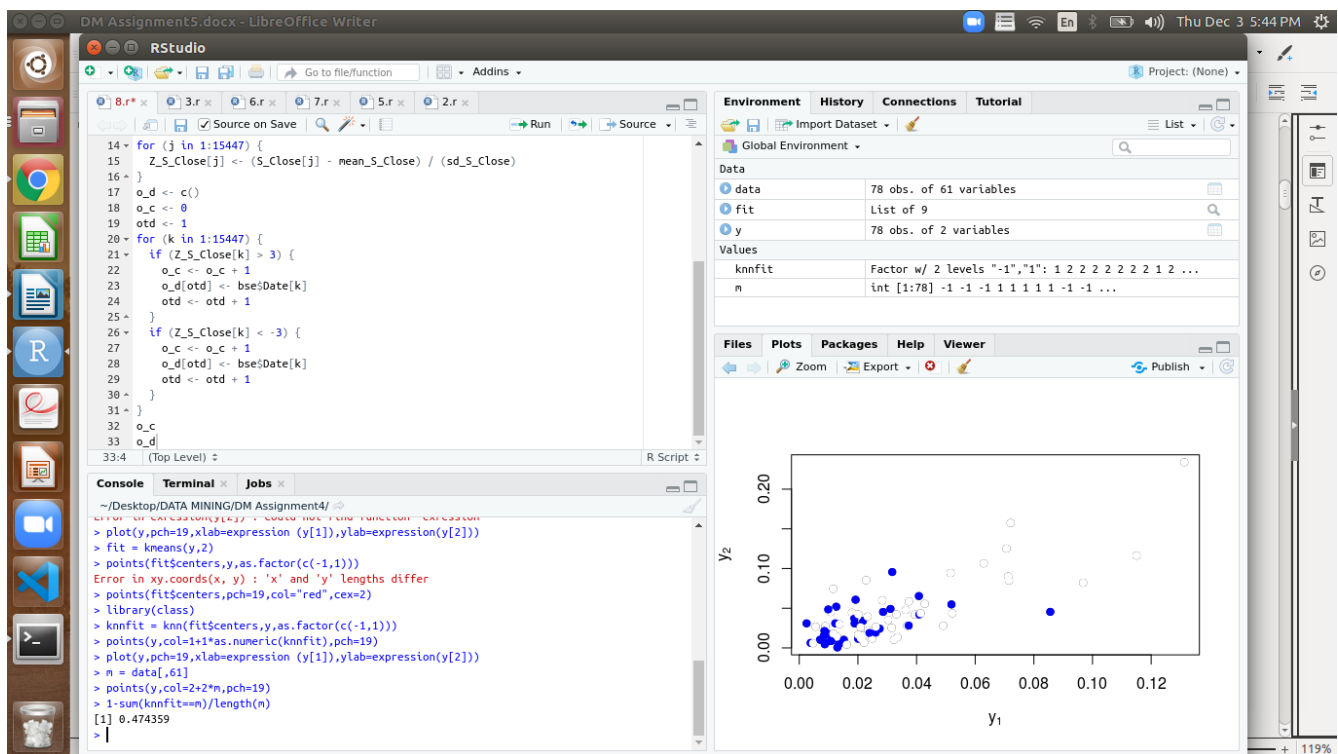
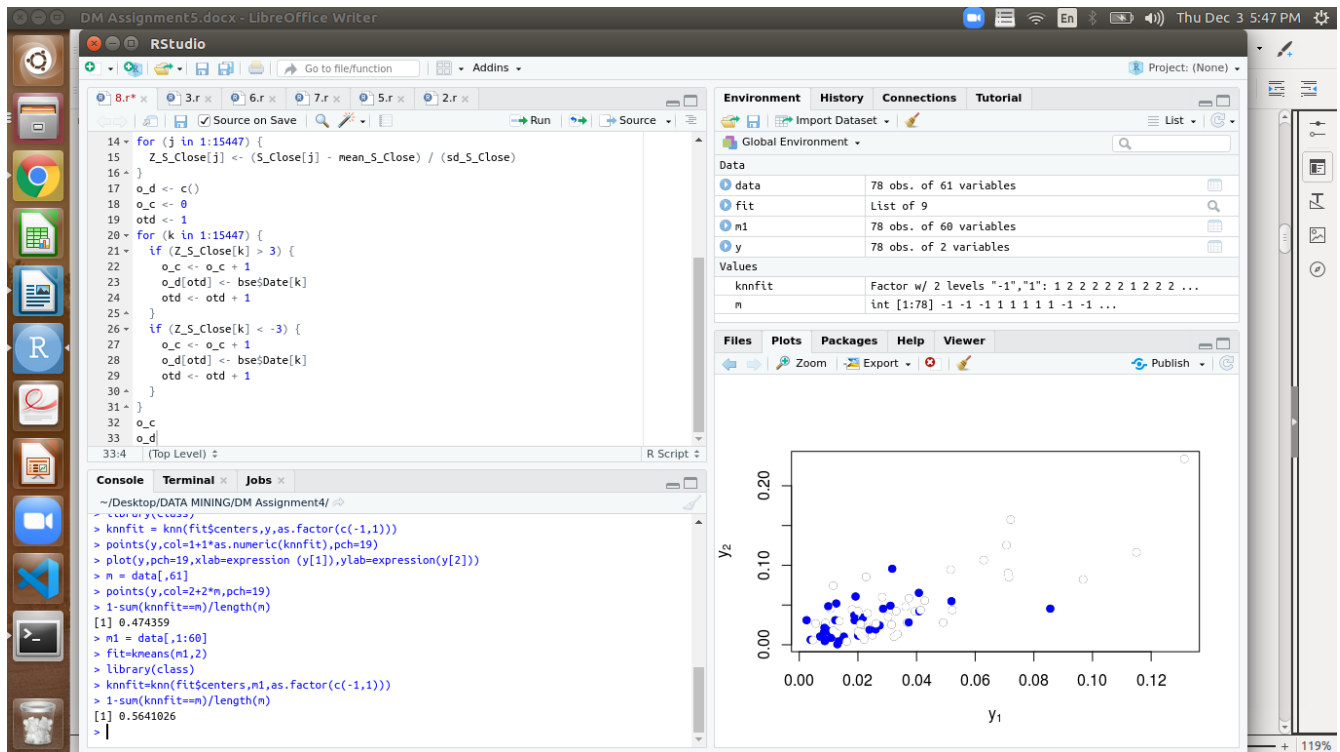


## Data Mining Assignment 5

- 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.

```

{r}
x<-data[,1:60]
fit<-kmeans(x, 2)
library(class)
knnfit<-knn(fit$centers,x,as.factor(c(-1,1)))

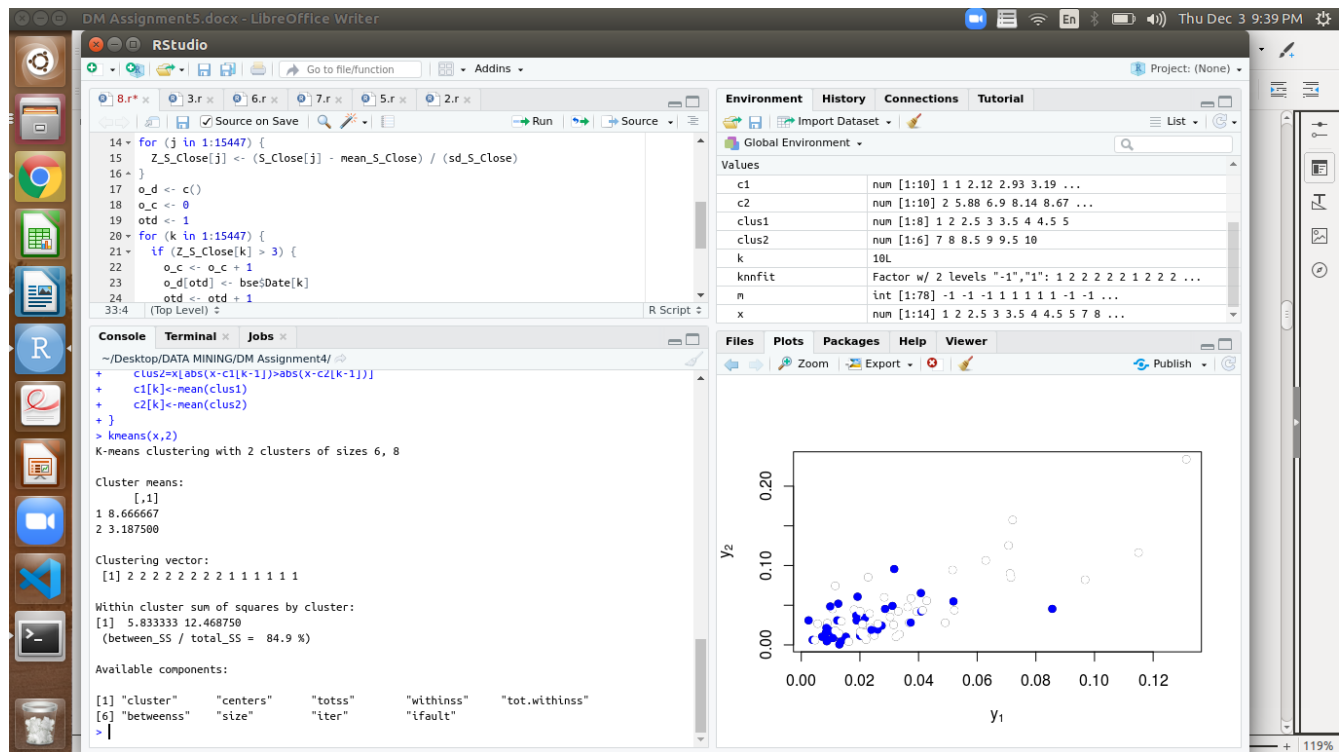
1-sum(knnfit==y)/length(y)

```

[1] 0.4358974

5) Consider the one dimensional data set given  $x \leftarrow c(1,2,2.5,3,3.5,4,4.5,5,7,8,8.5,9,9.5,10)$ . 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.

DM Assignment5.docx - LibreOffice Writer

RStudio

Go to file/function Addins

Project: (None)

Environment History Connections Tutorial

Global Environment

Values

|        |                              |                               |
|--------|------------------------------|-------------------------------|
| c1     | num [1:10]                   | 1 1 2.12 2.93 3.19 ...        |
| c2     | num [1:10]                   | 2 5.88 6.9 8.14 8.67 ...      |
| clus1  | num [1:8]                    | 1 2 2.5 3 3.5 4 4.5 5         |
| clus2  | num [1:6]                    | 7 8 8.5 9 9.5 10              |
| k      |                              | 10L                           |
| knnfit | Factor w/ 2 levels "-1","1": | 1 2 2 2 2 1 2 2 2 ...         |
| m      | int [1:78]                   | -1 -1 -1 1 1 1 1 1 -1 -1 ...  |
| x      | num [1:14]                   | 1 2 2.5 3 3.5 4 4.5 5 7 8 ... |

Files Plots Packages Help Viewer

Zoom Export Publish

```
14 ~ for (j in 1:15447) {
15   Z_5_close[j] <- (S_close[j] - mean_S_close) / (sd_S_close)
16 ~ }
17 o_d <- c()
18 o_c <- 0
19 otd <- 1
20 ~ for (k in 1:15447) {
21   if (Z_5_close[k] > 3) {
22     o_c <- o_c + 1
23     o_d[otd] <- bse$Date[k]
24     otd <- otd + 1
33:4 (Top Level)
```

R Script

Console Terminal Jobs

~/Desktop/DATA MINING/DM Assignment4/

```
+ c1us2=x[abs(x-c1[k-1])>abs(x-c2[k-1])]
+ c1[k]<-mean(c1us1)
+ c2[k]<-mean(c2us2)
+ }
> kmeans(x,2)
K-means clustering with 2 clusters of sizes 6, 8

Cluster means:
[,1]
1 8.666667
2 3.187500

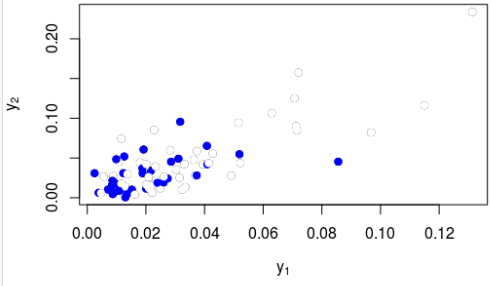
Clustering vector:
[1] 2 2 2 2 2 2 2 1 1 1 1 1 1 1

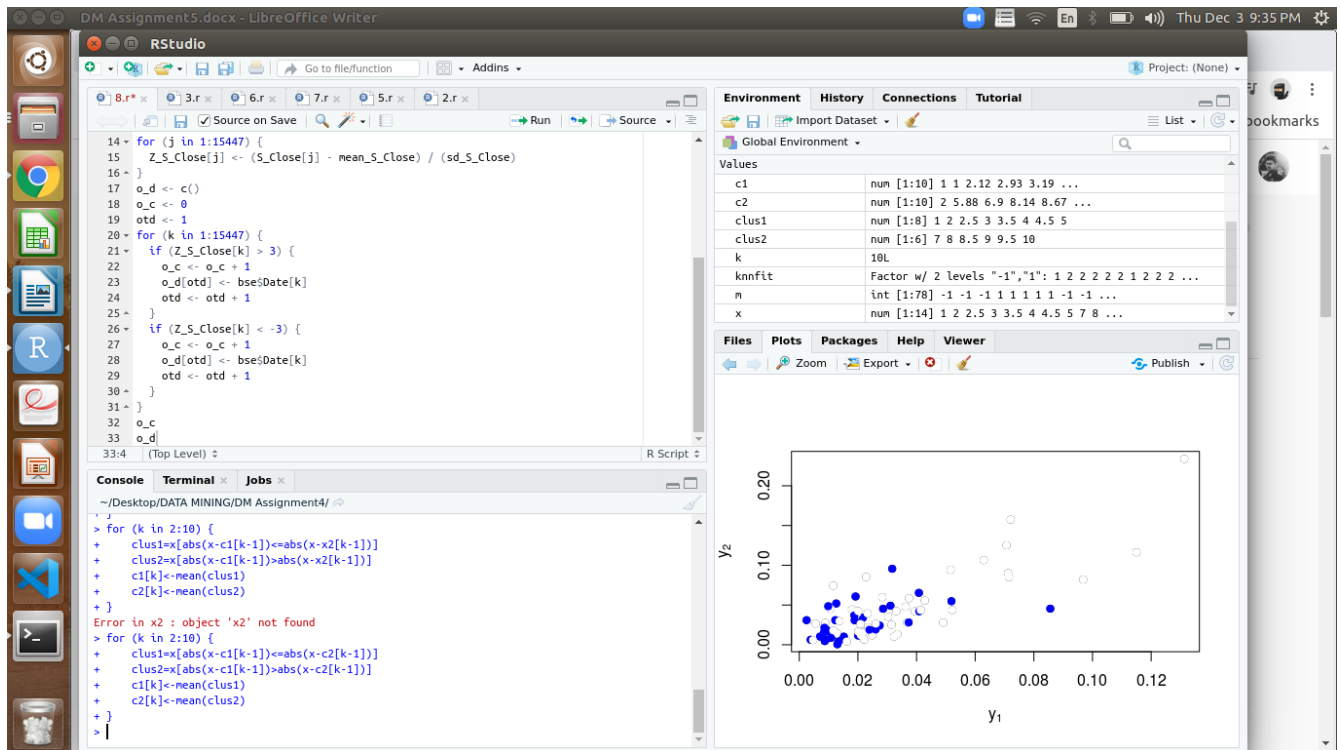
Within cluster sum of squares by cluster:
[1] 5.833333 12.468750
(between_SS / total_SS = 84.9 %)
```

Available components:

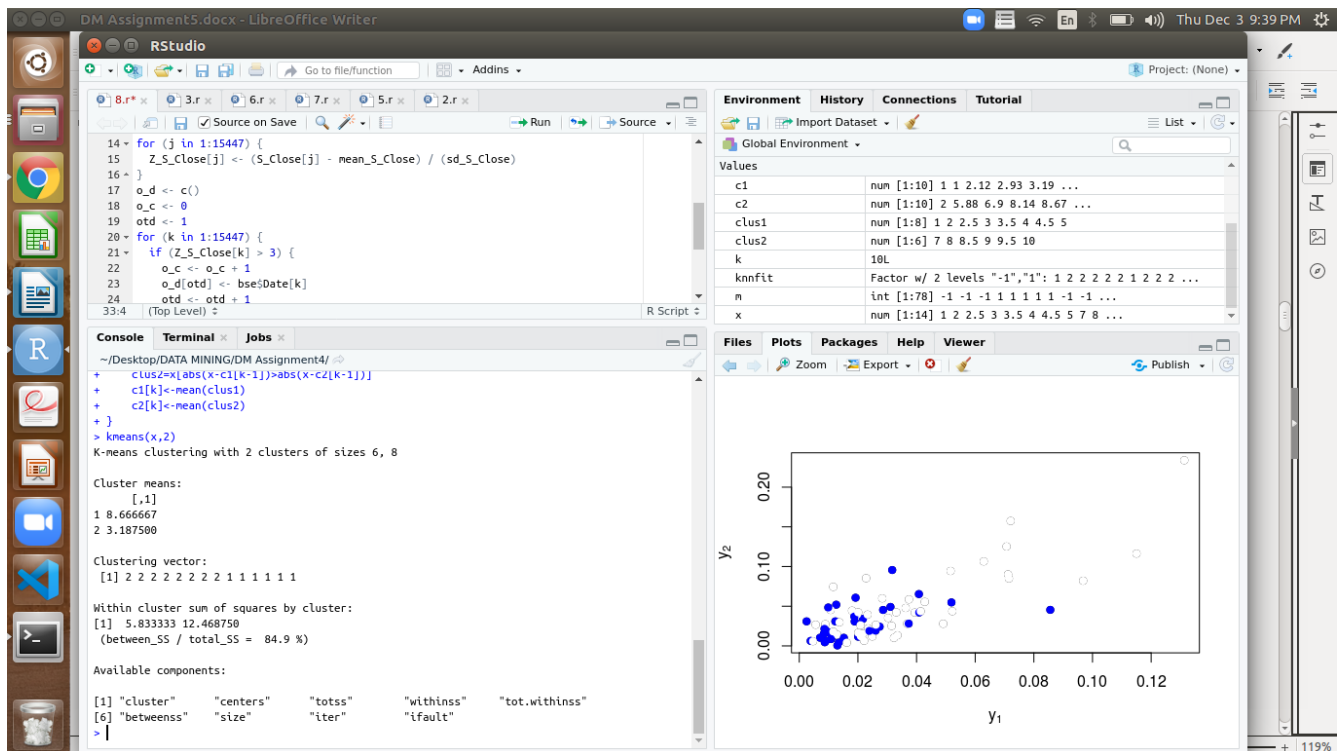
|                 |           |         |            |                |
|-----------------|-----------|---------|------------|----------------|
| [1] "cluster"   | "centers" | "totss" | "withinss" | "tot.withinss" |
| [6] "betweenss" | "size"    | "iter"  | "ifault"   |                |

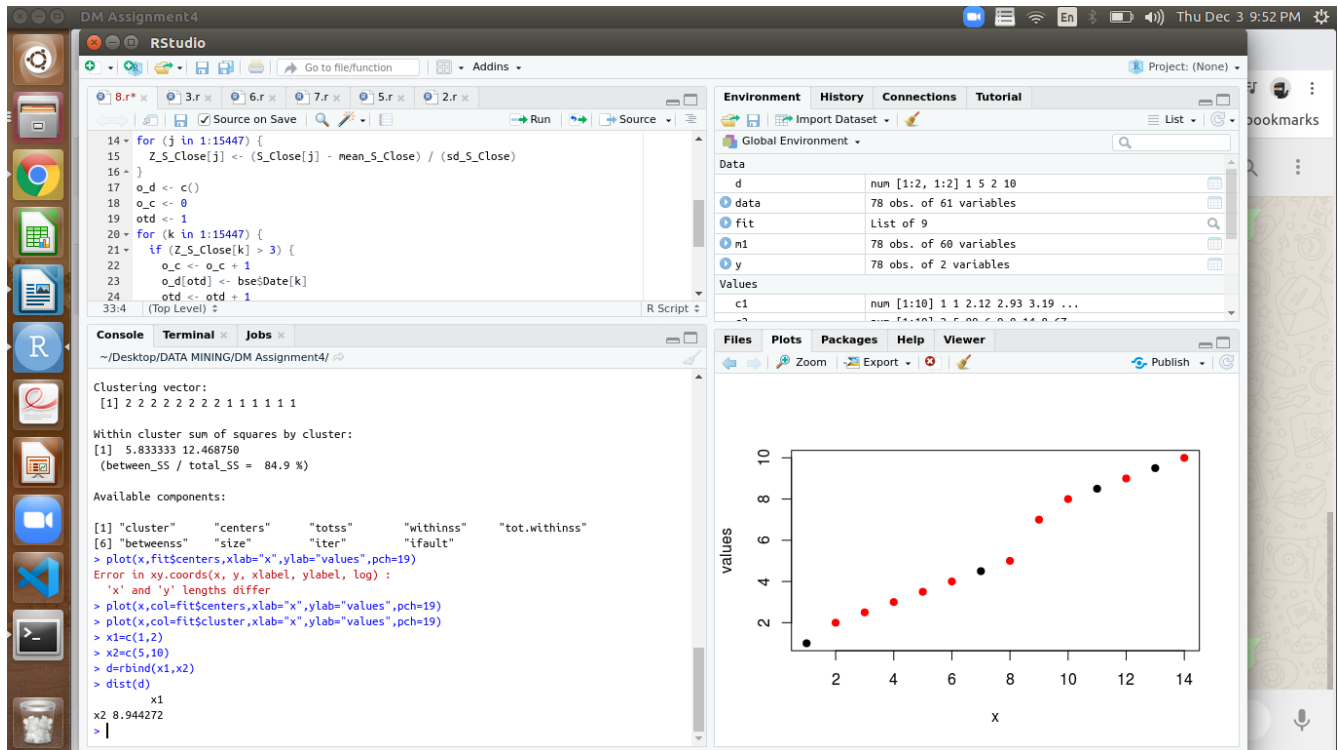
> |





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.





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.

e) Consider the points  $x_1 = C(1,2)$  and  $x_2 = C(5,10)$   
Compute euclidean distance

Sol: given points

$$x_1 = (1,2)$$

$$x_2 = (5,10)$$

$$d(x_1, x_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

here

$$x_1 = 1 \quad x_2 = 5$$

$$y_1 = 2 \quad y_2 = 10$$

$$\begin{aligned} d(x_1, x_2) &= \sqrt{(5-1)^2 + (10-2)^2} \\ &= \sqrt{4^2 + 8^2} \\ &= \sqrt{16 + 64} \\ &= \sqrt{80} \end{aligned}$$

$$\approx 8.944271$$

By Pythagorus theorem

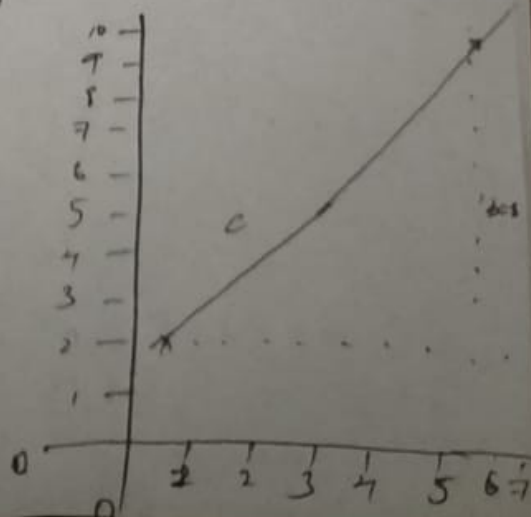
$$c^2 = a^2 + b^2$$

$$c^2 = 4^2 + 8^2$$

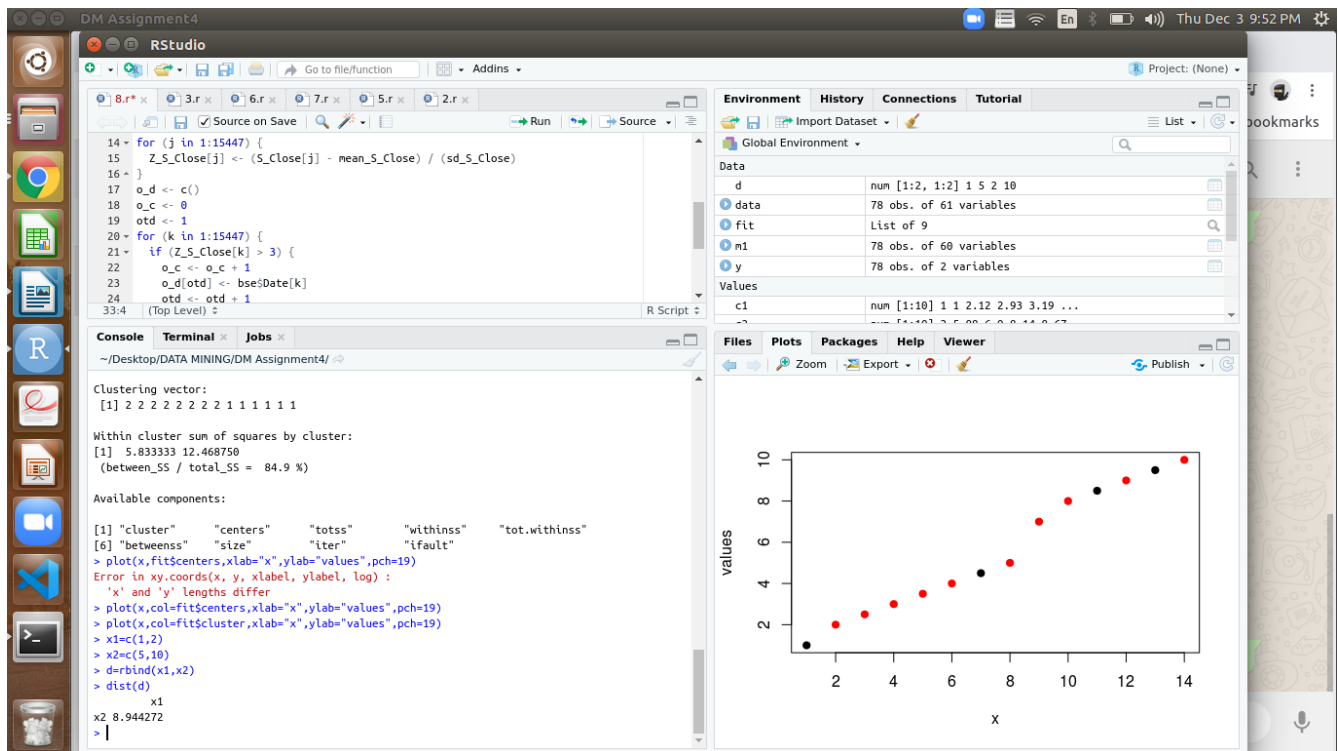
$$c^2 = 16 + 64$$

$$c = \sqrt{80}$$

$$\approx 8.94427$$



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  $x_1 \sim c(1, 2, 3, 6)$  and  $x_2 \sim c(5, 10, 4, 12)$ .

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



9) Consider points  $x_1 = c(1, 2, 3, 6)$ ,  $x_2 = c(5, 10, 4, 12)$

sol: given  $x_1 = (1, 2, 3, 6)$

$x_2 = (5, 10, 4, 12)$

$$\text{Euclidean distance} = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (x_3 - x_1)^2 + (y_3 - y_1)^2}$$

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

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

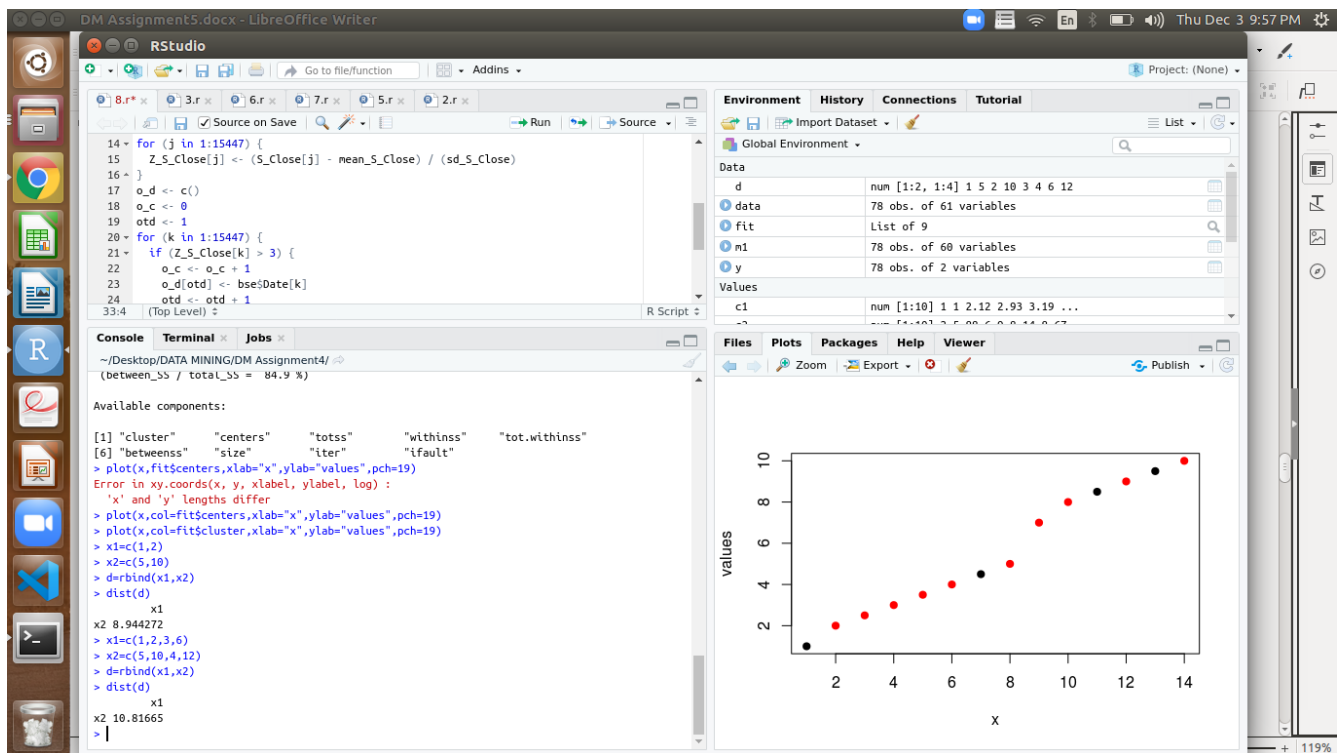
$$= \sqrt{16 + 64 + 1 + 36}$$

$$= \sqrt{81 + 36}$$

$$= \sqrt{117}$$

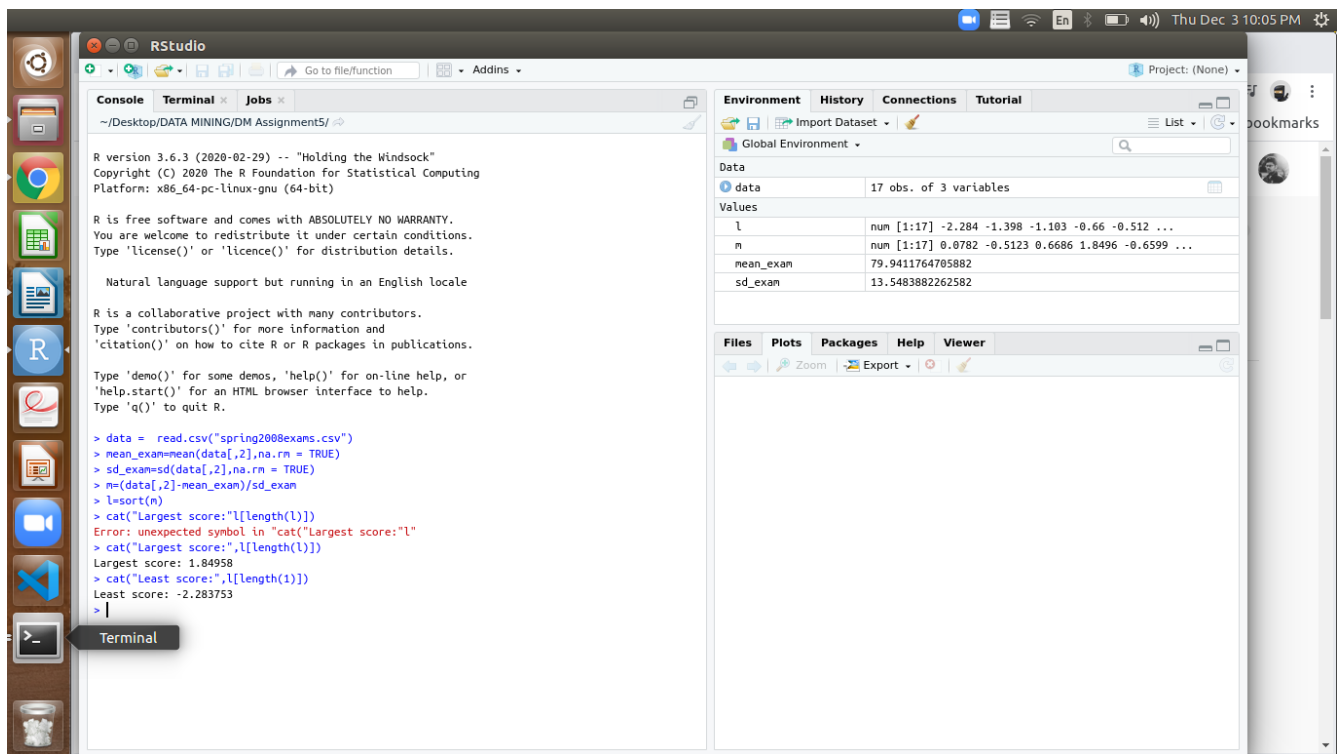
$$\approx 10.8167$$

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](http://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.



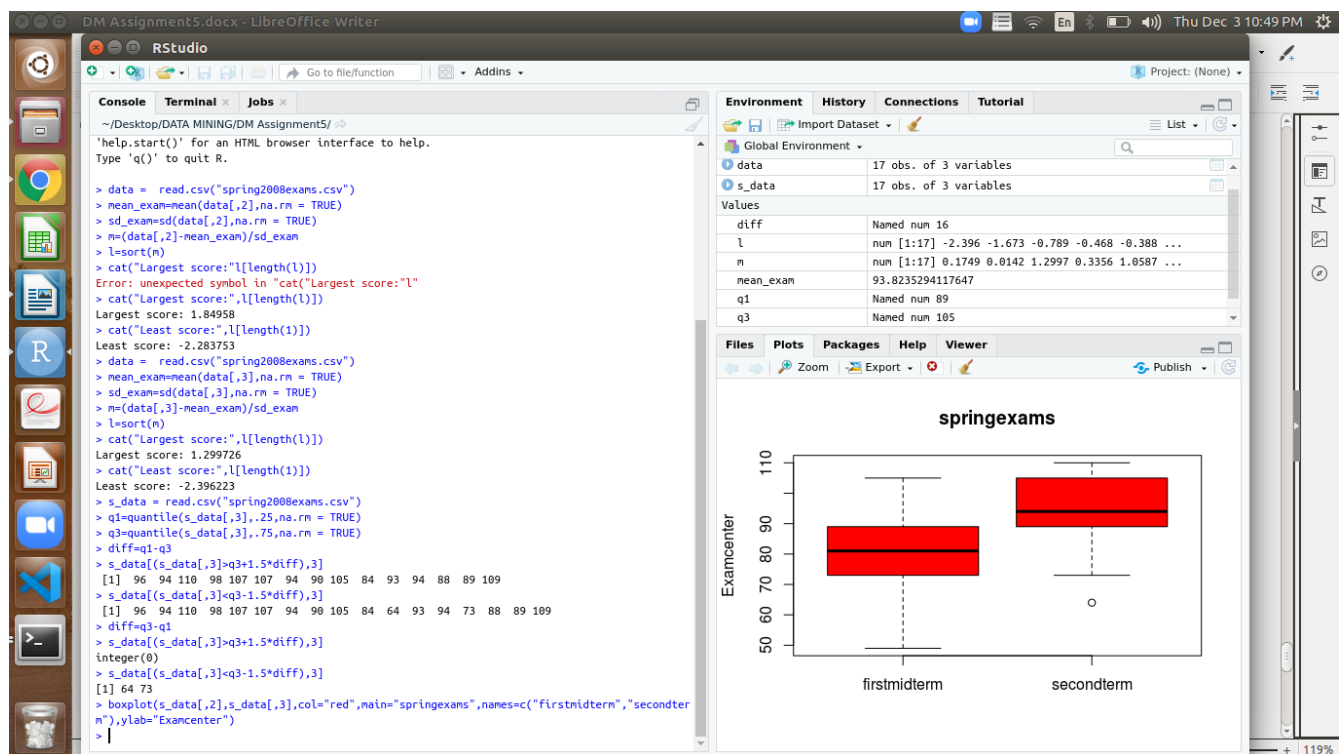
12) Use a z score cut off of 3 to identify any outliers using the grades for the second midterm at [www.stats202.com/spring2008exams.csv](http://www.stats202.com/spring2008exams.csv). Are there any outliers according to the  $z = \pm 3$  rule? What is the value of the largest z score and what is the value of the smallest (most negative) z score? Show your R commands.

13) Compute the count of each ip address (1<sup>st</sup> column) in the data stats202log.txt, then use a z score cut off of 3 to identify any outliers for these counts 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  $z = \pm 3$  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.)

There is 0 outlier.

| Row Labels      | Count of ip | z score    | outlier |      |            |
|-----------------|-------------|------------|---------|------|------------|
| 12.105.244.10   | 2           | -0.0485597 | 0       |      |            |
| 12.108.24.162   | 8           | -0.0348531 | 0       |      |            |
| 12.151.36.202   | 2           | -0.0485597 | 0       |      |            |
| 12.163.145.67   | 4           | -0.0439908 | 0       |      |            |
| 12.214.178.123  | 2           | -0.0485597 | 0       |      |            |
| 12.217.249.223  | 8           | -0.0348531 | 0       |      |            |
| 12.37.213.243   | 2           | -0.0485597 | 0       |      |            |
| 12.72.157.168   | 2           | -0.0485597 | 0       | mean | 23.2568735 |
| 121.108.70.25   | 4           | -0.0439908 | 0       | std  | 437.747565 |
| 121.247.76.197  | 3           | -0.0462752 | 0       | max  | 11.4100992 |
| 121.44.209.154  | 24          | 0.00169761 | 0       | min  | -0.0508441 |
| 122.167.140.219 | 9           | -0.0325687 | 0       |      |            |
| 122.167.153.52  | 1           | -0.0508441 | 0       |      |            |
| 122.167.158.81  | 1           | -0.0508441 | 0       |      |            |
| 122.167.233.70  | 1           | -0.0508441 | 0       |      |            |
| 122.167.240.97  | 12          | -0.0257154 | 0       |      |            |
| 124.120.143.180 | 28          | 0.0108353  | 0       |      |            |

14) Identify any outliers more than 1.5 IQR's above the 3<sup>rd</sup> quartile or below the 1<sup>st</sup> quartile. Verify that these are the same outliers found by the boxplot function 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) Use functions to fit a least squares regression model which predicts the exam 2 score as a function of the exam 1 score for the data spring2008exams.csv. Plot the fitted line and determine for which points the fitted exam 2 values are the furthest from the actual values using the model residuals 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.

