HW2 - STAT 4510/7510 - Spring 2024

Yang, Anton - #14405729

Due Wednesday, Feb. 7, 11:30 pm (upload PDF to Canvas)

**Instructions:** Please list your name and student number clearly. In order to receive credit for a problem, your solution must show sufficient detail so that the grader can determine how you obtained your answer.

Use R Markdown to create a WORD file. Before submitting, make sure you convert the WORD file to a PDF. All R code should be included, as well as all output produced. Upload your work to the Canvas course site.

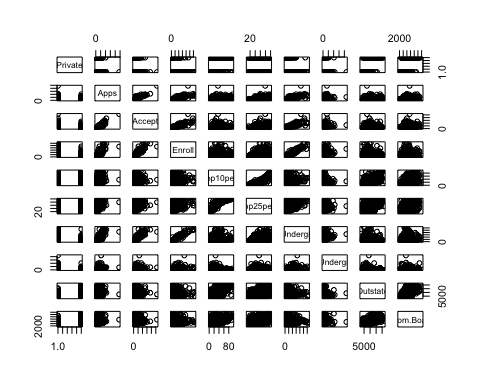
## Problem 1

Complete Chapter 2, problem 8 (p. 54), parts (a), (b), (c.i), (c.ii), and (c.iii). You need not complete the remaining parts or beyond sub-part (iii) of part (c).

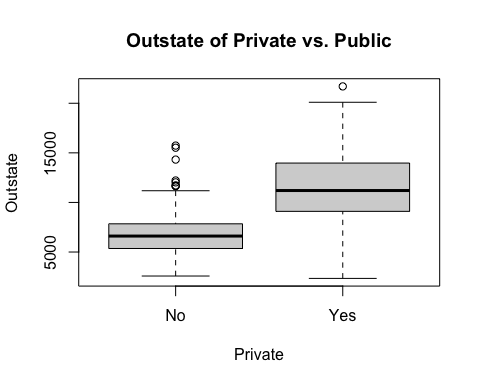
library(e1071)   
library(caTools)   
library(class)  
  
college<-read.csv("College.csv")  
rownames(college)<-college[,1]  
View(college)  
college<-college[,-1]  
View(college)  
summary(college)

## Private Apps Accept Enroll   
## Length:777 Min. : 81 Min. : 72 Min. : 35   
## Class :character 1st Qu.: 776 1st Qu.: 604 1st Qu.: 242   
## Mode :character Median : 1558 Median : 1110 Median : 434   
## Mean : 3002 Mean : 2019 Mean : 780   
## 3rd Qu.: 3624 3rd Qu.: 2424 3rd Qu.: 902   
## Max. :48094 Max. :26330 Max. :6392   
## Top10perc Top25perc F.Undergrad P.Undergrad   
## Min. : 1.00 Min. : 9.0 Min. : 139 Min. : 1.0   
## 1st Qu.:15.00 1st Qu.: 41.0 1st Qu.: 992 1st Qu.: 95.0   
## Median :23.00 Median : 54.0 Median : 1707 Median : 353.0   
## Mean :27.56 Mean : 55.8 Mean : 3700 Mean : 855.3   
## 3rd Qu.:35.00 3rd Qu.: 69.0 3rd Qu.: 4005 3rd Qu.: 967.0   
## Max. :96.00 Max. :100.0 Max. :31643 Max. :21836.0   
## Outstate Room.Board Books Personal   
## Min. : 2340 Min. :1780 Min. : 96.0 Min. : 250   
## 1st Qu.: 7320 1st Qu.:3597 1st Qu.: 470.0 1st Qu.: 850   
## Median : 9990 Median :4200 Median : 500.0 Median :1200   
## Mean :10441 Mean :4358 Mean : 549.4 Mean :1341   
## 3rd Qu.:12925 3rd Qu.:5050 3rd Qu.: 600.0 3rd Qu.:1700   
## Max. :21700 Max. :8124 Max. :2340.0 Max. :6800   
## PhD Terminal S.F.Ratio perc.alumni   
## Min. : 8.00 Min. : 24.0 Min. : 2.50 Min. : 0.00   
## 1st Qu.: 62.00 1st Qu.: 71.0 1st Qu.:11.50 1st Qu.:13.00   
## Median : 75.00 Median : 82.0 Median :13.60 Median :21.00   
## Mean : 72.66 Mean : 79.7 Mean :14.09 Mean :22.74   
## 3rd Qu.: 85.00 3rd Qu.: 92.0 3rd Qu.:16.50 3rd Qu.:31.00   
## Max. :103.00 Max. :100.0 Max. :39.80 Max. :64.00   
## Expend Grad.Rate   
## Min. : 3186 Min. : 10.00   
## 1st Qu.: 6751 1st Qu.: 53.00   
## Median : 8377 Median : 65.00   
## Mean : 9660 Mean : 65.46   
## 3rd Qu.:10830 3rd Qu.: 78.00   
## Max. :56233 Max. :118.00

college$Private<-as.factor(college$Private)  
pairs(college[,1:10])



plot(college$Private,college$Outstate,  
 main = "Outstate of Private vs. Public",  
 xlab = "Private",  
 ylab = "Outstate")



## Problem 2

Continue working with the College.csv data set from problem 1.

1. Split the data into a 80% training and 20% test set. Set a seed of 10 for consistent results. How many observations are in each of the two sets?

set.seed(10)  
split<-sample.split(college,SplitRatio = 0.8)  
  
training\_set<-subset(college, split == TRUE)  
test\_set<-subset(college, split == FALSE)  
  
nrow(training\_set)

## [1] 605

nrow(test\_set)

## [1] 172

There is a total of 605 of observations in the training set and 172 observations in the test set.

1. We want to try to predict whether a college is private using K nearest neighbors. Install the class package (if you haven’t already done so), and remember to run library(class), which contains the knn() function. Change Private to a factor variable. Then predict the classes of your test set using the knn() function with k=8. What is the misclassification rate?

train\_scale<-scale(training\_set[,-1])  
test\_scale<-scale(test\_set[,-1])  
  
classifier\_knn<-knn(train=train\_scale,  
 test=test\_scale,  
 cl=training\_set$Private,  
 k=8)  
  
cm<-table(test\_set$Private, classifier\_knn)  
  
  
misClassError<- 1 - (sum(diag(cm)) / sum(cm))  
print(paste('Misclassification Rate = ',misClassError))

## [1] "Misclassification Rate = 0.0813953488372093"

The misclassification rate for k=8 is approximately 0.08.

1. Repeat the KNN analysis using a values of k = 5, k = 10, k = 15, and k = 20. Find the misclassification rate for each value of k and comment on your results.

k\_values <- c(5, 10, 15, 20)  
  
knn\_classifier <- sapply(k\_values, function(k) {  
 classifier\_knn <- knn(train = train\_scale,  
 test = test\_scale,  
 cl = training\_set$Private,  
 k = k)  
 cm <- table(test\_set$Private, classifier\_knn)  
   
 misClassError <- 1 - (sum(diag(cm)) / sum(cm))  
})  
  
accuracy\_data <- data.frame(K = k\_values, Accuracy = 1 - knn\_classifier)  
  
print(accuracy\_data)

## K Accuracy  
## 1 5 0.9244186  
## 2 10 0.9302326  
## 3 15 0.9244186  
## 4 20 0.9244186

Out of the 4 K values, highest misclassification rate are when K = 10 and k = 5, 15, 20 have the same misclassification rate.