ISLR Notes and Exercises

Nathaniel Lai

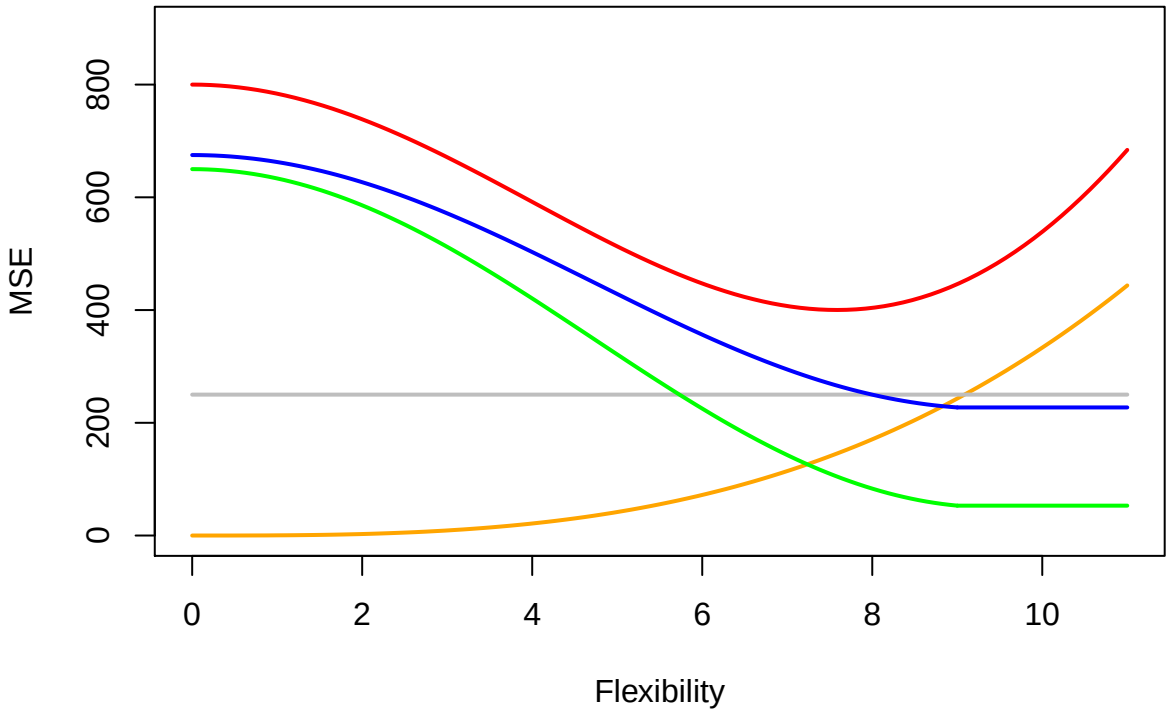
Updated March 2018

Chapter 2: Statistical Learning

Bias-Variance Tradeoff

One of the key concepts in statistical learning is the bias-variance tradeoff (diagram from Weatherwax).

- `red` = test error
- `orange` = estimator variance
- `green` = model bias
- `gray` = irreducible error
- `blue` = training error



where

$$MSE = \frac{1}{n}\sum (y_i - \hat{f}(x_i)^2$$

(for regression models) and, for classification, the training error rate is :

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i)$$

The bias-variance tradeoff decomposes the *expected test MSE* into:

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0) + [Bias(\hat{f}(x_0))]^2 + Var(\varepsilon)$$

$$\textit{Expected MSE} = \underbrace{\textit{variance} + \textit{bias}}_{\textit{reducible error}} + \textit{irreducible error}$$

"When a given method yields a small training MSE but a large test MSE, we are said to be **overfitting** the data. This happens because our statistical learning procedure is working too hard to find patterns in the training data, and may be picking up some patterns that are just caused by random chance rather than by true properties of the unknown function `f`. When we overfit the training data, the test MSE will be very large because the supposed patterns that the method found in the training data simply do not exist in the test data. Note that regardless of whether or not overfitting has occurred, we almost always expect the training MSE to be smaller than the test MSE because most statistical learning methods either directly or indirectly seek to minimize the training MSE. Overfitting refers specifically to the case in which a less flexible model would have yielded a smaller test MSE." (ISLR P.32)

It is possible to show that the test error rate is minimized, on average, by a very simple classifier that assigns each observation to the most likely class, given its predictor values. In other words, we should simply assign a test observation with predictor vector x_0 to the class j for which the **conditional probability** (ISLR P.38)

$$Pr(Y = j|X = x_0)$$

is the largest with the error rate:

$$1 - E\Big(\max_j Pr(Y = j|X\Big)$$

Prediction vs. Model Accuracy

Question 5

Advantages of a very flexible model include better fit to data and fewer prior assumptions.

Disadvantages are the increased difficulties to interpret and the danger of overfitting.

A more flexible approach might be preferred is the underlying data is very complex (simple linear fit does not suffice) or if we mainly care about the result and not inference, provided that sample size is large enough.

A less flexible model is preferred is the underlying data has a simple shape or if inference and interpretability are important.

Parametric and Non-parametric Methods

Question 6

For parametric methods, we make an assumption about the shape of the underlying data, select a model form, and fit the data to our selected form. The advantage is that we can incorporate any prior/expert knowledge and do not tend to have too many parameters that need to be fit. To the extent that our prior/expert assumptions are wrong, then that would be a disadvantage.

Non-parametric methods do not make explicit assumptions on the shape of the data, which could be an advantage. The key disadvantage is that they need a large number of observations to fit an accurate estimate.

Applied Questions

Question 8

This question is on standard regrassion prcedures and the use of `ggplot`.

Part a)

```
require(ISLR);
```

```
## Loading required package: ISLR
```

```
data(College)
str(College)
```

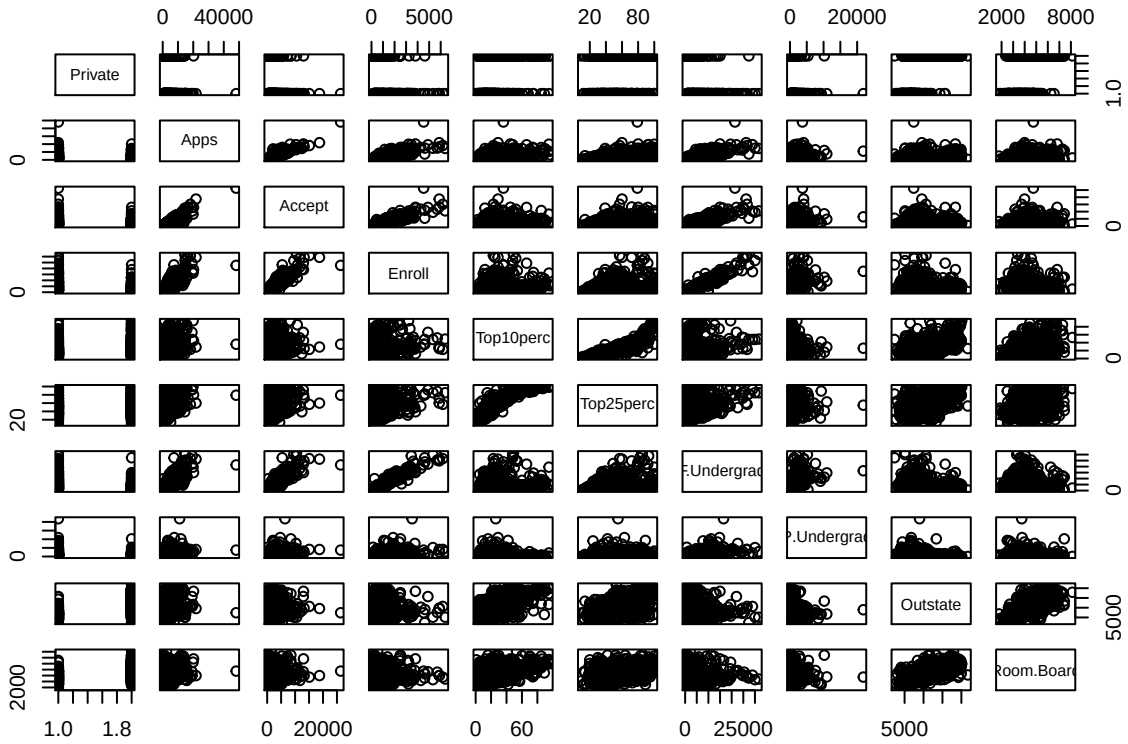
```
## 'data.frame':    777 obs. of  18 variables:
##  $ Private    : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 2 2 2 2 2 ...
##  $ Apps       : num  1660 2186 1428 417 193 ...
##  $ Accept     : num  1232 1924 1097 349 146 ...
##  $ Enroll     : num  721 512 336 137 55 158 103 489 227 172 ...
##  $ Top10perc  : num  23 16 22 60 16 38 17 37 30 21 ...
##  $ Top25perc  : num  52 29 50 89 44 62 45 68 63 44 ...
##  $ F.Undergrad: num  2885 2683 1036 510 249 ...
##  $ P.Undergrad: num  537 1227 99 63 869 ...
##  $ Outstate   : num  7440 12280 11250 12960 7560 ...
##  $ Room.Board : num  3300 6450 3750 5450 4120 ...
##  $ Books      : num  450 750 400 450 800 500 500 450 300 660 ...
##  $ Personal   : num  2200 1500 1165 875 1500 ...
##  $ PhD        : num  70 29 53 92 76 67 90 89 79 40 ...
##  $ Terminal   : num  78 30 66 97 72 73 93 100 84 41 ...
##  $ S.F.Ratio  : num  18.1 12.2 12.9 7.7 11.9 9.4 11.5 13.7 11.3 11.5 ...
##  $ perc.alumni: num  12 16 30 37 2 11 26 37 23 15 ...
##  $ Expend     : num  7041 10527 8735 19016 10922 ...
##  $ Grad.Rate  : num  60 56 54 59 15 55 63 73 80 52 ...
```

Part b)

```
# these steps were already taken on College data in the ISLR package
rownames(College) <- College[,1]  # set row names
College <- College[,-1]  # drop first col
# i.
summary(College)
```

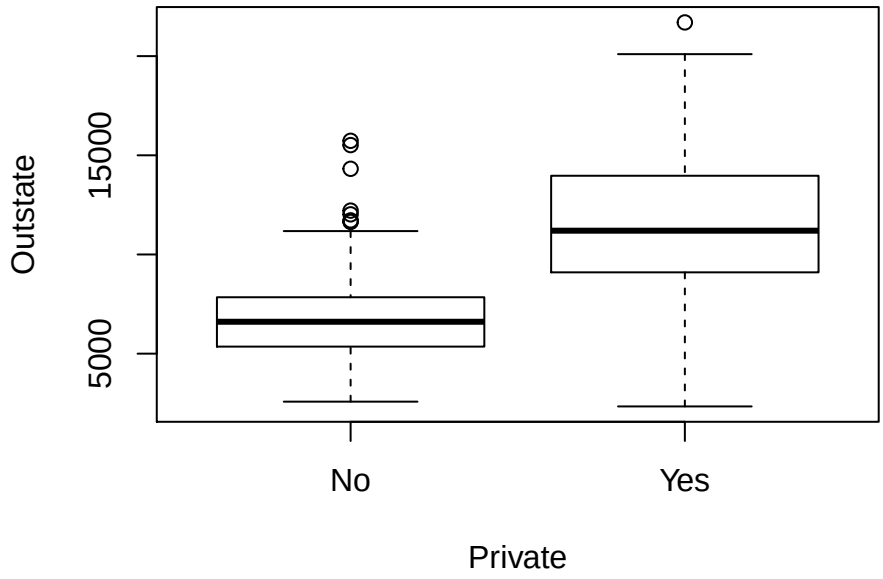
Part c)

```
attach(College)
# ii.
pairs(College[,1:10])
```




```
# iii.
```

```
boxplot(Outstate~Private, data=College, xlab="Private", ylab="Outstate")
```

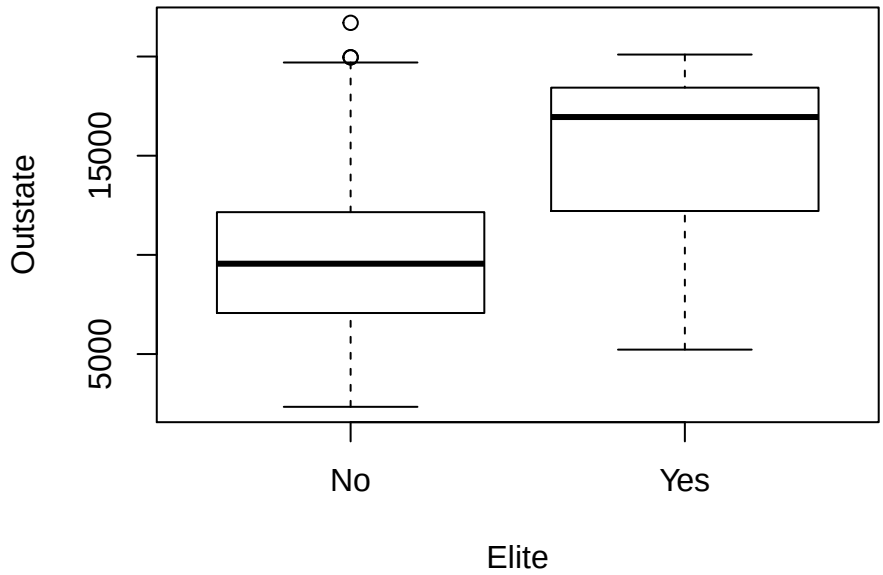



```
# iv.
Elite <- rep("No", nrow(College))
Elite[Top10perc>50] <- "Yes"
College <- data.frame(College, Elite)
summary(College)  # 78 Elite
```

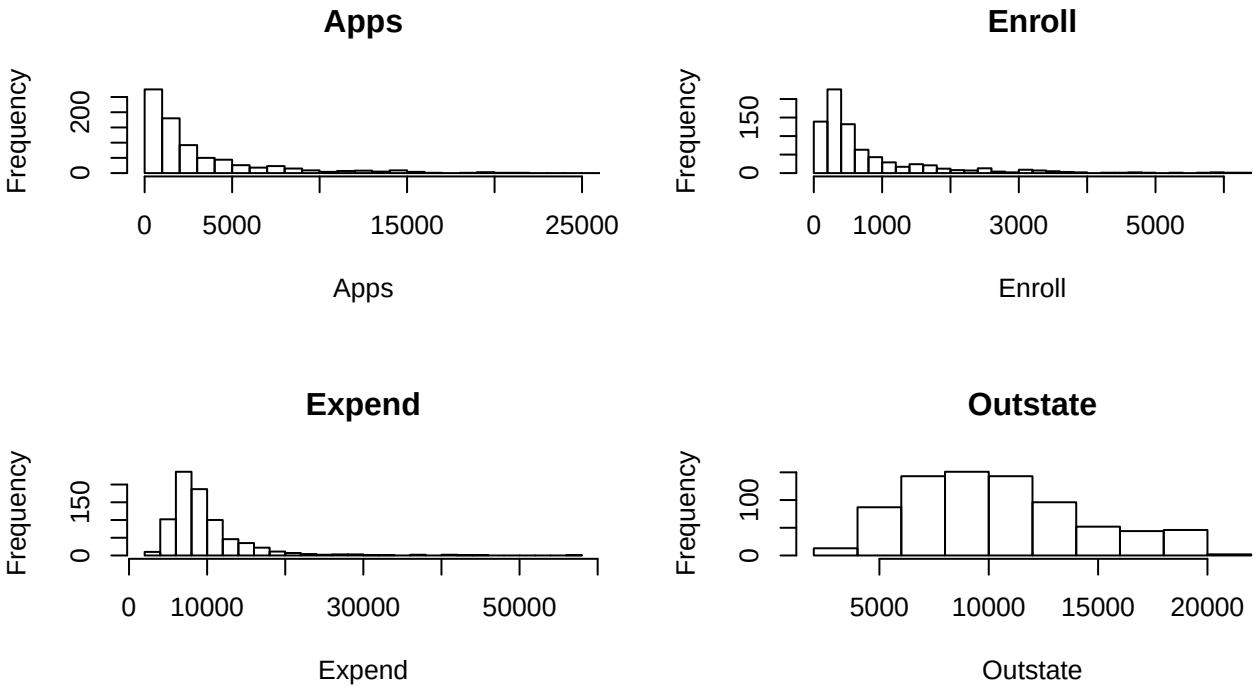
```
##  Private        Apps           Accept          Enroll       Top10perc
##  No :212   Min.   :   81   Min.   :   72   Min.   :  35   Min.   : 1.00
##  Yes:565   1st Qu.:  776   1st Qu.:  604   1st Qu.: 242   1st Qu.:15.00
```

```
##            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
```

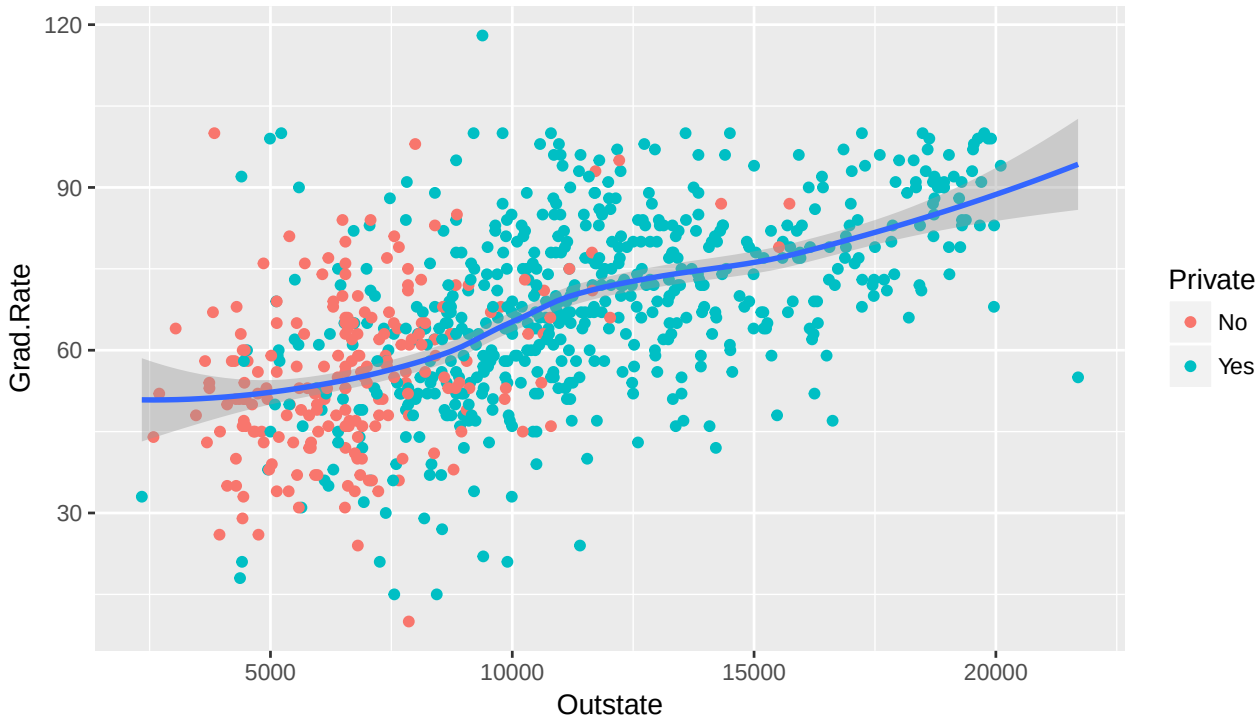
```
boxplot(Outstate~Elite, data=College, xlab="Elite", ylab="Outstate")
```

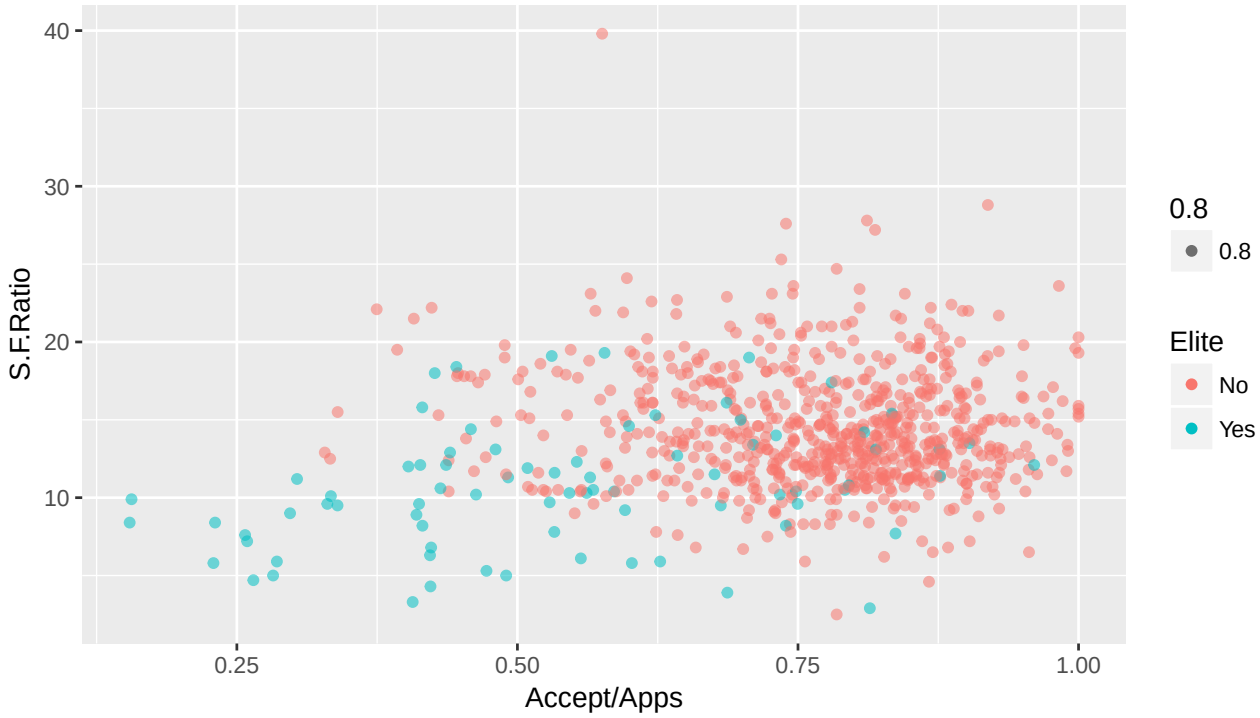


```r
# v.
par(mfrow=c(2,2))
hist(Apps, breaks=50, xlim=c(0,25000), main="Apps")
hist(Enroll, breaks=25, main="Enroll")
hist(Expend, breaks=25, main="Expend")
hist(Outstate, main="Outstate")
```




```
## `geom_smooth()` using method = 'loess'
```


120
90
60
30
Grad.Rate
5000
10000
15000
20000
Outstate
Private
No
Yes


40
30
20
10
S.F.Ratio
0.25
0.50
0.75
1.00
Accept/Apps
0.8
0.8
Elite
No
Yes

Colleges with the most students from top 10% perc do not necessarily have the highest graduation rate. Also, rate over 100 is erroneous.

References

G.James, D.Witten, T.Hastie and R.Tibshirani (2013), An Introduction to Statistical Learning, with applications in R (ISLR), Springer