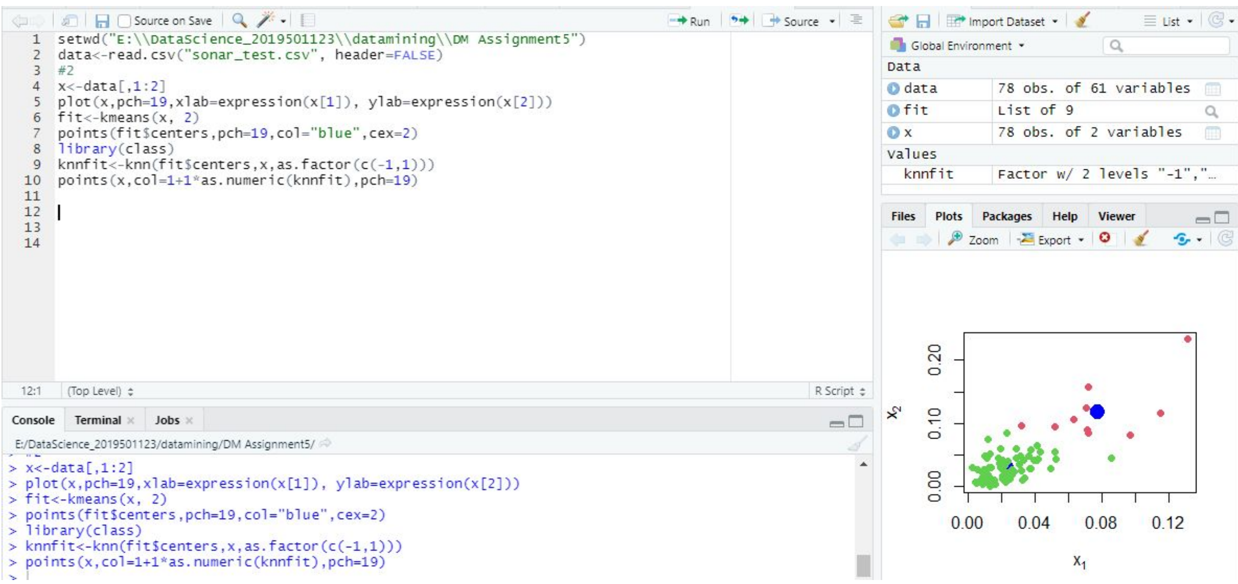
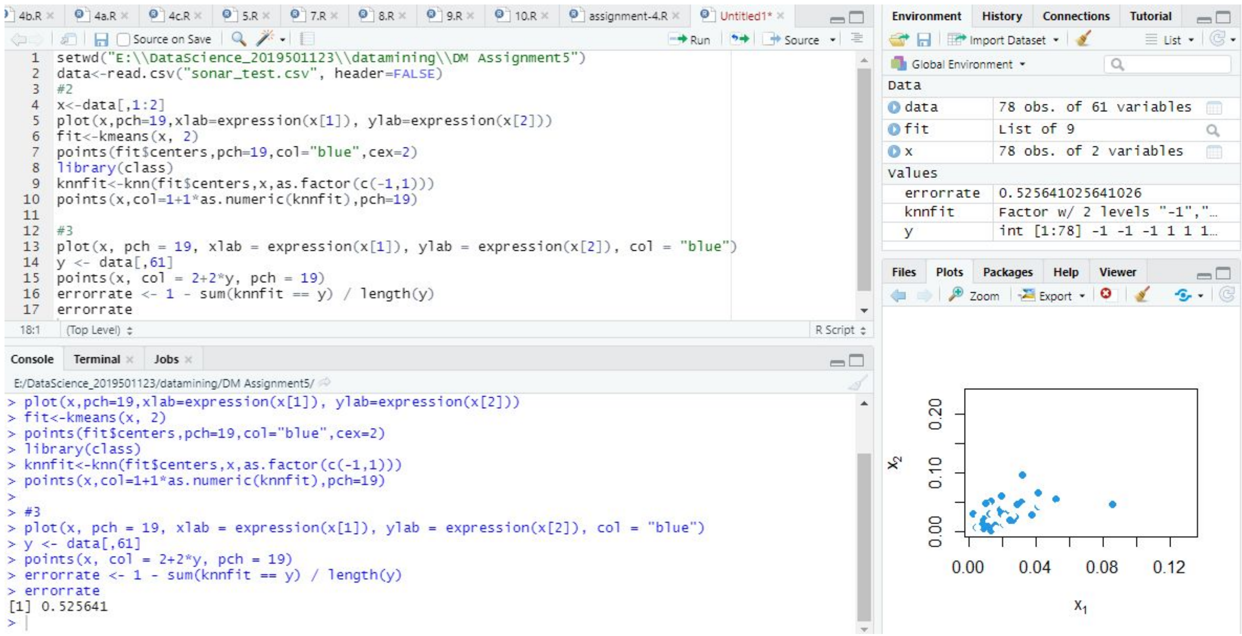
Data Mining Assignment 5

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

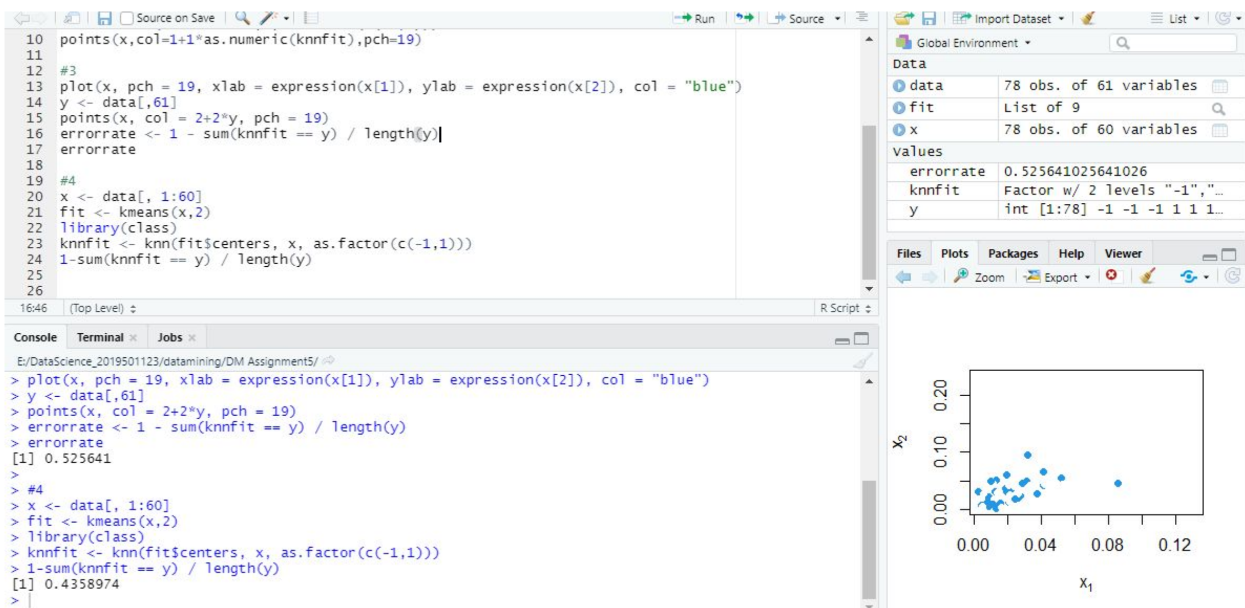
2. Use Kmeans() with all `the default values to find the k=2 solution for the first two columns of the sonar test data. Plot these two columns. Also plot the fitted cluster centers using a different color. Finally use the knn() function to assign the cluster membership for the points to the nearest cluster center. Color the points according to their cluster membership. Show your R commands for doing so.



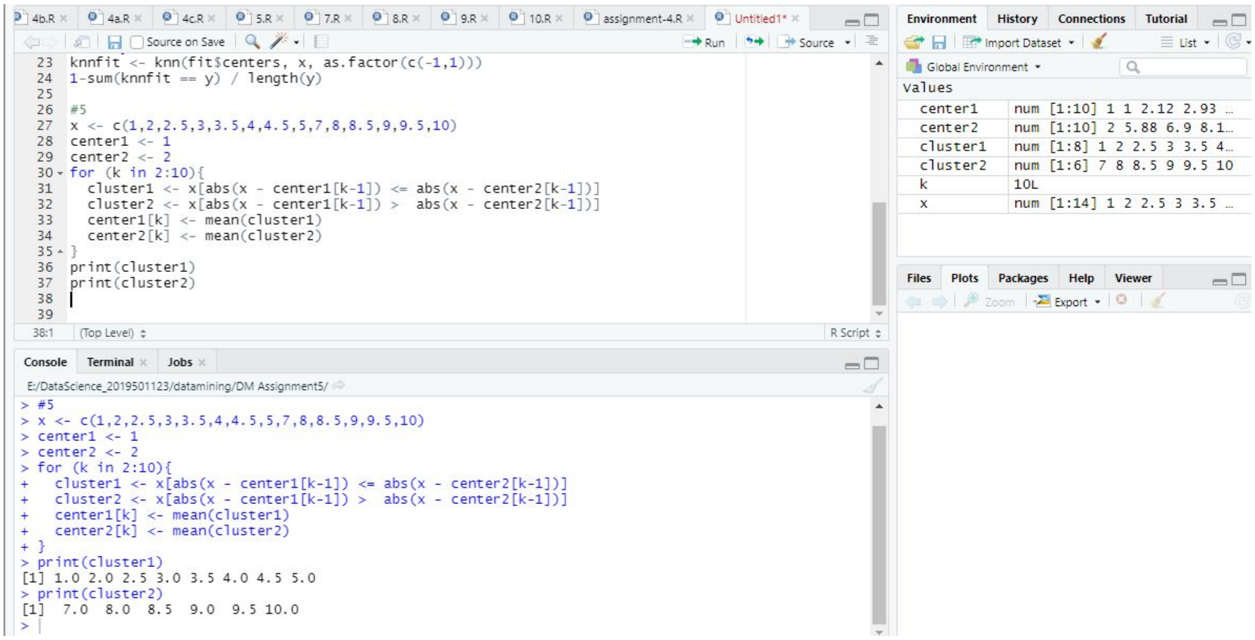
3. Graphically compare the cluster memberships from the previous problem to the actual labels in the test data. Also, compute the misclassification error that would result if you used your clustering rule to classify the data. Show your R commands for doing so.



4. **Repeat the previous problem using all 60 columns. Show your R commands for doing so.**



5. **Consider the one dimensional data set given x←c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10). Starting with initial cluster center values of 1 and 2 carry out algorithm 10 until convergence by hand for k=2 clusters. Show all your work for each step and be sure to say specifically which points are in each cluster at each step.**



6. Repeat the previous problem by writing a loop and verify that the final answer is the same and show your R commands for doing so

```r
x <- data[, 1:60]
fit <- kmeans(x,2)
library(class)
knnfit <- knn(fit$centers, x, as.factor(c(-1,1)))
1-sum(knnfit == y) / length(y)

#5
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)
```

```
E:/DataScience_2019501123/datamining/DM Assignment5/
> setwd("E:\\DataScience_2019501123\\datamining\\DM Assignment5")
> #5
> 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)
[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. Verify that the kmeans function gives the same solution for the previous problem when you use all of the default values and show your R commands for doing so.

```r
#7
km <- kmeans(x,2)
print(km)
```

42:1 (Top Level) ÷ R Scr

Console Terminal × Jobs ×

E:/DataScience_2019501123/datamining/DM Assignment5/

```
Cluster means:
      [,1]
1 3.187500
2 8.666667

Clustering vector:
 [1] 1 1 1 1 1 1 1 1 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 12.468750  5.833333
 (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<-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.

a(a)

Given points are $x_1=(1,2)$ and $x_2=(5,10)$

Euclidean distance formula is $\sqrt{(x-a)^2+(y-b)^2}$ where there are two points (x,y) and (a,b)

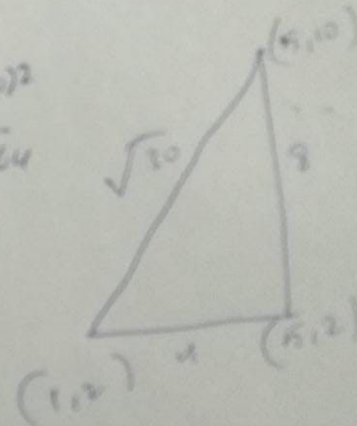
Here $x=1$, $y=2$, $a=5$, $b=10$

Euclidean distance $=\sqrt{(1-5)^2+(2-10)^2}$

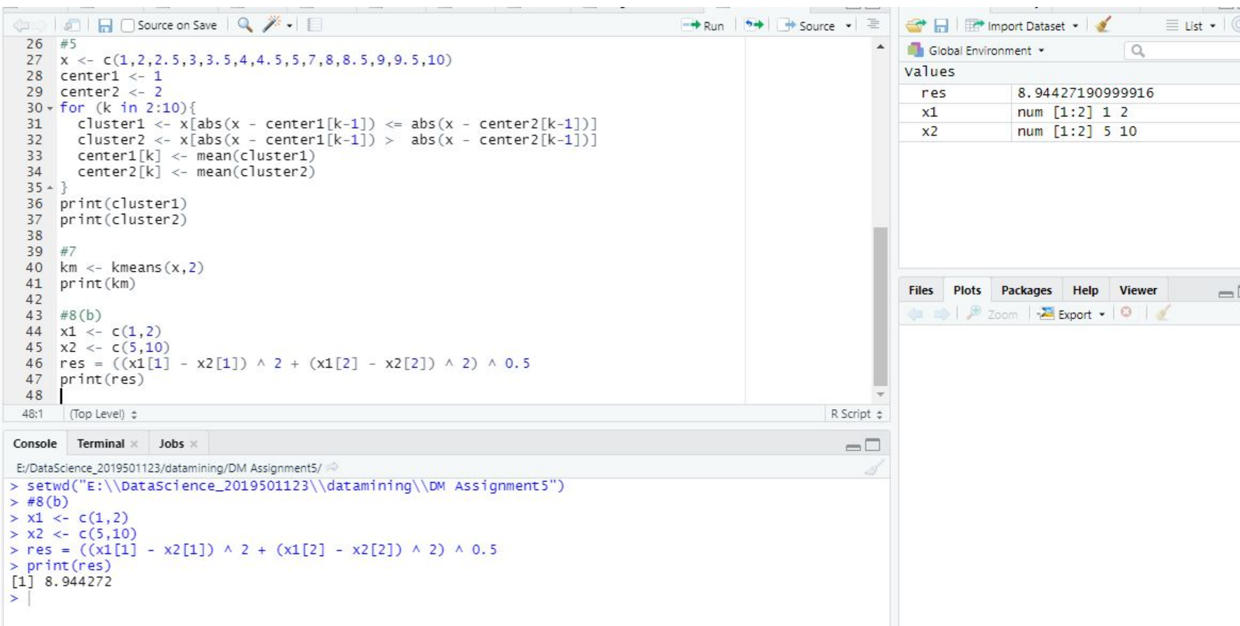
$=\sqrt{(-4)^2+(-8)^2}=\sqrt{16+64}$

$=\sqrt{80}=8.9442719$

$\cong 8.944272$



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.

(a)

→ Given points are $x_1 = (1,2,3,6)$ and $x_2 = (5,10,4,12)$

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

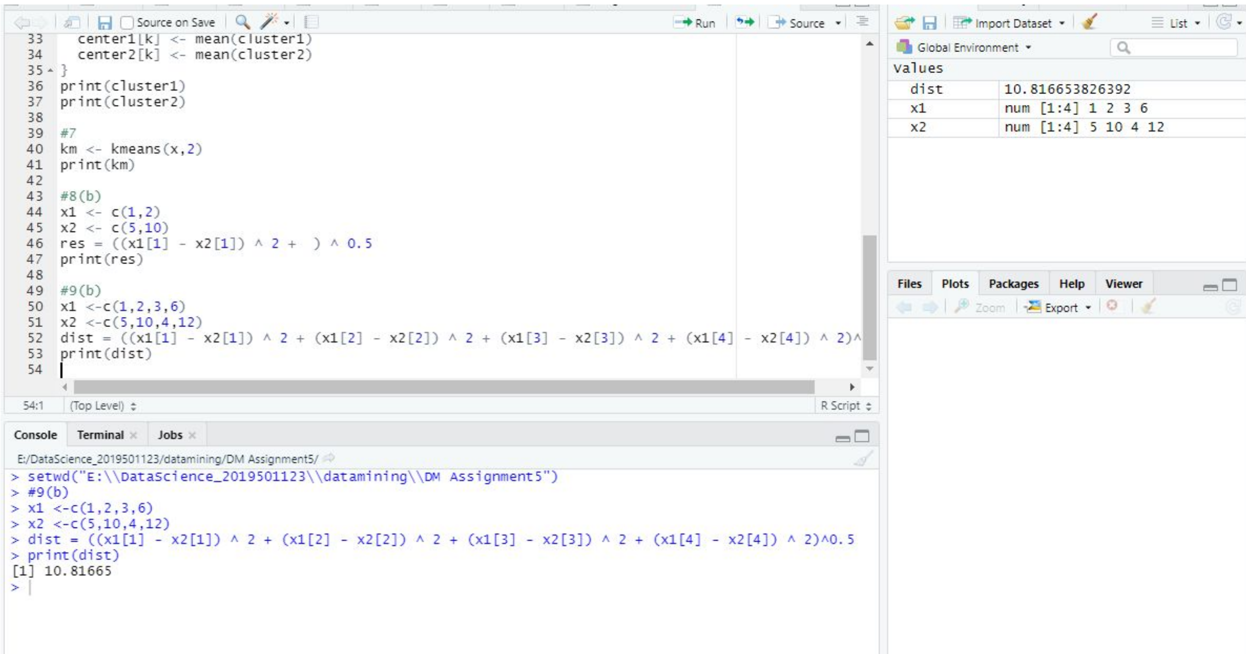
Euclidian distance $= \sqrt{(1-5)^2+(2-10)^2+(3-4)^2+(6-12)^2}$

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

$= \sqrt{16+64+1+36} = \sqrt{117} = 10.816653826391$

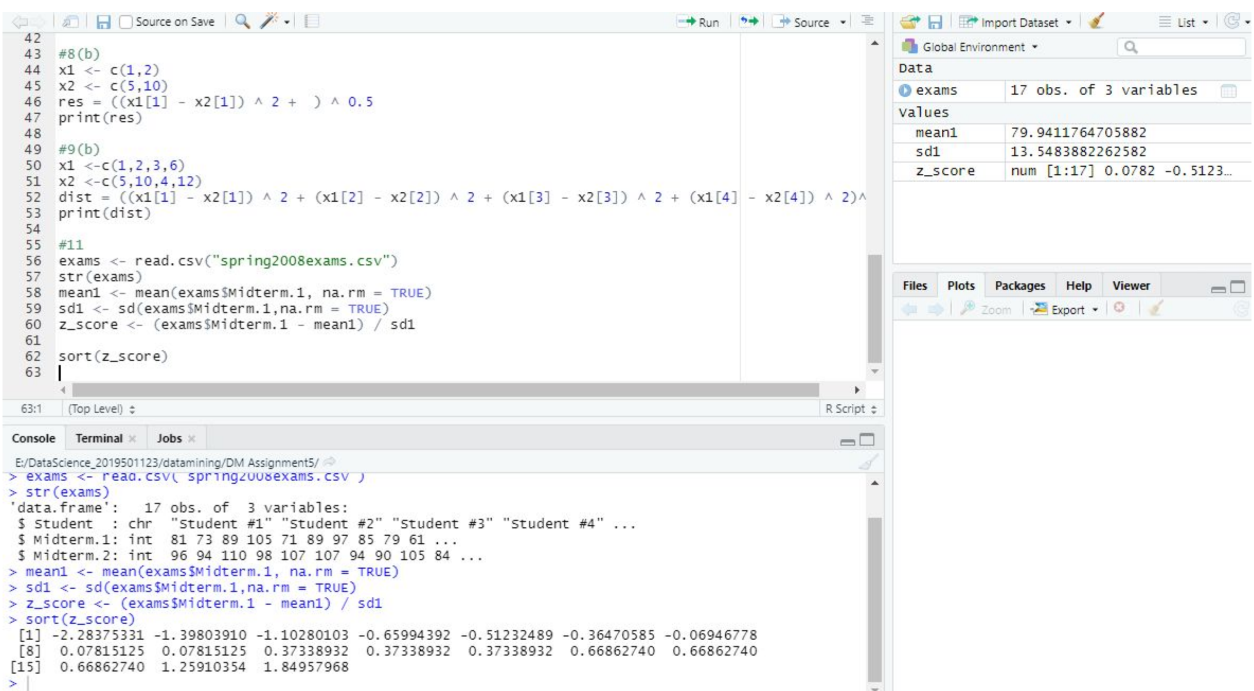
$\cong 10.8166538264$

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.



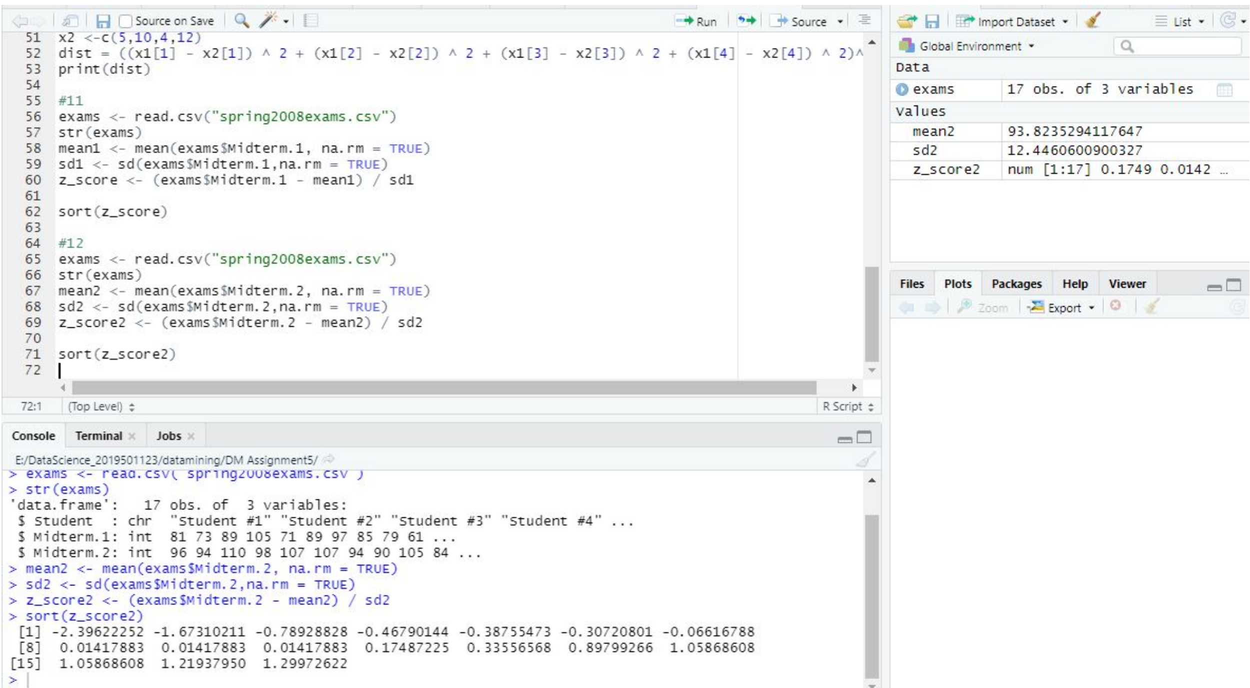
10) Read Chapter 10.

11. Use a z score cut off of 3 to identify any outliers 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.



Largest = 1.84957968 and smallest = -2.28375331 and No outliers are found.

12)Use a z score cut off of 3 to identify any outliers 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.



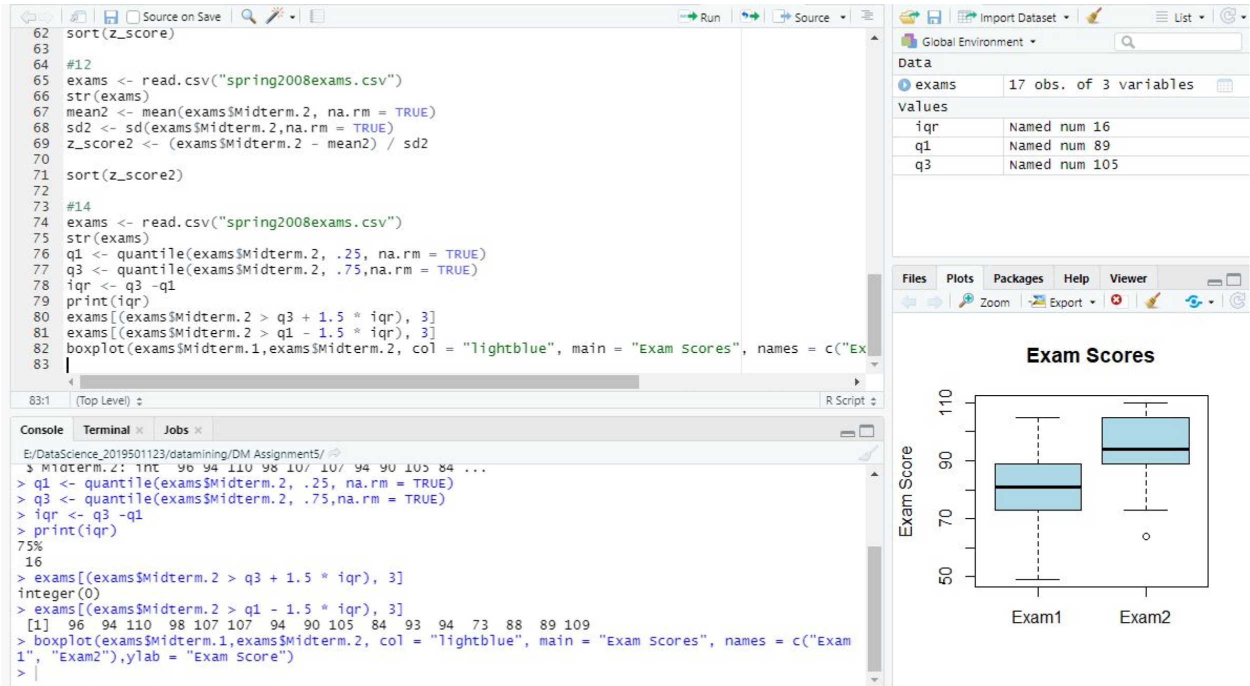
Largest = 1.29972622 and smallest = -2.39622252 and No outliers found

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.)

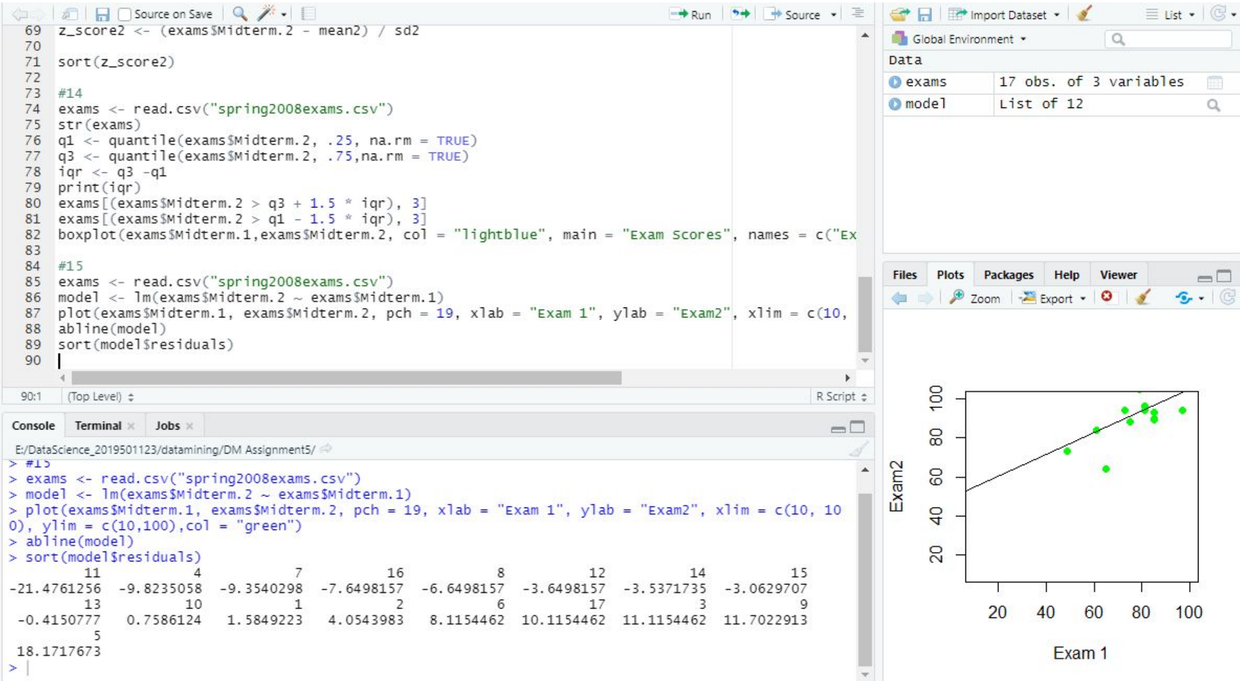
The only value which has outlier is with IP address 65.57.245.11 with a z-score of 8.426135321

Filename: [Question13_stats202Zscores.xlsx]()

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.


```r
z_score2 <- (exams$Midterm.2 - mean2) / sd2

sort(z_score2)

#14
exams <- read.csv("spring2008exams.csv")
str(exams)
q1 <- quantile(exams$Midterm.2, .25, na.rm = TRUE)
q3 <- quantile(exams$Midterm.2, .75,na.rm = TRUE)
iqr <- q3 -q1
print(iqr)
exams[(exams$Midterm.2 > q3 + 1.5 * iqr), 3]
exams[(exams$Midterm.2 > q1 - 1.5 * iqr), 3]
boxplot(exams$Midterm.1,exams$Midterm.2, col = "lightblue", main = "Exam Scores", names = c("Ex

#15
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,
abline(model)
sort(model$residuals)
```

```
> #15
> 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),col = "green")
> abline(model)
> sort(model$residuals)
         11          4          7         16          8         12         14         15 
-21.4761256  -9.8235058  -9.3540298  -7.6498157  -6.6498157  -3.6498157  -3.5371735  -3.0629707 
         13         10          1          2          6         17          3          9 
 -0.4150777   0.7586124   1.5849223   4.0543983   8.1154462  10.1154462  11.1154462  11.7022913 
          5 
 18.1717673 
>
```


5th student has highest residual = 18.1717673