Homework 1

Harry Wang

2024-01-22

Problem 1:

Part i

```
x = c(2.7,4.0,2.3,5.4,-5.3,1.8,-1.3,-2.9,2.1,3.9,-1.8,0.4,-4.2,0.5,-0.1,1.5,-0.7)
y = c(1.4,2.5,2.6,5.6,-2.2,0.4,0.1,-3.0,2.2,0.9,-2.4,1.6,-2.5,0.1,-9.9,1.1,-1.7)
cbind(x,y)
```

```
##          x    y
##  [1,]  2.7  1.4
##  [2,]  4.0  2.5
##  [3,]  2.3  2.6
##  [4,]  5.4  5.6
##  [5,] -5.3 -2.2
##  [6,]  1.8  0.4
##  [7,] -1.3  0.1
##  [8,] -2.9 -3.0
##  [9,]  2.1  2.2
## [10,]  3.9  0.9
## [11,] -1.8 -2.4
## [12,]  0.4  1.6
## [13,] -4.2 -2.5
## [14,]  0.5  0.1
## [15,] -0.1 -9.9
## [16,]  1.5  1.1
## [17,] -0.7 -1.7
```

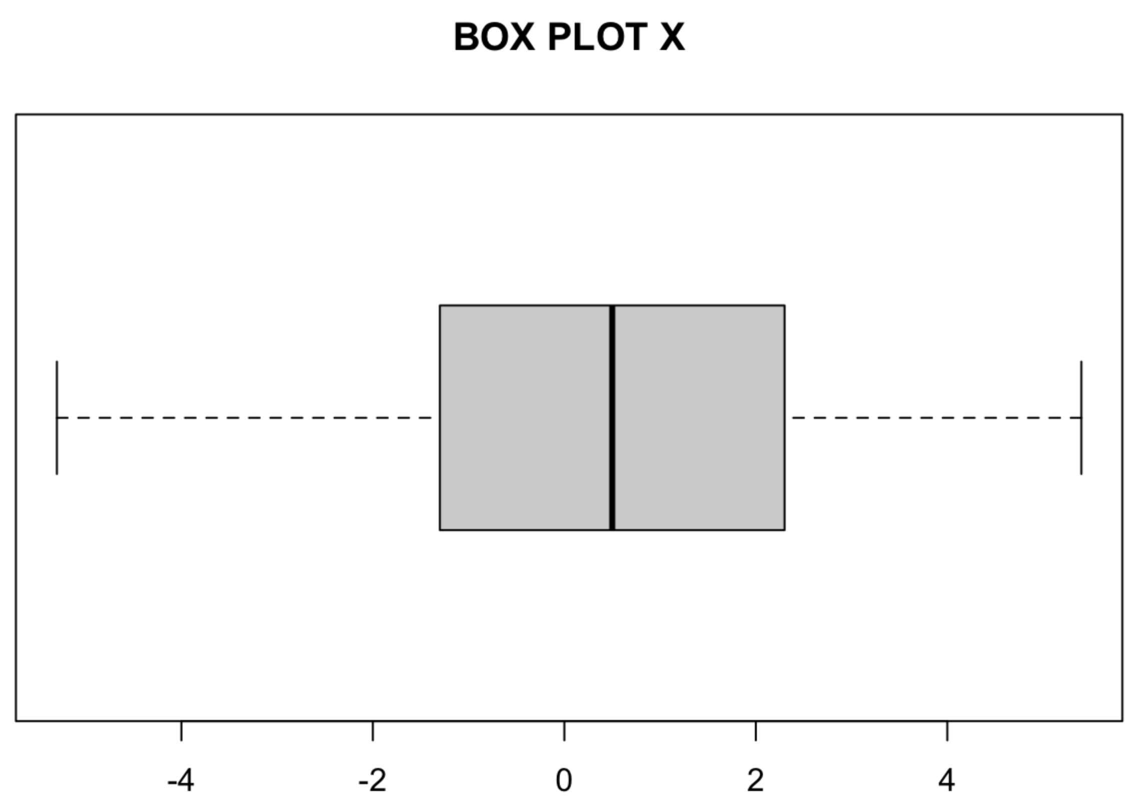
```
summary(x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -5.3000 -1.3000  0.5000  0.4882  2.3000  5.4000
```

```
var(x)
```

```
## [1] 8.673603
```

```
boxplot(x,horizontal = TRUE, main = "BOX PLOT X")
```



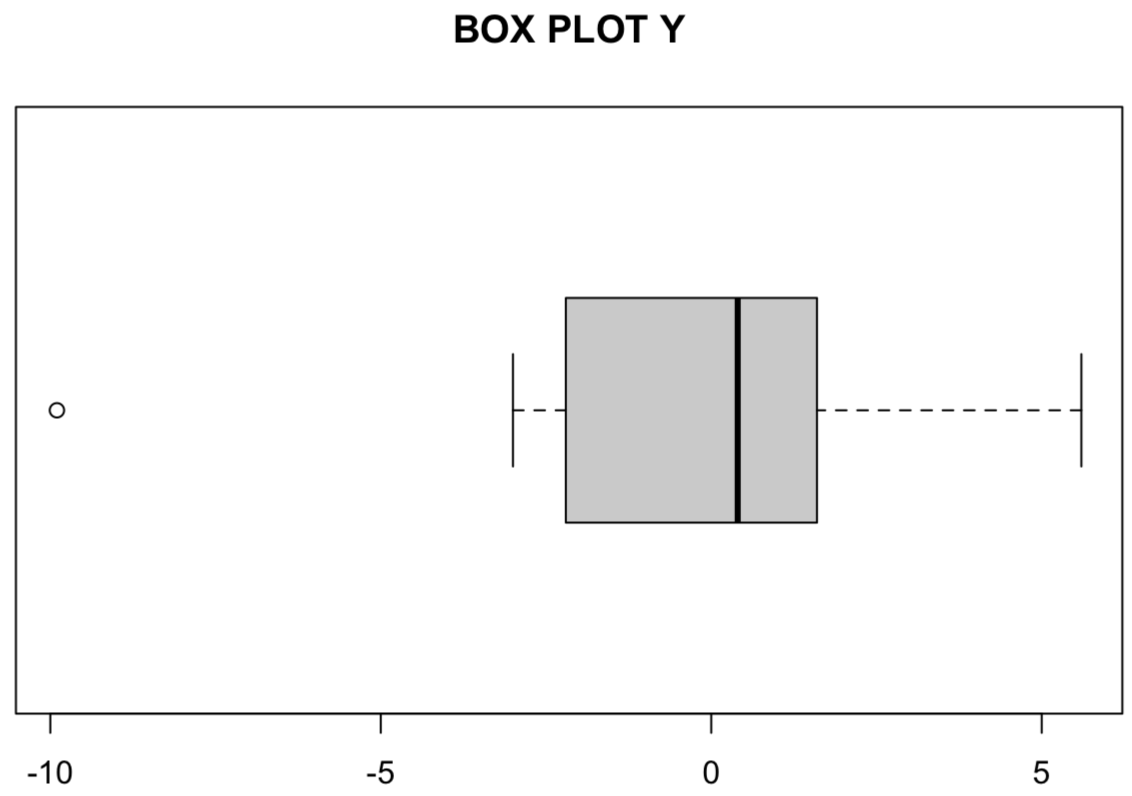
```
summary(y)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
## -9.9000 -2.2000  0.4000 -0.1882  1.6000  5.6000
```

```
var(y)
```

```
## [1] 11.37985
```

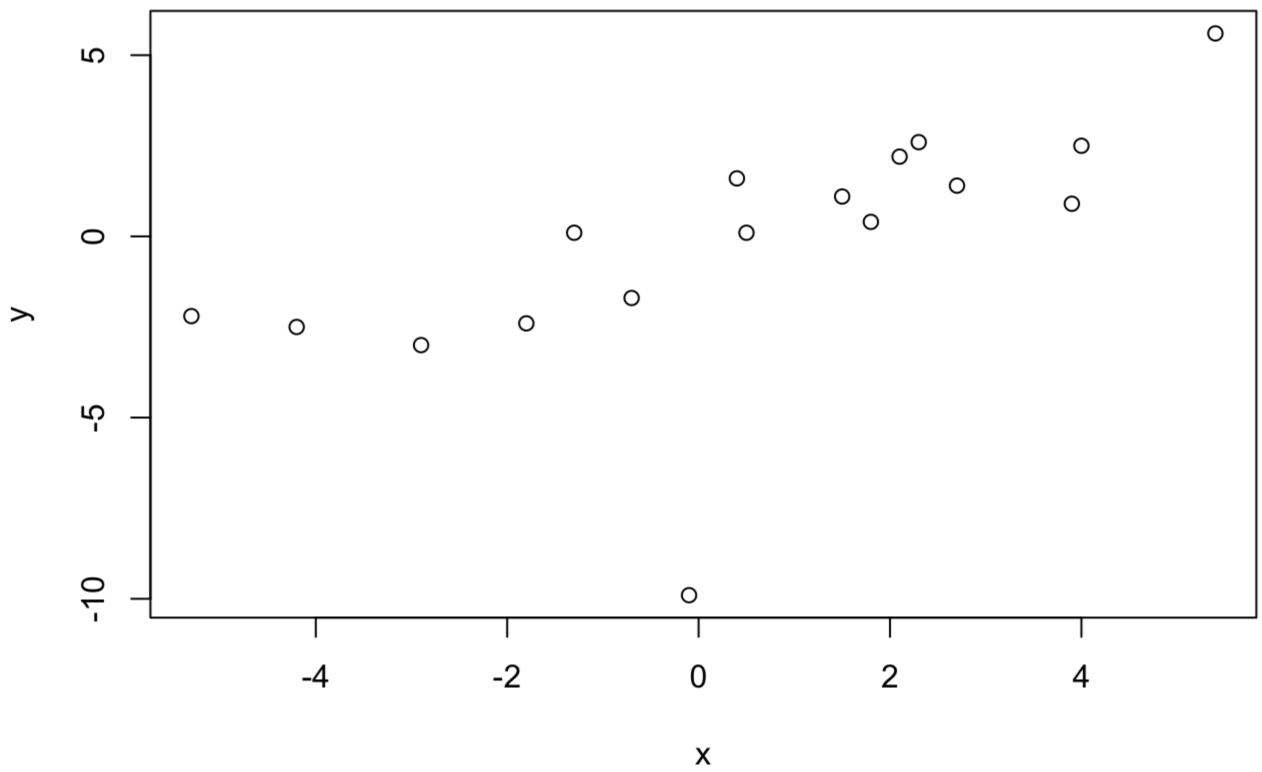
```
boxplot(y,horizontal = TRUE, main = "BOX PLOT Y")
```



For x, the box-plot is skewed left. There are no outliers.
For y, the box-plot is skewed right. There is one outlier which is -9.9.

Part ii

```
plot(x,y)
```




```
cor(x,y)
```

```
## [1] 0.6289777
```

The scatter plot for x and y, the linear association between x and y is

Part iii

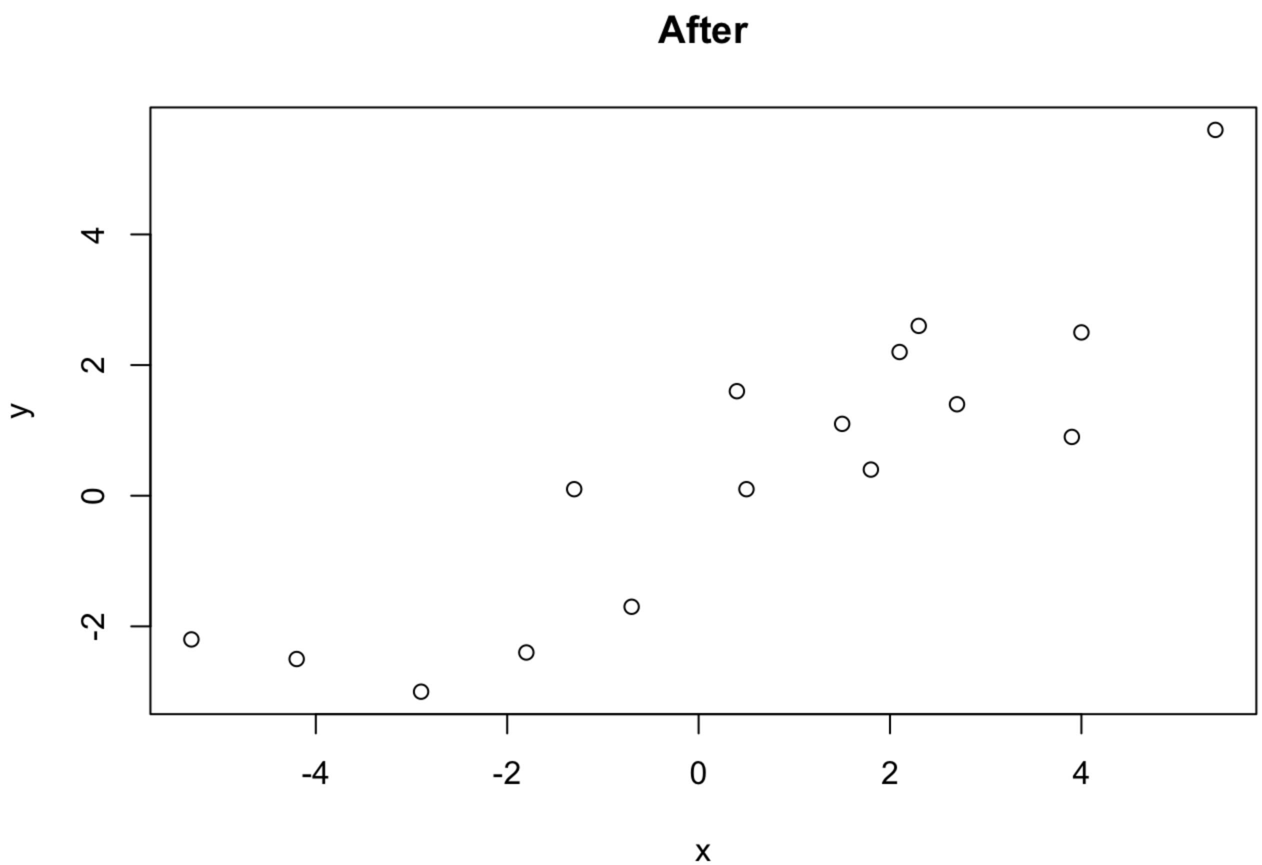
Since a paired observation is usually considered to be an outlier if one if its two coordinators is an outlier. Since there is one in y, the outlier is (-0.1,-9.9). This is the correlation coefficient after removing the outlier.

```
x = c(2.7,4.0,2.3,5.4,-5.3,1.8,-1.3,-2.9,2.1,3.9,-1.8,0.4,-4.2,0.5,1.5,-0.7)
y = c(1.4,2.5,2.6,5.6,-2.2,0.4,0.1,-3.0,2.2,0.9,-2.4,1.6,-2.5,0.1,1.1,-1.7)
cor(x,y)
```

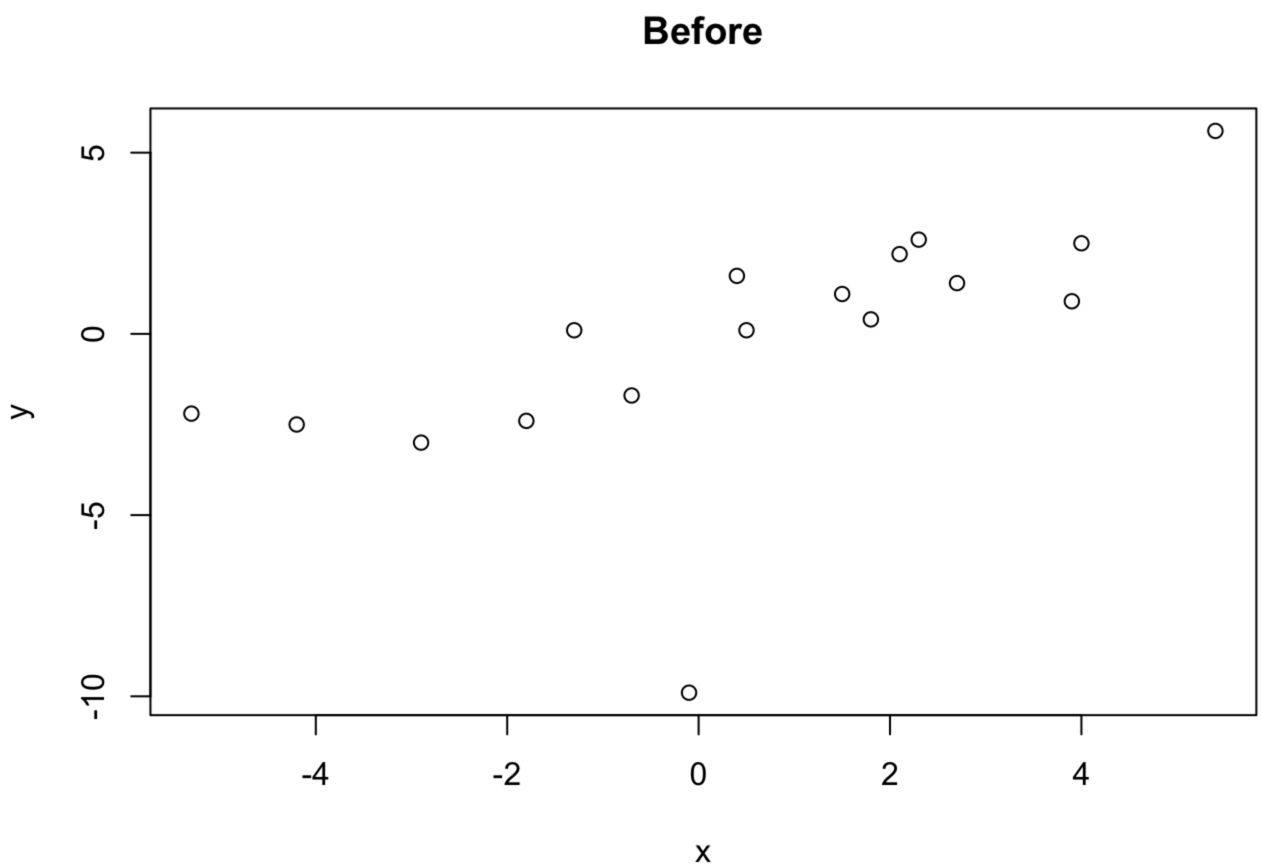
```
## [1] 0.8822511
```

Part iv

```
plot(x,y, main = "After")
```



```
x = c(2.7,4.0,2.3,5.4,-5.3,1.8,-1.3,-2.9,2.1,3.9,-1.8,0.4,-4.2,0.5,-0.1,1.5,-0.7)
y = c(1.4,2.5,2.6,5.6,-2.2,0.4,0.1,-3.0,2.2,0.9,-2.4,1.6,-2.5,0.1,-9.9,1.1,-1.7)
plot(x,y, main = "Before")
```

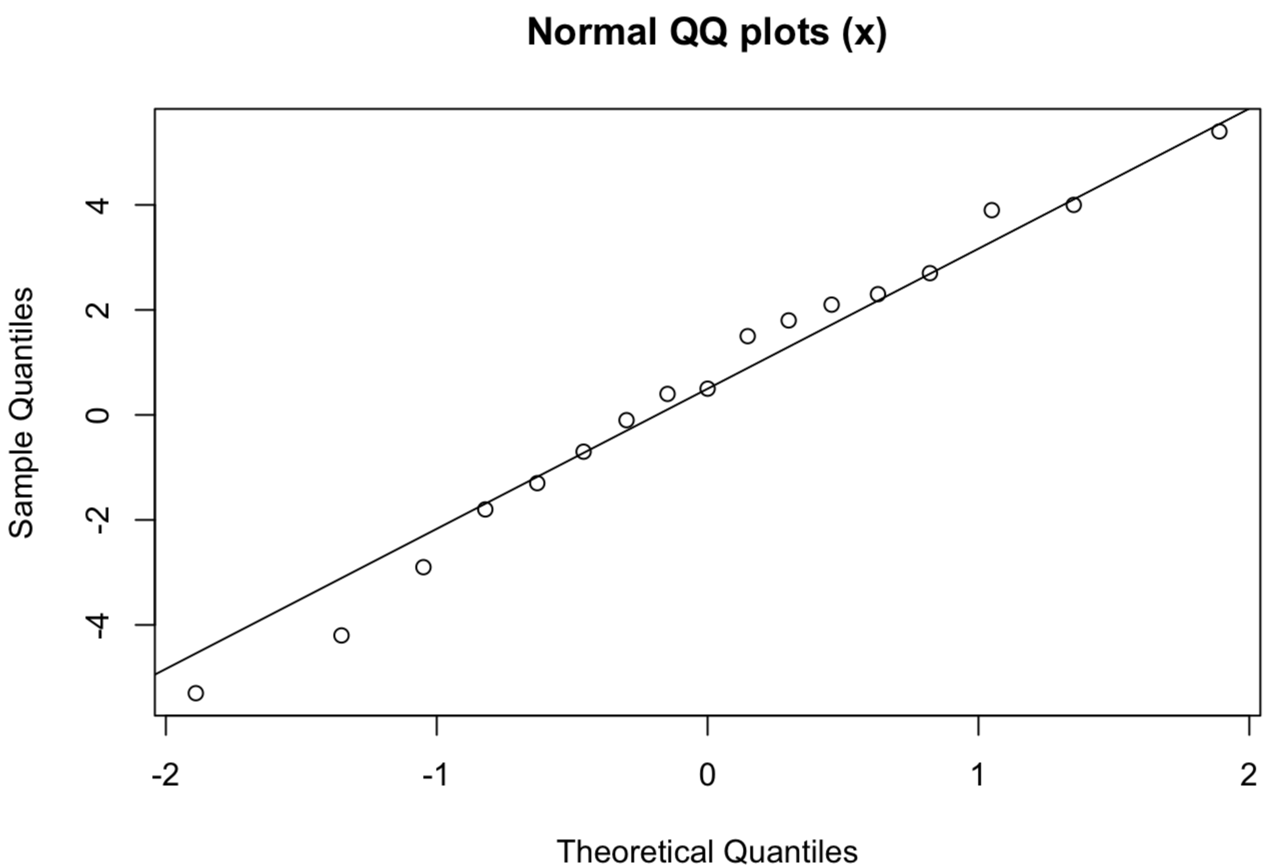


You can visually see the “Before” graph has an outlier on the bottom which causes a huge decrease on the correlation coefficient. But the “After”, you can not visually see any outliers.

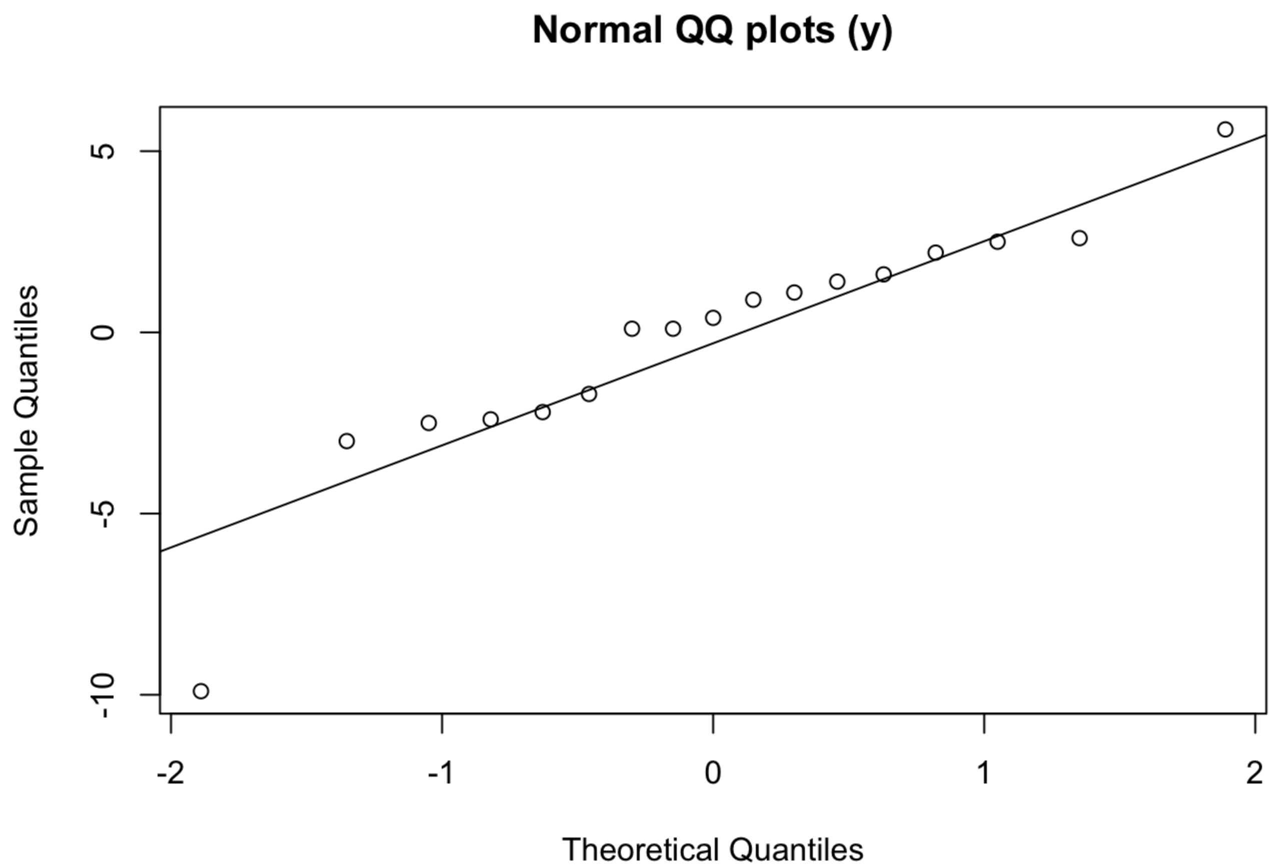
Part v The correlation coefficient after the removal of the outlier went from 0.6289777 to 0.8822511. The correlation coefficient measures the degree of linear association between vectors x and y. So removing of the outlier increased the linear association between vectors x and y.

Part vi

```
qqnorm(x, main = "Normal QQ plots (x)")
qqline(x)
```

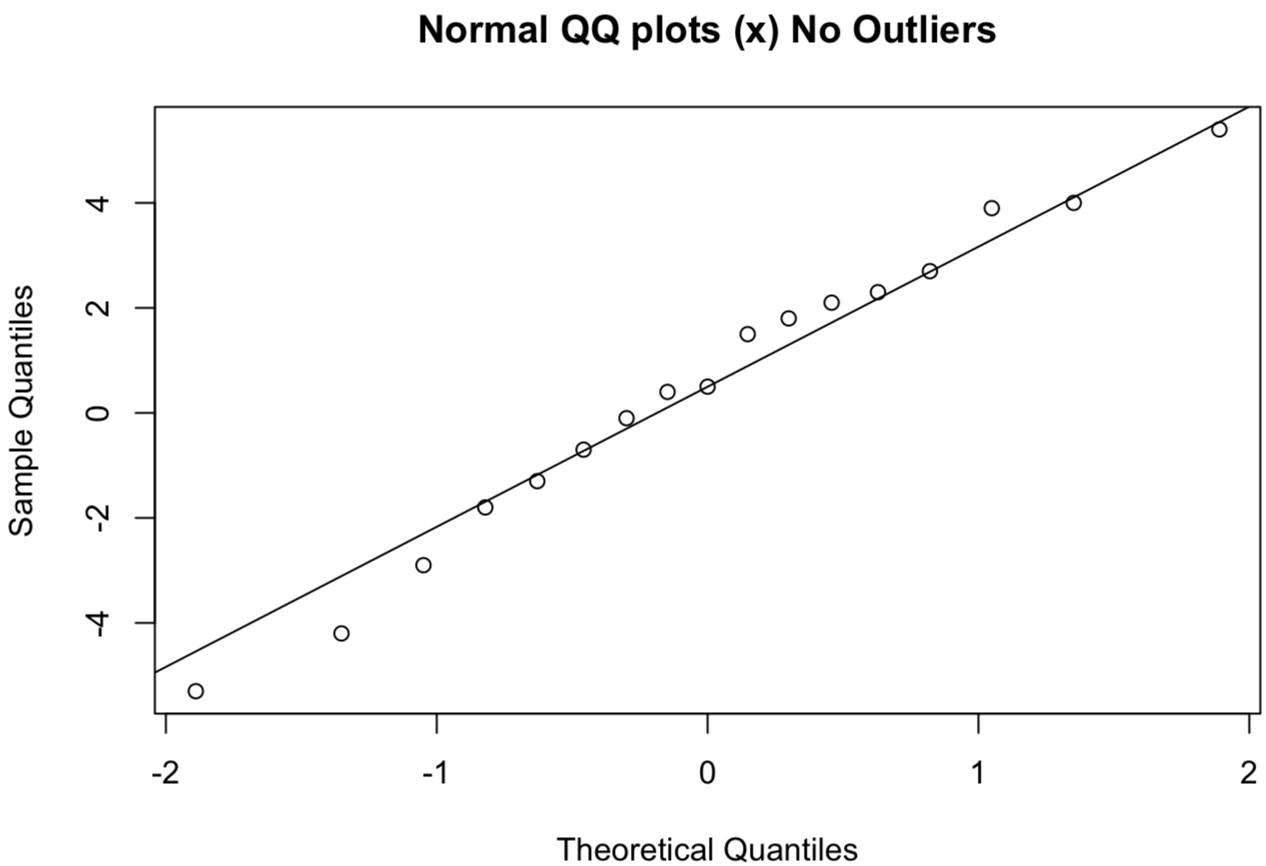



```
qqnorm(y, main = "Normal QQ plots (y)")
qqline(y)
```

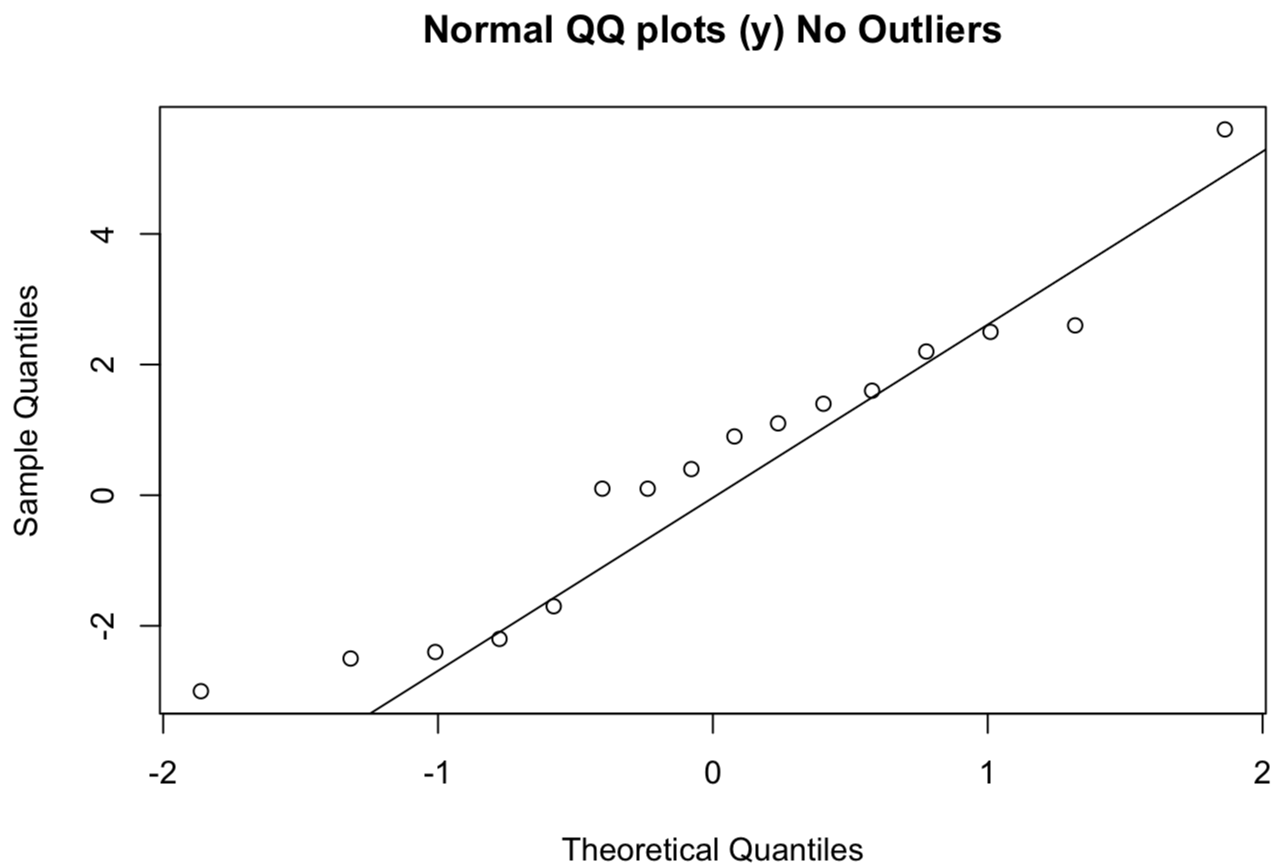


The one that is more likely to be of normal distribution is x because it is more linear. Normally distributed data appears as roughly a straight line.

```
x = c(2.7,4.0,2.3,5.4,-5.3,1.8,-1.3,-2.9,2.1,3.9,-1.8,0.4,-4.2,0.5,1.5,-0.7,-0.1)
y = c(1.4,2.5,2.6,5.6,-2.2,0.4,0.1,-3.0,2.2,0.9,-2.4,1.6,-2.5,0.1,1.1,-1.7)
qqnorm(x, main = "Normal QQ plots (x) No Outliers")
qqline(x)
```




```
qqnorm(y, main = "Normal QQ plots (y) No Outliers")
qqline(y)
```



After removing outliers, it seems like x still the one that is more likely to be normal distributed.

Problem 2

$P(|Z|>1) = 0.3173105$
$P(|Z|>2) = 0.04550026$
$P(|Z|>3) = 0.002699796$
$P(Z \leq z0.1/2) = 0.05$
$P(Z \leq z1-0.1/2) = 0.950$
$P(z0.1/2 \leq Z \leq z1-0.1/2) = 0.900$

```
pnorm(-1) * 2
```

```
## [1] 0.3173105
```

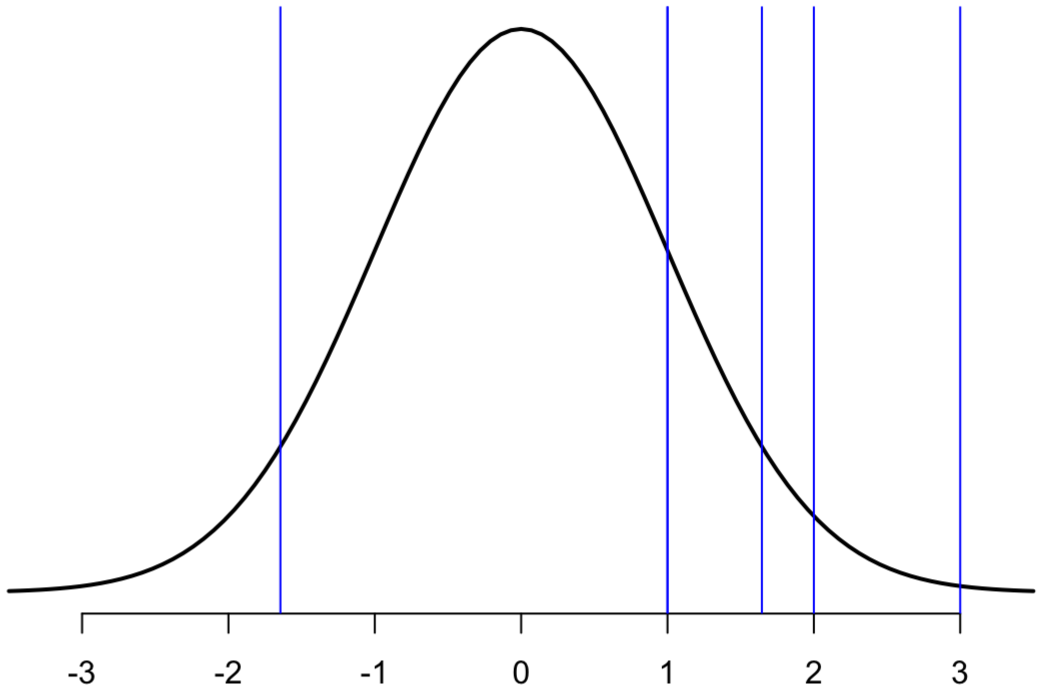
```
pnorm(-2) * 2
```

```
## [1] 0.04550026
```

```
pnorm(-3) * 2
```

```
## [1] 0.002699796
```

```
curve(dnorm, -3.5, 3.5, lwd=2, axes = FALSE, xlab = "", ylab = "")
axis(1, at = -3:3, labels = c("-3", "-2", "-1", "0", "1", "2", "3"))
abline(v= 1, col="blue")
abline(v= 2, col="blue")
abline(v= 3, col="blue")
abline(v= -1.645, col="blue")
abline(v= 1.645, col="blue")
abline(v= 1, col="blue")
```



Problem 3

P(X ≤ F −1(α/2))

Meaning: This probability represents the likelihood that the random variable is less than or equal to the value at which the cumulative distribution function F(X) reaches (α/2). Essentially, it's the probability of X being in the lower tail of its distribution, up to the α/2 quantile.

Numerical Value: F(F^(-1)(α/2)) = α/2.

P(X > F −1(1 − α/2))

Meaning: This is the probability that X is greater than the value at which the CDF F(X) reaches 1- α/2 . It represents the probability of X being in the upper tail of its distribution, beyond the 1 - α/2 quantile.

Numerical Value: 1 - F(F^(-1)(1 − α/2)) = 1 - (1 - α/2) = α/2.

PF −1(α/2) ≤ X ≤ F −1(1 − α/2) = This probability indicates the likelihood that X falls between the α/2 and 1 - α/2 quantiles of its distribution. It essentially measures the probability of X being within the central 1-α portion of its distribution.

Numerical Value: F(F^(-1)(1 − α/2)) - F(F^(-1)(α/2)) = (1 - α/2) - (α/2) = 1 - α.

In summary, as α decreases, the probabilities of X being in the extreme tails of its distribution decrease, while the probability of X falling within a wider central interval increases.

Problem 4

Centralization:

$\sum_{i=1}^{n} (x_i - x) = \sum_{i=1}^{n} ((x_i) - nx)$
$\sum_{i=1}^{n} (x_i - x) = \sum_{i=1}^{n} x_i - n$ (1/n $\sum_{i=1}^{n} x_i$) = $\sum_{i=1}^{n} x_i$ - $\sum_{i=1}^{n} x_i$ = 0

Square of the Sum:

$(\sum_{i=1}^{n} x_i)^2 = \sum_{i=1}^{n} x_i$ x $\sum_{i=1}^{n} x_i^2 = (\sum_{i=1}^{n} x_i)^2$ + 2 $\sum_{1<=i<=j<=n}^{n} x_i x_j$

So, the left side is equal to the right side, and the equation is proven.

Sum of Squares:

($\sum_{i=1}^{n} x_i^2$)/n = (n $\sum_{i=1}^{n} x_i^2$)/n^2 = 1/n $\sum_{i=1}^{n} x_i^2$)
($\sum_{i=1}^{n} x_i$)/n)^2 = (x)^2
1/n ($\sum_{i=1}^{n} x_i^2$) >= (x)^2

So, the inequality is proven.

Sum of Squared Distances:

$\sum_{i=1}^{n} (x_i - x)^2 = \sum_{i=1}^{n} (x_i^2 - 2x_i x + x^2) = \sum_{i=1}^{n} x_i^2$ - 2x $\sum_{i=1}^{n} x_i + \sum_{i=1}^{n} (x^2) = \sum_{i=1}^{n} (x_i^2 - 2nx^2 + nx^2) = \sum_{i=1}^{n} (x_i^2 - nx^2)$

So, the left side is equal to the right side, and the equation is proven.