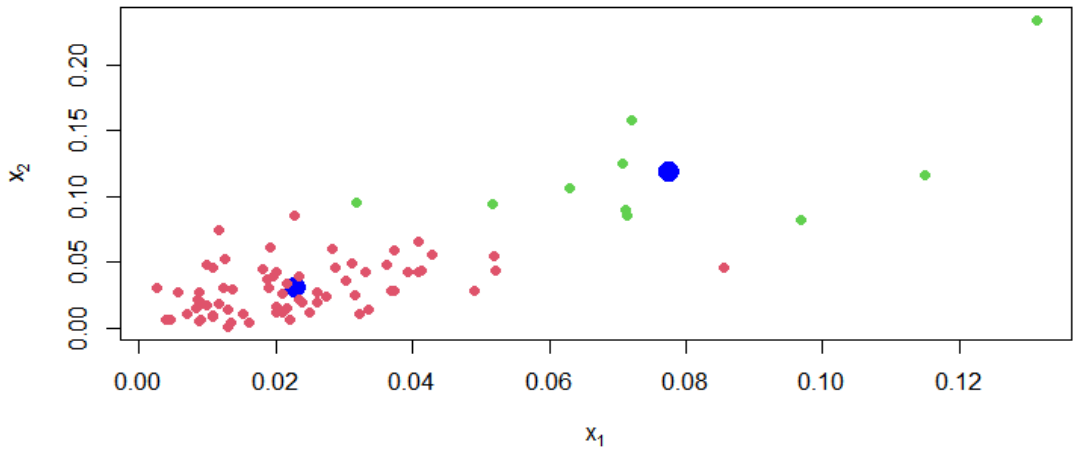
Data Mining Assignment 5

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

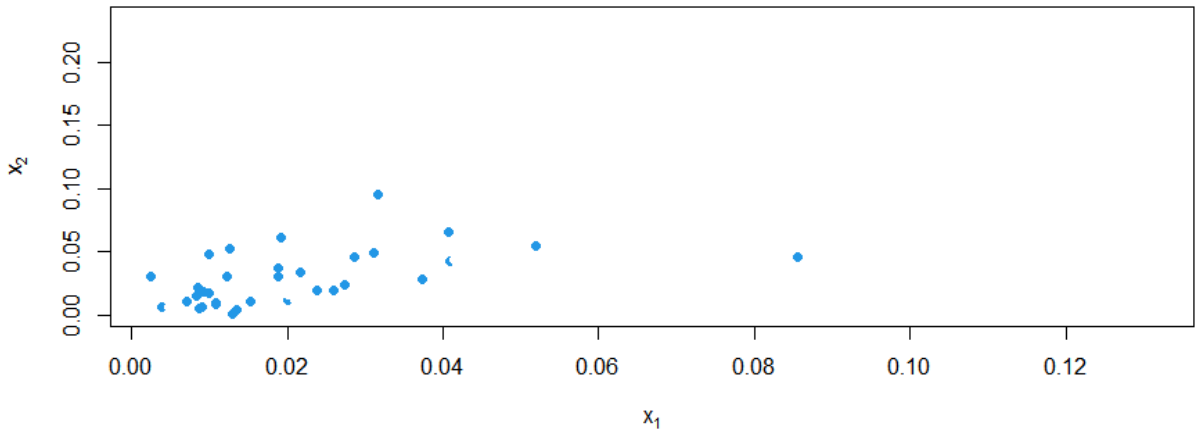
2) Repeat In Class Exercise #50 using the sonar test data instead of the sonar training data and show your R commands for doing so.

```
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)

```



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.

```
>
> plot(x,pch=19,xlab=expression(x[1]), ylab=expression(x[2]))
> y<-data[,61]
> points(x,col=2+2*y,pch=19)
> 1-sum(knnfit==y)/length(y)
[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.

```
> #4) Repeat In Class Exercise #53 using the sonar test data inste
ining 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)))
> 1-sum(knnfit==y)/length(y)
[1] 0.5641026
>
```

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.

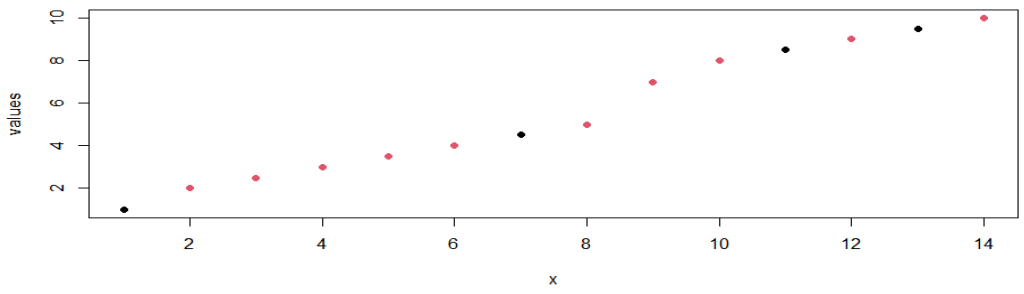
```
> #5) Repeat In Class Exercise #54 using the data x<-c(1,2,2.5,3,3.5,4
7,8,8.5,9,9.5,10) instead. Show all your work for each step and be sur
y specifically which points are in each cluster at each step.
> x<-c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10)
> center1<-1
> center2<-2
> for (k in 2:10){
+   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)
+ }
>
```

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.

```
> #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.
>
> 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)
+ }
```

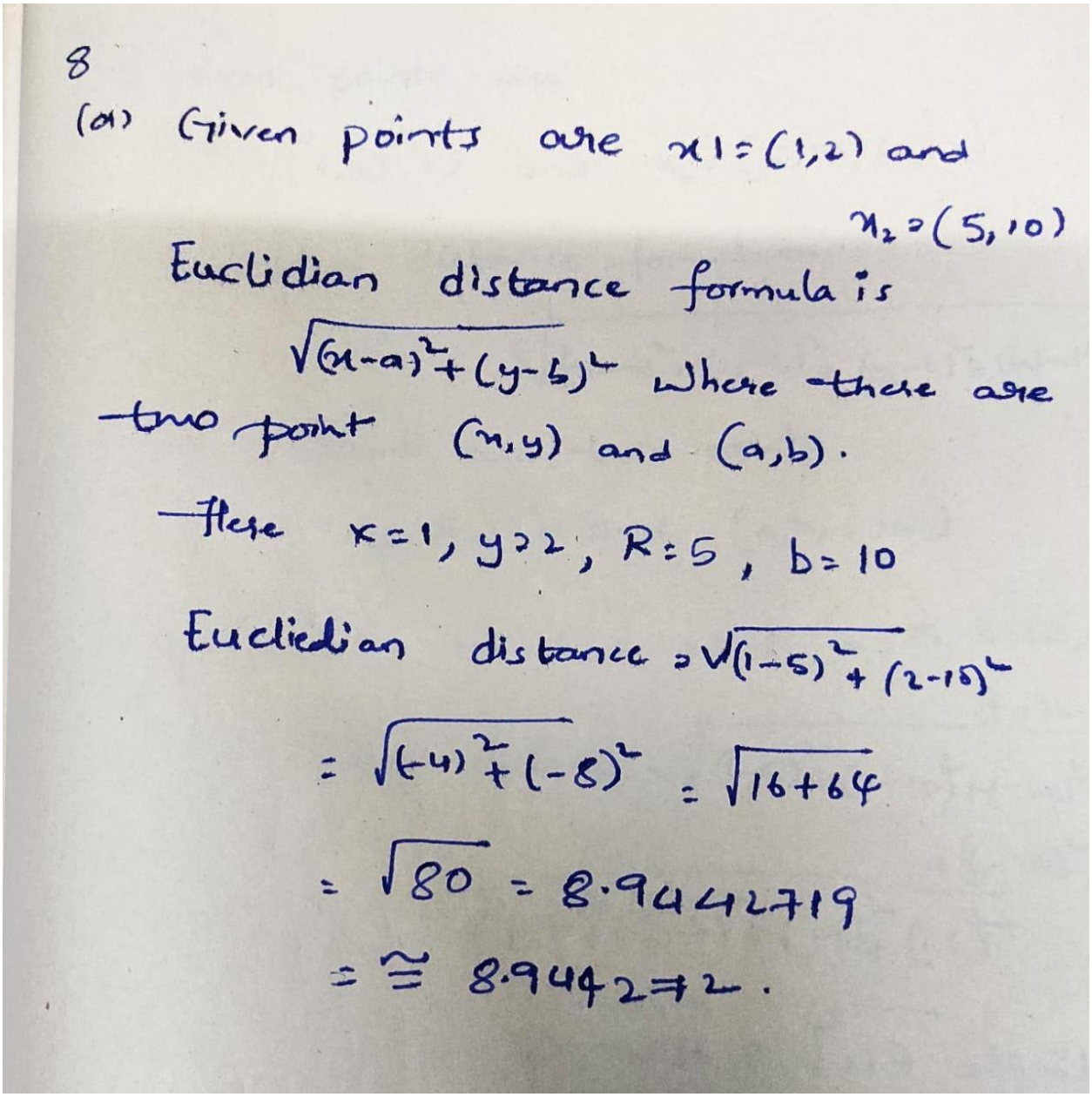
7) Repeat In Class Exercise #56 using the data x<-c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10) instead and show your R commands for doing so.

```
> #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.
>
> 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"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
> plot(x, col=fit$cluster, xlab = 'x', ylab='values', pch=19)
>
```



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.



8

(a) Given points are $x1=(1,2)$ and $x_2=(5,10)$

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

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

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

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

$=\sqrt{80}=8.94427191$

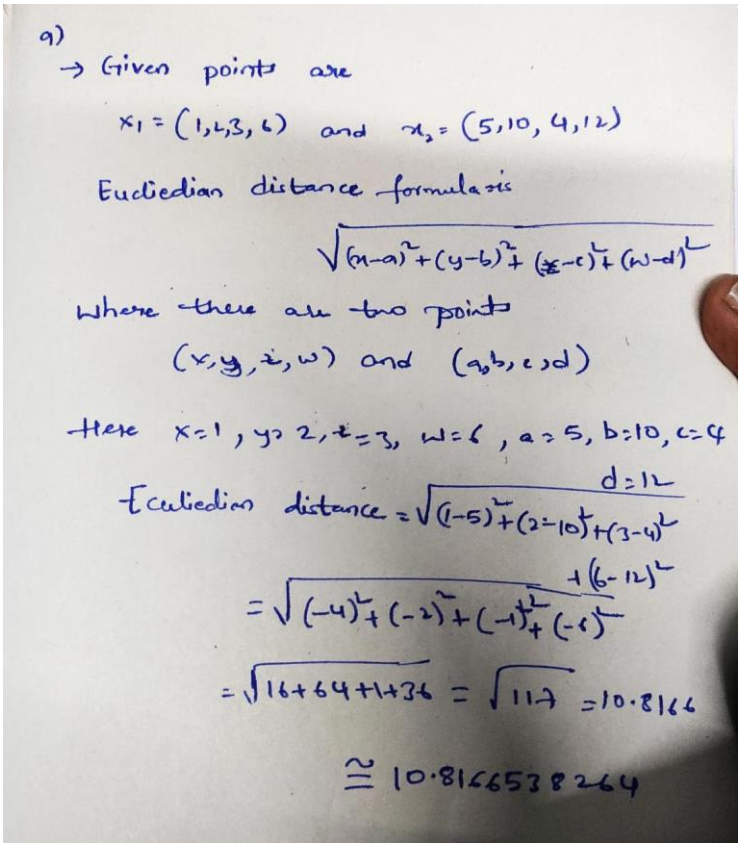
$=\cong 8.94427\overline{2}$.

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.

```
>
> #8) Consider the points x1<-c(1,2) and x2<-c(5,10).
> #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)
> data<-(rbind(x1,x2))
> dist(data)
         x1
x2 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.



a)

→ Given points are

$x_1 = (1,2,3,6)$ 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 $x=1, y=2, z=3, w=6, a=5, b=10, c=4, d=12$

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

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

$= \sqrt{16+64+1+36} = \sqrt{117} = 10.8166$

$\cong 10.816653826$

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.

```
>
> #9) Consider the points x1<-c(1,2,3,6) and x2<-c(5,10,4,12).
>
> #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)
> data<-(rbind(x1,x2))
> dist(data)
          x1
x2 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.

```
D:/Courses/2nd Year/4.Data Science/-Sp-Data_Science_2019501125/Intro to Data Mining/DM Assignmen
> data<-read.csv("spring2008exams.csv")
> mean_exam<-mean(data[,2],na.rm=TRUE)
> sd_exam<-sd(data[,2],na.rm=TRUE)
> z<-(data[,2]-mean_exam)/sd_exam
> li=sort(z)
> cat("largest z score:",li[length(li)])
largest z score: 1.84958> cat("\nSmallest z score:",li[1])

Smallest z score: -2.283753
>
```

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.

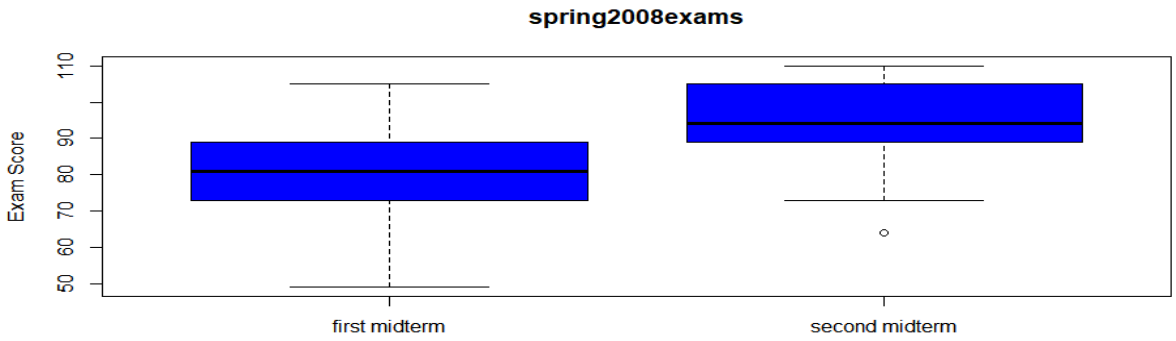
```
> #12) Repeat In Class Exercise #59 using the grades for the second midterm at www.stats202.com/
spring2008exams.csv. Are there any outliers according to the z=+/-3 rule? What is the value of t
he largest z score and what is the value of the smallest (most negative) z score? Show your R co
mmands.
>
> spring_data<-read.csv("spring2008exams.csv")
> mean_exam<-mean(spring_data[,3],na.rm=TRUE)
> sd_exam<-sd(spring_data[,3],na.rm=TRUE)
> z<-(spring_data[,3]-mean_exam)/sd_exam
> li=sort(z)
> cat("largest z score:",li[length(li)])
largest z score: 1.299726> cat("\nSmallest z score:",li[1])

Smallest z score: -2.396223
>
```

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?
> spring_data<-read.csv("spring2008exams.csv")
> q1<-quantile(spring_data[,3],.25,na.rm=TRUE)
> q3<-quantile(spring_data[,3],.75,na.rm=TRUE)
> iqr<-q3-q1
> spring_data[(spring_data[,3]>q3+1.5*iqr),3]
integer(0)
> spring_data[(spring_data[,3]<q1-1.5*iqr),3]
[1] 64
> boxplot(spring_data[,2],spring_data[,3],col="blue",
+         main="spring2008exams",
+         names=c("first midterm","second midterm"),ylab="Exam Score")
>
```

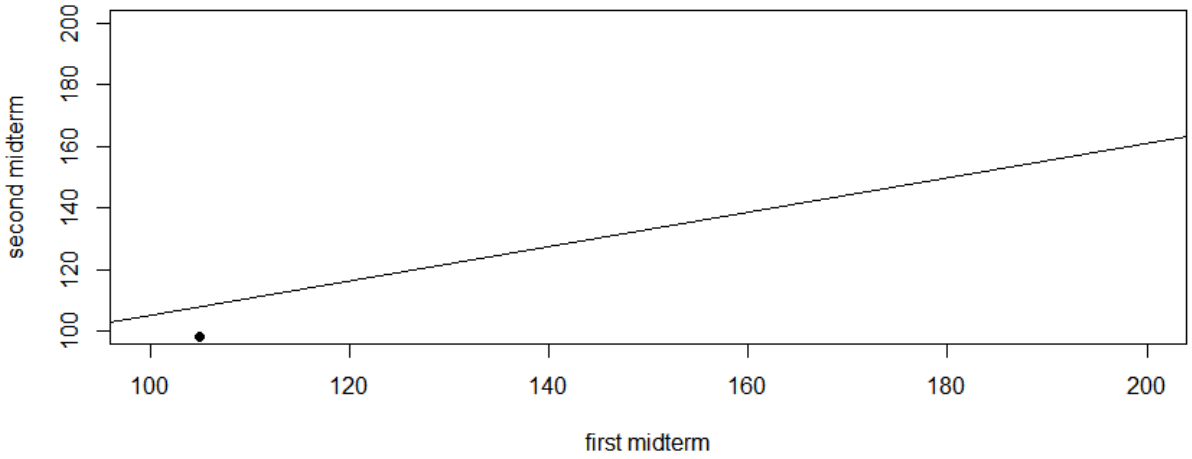
14) Repeat In Class Exercise #61 using the grades for the second midterm at www.stats202.com/spring2008exams.csv. Show your R commands and include the boxplot. Are any of the grades for the second midterm outliers by this rule? If so, which ones?

```
> #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?
> spring_data<-read.csv("spring2008exams.csv")
> q1<-quantile(spring_data[,3],.25,na.rm=TRUE)
> q3<-quantile(spring_data[,3],.75,na.rm=TRUE)
> iqr<-q3-q1
> spring_data[(spring_data[,3]>q3+1.5*iqr),3]
integer(0)
> spring_data[(spring_data[,3]<q1-1.5*iqr),3]
[1] 64
> boxplot(spring_data[,2],spring_data[,3],col="blue",
+         main="spring2008exams",
+         names=c("first midterm","second midterm"),ylab="Exam Score")
>
```



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.

```
Console  Terminal  Jobs
D:/Courses/2nd Year/4.Data Science/-Sp-Data_Science_2019501125/Intro to Data Mining/DM Assignment5/
> #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.
>
> spring_data<-read.csv("spring2008exams.csv")
> model<-lm(spring_data[,3]~spring_data[,2])
> plot(spring_data[,2],spring_data[,3],pch=19,xlab="first midterm",ylab="second midterm",xlim=c(100,200),
ylim=c(100,200))
> abline(model)
> max(model$residuals)
[1] 18.17177
>
```


second midterm
100
120
140
160
180
200
100
120
140
160
180
200
first midterm