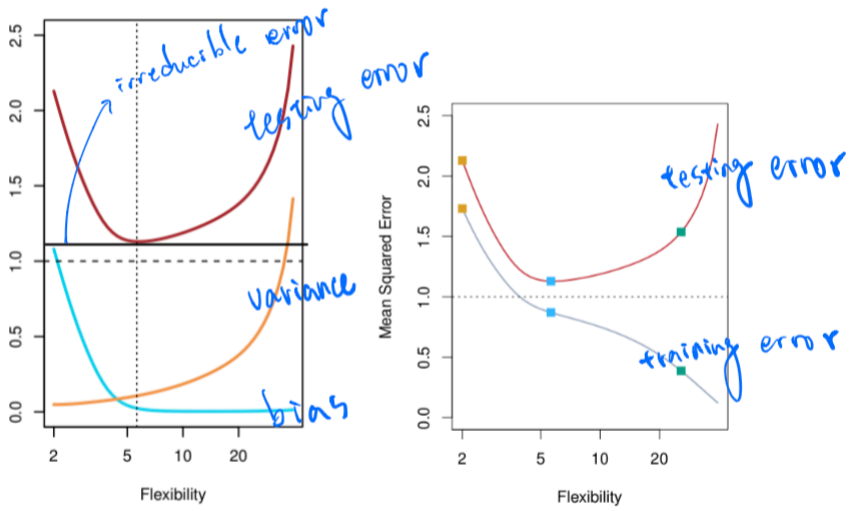
1. Exercise 1
   (a) Flexible approach will be better. Since the variance of the data is small (noise is small), Flexible approach can fit the data well. In addition, the flexible approach can benefit from a large sample size, which can better predict the true shape of the data.
   (b) Inflexible approach will be better. Since the variance of the data is big (noise is large), a flexible approach will overfit the data.
   (c) Flexible is better. The more flexible the approach is, the less bias will be obtained, which is better for non-linear relationship.
   (d) Inflexible approach will be better since the noise is large
2. Exercise 2
   (a) Regression, inference. Since it needs the quantitative output of salary, it need regression to calculate it. In addition, the inference is better for the firm to understand the relationship of each factor and salary.
   N= 500, P= profit, number of employees, industry.
   (b) Classification, prediction. The category is success or failure.
   N=20, P= price, marketing budget, competition price, and ten other variables.
   (c) Regression, prediction. % change is a quantitative output.
   N= weekly data of 2012, P= % change in the US, British, and German market.
3. Exercise 3



Orange line: variance, as flexibility increases, the line will overfit the data, causing variance increases.
Blue line: bias, as flexibility increases, the line will be a closer fit to the data, the bias decreases.
Grey line: training error, as flexibility increases, the training error will decrease since the line overfit the data.
Red line: testing error, as flexibility increases, the testing error will also increase due to the overfitting of the data.
Black line: irreducible error, at the lowest point of testing error, when the variance and the bias reach a minimum balance, the only error left will be the irreducible

error, and the testing error can't be lower than this error. If testing error is lower than the irreducible error, overfitting occurs.

4. Exercise 4
   (a) 1. car selection, response: Benz, Toyota, etc., predictors: car speed, size, design, price, etc., prediction. Use the car speed, size, and design to classify the brand of car, and further predict the possible brand of car.
   2. course classification, response: data mining or English comprehension, predictors: class schedule, course material, course number. Prediction.
   3. phone classification, response: Apple, Samsung, predictors: price, design, system, prediction.
   (b) 1. % of body fat, response: % of body fat, predictors: height, weight. Inference.
   2. salary, response: salary, predictors: age, industry experience. Inference.
   3. % change in the USD/Euro exchange rate, response: % change, predictors: % in the other market. Inference.
   (c) the race of human being, the courses of the department

5. Exercise 5
   a) The advantages of the flexible approach for regression or classification are decreasing bias and better fitting to the non-linear data.
   b) The disadvantages of the flexible approach for regression or classification are increasing variance, overfitting.
   c) When there are less variance, non-linear data, and needing for prediction, the flexible approach is preferred.
   d) When there are more variance, needing for inference, the inflexible approach is preferred.

6. Exercise 6
   a) Parametric approach, which is also the inflexible approach, assume the predicted y to a set of parameters. Nonparametric approach, which is also the flexible approach, doesn't have any formula for predicted y (like black box), needing a lot of observations.
   b) The advantages of a parametric approach to regression or classification are requiring less observations, and will not following too much noise, and will be simpler to interpret.
   c) The disadvantage of a parametric approach to regression or classification is having a limit of fit quality.

7. Exercise 7
   (a)

Obs.	Distance	Y
1	3	Red
2	2	Red
3	3.2	Red
4	2.2	Green
5	1.4	Green
6	1.7	Red

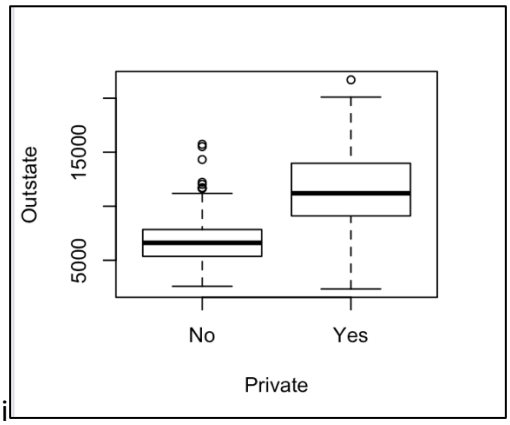
(b) Green, Obs.5 is the closest data of K=1, which is green.
(c) Red, Obs.2, 5, 6 are the closest data of K=3, two of which are red.
(d) Small, a small K is more flexible, better fitting for highly non-linear data.

8. Exercise 8

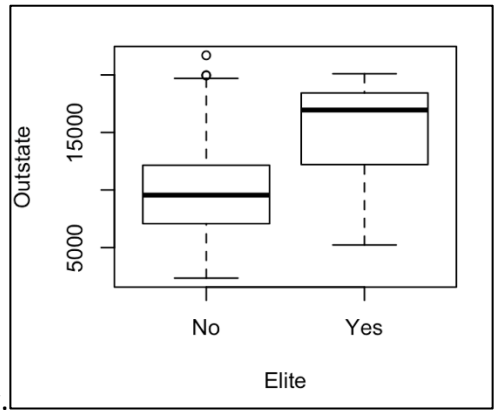
```
install.packages("ISLR")
require(ISLR)
data(College)
str(College)

# i
summary(College)
# ii
pairs(College[,1:10])
# iii
boxplot(Outstate~Private, data=College, xlab="Private", ylab="Outstate")
# iv
Elite <- rep("No", nrow(College))
Elite[College$Top10perc>50] <- "Yes"
College <- data.frame(College, Elite)
summary(College)  # 78 Elite
boxplot(Outstate~Elite, data=College, xlab="Elite", ylab="Outstate")
# v
par(mfrow=c(2,2))
hist(College$Apps, breaks=20, xlim=c(0,20000), main="Apps")
hist(College$Enroll, breaks=20, main="Enroll")
hist(College$Expend, breaks=20, main="Expend")
hist(College$Outstate, breaks=20, main="Outstate")
```

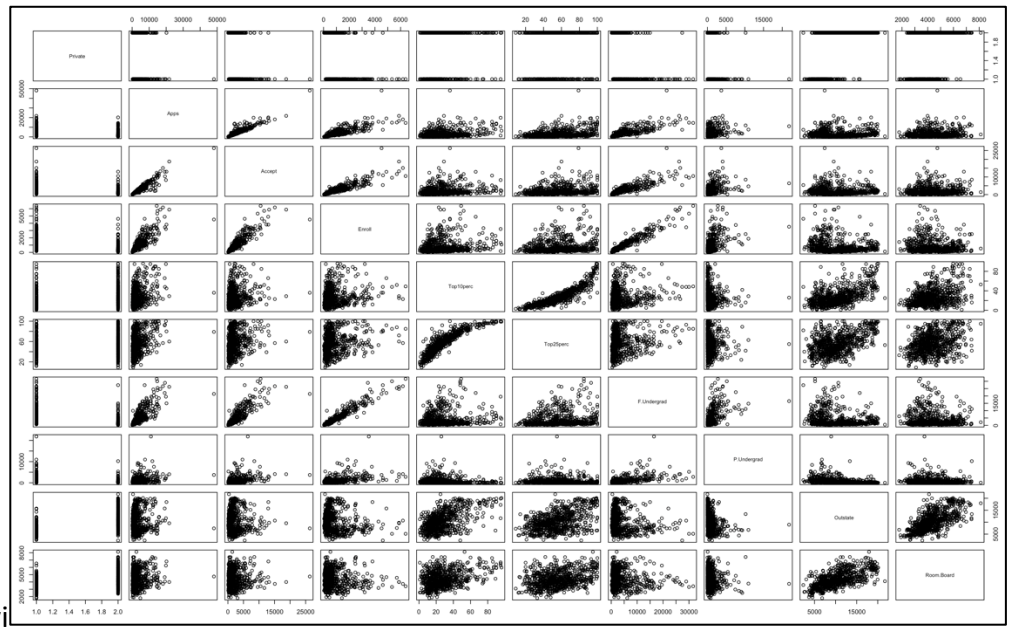
Vi. There is an interesting phenomenon, when the number of applicants accepted increases, the new students from top 10 or 25% school usually won't increase, which means for school, even they increase the number of accepted student, the number of good student from 10 or 25% school is still the same, it won't increase the chance of getting more good student for school. Funny fact!

iii 

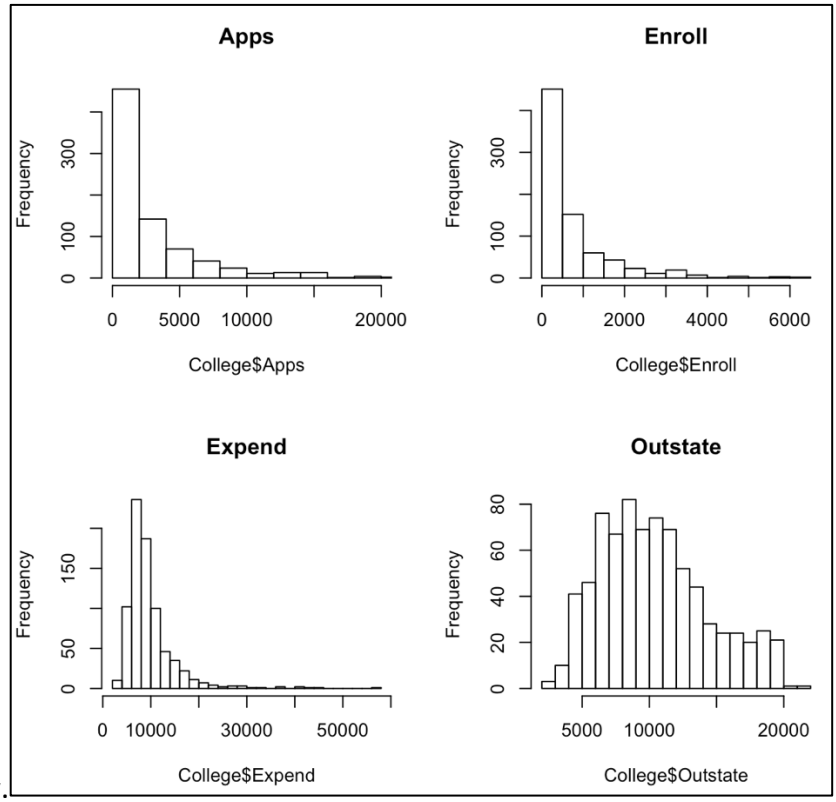
iv.


Outstate
5000
15000
No
Yes
Elite

vi


Private
Apps
Accept
Enroll
Top10perc
Top25perc
F.Undergrad
P.Undergrad
Outstate
Room.Board

v.


Apps
Frequency
0
100
300
0
5000
10000
20000
College$Apps
Enroll
Frequency
0
100
300
0
2000
4000
6000
College$Enroll
Expend
Frequency
0
50
150
0
10000
30000
50000
College$Expend
Outstate
Frequency
0
20
40
60
80
5000
10000
20000
College$Outstate