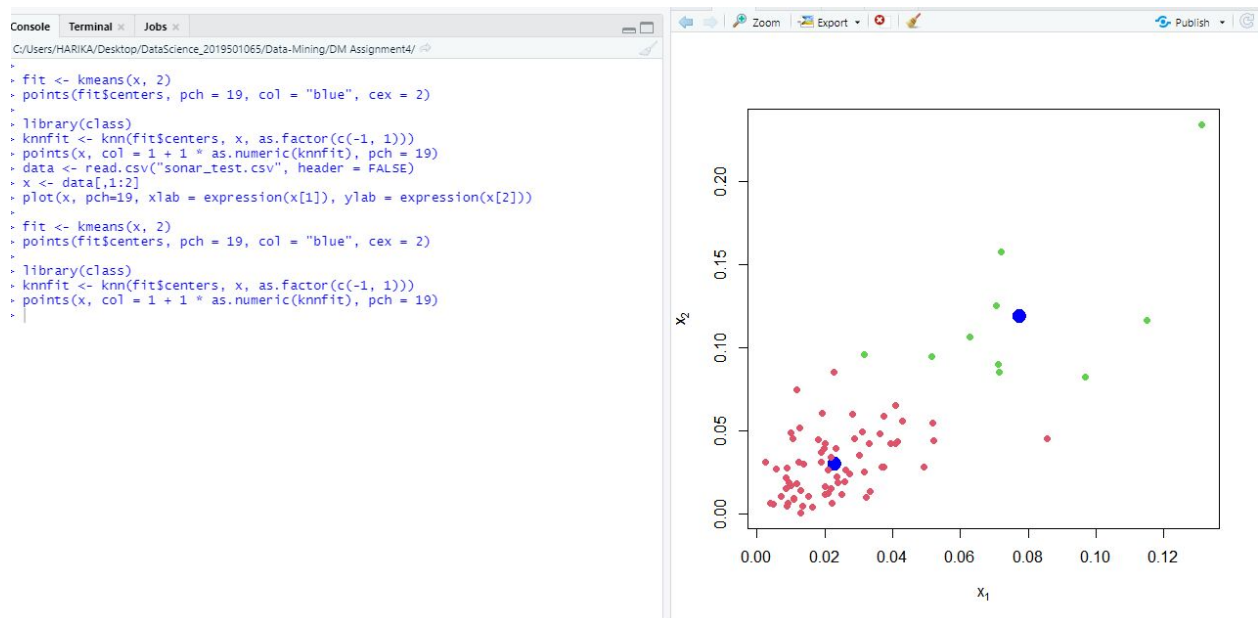
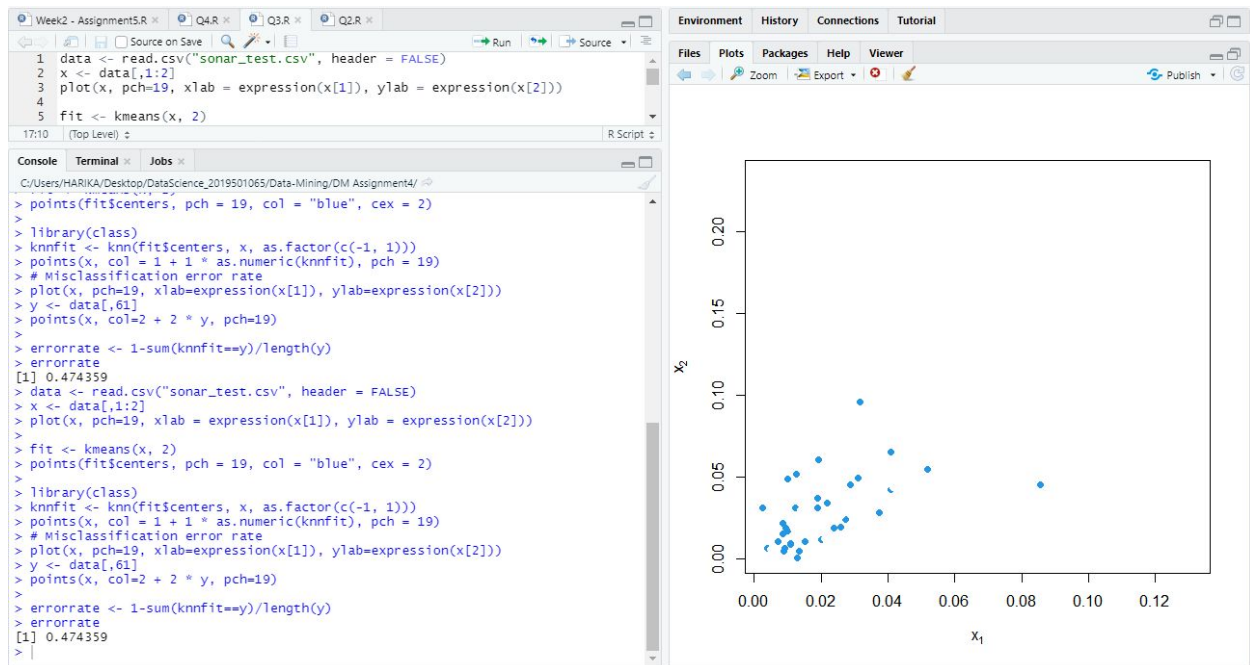


Data Mining Assignment 5

- 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.

```

> data <- read.csv("sonar_test.csv", header = FALSE)
> x <- data[,1:2]
> plot(x, pch=19, xlab = expression(x[1]), ylab = expression(x[2]))
>
> fit <- kmeans(x, 2)
> points(fit$centers, pch = 19, col = "blue", cex = 2)
>
> library(class)
> knnfit <- knn(fit$centers, x, as.factor(c(-1, 1)))
> points(x, col = 1 + 1 * as.numeric(knnfit), pch = 19)
> # Misclassification error rate
> plot(x, pch=19, xlab=expression(x[1]), ylab=expression(x[2]))
> y <- data[,61]
> points(x, col=2 + 2 * y, pch=19)
>
> errorrate <- 1-sum(knnfit==y)/length(y)
> errorrate
[1] 0.525641
> x <- data[,1:60]
> fit <- kmeans(x, 2)
> library(class)
> knnfit <- knn(fit$centers,x,as.factor(c(-1,1)))
> errorrate1 = 1 - sum(knnfit==y)/length(y)
> errorrate1
[1] 0.4358974
>

```

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

```
> print(cluster1)
[1] 1.0 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> print(cluster2)
[1] 7.0 8.0 8.5 9.0 9.5 10.0
> |
```

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

```
> print(cluster1)
[1] 1.0 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> print(cluster2)
[1] 7.0 8.0 8.5 9.0 9.5 10.0
> |
```

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

```
> print(kmeans(x, 2))
K-means clustering with 2 clusters of sizes 6, 8

cluster means:
      [,1]
1 8.666667
2 3.187500

clustering vector:
[1] 2 2 2 2 2 2 2 2 1 1 1 1 1 1

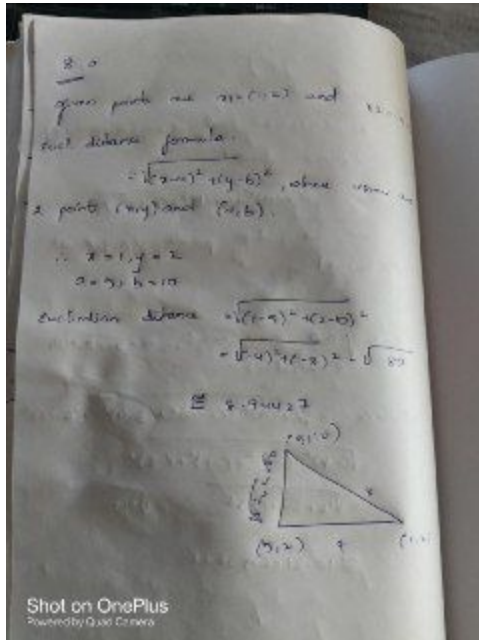
within cluster sum of squares by cluster:
[1] 5.833333 12.468750
(between_SS / total_SS = 84.9 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> |
```

8) Consider the points $x1 \leftarrow c(1,2)$ and $x2 \leftarrow c(5,10)$.

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.

```
> x1<-c(1,2)
> x2<-c(5,10)
> result = ((x1[1]-x2[1])^2 + (x1[2]-x2[2])^2)^0.5
> print(result)
[1] 8.944272
> |
```

9) Consider the points $x1 \leftarrow c(1,2,3,6)$ and $x2 \leftarrow c(5,10,4,12)$.

a) Compute the (Euclidean) distance by hand. Show your work.

Qa.

Given points are

$$x_1 = (1, 2, 3, 6) \text{ and } x_2 = (5, 10, 4, 12)$$

Euclidian distance formula is

$$\sqrt{(x-a)^2 + (y-b)^2 + (z-c)^2 + (w-d)^2}$$

where there are two points (x, y, z, w)
and (a, b, c, d)

Here

$$x=1, y=2, z=3, w=6; a=5, b=10, \\ c=4, d=12.$$

$$\text{Eucl. distance} = \sqrt{(1-5)^2 + (2-10)^2 + (3-4)^2 + (6-12)^2}$$

$$= \sqrt{(-4)^2 + (-8)^2 + (-1)^2 + (-6)^2}$$

$$= \sqrt{16 + 64 + 1 + 36} = \sqrt{117}$$

$$= \cong 10.816653826.$$

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.

```
> x1<-c(1,2,3,6)
> x2<-c(5,10,4,12)
> result = ((x1[1]-x2[1])^2 + (x1[2]-x2[2])^2 + (x1[3]-x2[3])^2 + (x1[4]-x2[4])^2)^0.5
> print(result)
[1] 10.81665
>
```

10) Read Chapter 10.

11) Repeat In Class Exercise #59 using the grades for the first midterm at www.stats202.com/spring2008exams.csv. Are there any outliers according to the $z=\pm 3$ 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.

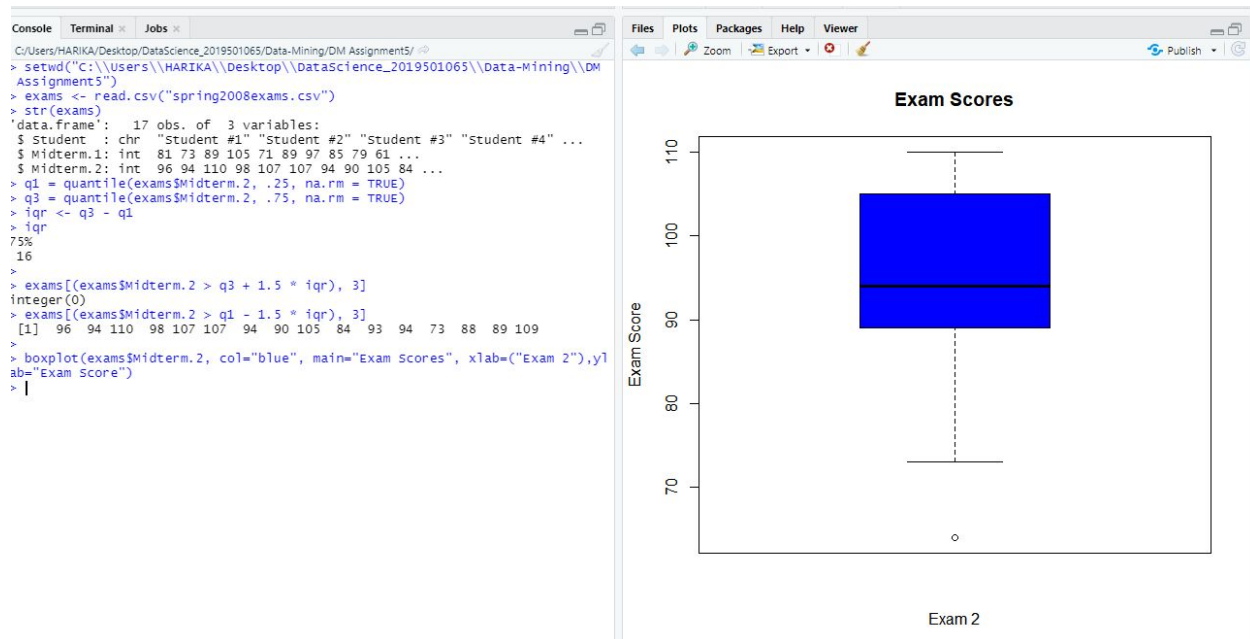
```
> setwd("C:\\Users\\HARIKA\\Desktop\\DataScience_2019501065\\Data-Mining\\DM Assignment5")
> exams <- read.csv("spring2008exams.csv")
> str(exams)
'data.frame': 17 obs. of 3 variables:
 $ Student : chr "Student #1" "Student #2" "Student #3" "Student #4" ...
 $ Midterm.1: int 81 73 89 105 71 89 97 85 79 61 ...
 $ Midterm.2: int 96 94 110 98 107 107 94 90 105 84 ...
> mean1 <- mean(exams$Midterm.1, na.rm = TRUE)
> sd1 <- sd(exams$Midterm.1, na.rm = TRUE)
> z_score <- (exams$Midterm.1 - mean1)/sd1
> sort(z_score)
[1] -2.28375331 -1.39803910 -1.10280103 -0.65994392 -0.51232489 -0.36470585 -0.06946778 0.07815125
[9] 0.07815125 0.37338932 0.37338932 0.37338932 0.66862740 0.66862740 0.66862740 1.25910354
[17] 1.84957968
>
```

12) Repeat In Class Exercise #59 using the grades for the second midterm at www.stats202.com/spring2008exams.csv. Are there any outliers according to the $z=\pm 3$ 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.

```
> mean2 <- mean(exams$Midterm.2, na.rm = TRUE)
> sd2 <- sd(exams$Midterm.2, na.rm = TRUE)
> z_score2 <- (exams$Midterm.2 - mean2)/sd2
> sort(z_score2)
[1] -2.39622252 -1.67310211 -0.78928828 -0.46790144 -0.38755473 -0.30720801 -0.06616788 0.01417883
[9] 0.01417883 0.01417883 0.17487225 0.33556568 0.89799266 1.05868608 1.05868608 1.21937950
[17] 1.29972622
>
```

13) Repeat In Class Exercise #60 using Excel for the user agent column of the data at 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 $z=\pm 3$ 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. 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. Be sure to include the plot. Which student # had the largest POSITIVE residual? Show your R commands.

