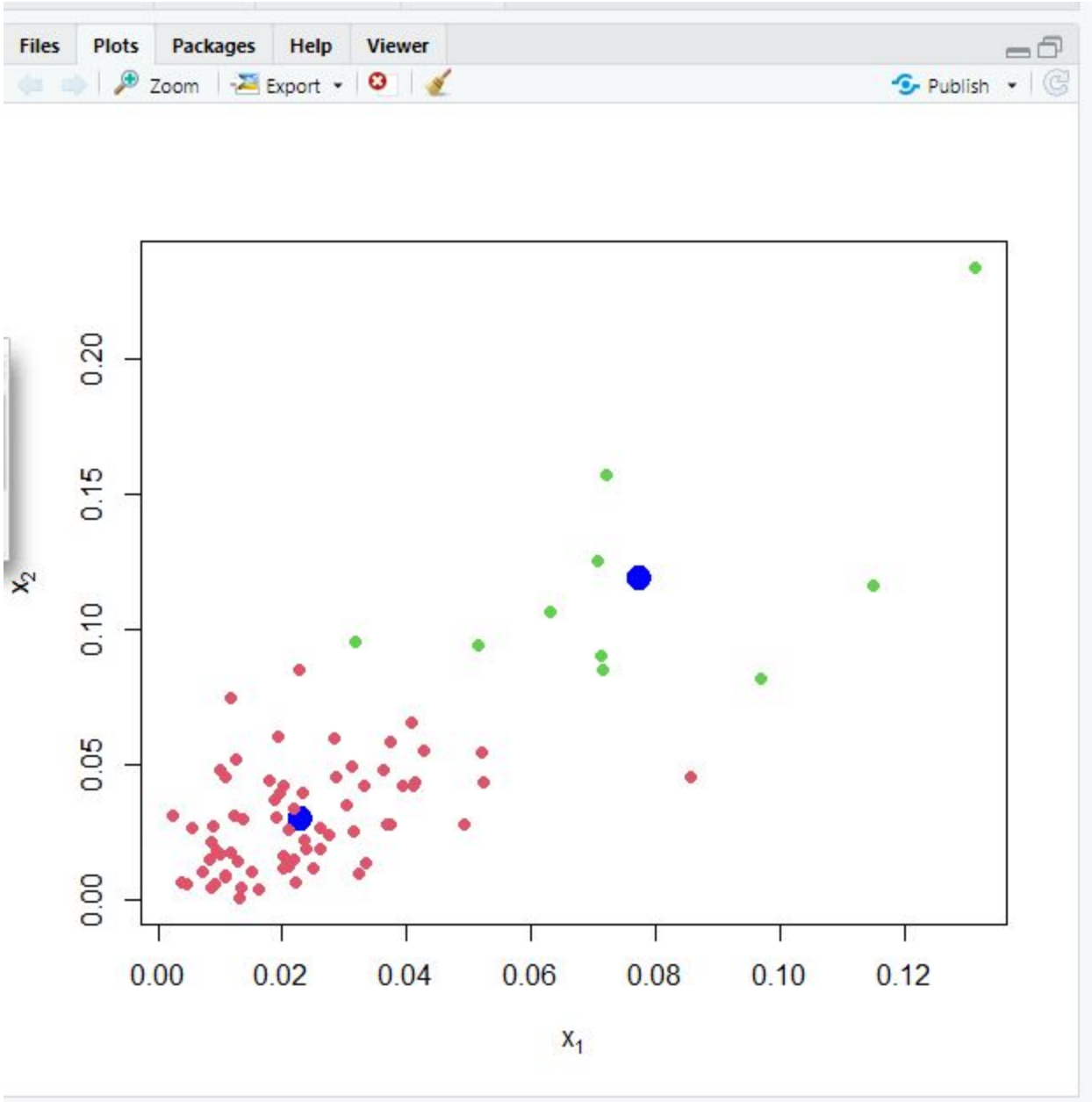
Data Mining Assignment 5

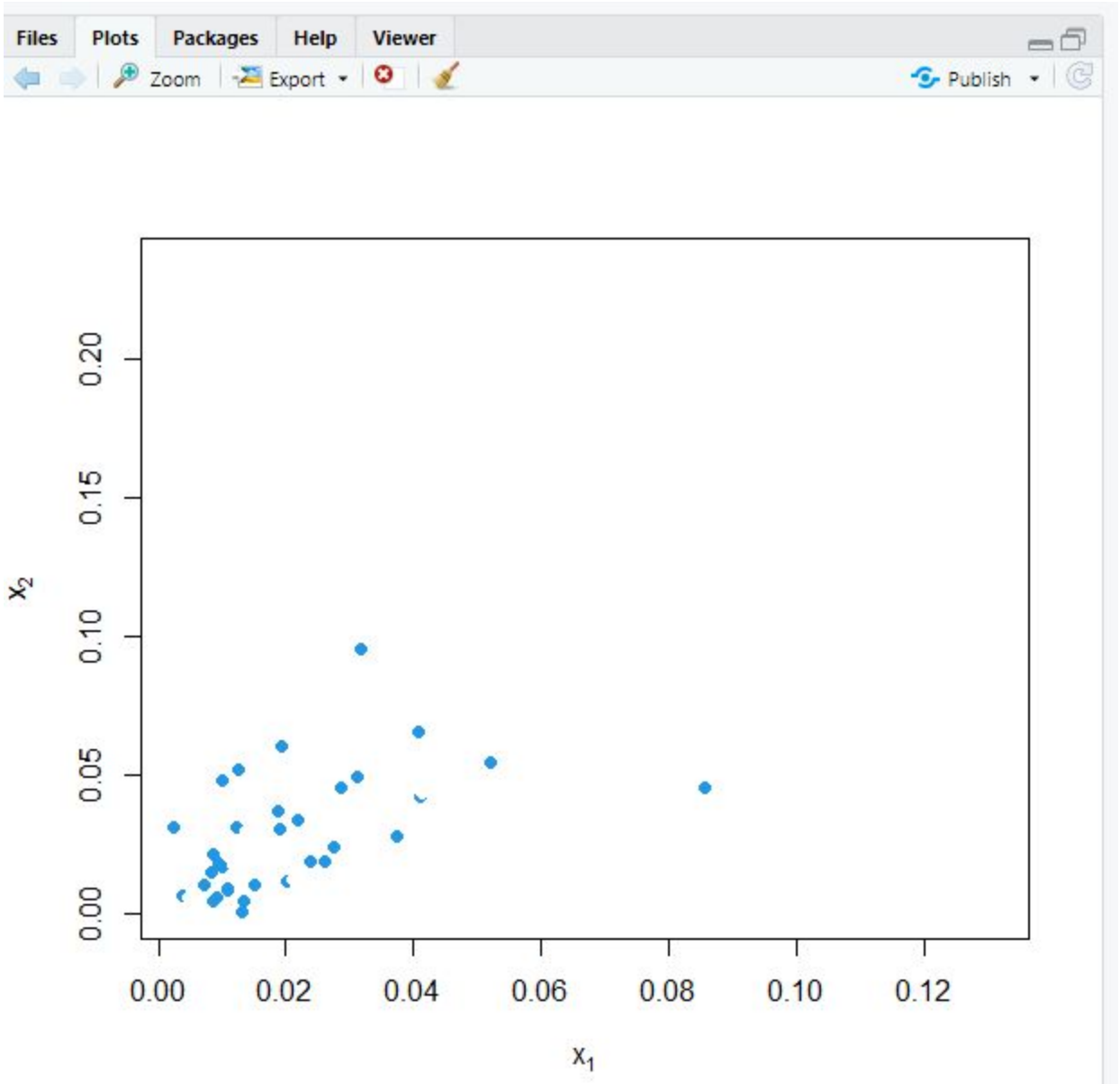
1) Read Chapter 8 (Sections 8.1 and 8.2) and Chapter 2 (Section 2.4).

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.

```
> library(class)
> knnfit <- knn(fit$centers, x, as.factor(c(-1, 1)))
> points(x, col = 1 + 1 * as.numeric(knnfit), pch = 19)
> 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
>
```



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.

```
> 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<-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.

```
> 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)
[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<-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.

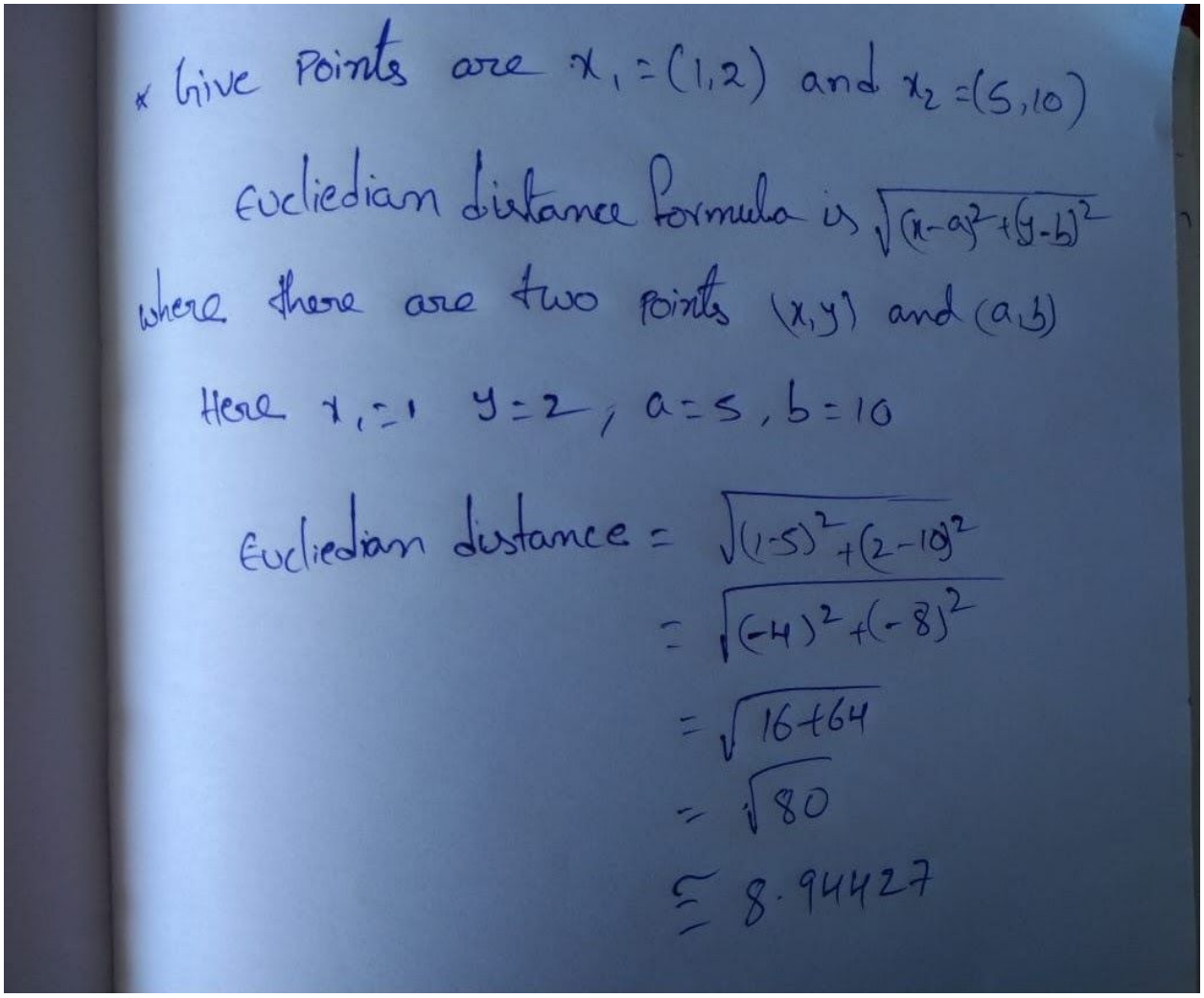
```
> center2 <- 2
>
> for (k in 2:10){
+   cluster1 <- x[abs(x-center1[k-1]) <= abs(x-center2[k-1])]
+   cluster2 <- x[abs(x-center1[k-1]) >  abs(x-center2[k-1])]
+   center1[k] <- mean(cluster1)
+   center2[k] <- mean(cluster2)
+ }
> print(cluster1)
[1] 1.0 2.0 2.5 3.0 3.5 4.0 4.5 5.0
> print(cluster2)
[1]  7.0  8.0  8.5  9.0  9.5 10.0
> |
```

7) Repeat In Class Exercise #56 using the data x<-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.

```
> x <- c(1, 2, 2.5, 3, 3.5, 4, 4.5, 5, 7, 8, 8.5, 9, 9.5, 10)
> 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"
[6] "betweenss"    "size"         "iter"         "ifault"
> |
```

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.

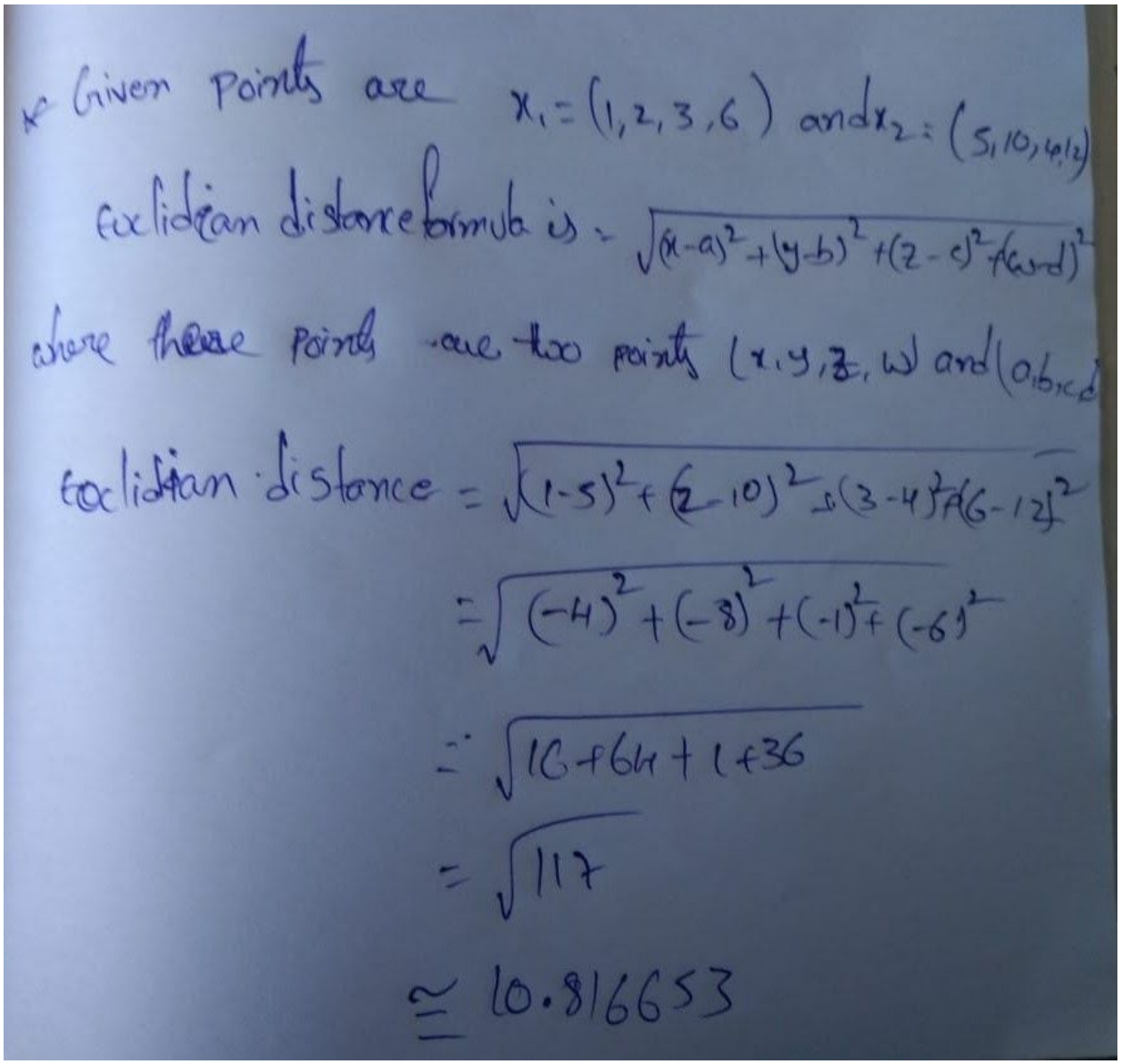


b) Verify that the dist function in R gives the same value as you got in part a. Show your R commands for doing so.

```
E:/2nd Year/DataScience_2019501107/Data Mining/DM Assignment4/
> x1 <- c(1, 2)
> x2 <- c(5, 10)
> data <- matrix(c(x1,x2),nrow=2,byrow=T)
> dist(data)
         1
2 8.944272
>
```

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.

```
Console  Jobs
E:/2nd Year/DataScience_2019501107/Data Mining/DM Assignment4/
> x1 <- c(1, 2, 3, 6)
> x2 <- c(5, 10, 4, 12)
> data <- matrix(c(x1,x2),nrow=2,byrow=T)
> dist(data)
         1
2 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=+/-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.

```
Console  Jobs
E:/2nd Year/DataScience_2019501107/Data Mining/DM Assignment5/
> setwd("E:\\2nd Year\\DataScience_2019501107\\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
 [7] -0.06946778  0.07815125  0.07815125  0.37338932  0.37338932  0.37338932
[13]  0.66862740  0.66862740  0.66862740  1.25910354  1.84957968
>
```

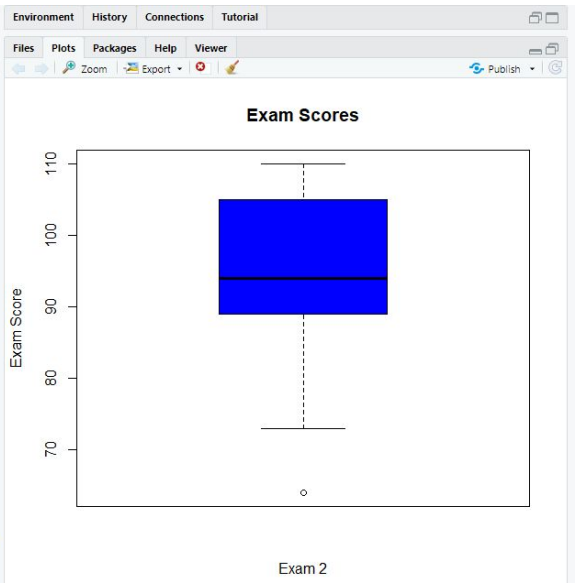
12) Repeat In Class Exercise #59 using the grades for the second midterm at www.stats202.com/spring2008exams.csv. Are there any outliers according to the z=+/-3 rule? What is the value of the largest z score and what is the value of the smallest (most negative) z score? Show your R commands.

```
Console  Jobs
E:/2nd Year/DataScience_2019501107/Data Mining/DM Assignment5/
> setwd("E:\\2nd Year\\DataScience_2019501107\\Data Mining\\DM Assignment5")
> exams2 <- read.csv("spring2008exams.csv")
> str(exams2)
'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 ...
> mean2 <- mean(exams$Midterm.2, na.rm = TRUE)
> sd2 <- sd(exams$Midterm.2, na.rm = TRUE)
> z_score2 <- (exams2$Midterm.2 - mean2)/sd2
> sort(z_score2)
 [1] -2.39622252 -1.67310211 -0.78928828 -0.46790144 -0.38755473 -0.30720801
 [7] -0.06616788  0.01417883  0.01417883  0.01417883  0.17487225  0.33556568
[13]  0.89799266  1.05868608  1.05868608  1.21937950  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=+/-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?

```
Source
Console  Jobs
E:/2nd Year/DataScience_2019501107/Data Mining/DM Assignment5/
> setwd("E:\\2nd Year\\DataScience_2019501107\\Data Mining\\DM Assignment5")
> exams2 <- read.csv("spring2008exams.csv")
> str(exams2)
'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 ...
> mean2 <- mean(exams$Midterm.2, na.rm = TRUE)
> sd2 <- sd(exams$Midterm.2, na.rm = TRUE)
> z_score2 <- (exams2$Midterm.2 - mean2)/sd2
> sort(z_score2)
 [1] -2.39622252 -1.67310211 -0.78928828 -0.46790144 -0.38755473 -0.30720801
 [7] -0.06616788  0.01417883  0.01417883  0.01417883  0.17487225  0.33556568
[13]  0.89799266  1.05868608  1.05868608  1.21937950  1.29972622
> setwd("E:\\2nd Year\\DataScience_2019501107\\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 ...
> q1 = quantile(exams$Midterm.2, .25, na.rm = TRUE)
> q3 = quantile(exams$Midterm.2, .75, na.rm = TRUE)
> iqr <- q3 - q1
> iqr
75%
 16
>
> exams[(exams$Midterm.2 > q3 + 1.5 * iqr), 3]
integer(0)
> exams[(exams$Midterm.2 < q1 - 1.5 * iqr), 3]
 [1]  96  94 110  98 107 107  94  90 105  84  93  94  73  88  89 109
>
> boxplot(exams$Midterm.2, col="blue", main="Exam Scores", xlab=("Exam 2"),ylab="Exam Sco
re")
>
```



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.

```
> setwd("E:\\2nd Year\\DataScience_2019501107\\Data Mining\\DM Assignment5")
> exams <- read.csv("spring2008exams.csv")
> model <- lm(exams$Midterm.2 ~ exams$Midterm.1)
> plot(exams$Midterm.1, exams$Midterm.2, pch=19,xlab="Exam 1", ylab="Exam2",xlim=c(10,10
0),ylim=c(10,100))
> abline(model)
> sort(model$residuals)
        11         4         7        16         8        12        14
-21.4761256  -9.8235058  -9.3540298  -7.6498157  -6.6498157  -3.6498157  -3.5371735
        15        13        10         1         2         6        17
 -3.0629707  -0.4150777   0.7586124   1.5849223   4.0543983   8.1154462  10.1154462
         3         9         5
 11.1154462  11.7022913  18.1717673
>
```