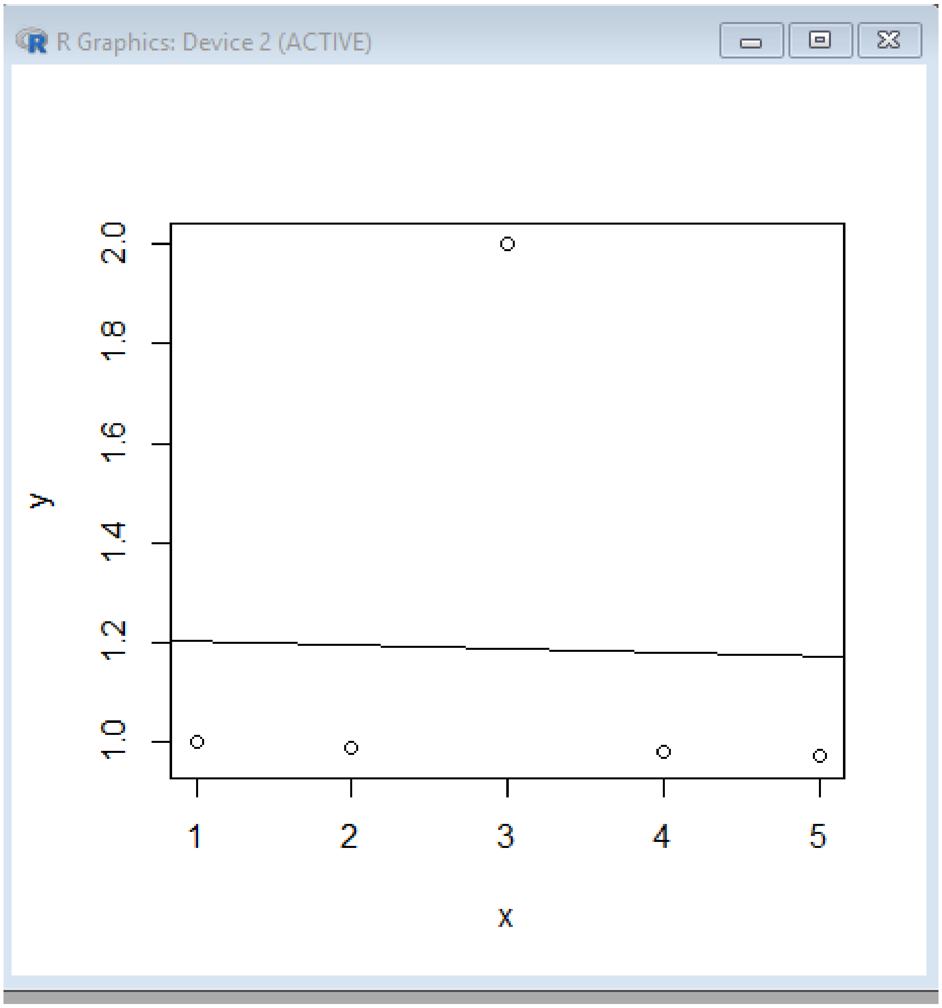
Exercise 1. (3 points)

Consider the data set of the following observations:

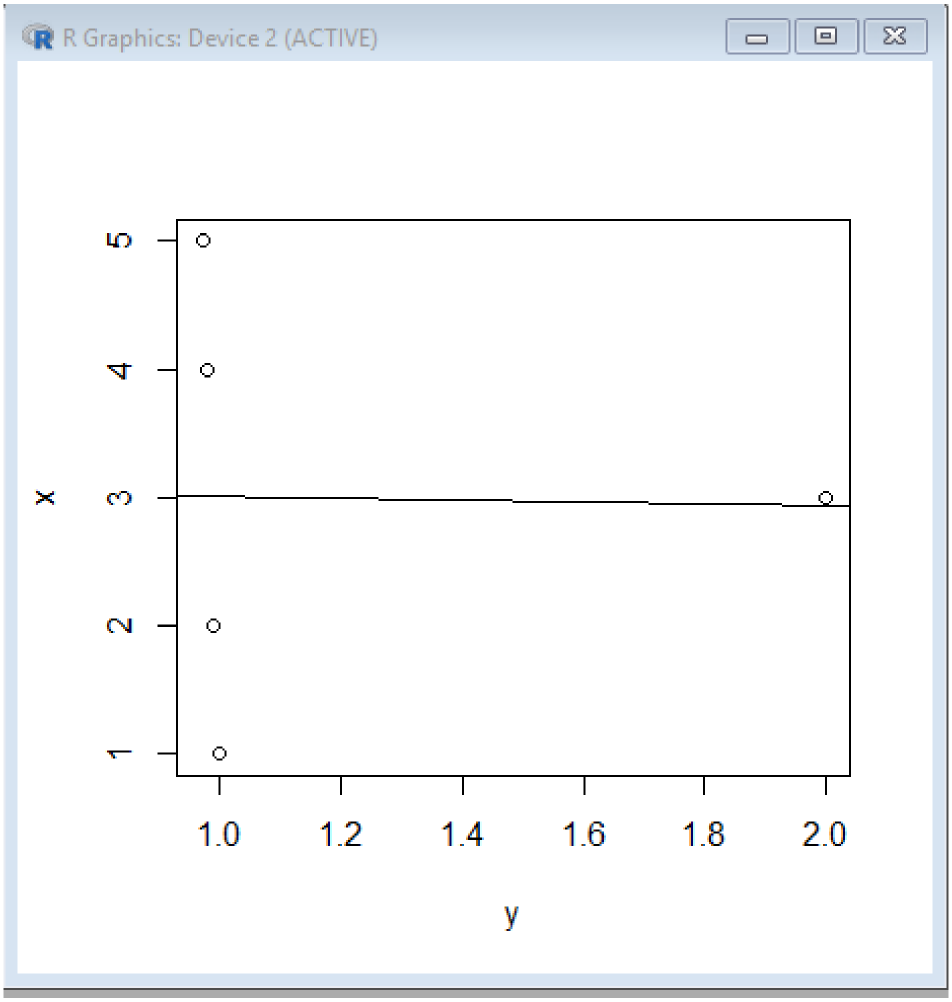
{ (1, 1), (2, 0.99), (3, 2), (4, 0.98), (5, 0.97) }.

PART 1: fit and plot the two regression lines & submit

```
> x <- c(1, 2, 3, 4, 5)

> y <- c(1, 0.99, 2, 0.98, 0.97)

> fit1 <- lm(y ~ x)

> fit1

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept)          x

     1.209      -0.007

> fit2 <- lm(x ~ y)

> fit2

Call:

lm(formula = x ~ y)

Coefficients:

(Intercept)          y

   3.10084    -0.08488

> plot(x,y)

> abline(fit1)

> plot(y,x)

> abline(fit2)

>
```

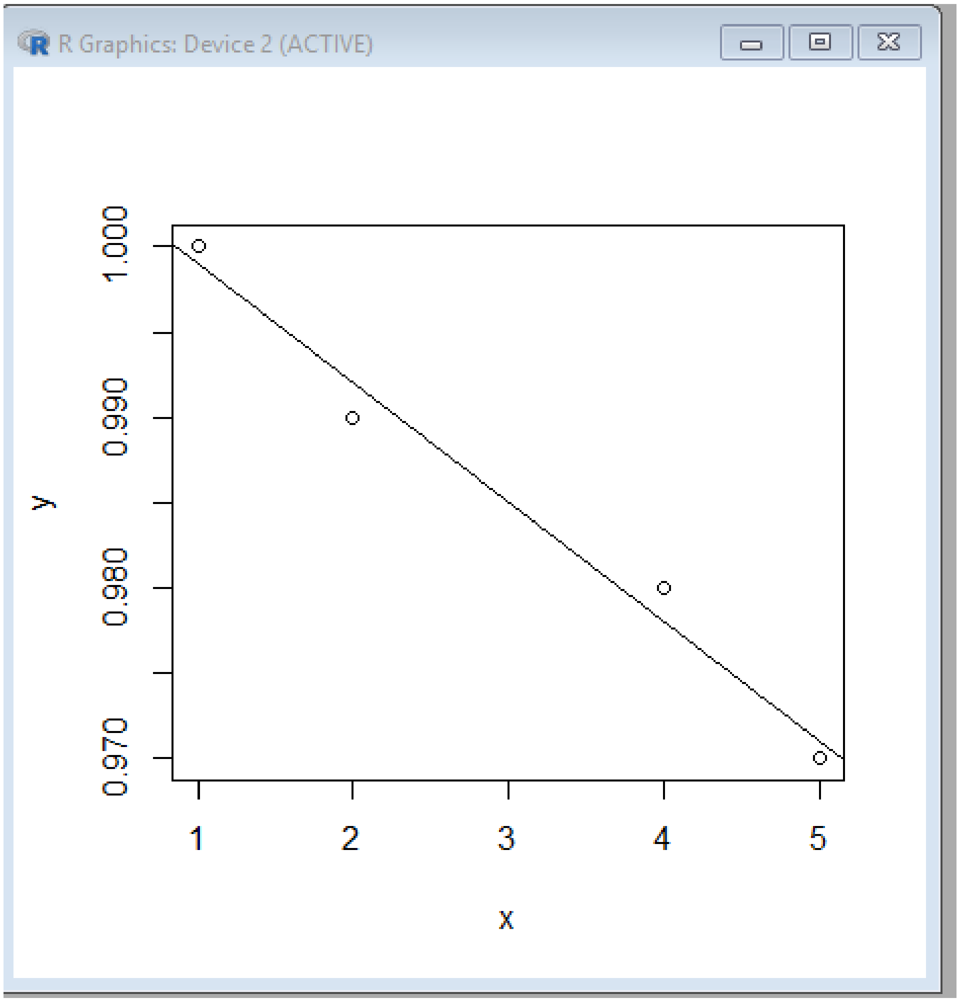


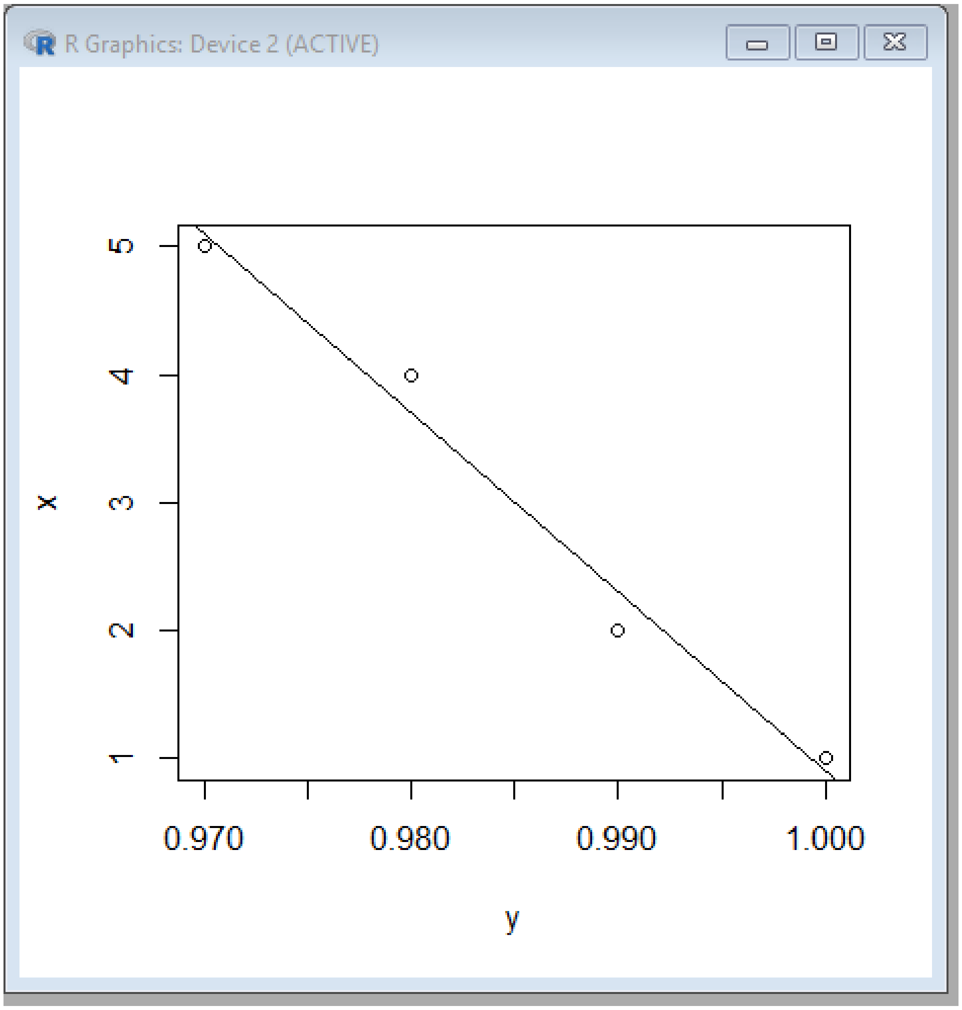
R Graphics: Device 2 (ACTIVE)
y
2.0
1.8
1.6
1.4
1.2
1.0
1
2
3
4
5
x



PART 2: Why are the regression lines so different? Which point should be removed to make the regression lines more similar to one another?

There is one outlier (3, 2) that causes the regression lines so different. So (3, 2) should be removed to make the regression lines more similar to one another.


R Graphics: Device 2 (ACTIVE)
y
1.000
0.990
0.980
0.970
1
2
3
4
5
x



Exercise 2. (4 points.)

A manufacturing company produces 2-dimensional square widgets, which are normally distributed with a length of 1 meter on each side, and a standard deviation of 0.01 meters.

PART 1: generate the two simulated datasets & submit the code

(a) Generate a data set with 100,000 widgets from this distribution.

```
> norm1 <- rnorm(100000, 1, 0.01)

> head(norm1, 50)

 [1] 0.9964152 0.9972576 1.0038010 1.0045504 1.0156240 1.0085583 0.9929304

 [8] 1.0027112 0.9831095 0.9989925 1.0042497 1.0118822 1.0038794 0.9986176

[15] 0.9779661 1.0005565 0.9939604 0.9949324 1.0113280 1.0023421 0.9869311

[22] 1.0030729 0.9908327 1.0168249 0.9911265 1.0158306 1.0014264 1.0100603
```

[29] 1.0159185 1.0005055 0.9901053 1.0056696 0.9991931 1.0048169 0.9719249

[36] 0.9957599 1.0155649 1.0000973 0.9985871 1.0137563 1.0151059 0.9946644

[43] 0.9848120 1.0079684 1.0129771 1.0159344 1.0002290 1.0009283 0.9865241

[50] 1.0006352

(b) The company produced 5 anomalous widgets, due a defect in the manufacturing process. Each such widget had a square length of 0.1 meters, and standard deviation of 0.001 meters. Generate these 5 anomalous points using the normal distribution assumption.

> norm2 <- rnorm(5, 0.1, 0.001)

> norm2

[1] 0.09911192 0.09881395 0.10253473 0.10087749 0.09790973

PART 2:

(c) Does a 1-NN approach find the anomalous widgets?

yes, 1-NN approach will help find the anomalous widgets.

> length(norm1)

[1] 100000

> length(norm2)

[1] 5

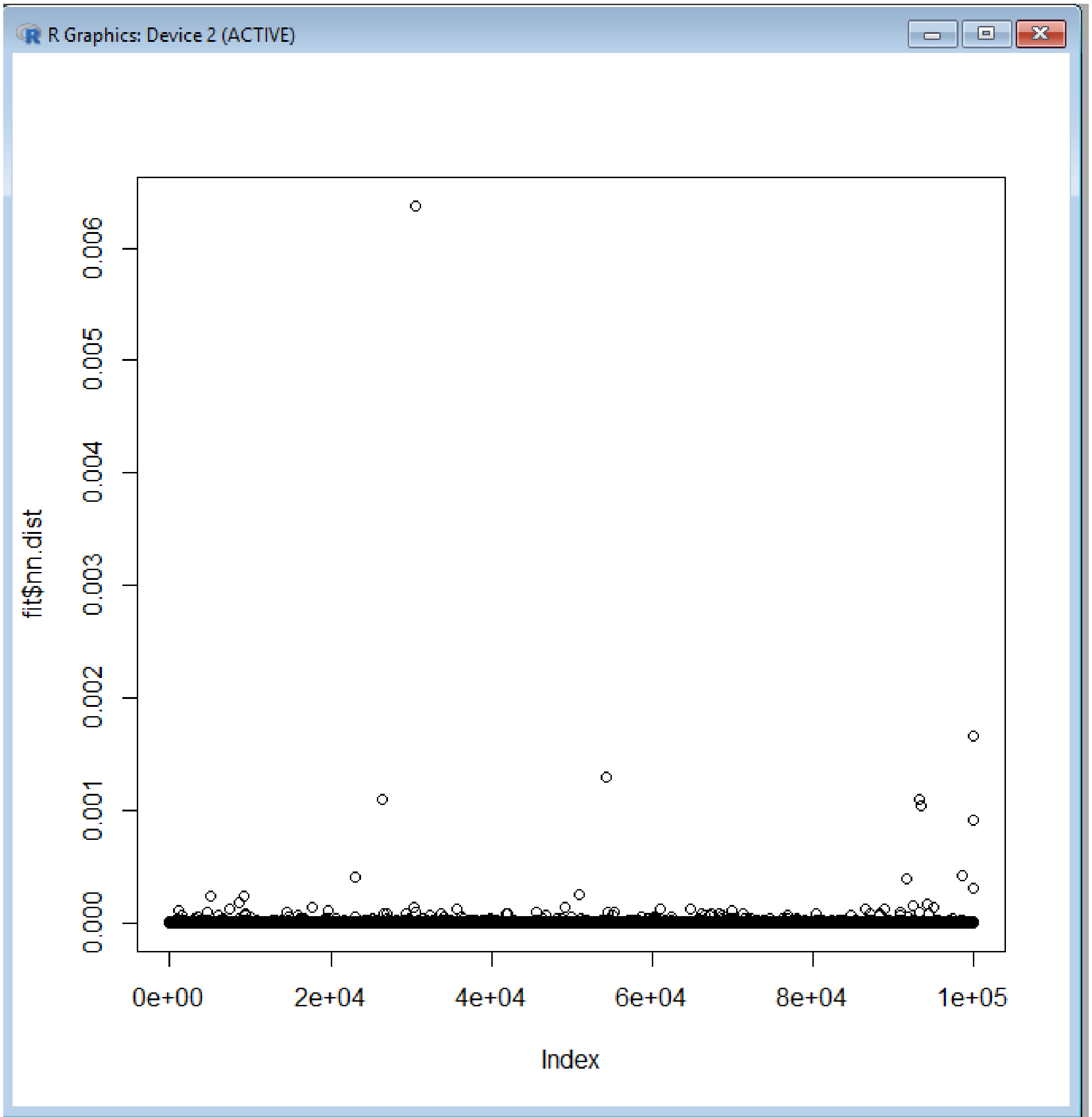
> norm<-c(norm1, norm2)

> length(norm)

[1] 100005

> fit<-get.knn(norm, k=1)

> plot(fit$nn.dist)



(d) Does a 10-NN approach find the anomalous widgets?

10-NN is not a good approach to find the anomalous widgets.