from the ACT test score and mid-semester exams. Response variable is the GPA at the end of the freshman year and predictor is the ACT test score. The goal of this model is prediction since the interest is to determine GPA based on predictors.

ii. A Tricade Office Equipment Corporation sells an imported three different types of copier on a franchise basis and performs preventive maintenance and repair service on this copier. 45 recent calls from users to perform routine preventive maintenance service were recorded; for each call we have, the number of copiers, copier type and the number of hours spent by the service person in servicing the copiers. The predictor variables are the number of copiers and the type of copier serviced and the response is the total number of minutes spent by the service person. The goal of the company is inference since the interest is find out which of the two predictors is associated with the response.

iii. A substance used in biological and medical research is shipped by airfreight to users in cartons of 1000 ampules. A data is collected, involving 100 shipments. The variables collected involves the number of times the carton was transferred from one aircraft to another over the shipment route and the type of courier (DHL, UPS, or FEDEX). Also, the number of ampules found to be broken upon arrival was recorded. The response variable is the number of ampules found broken upon arrival and the predictors are the type of courier and the number of times the carton was transferred from one aircaft to another over the shipment route. The goal of this scenario is inference since the interest is the find out which of the predictors associated with response variables.

Reference

Kutner, M.H., Neter, J., Nachtsheim, C.J. and Li, W. (2004) Applied linear statistical models, 5th Edition. McGraw- Hill Irwin, Boston.

(c) **Cluster Analysis**

i. With 288 specimen of voles skulls collected from various places in central Europe and eight differenct kind of measurements taken from each specimen, *cluster analysis* may useful here to find out the possible number of types (species) of voles within the 288 specimen.

ii. Given 4189 RNA gene levels from 180 samples of cancer tumor cells, *cluster analysis* may be useful in determining the possible number of types(clusters) of cancer tumor within the 180 samples.

iii. *Cluster analysis* is useful to identify how many possible groupings of spending habits of customers given some demographic information such as zip code, family income, shopping habits (how much spent in the last month, how many items bought in the past two weeks, what kind of items bought, etc.)

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 nonparametric approach)? What are its disadvantages?

Solution to Question 3

Parametric statistical learning makes strict assumption about the functional form, or shape of f and reduces the problem of estimating f down to one of estimating **a finite set of parameters** while non-parametric statistical learning do not make explicit assumptions about the functional form of f and gives room for **infinitly many parameters**.

Parametric approach is *less flexible* and *more restrictive* which is more appropriate in **inference** is the main interest of the model because it is much more interpretable while non-parametric models are *more flexible* can lead to a complicated estimate of f that it is difficult to understand how any individual predictor is associated with the response.

Parametric approaches are simplier and faster to learn from the given training data set while the non-parametric models are more flexible, complicated, slower to learn from the given training data set which may sometimes lead to overfitting.

Disadvantages of parametric approach are:

- prediction inaccuracy/poor fit

  The model we choose (functional form) will usually not match the true unknown form of *f*. If the chosen model is too far from the true *f*, then our estimate will be poor.

- estimated f is contrained by the assumption of a finite set of paramters.

  By choosing a functional form which determines a finite set of parameters, the model is highly stricted by the specified functional form and only small amount of changes occur to the model fit even if more and more data is collected for training and testing.

Reference

https://machinelearningmastery.com/parametric-and-nonparametric-machine-learning-algorithms/

http://mlss.tuebingen.mpg.de/2015/slides/ghahramani/gp-neural-nets15.pdf

4. Question 2.4.8 pg 54-55

Solution (a) - (b)

(a) The data has been read into the working directory using `read.csv()` function.

(b) The `fix()` and the `row.names()` functions have been used to set the row names of the `college` data set.

Solution (c)

Solution (c) i

Private	Apps	Accept	Enroll	Top10perc	Top25perc	F.Undergrad
No :212	Min. : 81	Min. : 72	Min. : 35	Min. : 1.00	Min. : 9.0	Min. : 139
Yes:565	1st Qu.: 776	1st Qu.: 604	1st Qu.: 242	1st Qu.:15.00	1st Qu.: 41.0	1st Qu.: 992
NA	Median : 1558	Median : 1110	Median : 434	Median :23.00	Median : 54.0	Median : 1707
NA	Mean : 3002	Mean : 2019	Mean : 780	Mean :27.56	Mean : 55.8	Mean : 3700
NA	3rd Qu.: 3624	3rd Qu.: 2424	3rd Qu.: 902	3rd Qu.:35.00	3rd Qu.: 69.0	3rd Qu.: 4005
NA	Max. :48094	Max. :26330	Max. :6392	Max. :96.00	Max. :100.0	Max. :31643

The number of public schools is less than half of the private schools representing 38% of the private schools.

P.Undergrad	Outstate	Room.Board	Books	Personal	PhD
Min. : 1.0	Min. : 2340	Min. :1780	Min. : 96.0	Min. : 250	Min. : 8.00
1st Qu.: 95.0	1st Qu.: 7320	1st Qu.:3597	1st Qu.: 470.0	1st Qu.: 850	1st Qu.: 62.00
Median : 353.0	Median : 9990	Median :4200	Median : 500.0	Median :1200	Median : 75.00
Mean : 855.3	Mean :10441	Mean :4358	Mean : 549.4	Mean :1341	Mean : 72.66
3rd Qu.: 967.0	3rd Qu.:12925	3rd Qu.:5050	3rd Qu.: 600.0	3rd Qu.:1700	3rd Qu.: 85.00
Max. :21836.0	Max. :21700	Max. :8124	Max. :2340.0	Max. :6800	Max. :103.00

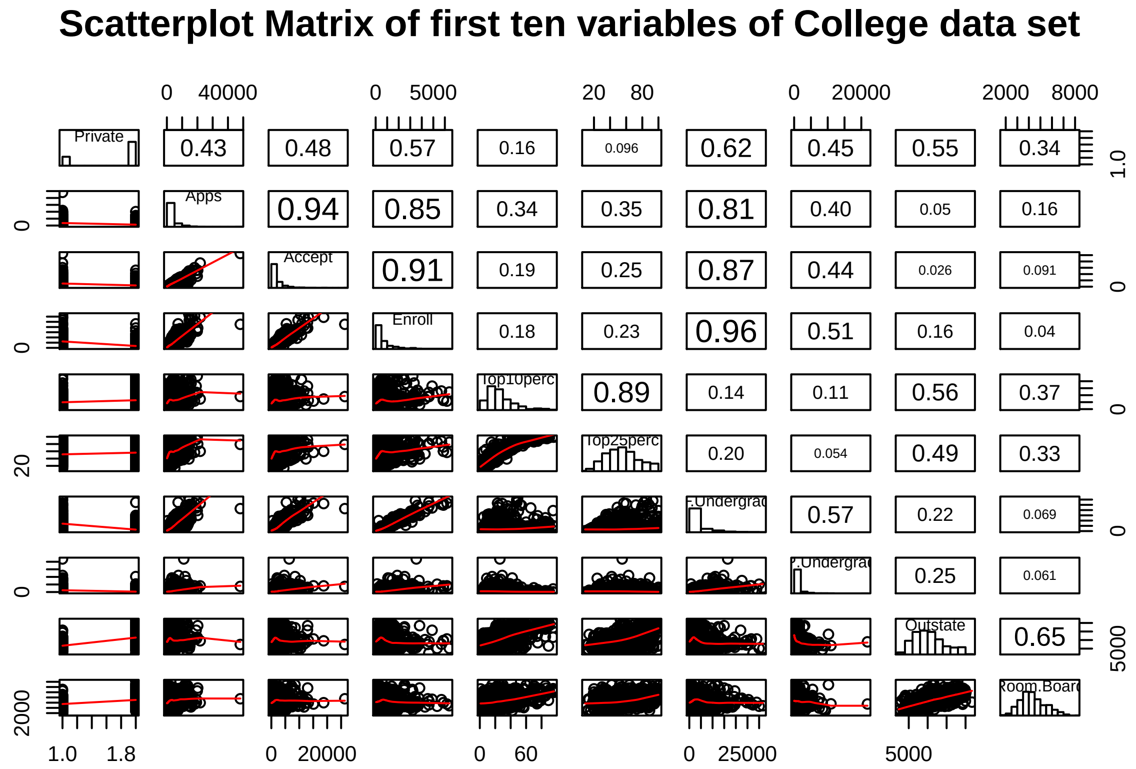
Terminal	S.F.Ratio	perc.alumni	Expend	Grad.Rate
Min. : 24.0	Min. : 2.50	Min. : 0.00	Min. : 3186	Min. : 10.00
1st Qu.: 71.0	1st Qu.:11.50	1st Qu.:13.00	1st Qu.: 6751	1st Qu.: 53.00
Median : 82.0	Median :13.60	Median :21.00	Median : 8377	Median : 65.00
Mean : 79.7	Mean :14.09	Mean :22.74	Mean : 9660	Mean : 65.46
3rd Qu.: 92.0	3rd Qu.:16.50	3rd Qu.:31.00	3rd Qu.:10830	3rd Qu.: 78.00
Max. :100.0	Max. :39.80	Max. :64.00	Max. :56233	Max. :118.00

From the summary it could be observed that the following variables are **fairly symmetrical** based on how close their mean and median values are.

- Outstate
- Room.Board
- S.F.Ratio
- Grad.Rate
- Top25perc

Solution (c) ii

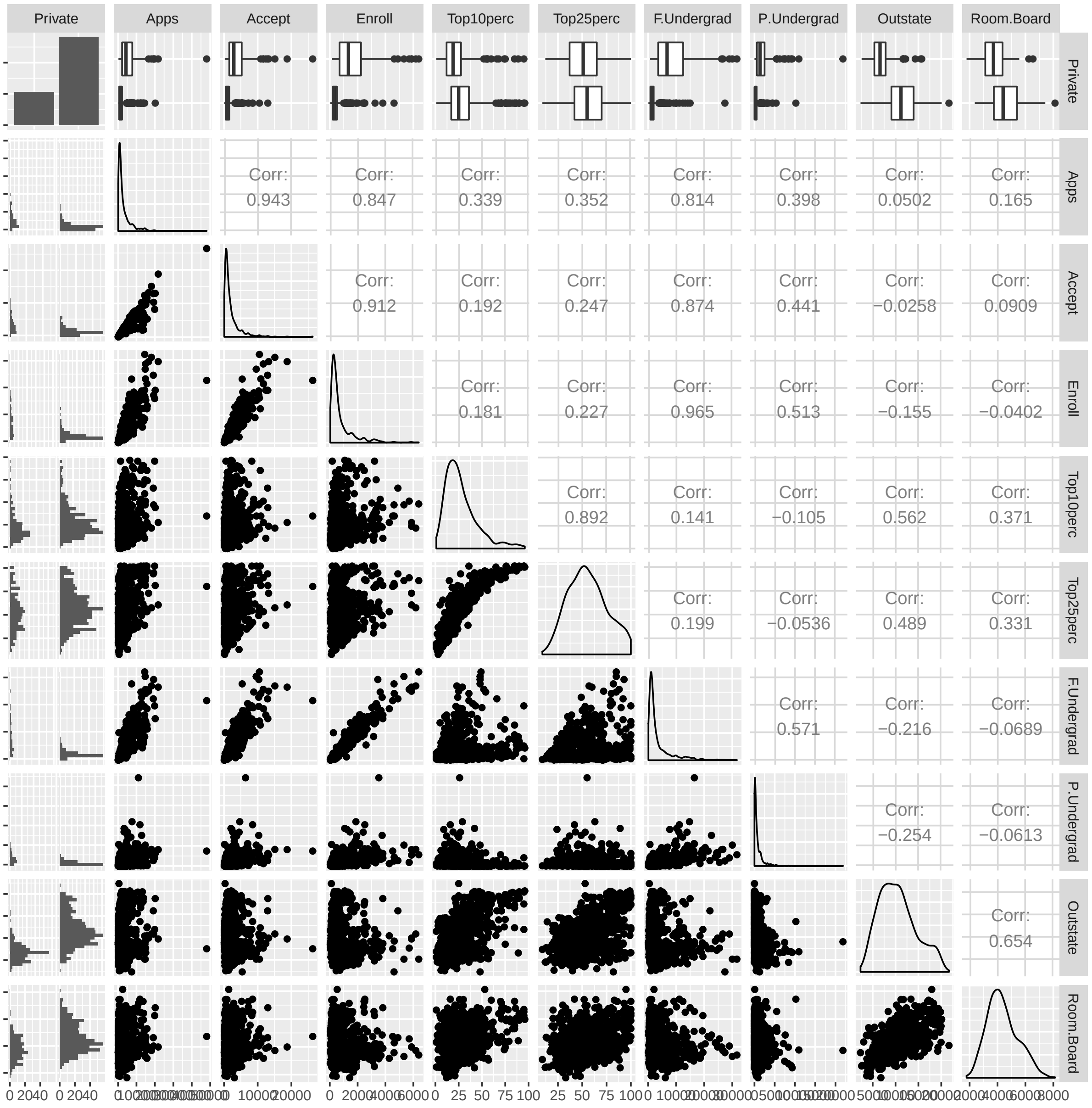
From the documentation of `pairs()` function, the following function has been used to modify the diagonal and upper panel of the scatter plot matrix.
`panel.cor()` function for `upper.panel` argument of `pairs()` function to get absolute correlations on the upper panels, with size proportional to the correlations.
`panel.hist()` function for `diag.panel` argument of `pairs()` function to get histograms on the diagonal.
`panel.smooth()` function for `lower.panel` argument of `pairs()` function to add LOWESS smoothed line to the scatter plots.


Scatterplot Matrix of first ten variables of College data set

There is high positive linear association or correlation between:
(a) Number of applications received and number of applications accepted.
(b) Number of applications accepted and number of new students enrolled.
(c) Number of new students enrolled and number of full-time undergraduates.

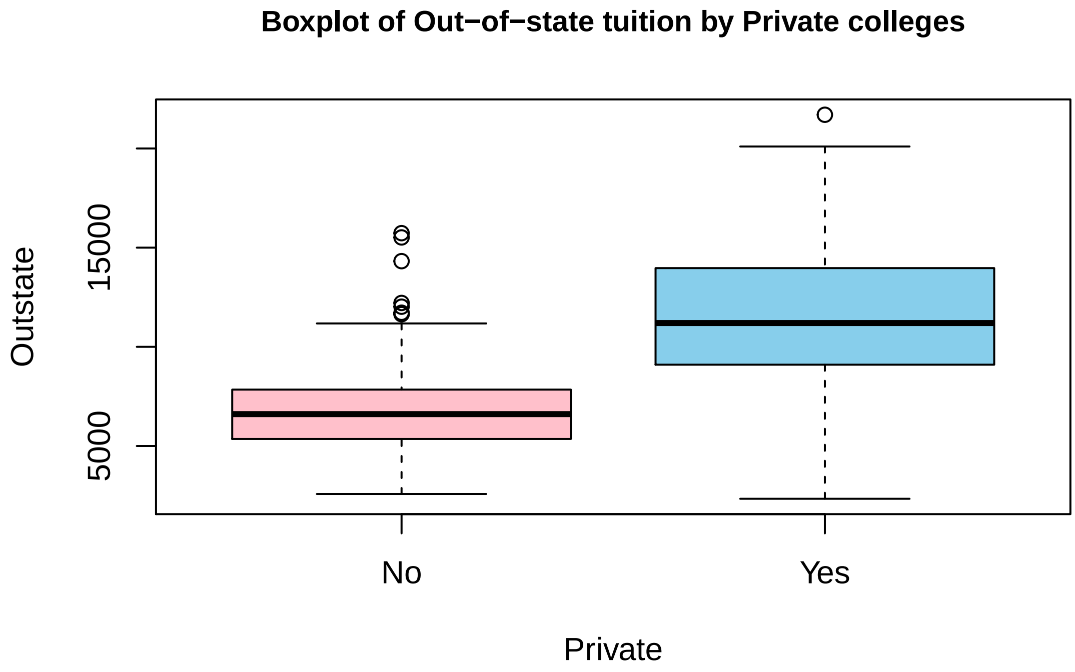
The following variables are fairly symmetrical.
(a) New students from top 25% of high school class
(b) Out-of-state tuition
(c) Room and board costs

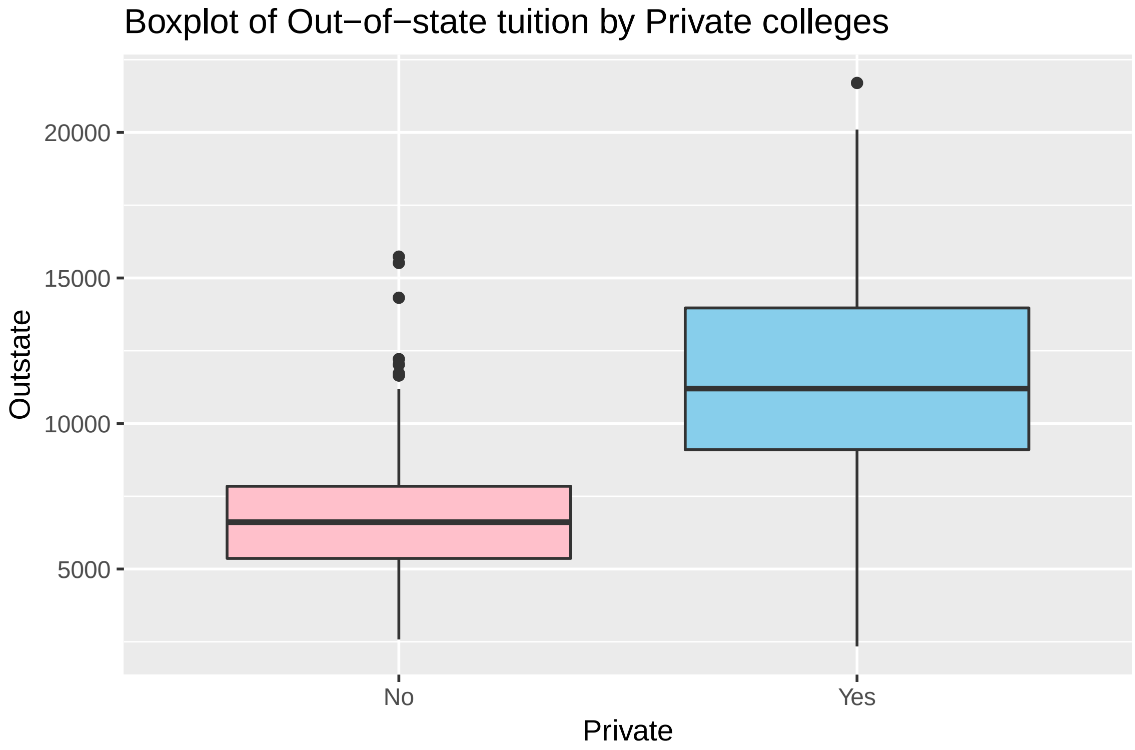
The following variables are highly skewed to the right.
(a) Number of applications received
(b) Number of applications accepted
(c) Number of new students enrolled
(d) New students from top 10% of high school class
(e) New students from top 25% of high school class



Scatterplot Matrix of first ten variables of College data set

Solution (c) iii



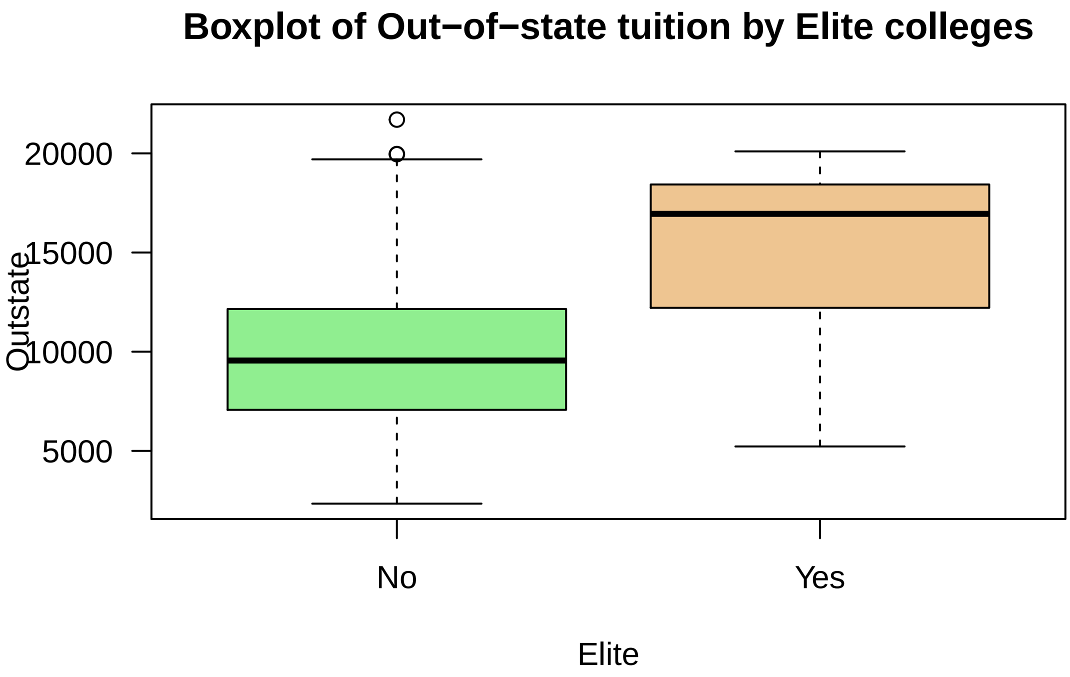


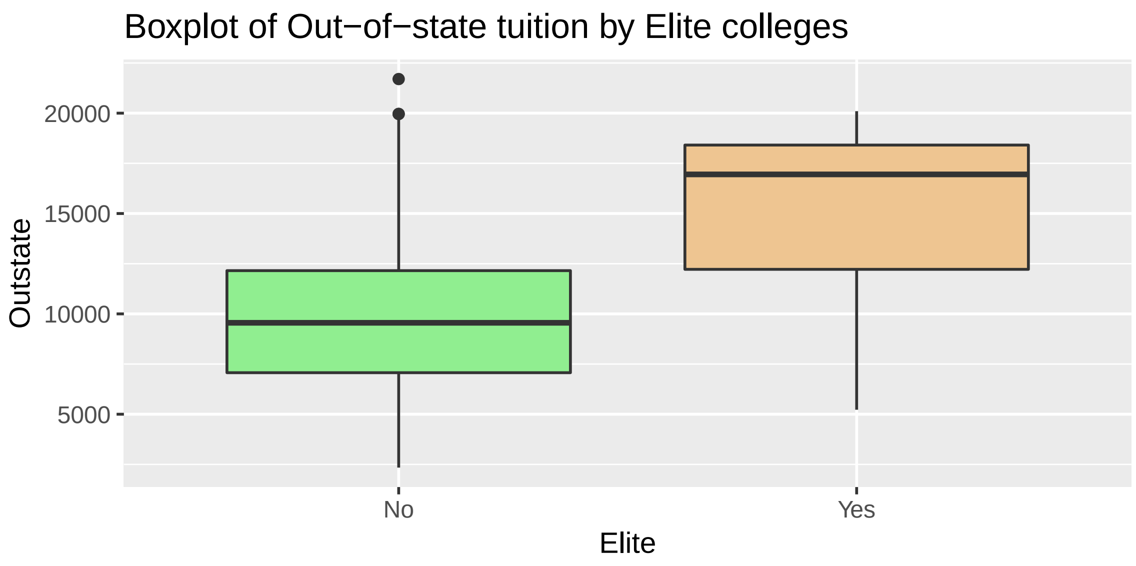
Out-of-state tuition for non-private colleges is lesser than private colleges. There is higher variance of out-of-state tuition for private colleges.

Solution iv

Summary of distribution of Elite colleges

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

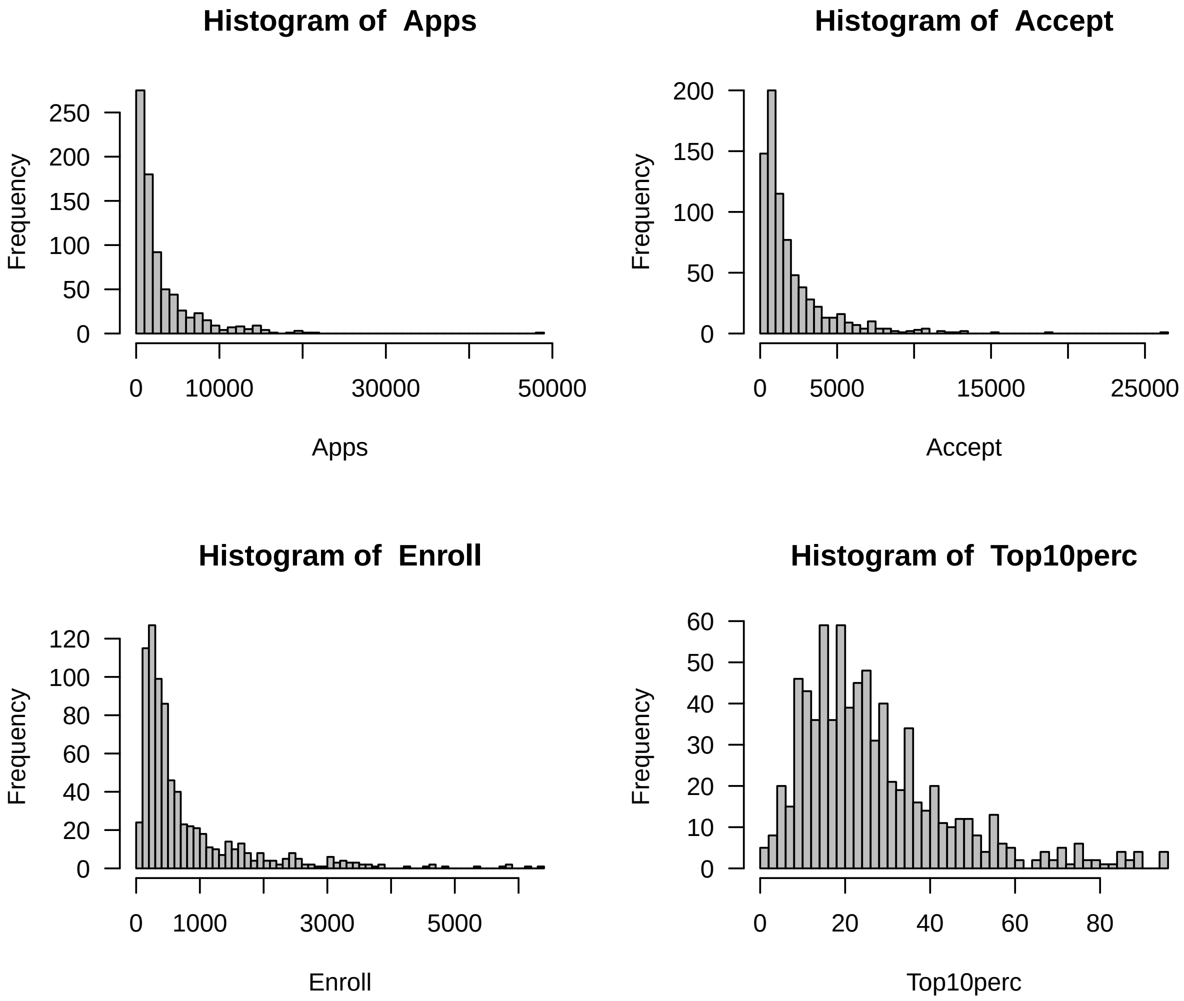




Out-of-state tuition for non-elite colleges is lesser than elite colleges. Out-of-state tuition for the elite colleges is a fairly skewed to the left.

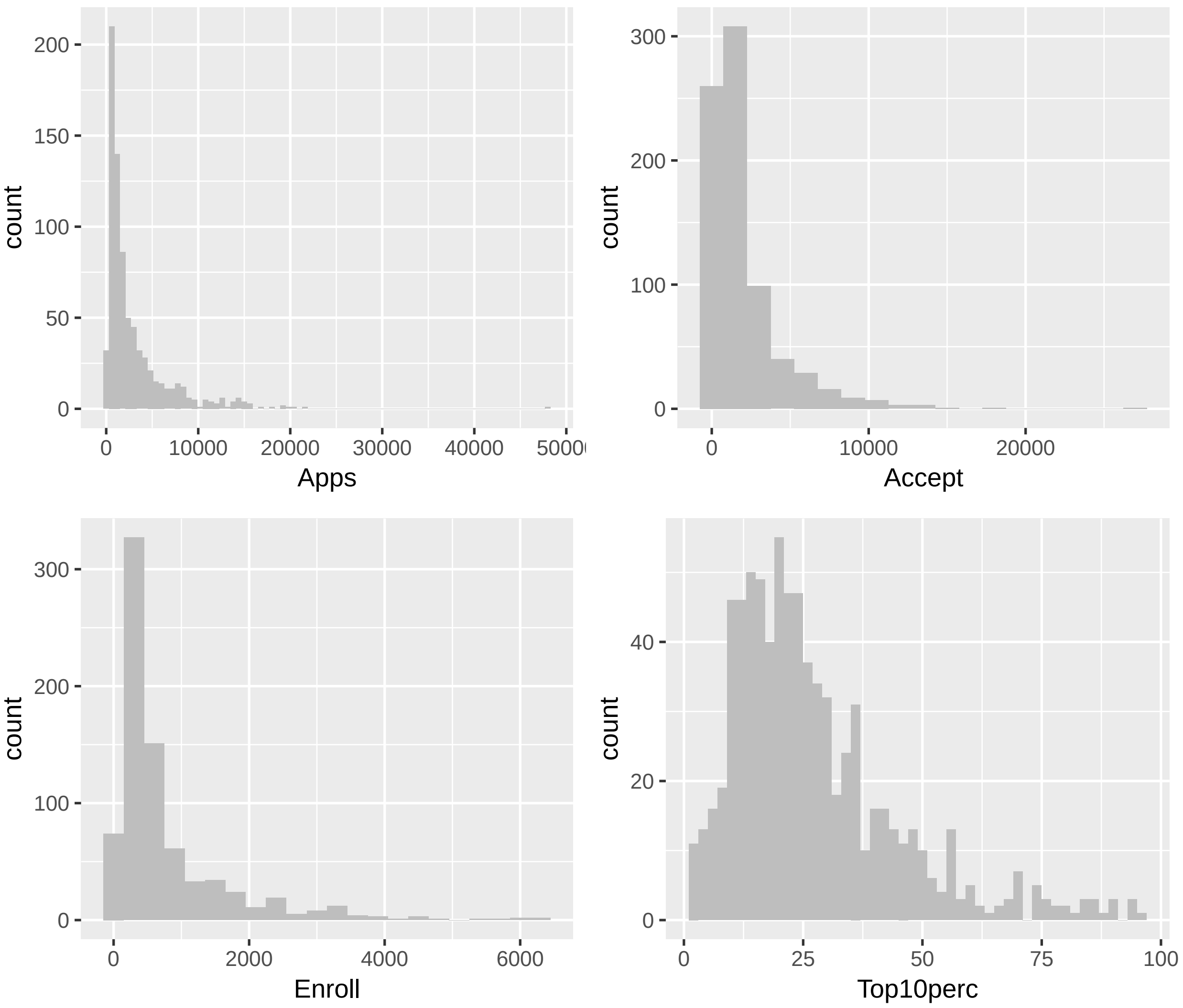
Solution v

Base R plots with smaller bandwidth

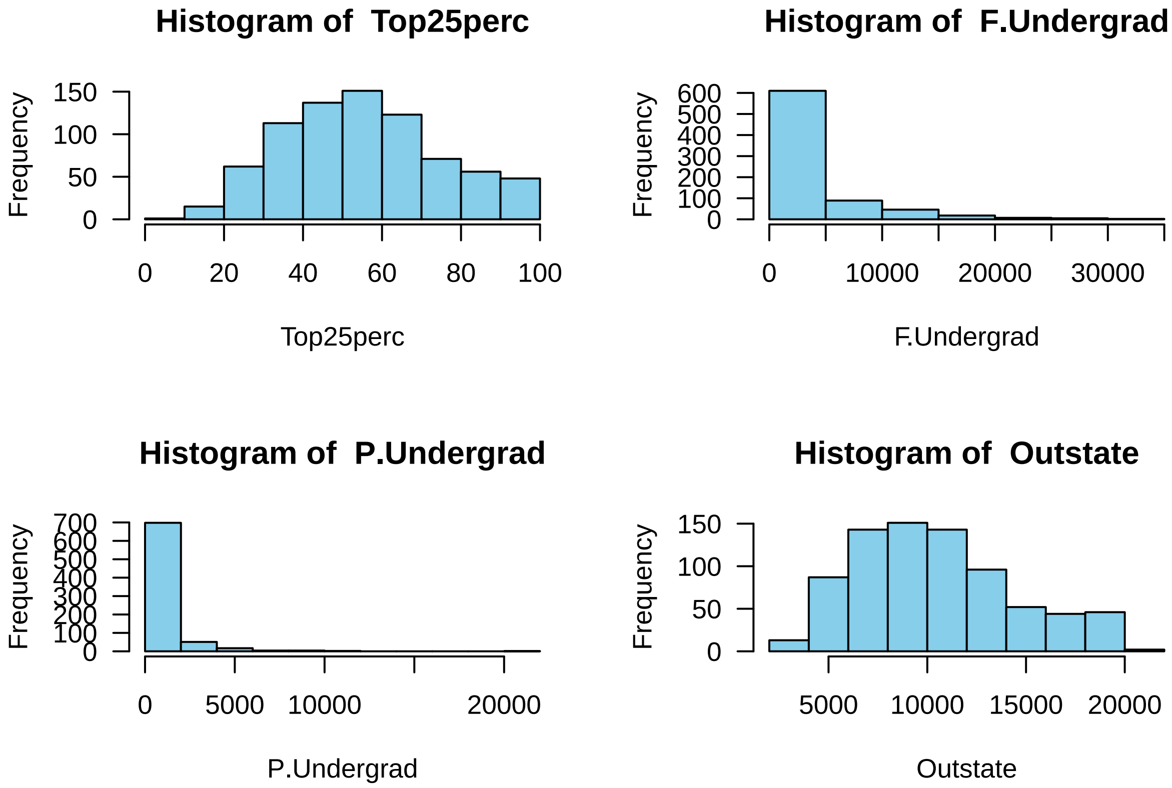


The Apps, Accept and Enroll variables are all higly skewed to the right as seen earlier. The top10perc variable is fairly symmetric but there are significant outlying values

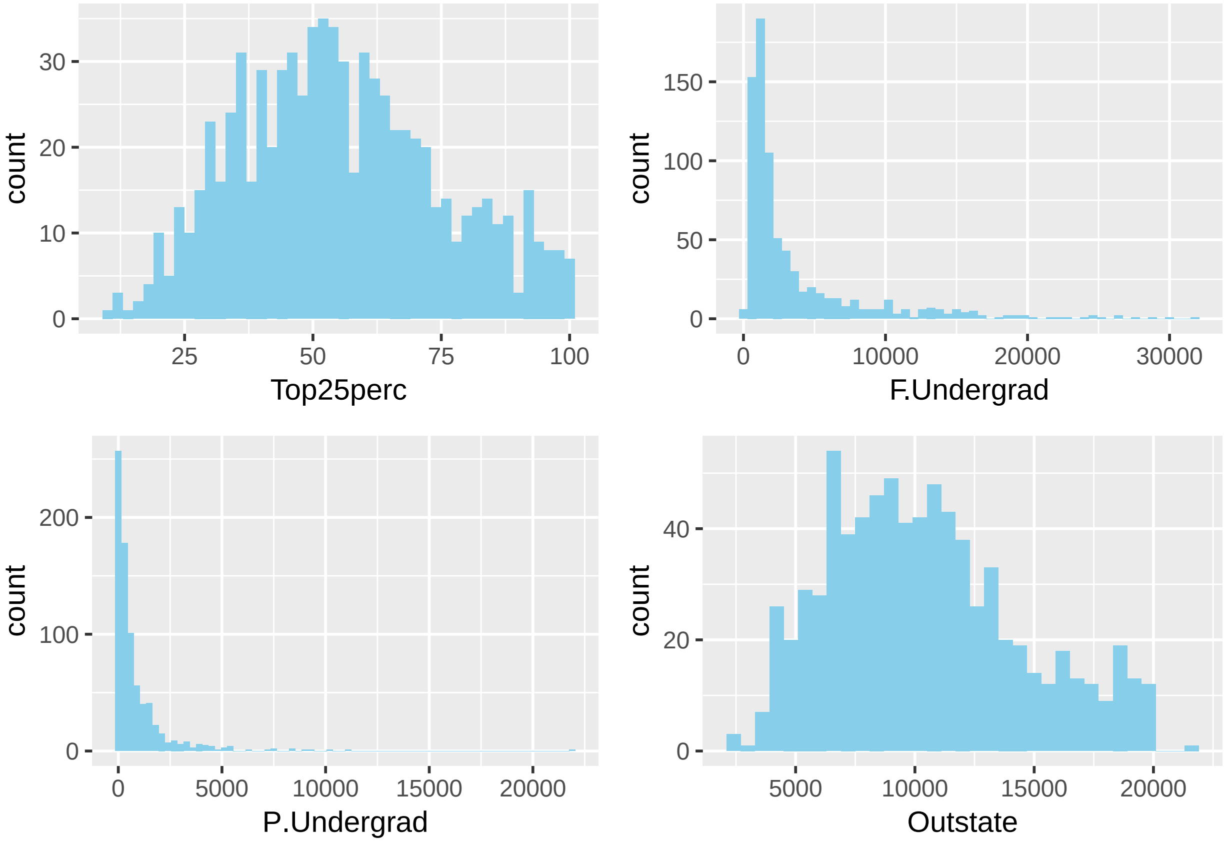
ggplot2 alternative varying bandwidth



Base R plots with bigger bandwidth



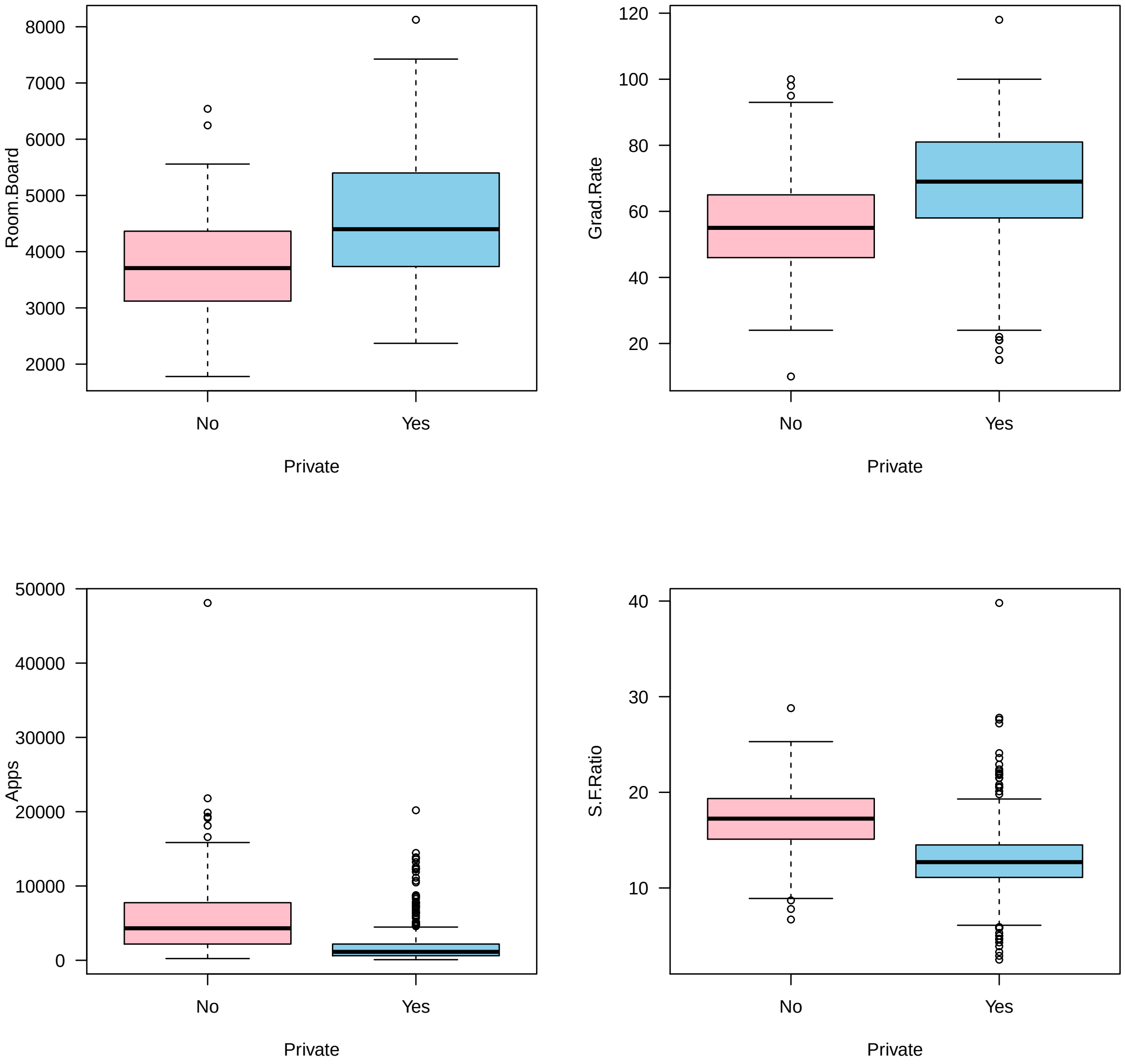
ggplot2 alternative with smaller bandwidth


count
30
20
10
0
25
50
75
100
Top25perc
count
150
100
50
0
0
10000
20000
30000
F.Undergrad
count
200
100
0
0
5000
10000
15000
20000
P.Undergrad
count
40
20
0
5000
10000
15000
20000
Outstate

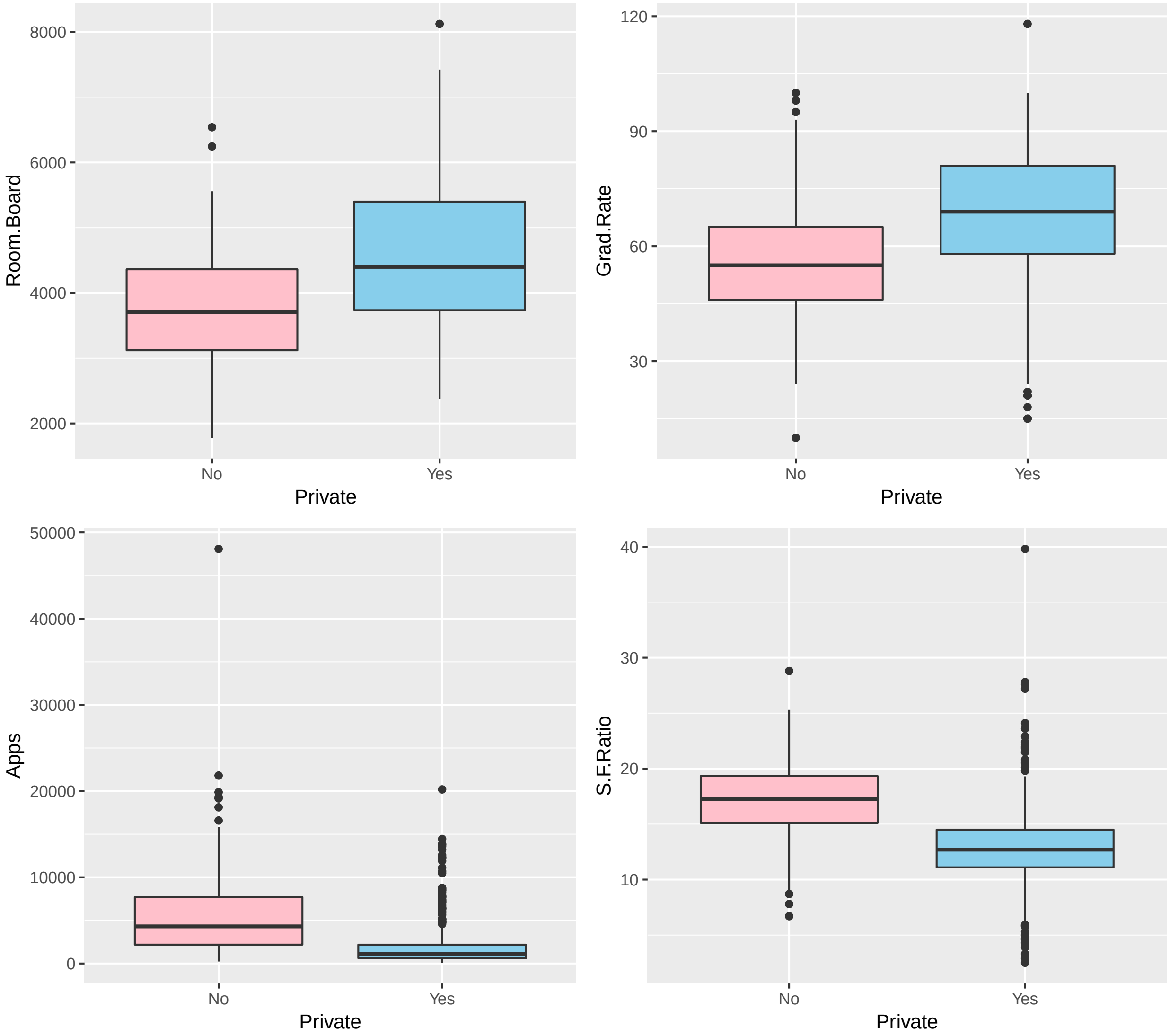
Solution vi

The following is further exploration of the data set using the boxplot to see if there is a distinction whether or not a colleges is Private for Room.Board, Grad.Rate, Outstate and S.F.Ratio.

Base R



ggplot2 alternative



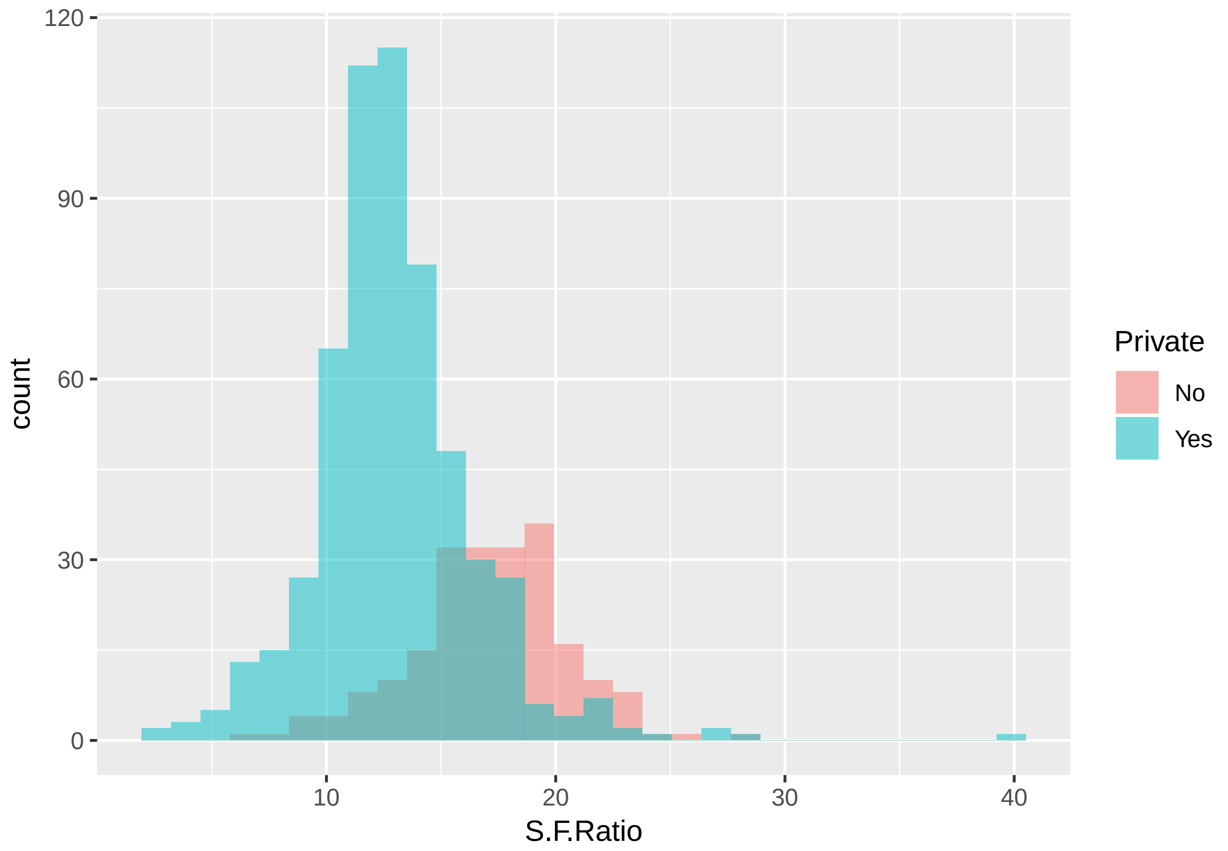


Table of Private and Elite variables

```
##        Elite
## Private  No Yes
##     No  199  13
##     Yes 500  65
```

Out of 565 private colleges, only 65 are elite.
Out of 212 non-private colleges, only 13 are elite.

Test of Independence between the Private and Elite variables

The hypothesis for the chi-squared test is:

$$H_0 = \text{Independent}$$

$$H_a = \text{Not Independent}$$

```
## 
##  Pearson's Chi-squared test with Yates' continuity correction
## 
## data:  table(Private = college$Private, Elite = college$Elite)
## X-squared = 4.3498, df = 1, p-value = 0.03701
```

The p-value = 0.03701 from the chi-squared shows that the null hypothesis is rejected at 5% level of significance. Hence, there is enough evidence to believe that there is association between the distribution of whether a college is Private or not and the distribution of whether or not a college is elite.

Brief Summary

(a) The distribution of whether a college is Private or not and the distribution of whether or not a college is elite not independent based on a chisquare test of independence.

(b) Room and board costs is on the average higher in private colleges than non-private colleges.

(c) Graduation rate is on the average higher in private colleges than non-private colleges.

(d) Number of applications received is on the average lower in private colleges than non-private colleges.

(e) Student/faculty ratio is on the average lower in private colleges than non-private colleges. However, there is an outlier in the distribution of student/faculty ratio of the private colleges.

(f) Out of 565 private colleges, only 65 are elite.

(g) Out of 212 non-private colleges, only 13 are elite.