DS502- HW1

Mahdi Alouane and Rahul Pande

1. 2.4 (question 1)

- (a) The sample size n is extremely large, and the number of predictors p is small.

As the sample size is extremely large the certainty of the veracity of the sample mean is high. Therefore the variance will be low, since an unknown sample will not deviate a lot from the sample mean of a large sample. The low number of predictors will prevent overfitting and thus reduce variance. Since we have two factors that reduce variance, in this case, a flexible statistical learning method is expected to be better because it will reduce the bias. The higher variance of more flexible learning method will be countered by the above two factors.

- (b) The number of predictors p is extremely large, and the number of observations n is small.

The large number of predictors will tend to overfit and perform poor on unseen data, thus increasing variance. Similarly with less number of samples there could be a lot of variablility in the least sqaures which would mean higher variance for unseen data. Therefore in this case an inflexible learning method is expected to perform better since it will counter the high variance.

- (c) The relationship between the predictors and response is highly non-linear.

Where the relationship between predictors and response is highly non-linear, an inflexible model will have high bias given that it won't be able to capture the complex relationship between the predictors and response. Therefore a flexible method is expected to perform better in this case since it will reduce the bias.

- (d) The variance of the error terms, i.e. $\mu^2 = \text{Var}()$, is extremely high.

As the variance of error terms is extremely high, the certainty of the veracity of the sample mean is very low. Therefore the model variance will be high, since an unknown sample could deviate a lot from the sample mean. Hence, in this case, an inflexible statistical learning method is expected to be better because it will reduce the variance.

2. 2.4 (question 3)

3. 2.4 (question 6)

4. 2.4 (question 8)

- (a) Reading `College.csv` into `college` variable

```
college <- read.csv("College.csv", stringsAsFactors = TRUE)
```

- (b) Set first column as row names and then remove that column from data

```
# fix(college)
rownames(college) = college[,1]
college = college[,-1]
# fix(college)
```

- (c)

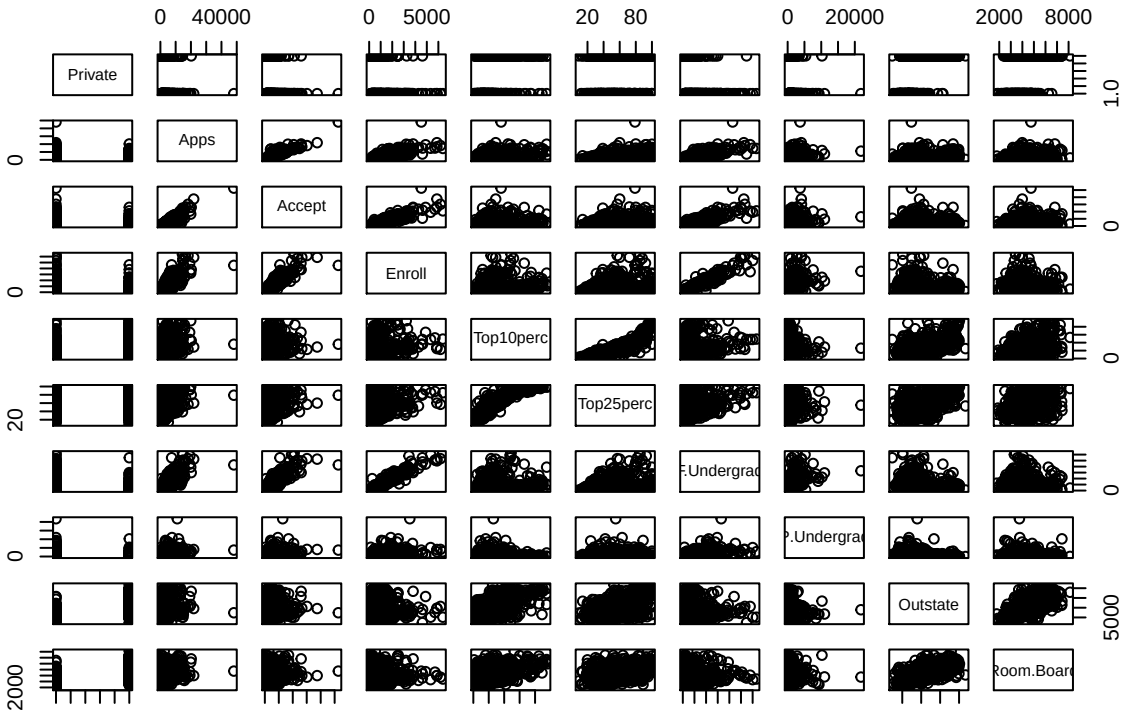
i. Summary of `college` variable

```
summary(college)
```

```
##  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
```

ii. Scatterplot matrix of first 10 variables

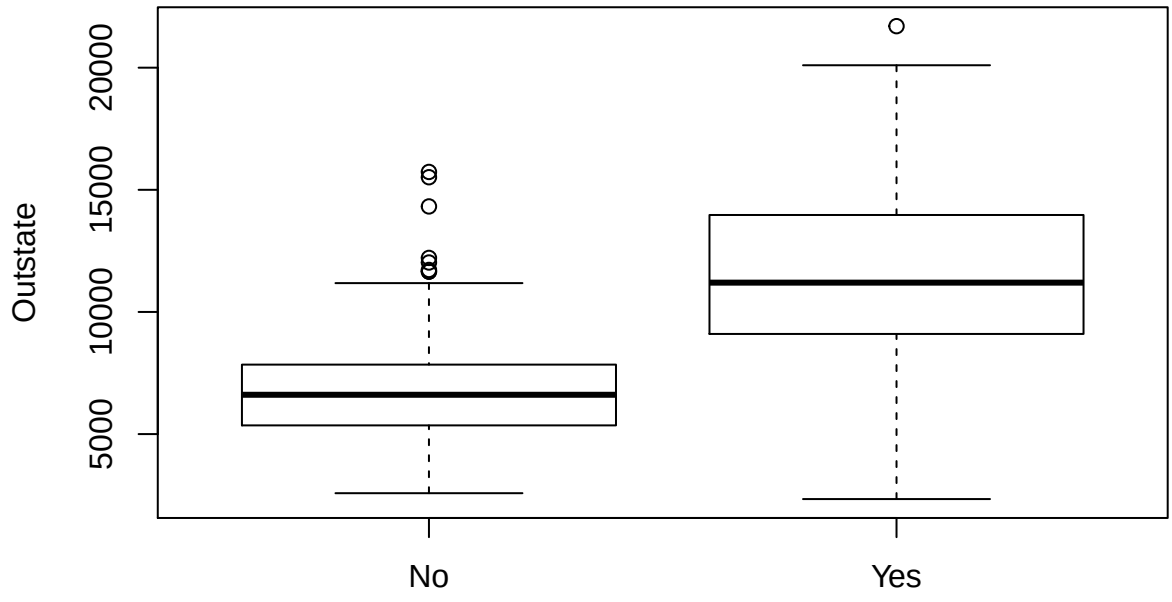
```
pairs(college[,1:10])
```



ii-1.bb

iii. Boxplot of `Outstate` versus `Private`

```
plot(Outstate~Private, data=college)
```



iii-1.bb

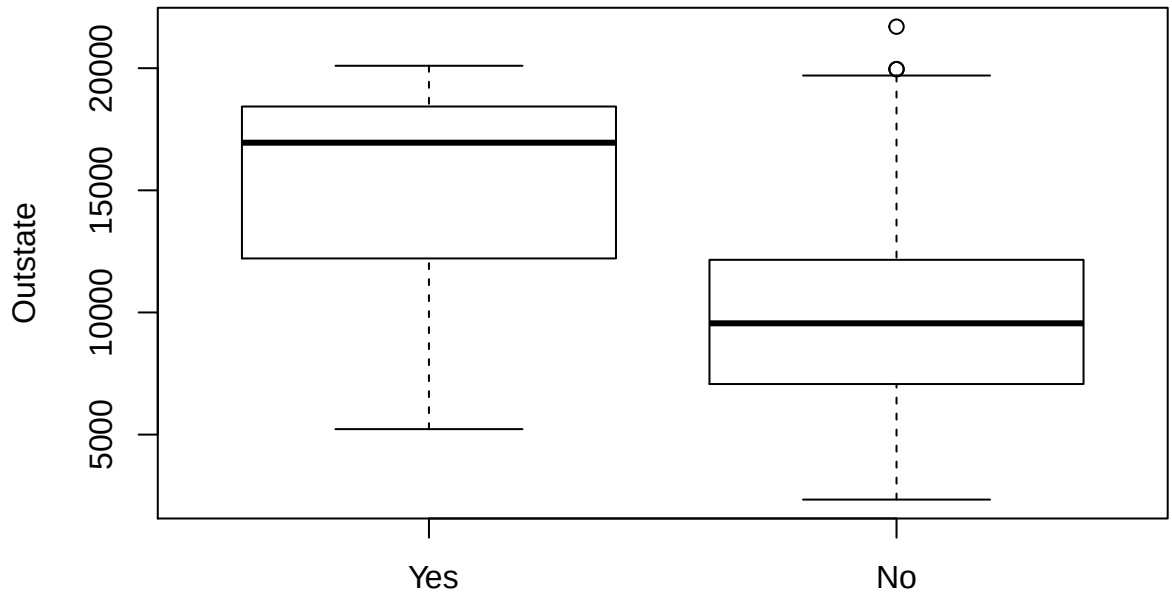
iv. Elite Universities

```
Elite = rep("No",nrow(college))
Elite[college$Top10perc > 50] = " Yes"
Elite = as.factor(Elite)
college = data.frame(college, Elite)
```

```r
summary(college)
```

```
##  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    Yes: 78
##  1st Qu.: 53.00   No  :699
##  Median : 65.00
##  Mean   : 65.46
##  3rd Qu.: 78.00
##  Max.   :118.00
```

```r
plot(Outstate~Elite, data=college)
```

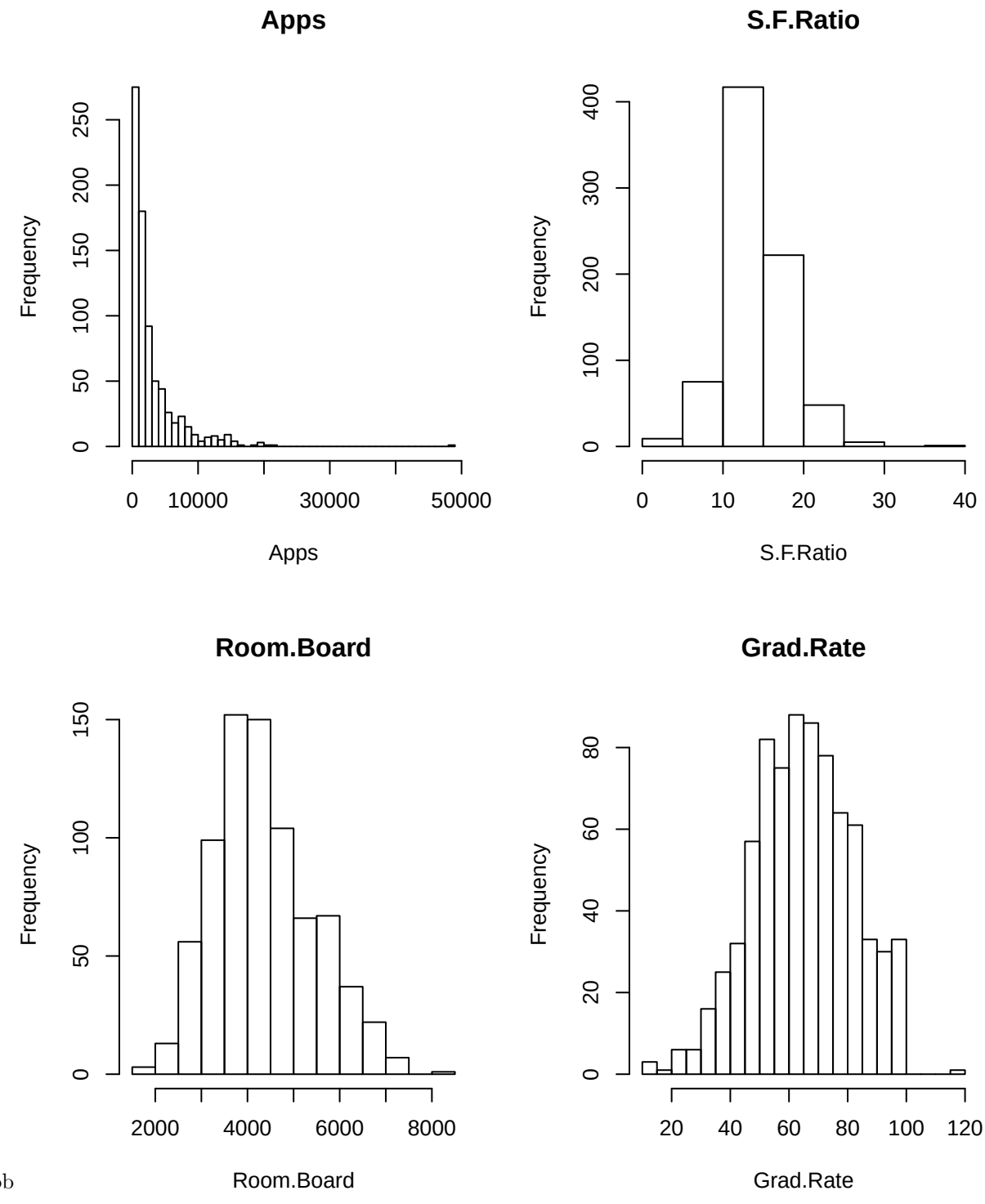


iv-1.bb

From the `summary`, we have **78** **`Elite`** universities.

v. Histograms

```
par(mfrow=c(2,2))

hist_vars   = c("Apps", "S.F.Ratio", "Room.Board", "Grad.Rate")
hist_breaks = c(50, 10, 20, 20)
hist_data = subset(college, select = hist_vars)

make_hist <- function(list.elem, names, breaks) {
  hist(list.elem, main = names, xlab = names, breaks = breaks)
}

mapply(make_hist, list.elem = hist_data, names = names(hist_data), breaks = hist_breaks)
```

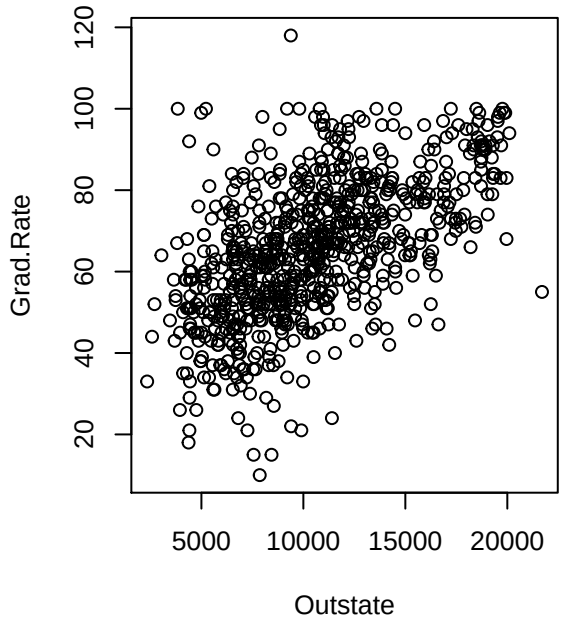


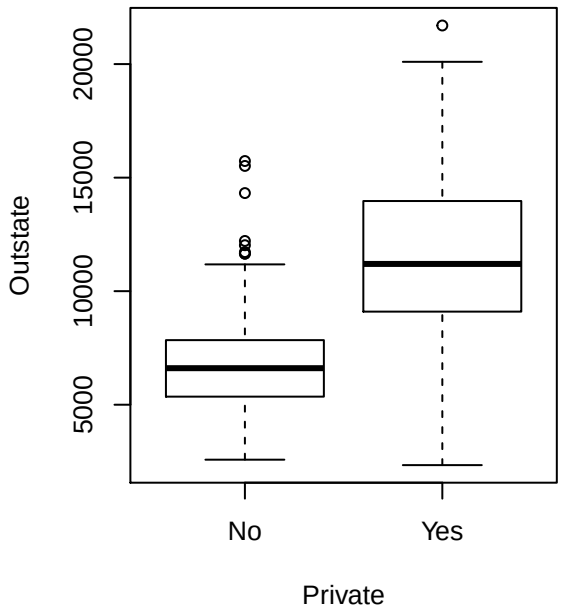
v-1.bb

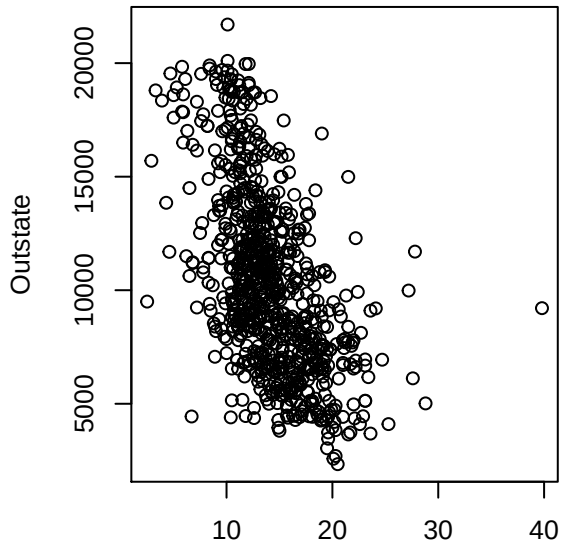
vi. Exploration

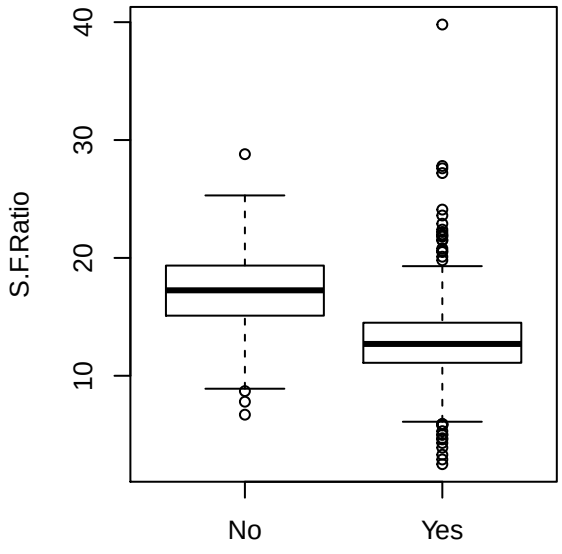
```
college$Accept.Rate = college$Accept / college$Apps * 100
college$Enroll.Rate = college$Enroll / college$Accept * 100

par(mfrow=c(2,2))
plot(Grad.Rate ~ Outstate, data = college)
plot(Outstate ~ Private, data = college)
```

```
plot(Outstate ~ S.F.Ratio, data = college)
plot(S.F.Ratio ~ Private, data = college)
```



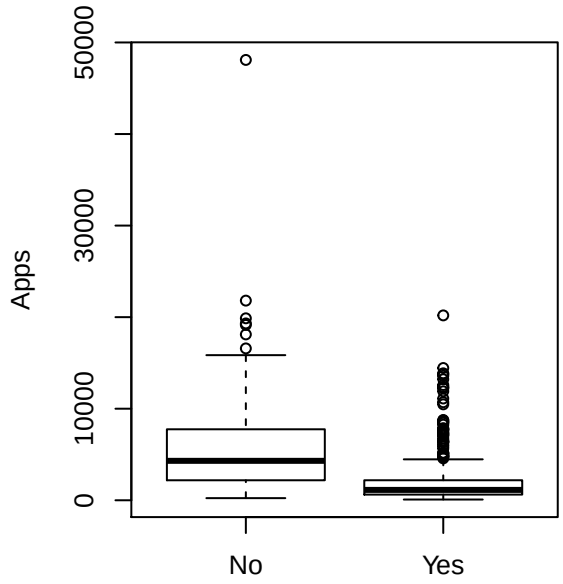


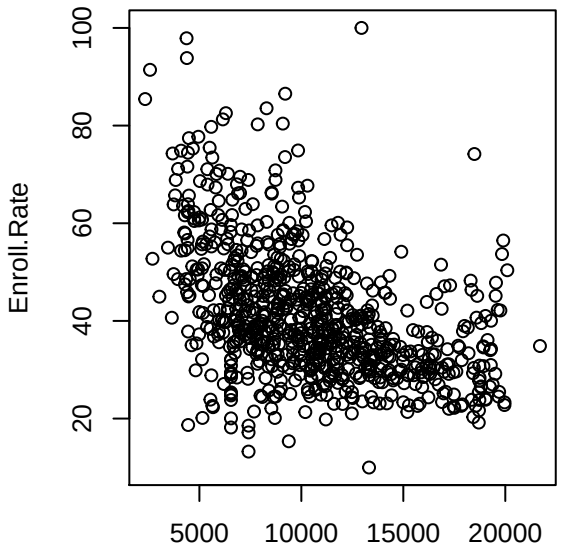




vi-1.bb

```
par(mfrow=c(2,2))
plot(Apps ~ Private, data = college)
plot(Enroll.Rate ~ Outstate, data = college)
```





vi-2.bb

Observations:

- Strong positive correlation between `Grad.Rate` and `Outstate` fees
- Significant difference in `Outstate` fees depending on if the university is `Private` and on the `S.F.Ratio`
- From the second point and above plot, we see that `Private` university colleges have a smaller `S.F.Ratio`
- `Private` university colleges tend to get lot more applications than public
- `Enroll.Rate` is negatively correlated with the college `Outstate` fees