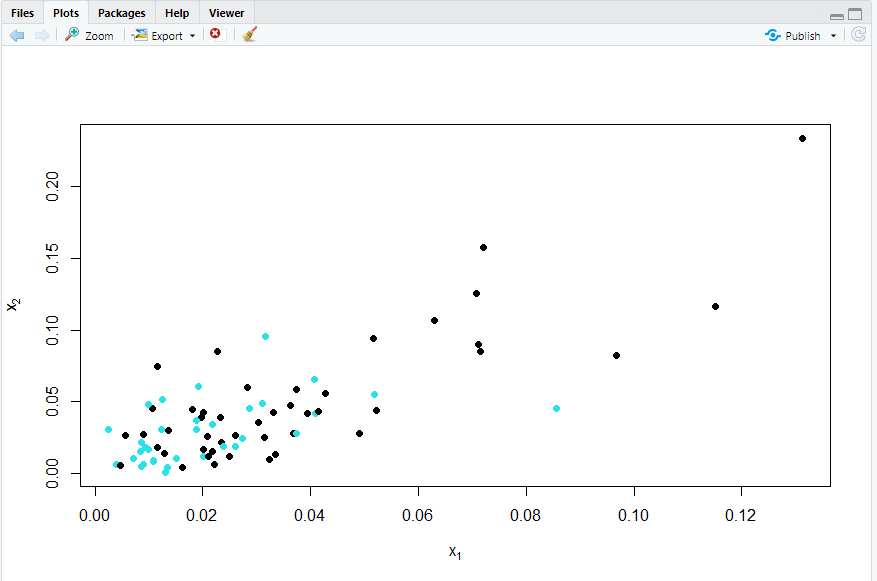
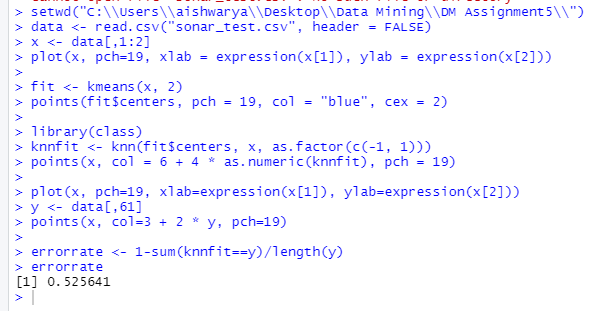
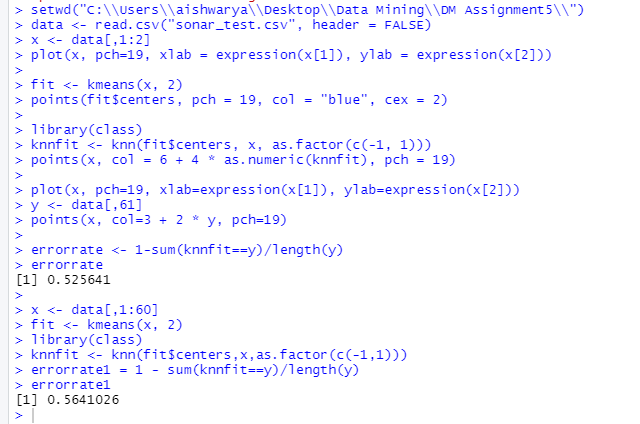
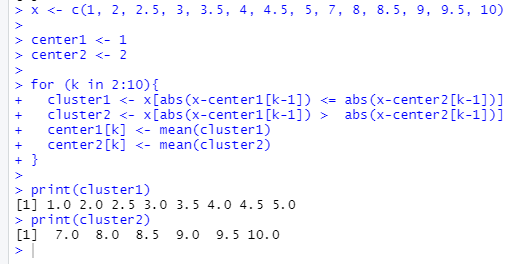
Data Mining Assignment 5

1) Read Chapter 8 (Sections 8.1 and 8.2) and Chapter 2 (Section 2.4).  
  
2) Repeat In Class Exercise #50 using the sonar test data instead of the sonar training data and show your R commands for doing so.  
  
3) Repeat In Class Exercise #52 using the sonar test data instead of the sonar training data and show your R commands for doing so.  
  
4) Repeat In Class Exercise #53 using the sonar test data instead of the sonar training data and show your R commands for doing so.

  
  
5) Repeat In Class Exercise #54 using the data x<-c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10) instead. Show all your work for each step and be sure to say specifically which points are in each cluster at each step.

  
  
6) Repeat In Class Exercise #55 using the data x<-c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10) instead and show your R commands for doing so.

Ans)

x <- c(1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 7, 8, 8.5, 9, 9.5, 10)

center1 <- 1

center2 <- 2

for (k in 2:10){

cluster1 <- x[abs(x-center1[k-1]) <= abs(x-center2[k-1])]

cluster2 <- x[abs(x-center1[k-1]) > abs(x-center2[k-1])]

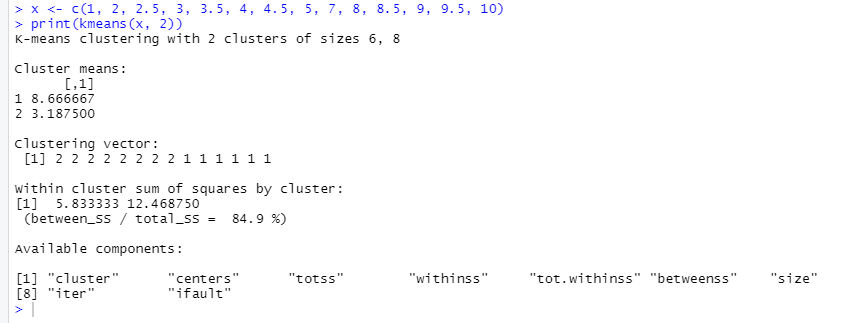
center1[k] <- mean(cluster1)

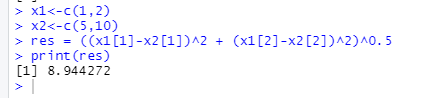
center2[k] <- mean(cluster2)

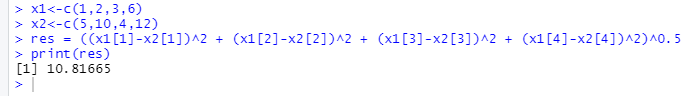
}

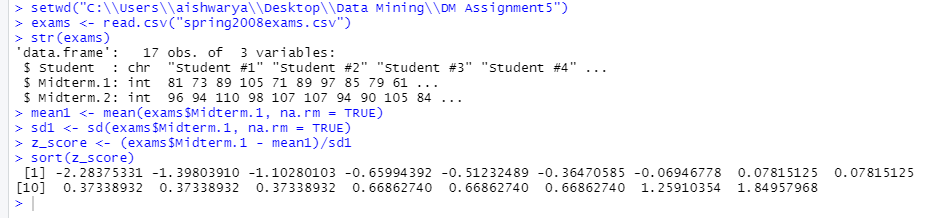
print(cluster1)

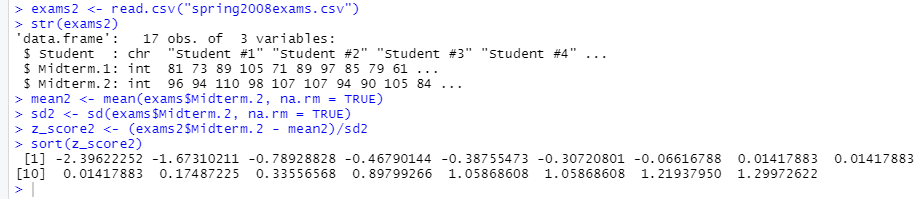
print(cluster2)  
  
7) Repeat In Class Exercise #56 using the data x<-c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10) instead and show your R commands for doing so.

  
  
8) Consider the points x1<-c(1,2) and x2<-c(5,10).  
  
a) Compute the (Euclidean) distance by hand. Show your work and include a picture of the triangle for the Pythagorean Theorem.  
  
b) Verify that the dist function in R gives the same value as you got in part a. Show your R commands for doing so.

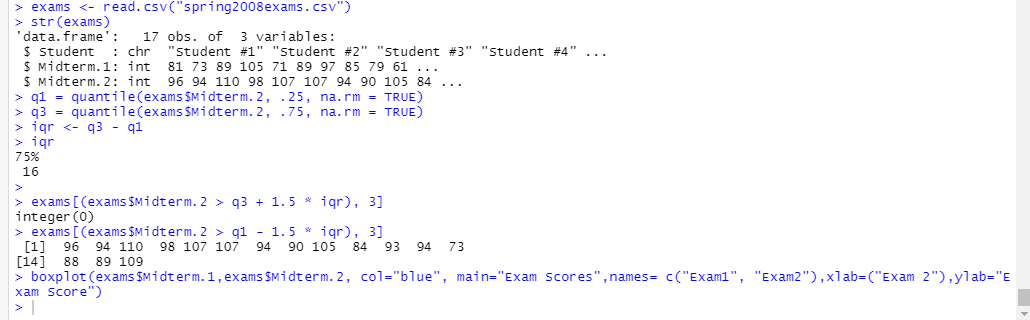
  
  
9) Consider the points x1<-c(1,2,3,6) and x2<-c(5,10,4,12).  
  
a) Compute the (Euclidean) distance by hand. Show your work.  
  
b) Verify that the dist function in R gives the same value as you got in part a. Show your R commands for doing so.

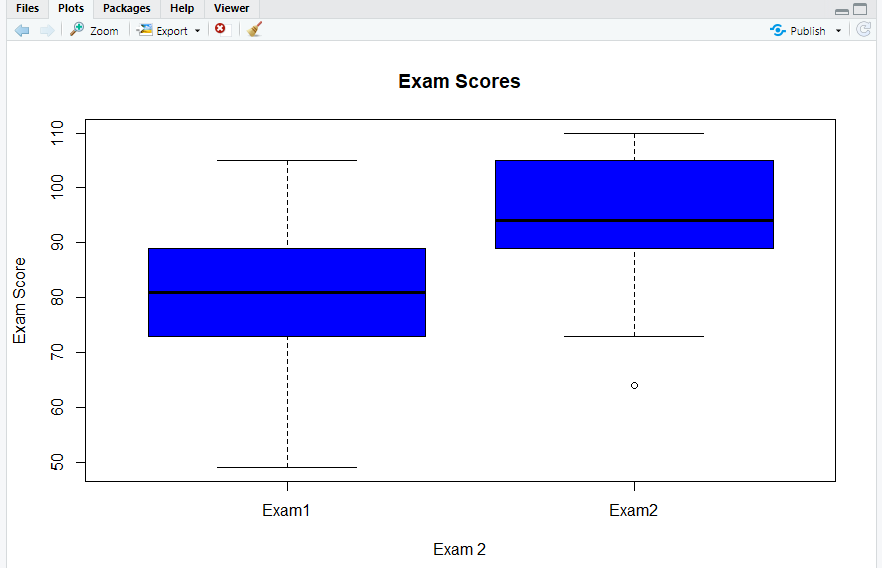
  
  
10) Read Chapter 10.  
  
11) Repeat In Class Exercise #59 using the grades for the first midterm at [www.stats202.com/spring2008exams.csv](http://www.stats202.com/spring2008exams.csv). Are there any outliers according to the z=+/-3 rule? What is the value of the largest z score and what is the value of the smallest (most negative) z score? Show your R commands.

  
  
12) Repeat In Class Exercise #59 using the grades for the second midterm at [www.stats202.com/spring2008exams.csv](http://www.stats202.com/spring2008exams.csv). Are there any outliers according to the z=+/-3 rule? What is the value of the largest z score and what is the value of the smallest (most negative) z score? Show your R commands.

  
  
13) Repeat In Class Exercise #60 using Excel for the user agent column of the data at [www.stats202.com/stats202log.txt](http://www.stats202.com/stats202log.txt). (The user agent column is the second to last column and the value for it in the first row is "Mozilla/4.0 (compatible; MSIE 7.0; Windows NT 5.1; .NET CLR 1.1.4322)"). What user agents are identified as outliers using the z=+/-3 rule on the counts of the user agents? What are the z scores for these outliers? (You do not need to show any work for this problem because you are using Excel.)

14) Repeat In Class Exercise #61 using the grades for the second midterm at [www.stats202.com/spring2008exams.csv](http://www.stats202.com/spring2008exams.csv). Show your R commands and include the boxplot. Are any of the grades for the second midterm outliers by this rule? If so, which ones?



  
  
15) Repeat In Class Exercise #62 using the midterm grades at [www.stats202.com/spring2008exams.csv](http://www.stats202.com/spring2008exams.csv). Be sure to include the plot. Which student # had the largest POSITIVE residual? Show your R commands.

