STAT8121 Assignment 2

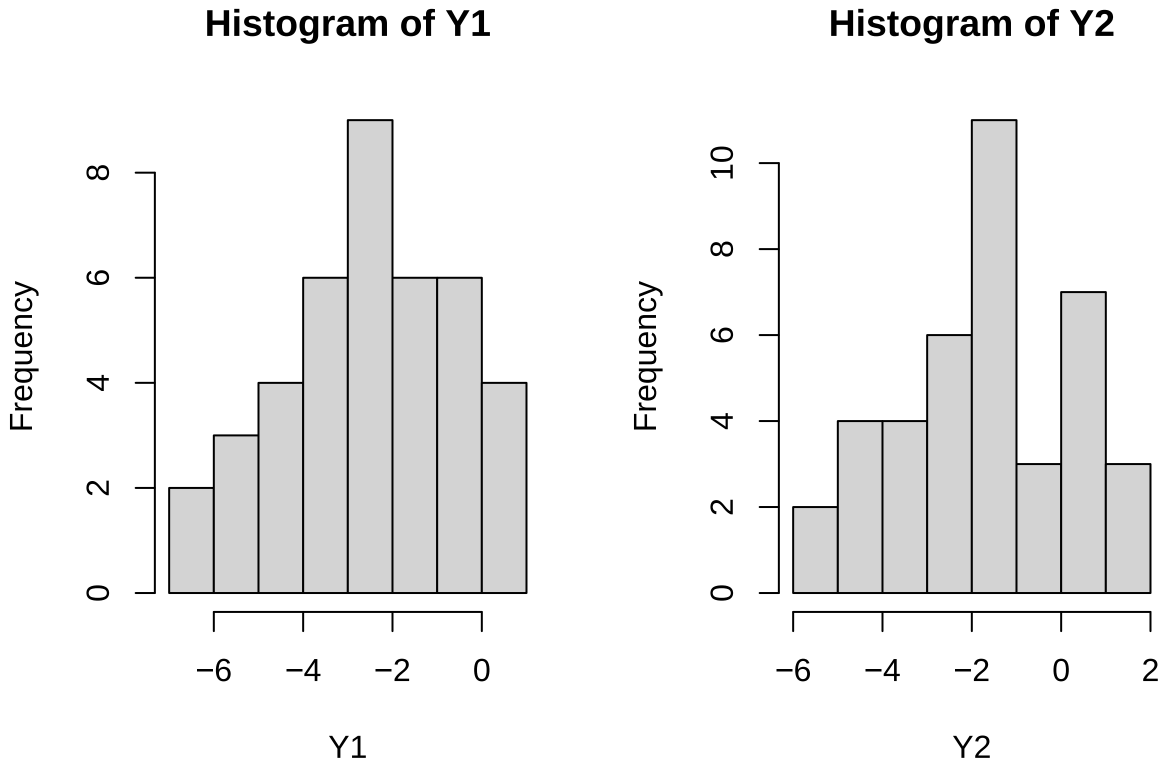
Aditya Tanaji Sagave

Student ID: 47541164

```
data <- read.table("./A2Data1.csv",header=T, sep=",")
```

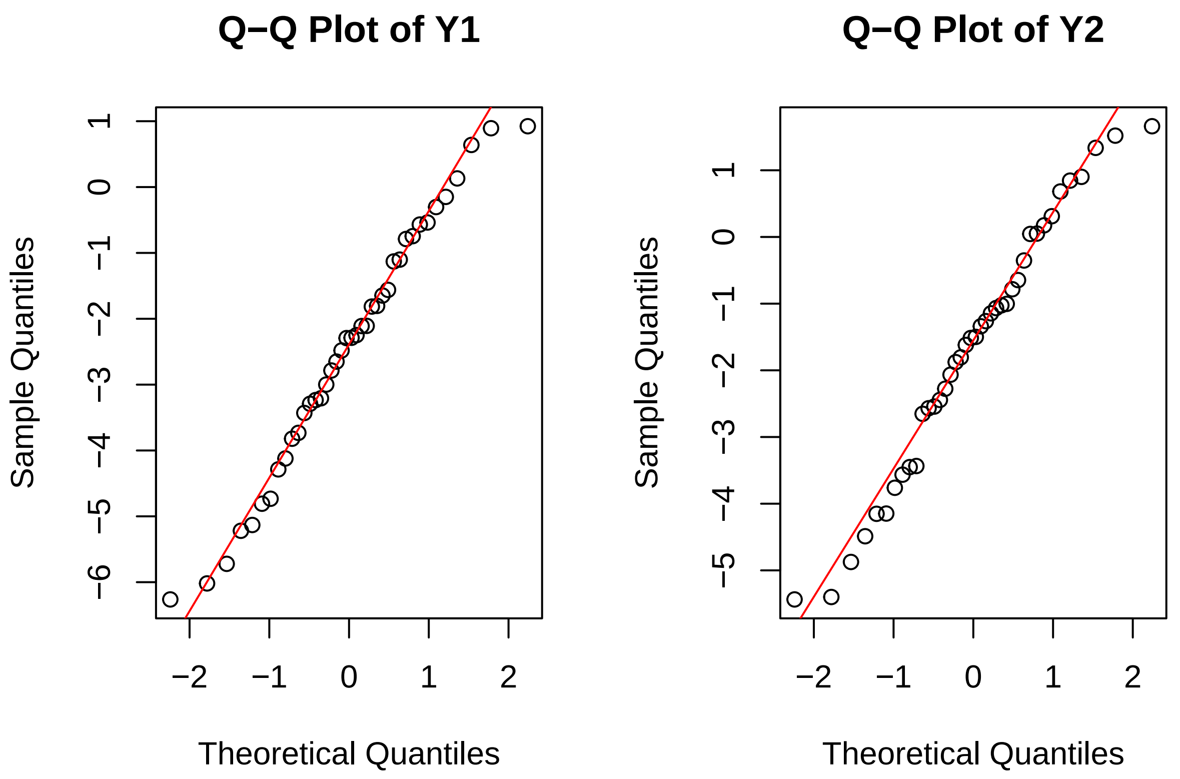
Q1

```
Y1 <- data$y1
Y2 <- data$y2

par(mfrow=c(1,2))

hist(Y1, main="Histogram of Y1", xlab="Y1")
hist(Y2, main="Histogram of Y2", xlab="Y2")
```



Answer 1.a:

```r
par(mfrow=c(1,2))  # Reset the plot layout

qqnorm(Y1, main="Q-Q Plot of Y1")
qqline(Y1, col="red")

qqnorm(Y2, main="Q-Q Plot of Y2")
qqline(Y2, col="red")
```



Histograms:

1. For Y1: The histogram of Y1 (Figure 1) display a roughly symmetric shape with a slight left skew. It does not seem to be perfect symmetrical.
2. For Y2: The histogram of Y2 (Figure 2) displays a somewhat symmetric shape, but it is slightly skewed to the left as well.

Q-Q Plots:

1. For Y1: The Q-Q plot of Y1 (Figure 3) shows points that points closely follow the red reference line, there are few points at the extreme upper and lower tails that deflect from the reference line, but that don't seem to influence other points.
2. For Y2: The Q-Q plot of Y3 (Figure 4) shows points that points from mid to lower tail don't follow the reference line which indicates a potential nonlinearity in the lower tail. However the rest point follow the reference line.

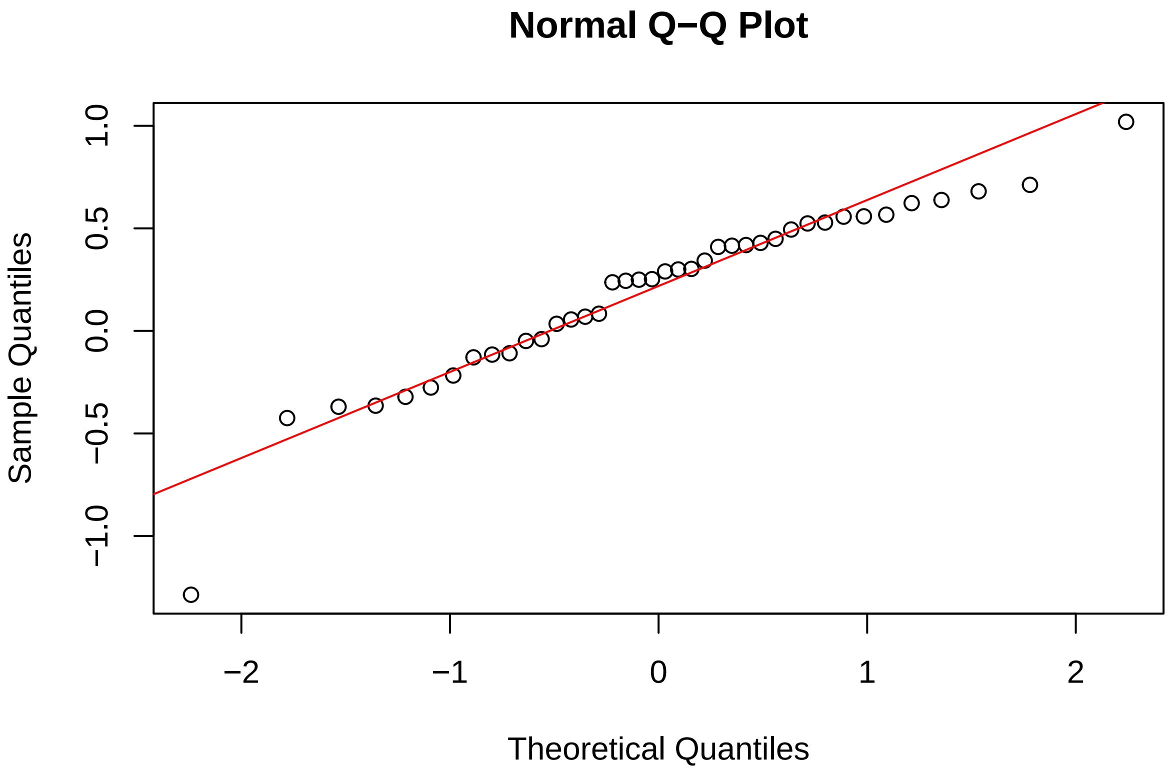
```r
data_matrix <- cbind(Y1, Y2)
```

```r
mean_vector <- colMeans(data_matrix)


cov_matrix <- cov(data_matrix)

n <- nrow(data_matrix)

gen_distances <- numeric(n)

for (i in 1:n) {
  gen_distances[i] <- sqrt(t(data_matrix[i,] - mean_vector) %*% solve(cov_matrix) %*% (data_matrix[i,] -
}

gen_distances
```

Answer 1.b:

```
##  [1] 1.2836644 0.8044013 1.3529006 0.9607154 1.2673011 0.8967455 0.9522374
##  [8] 0.2762041 1.5148188 0.7586795 1.7454639 1.3492415 1.5062157 2.7714688
## [15] 1.4085402 1.0350997 1.0873280 0.6537737 1.6393033 1.0571576 1.2767992
## [22] 2.0382887 0.8789156 1.9748716 1.5359484 1.3365682 0.8908617 0.6907763
## [29] 0.7250478 1.2869708 1.6958207 1.5198422 1.8932196 0.6944016 1.8645906
## [36] 1.7482318 1.7624329 1.5664691 1.0716687 1.6887848
```

```r

qqnorm(log(gen_distances))
qqline(log(gen_distances), col = "red")
```



Answer 1.c:

By looking at the qqplot it seems that the generalized distance points roughly follow the reference line suggesting that the data points do possess normality. There are 1 to 2 outliers but they don't influence any other data points. But to assess whether the distances follow bivariate normal distribution or not we have to use perspective and contour plots.

```
n <- nrow(data)
p <- ncol(data)
xbar <- colMeans(data)
xbar
```

Answer 1.d

```
##        y1        y2
## -2.463679 -1.666963
```

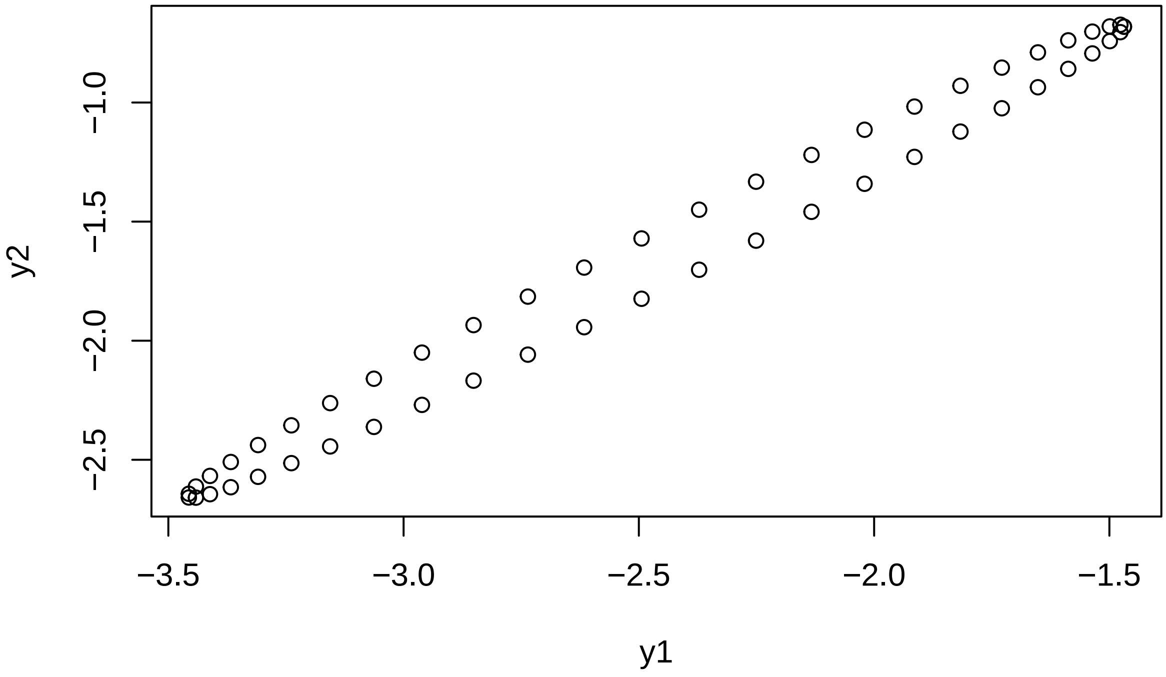
```
S <- cov(data)
S
```

```
##          y1       y2
## y1 3.699406 3.665310
## y2 3.665310 3.691522
```

```
tconst <- sqrt((p/n)*((n-1)/(n-p)) * qf(0.99,p,n-p))
tconst
```

```
## [1] 0.5171249
```

```
id <- c(1,2)
plot(ellipse(center=xbar[id], shape=S[id,id], radius=tconst, draw=F), xlab="y1", ylab="y2")
```



```
n <- nrow(data)
p <- ncol(data)
xbar <- colMeans(data)
xbar
```

Answer 1.e

```
##        y1        y2
## -2.463679 -1.666963
```

```
S <- cov(data)
S
```

```
##          y1       y2
## y1 3.699406 3.665310
## y2 3.665310 3.691522
```

```
a=0.01
```

```
critical_vb = qt(1-(a/p)/2, n-1)
B_lower = xbar-critical_vb*sqrt(diag(S)/n)
B_upper = xbar+critical_vb*sqrt(diag(S)/n)
B_lower
```

```
##        y1        y2
## -3.368603 -2.570921
```

```
B_upper
```

```
##         y1         y2
## -1.5587560 -0.7630039
```

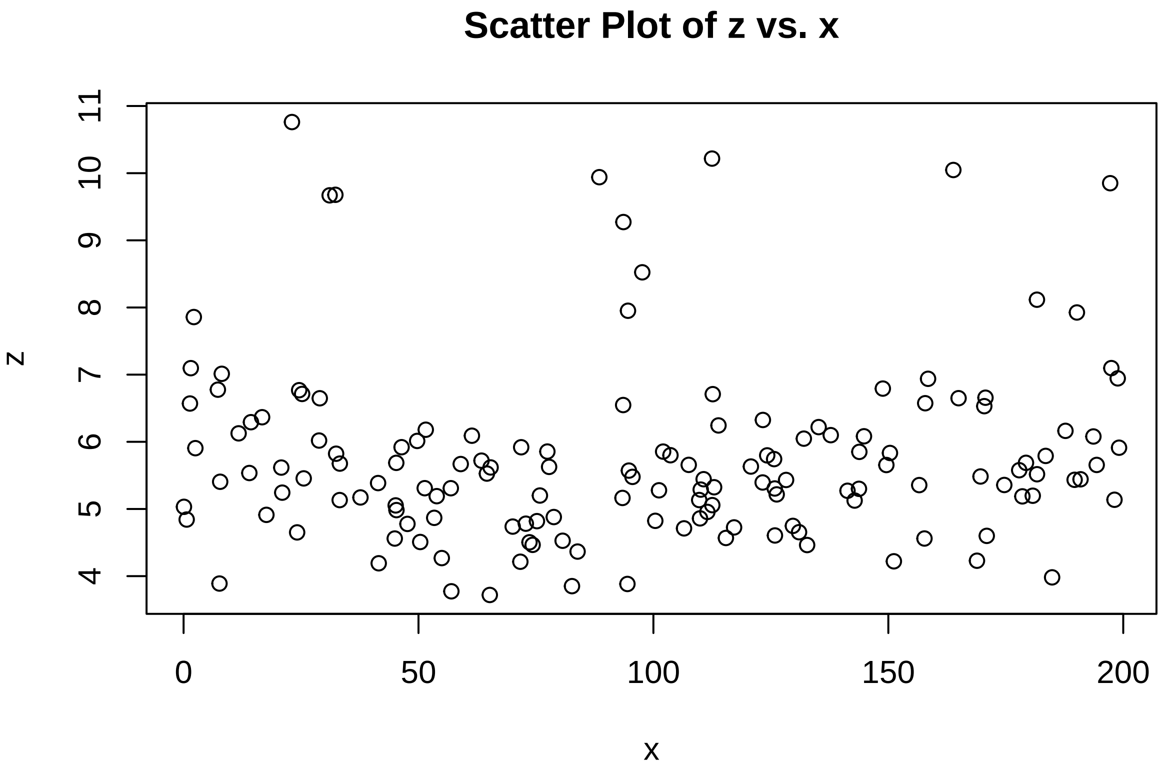
Interpretation:

1. We are 99% confident that the true mean of variable y1 (µ1) falls within the range of approximately -3.368 to -1.558.
2. We are 99% confident that the true mean of variable y2 (µ2) falls within the range of approximately -2.570 to -0.763.

Question 2

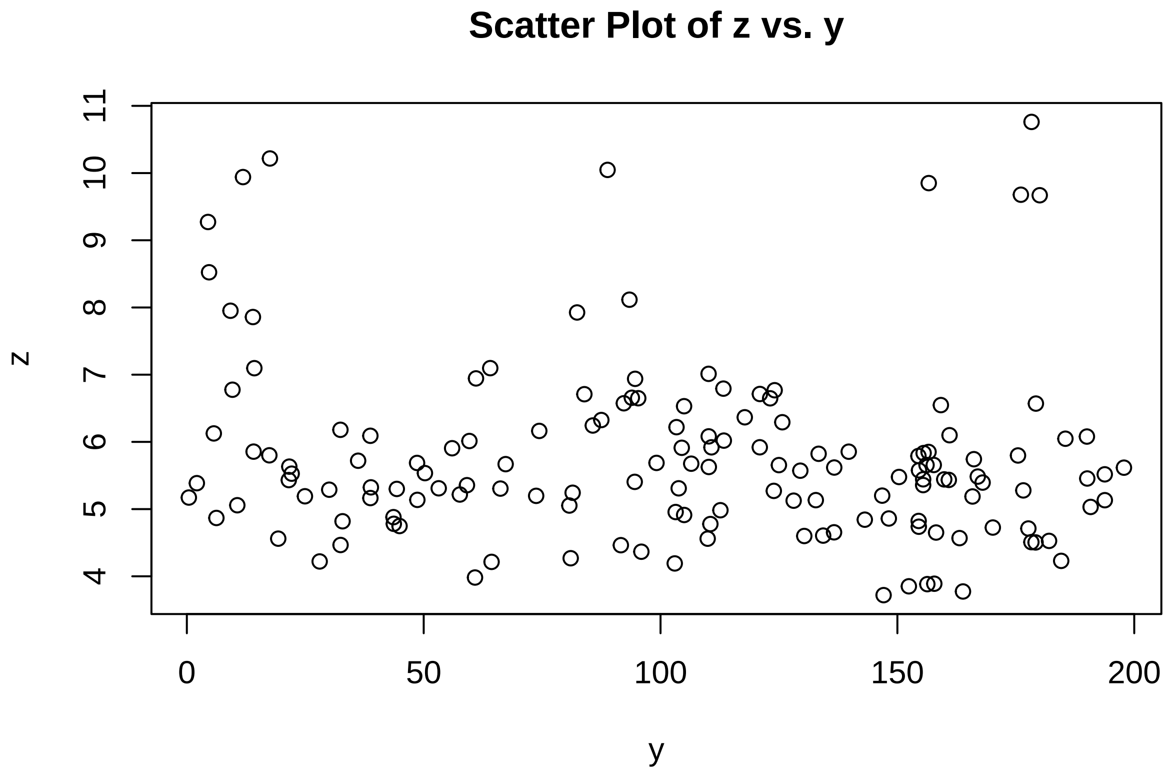
```
data <- read.table("./A2Data2.csv",header=T, sep=",")
```

```
# Create a 2D scatter plot of z vs. x and z vs. y
plot(data$x, data$z, xlab = "x", ylab = "z", main = "Scatter Plot of z vs. x")
```



Answer 1.a

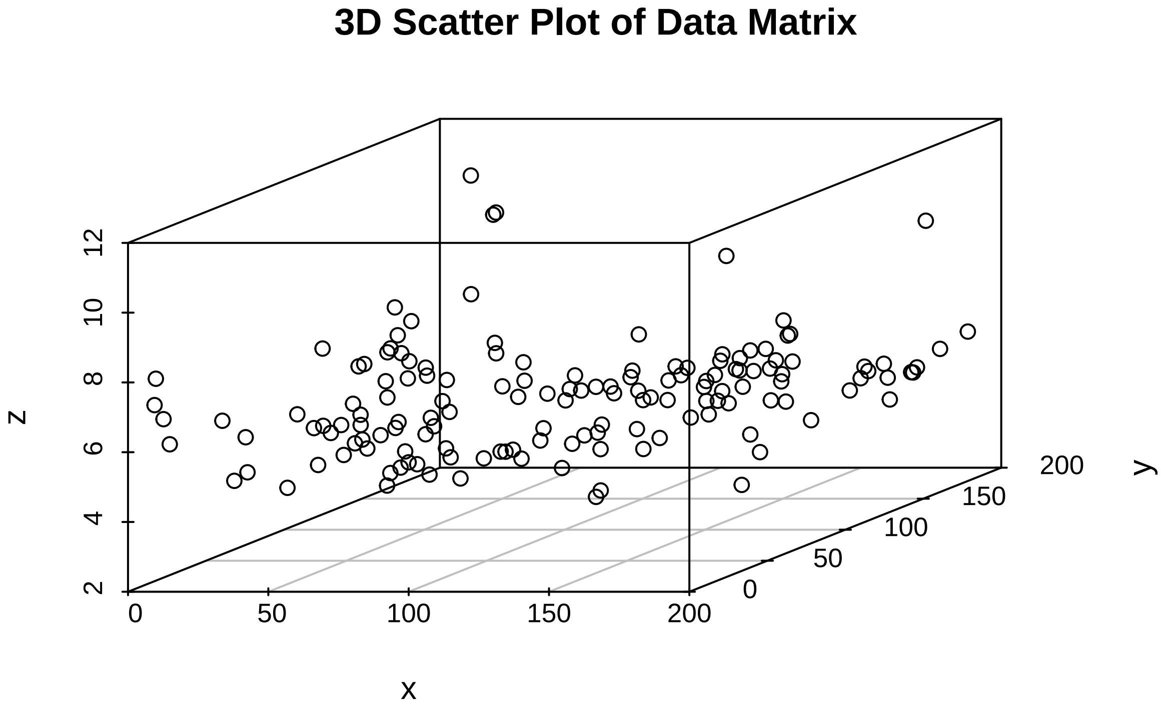
```
plot(data$y, data$z, xlab = "y", ylab = "z", main = "Scatter Plot of z vs. y")
```



Similarities: The scatter plots (z, x) and (z, y) share common characteristics in that both exhibit randomly scattered points, with a notable concentration observed under the z = 8 value in both plots. These commonalities suggest a consistent behavior of the variable z in relation to x and y across the data points.

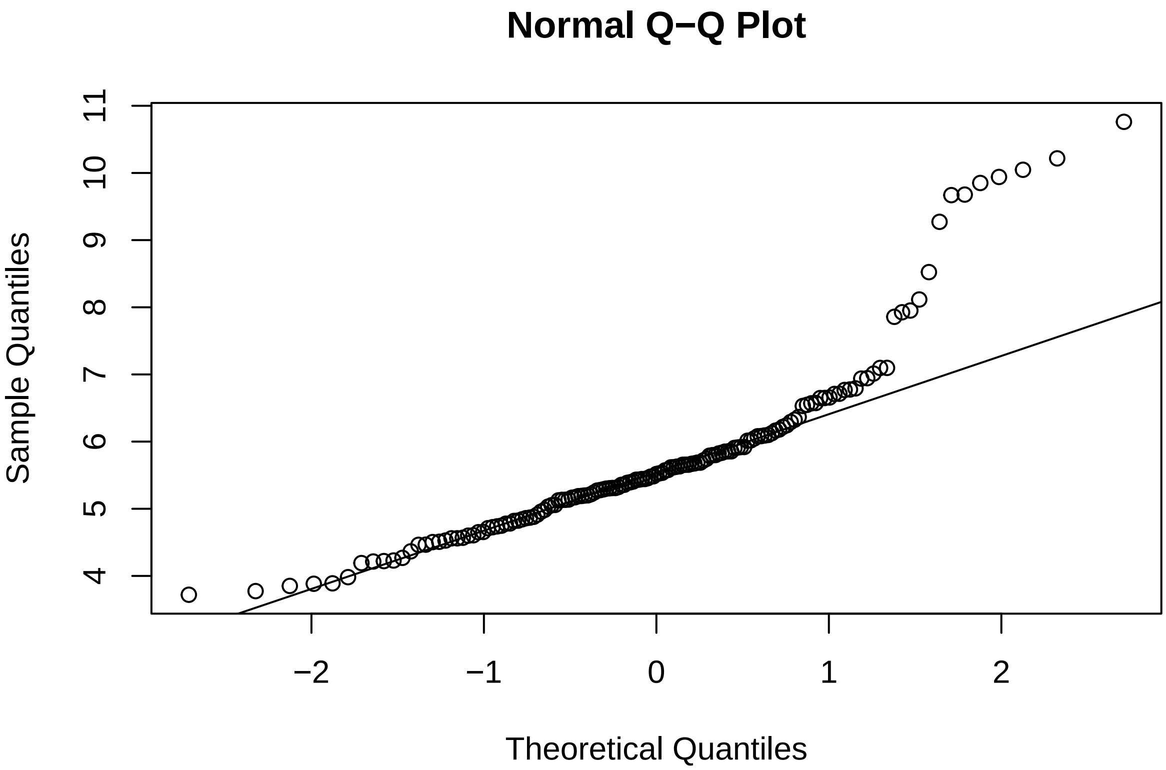
Differences: There are no noticeable differences in both plot since both are randomly scattered.

```
# Load the scatterplot3d library
library(scatterplot3d)

# Create a 3D scatter plot of the data matrix
scatterplot3d(data$x, data$y, data$z,
              xlab = "x", ylab = "y", zlab = "z",
              main = "3D Scatter Plot of Data Matrix")
```



Answer 2.b normality of z

```
qqnorm(data$z)
qqline(data$z)
```



The z col does not fully shows a normality because the data point in the extreme lower and upper 25% deviate from the qqline heavily suggesting that only mid tail poits show normality. This could be due to several reasons. maybe the data itself belongs to other distributions such as chi-squared or there might be some outliers who influence other data points.

Definitions for skewness and kurtosis

Skewness: Skewness evaluates the distribution's asymmetry. A right-skewed distribution is shown by a positive skewness, whereas a left-skewed distribution is shown by a negative skewness. The skewness of a normal distribution is 0.

Kurtosis: The distribution's tail heaviness is measured by kurtosis. It is common to report extra kurtosis, which subtracts 3 from the standard kurtosis. An excess kurtosis in a normal distribution is equal to 0.

```
skew <- skewness(data$z)
kurt <- kurtosis(data$z)

skew
```

```
## [1] 1.643386
```

```
kurt
```

```
## [1] 6.185303
```

The skewness value of 1.643386 indicates that the distribution of the variable z is positively skewed. Positive skewness means that the tail of the distribution extends to the right, and the majority of data points are concentrated on the left side of the distribution, causing the mean to be greater than the median.

The kurtosis value of 6.185303 suggests that the distribution of z has heavy tails or is leptokurtic. Leptokurtic distributions have a higher peak and fatter tails than a normal distribution. In this case, the distribution has more extreme values than a normal distribution, resulting in a higher kurtosis value.

Jarque - Bera

Hypothesis

Null Hypothesis (H_0) : The data z follows a normal distribution.

Alternative Hypothesis (H_a) : The data z does not follow a normal distribution.

```
# Jarque-Bera test
jb_test <- jarque.test(data$z)

jb_test
```

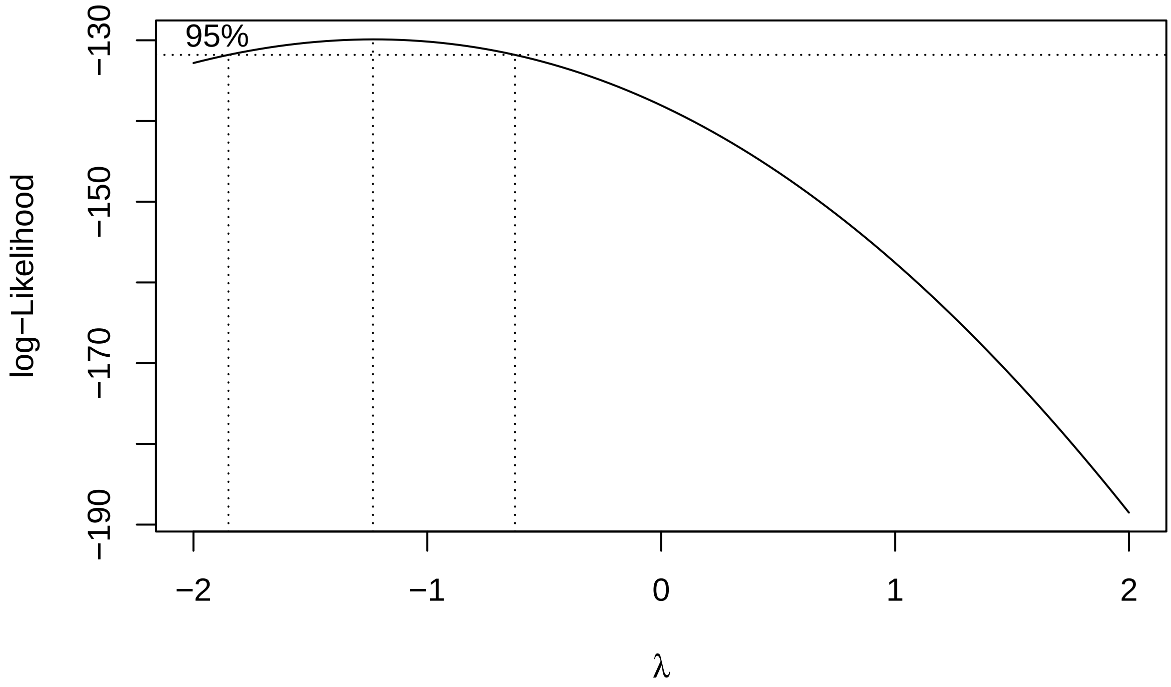
```
##
##  Jarque-Bera Normality Test
##
## data:  data$z
## JB = 130.06, p-value < 2.2e-16
## alternative hypothesis: greater
```

Interpretation: Since the p-value is much smaller than the typical significance level of 0.05, we can reject the null hypothesis. The small p-value suggests that the data significantly deviates from a normal distribution. In summary, based on the Jarque-Bera test, you have strong evidence to conclude that the variable z is not normally distributed.

```

library(MASS)


bc = boxcox(data$z ~ 1, lamda = seq(-3,3, 0.0001))
```

Answer 2.c

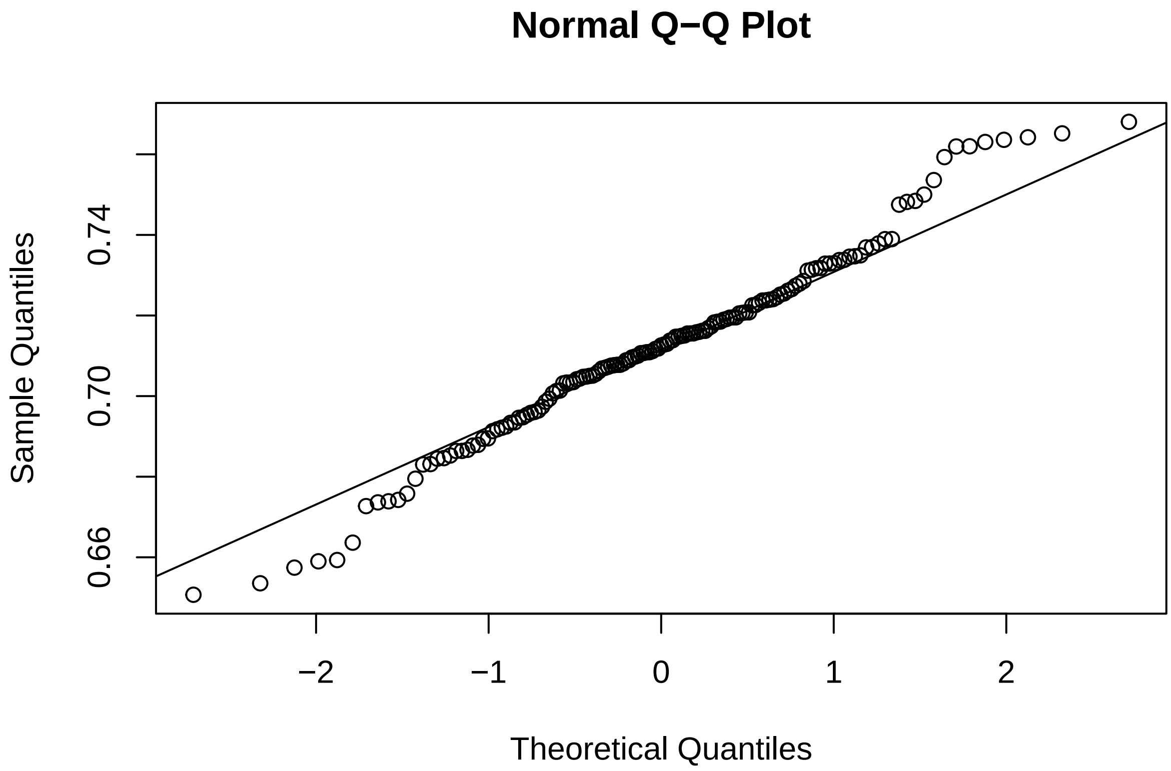
```
## Warning: In lm.fit(x, y, offset = offset, singular.ok = singular.ok, ...) :
##  extra argument 'lamda' will be disregarded
```



```
best.lam  = bc$x[which.max(bc$y)]

best.lam
```

```
## [1] -1.232323
```

```
#tz <- ifelse(optimal_lambda == 0, log(data$z), (data$z^optimal_lambda - 1) / optimal_lambda)

tz <- if (best.lam == 0){
  log(data$z)
} else {
  (data$z^best.lam - 1) / best.lam
}


qqnorm(tz)
qqline(tz)
```



(i) The QQ plot for tz indicates non-normality in its distribution. Data points in both tails deviate from the reference line, suggesting heavier tails and more pronounced central peak (leptokurtosis) compared to a normal distribution.

(ii)

```
skewness_tz <- skewness(tz)
kurtosis_tz <- kurtosis(tz)

skewness_tz
```

```
## [1] -0.02501491
```

```
kurtosis_tz
```

```
## [1] 3.358526
```

The data has a minor leftward skew (skewness = -0.03), and it shows moderate peakedness or heavy tails (kurtosis = 3.36) compared to a normal distribution, as observed in the QQ plot.

(iii)

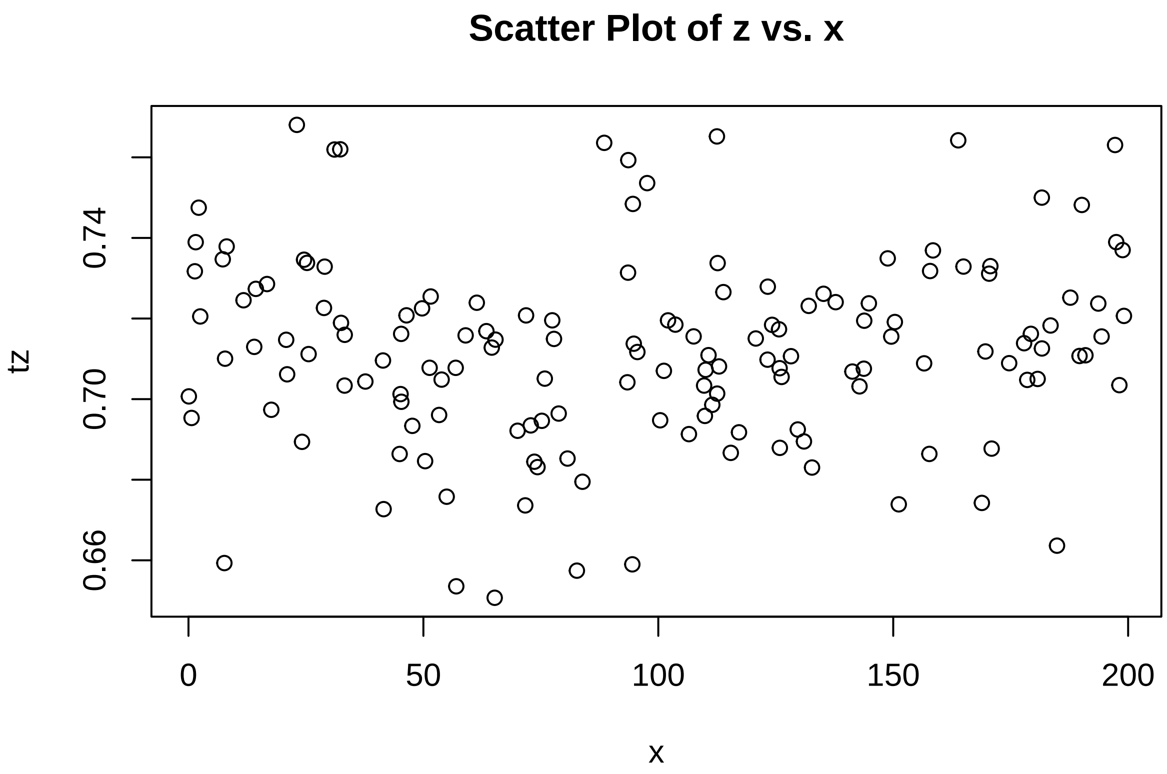
```
jb_test <- jarque.test(tz)

jb_test
```

```
##
##  Jarque-Bera Normality Test
##
## data:  tz
## JB = 0.81356, p-value = 0.6658
## alternative hypothesis: greater
```

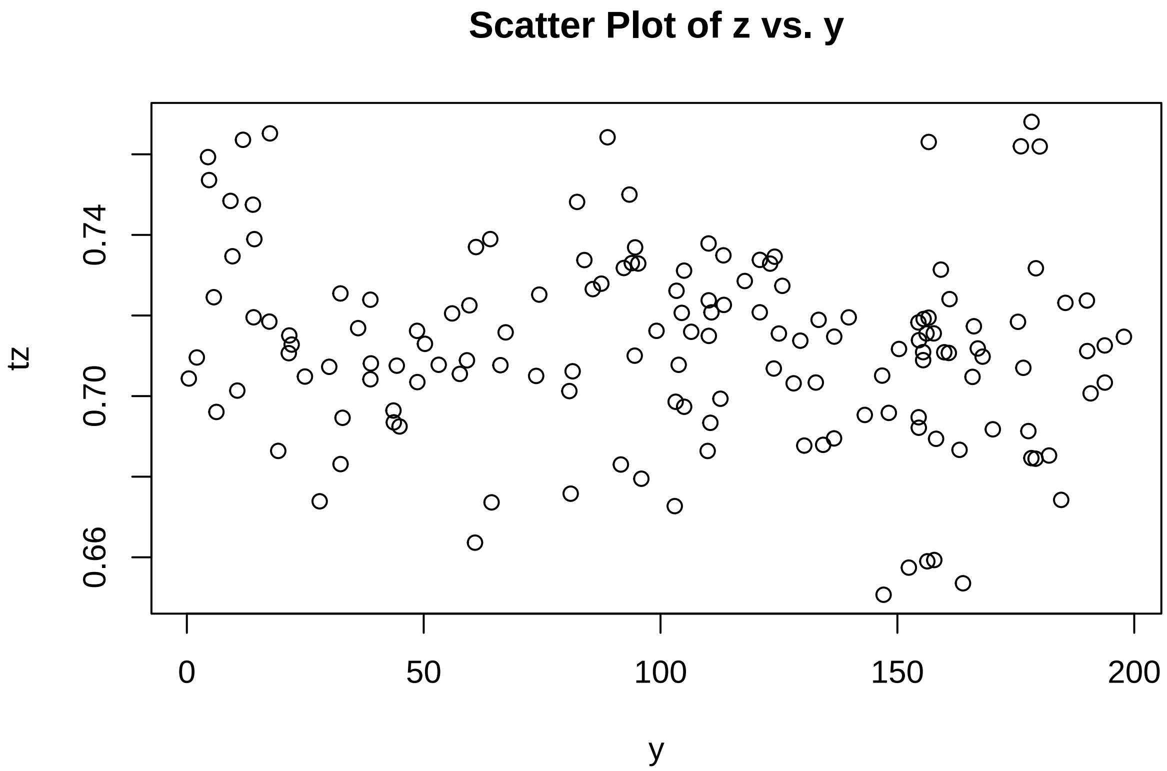
The Jarque-Bera test for normality on the transformed variable 'tz' yields a p-value of 0.6658. Since this p-value is greater than the significance level $\alpha = 0.05$, we fail to reject the null hypothesis. Thus, there is insufficient evidence to conclude that 'tz' significantly departs from a normal distribution.

(iv)

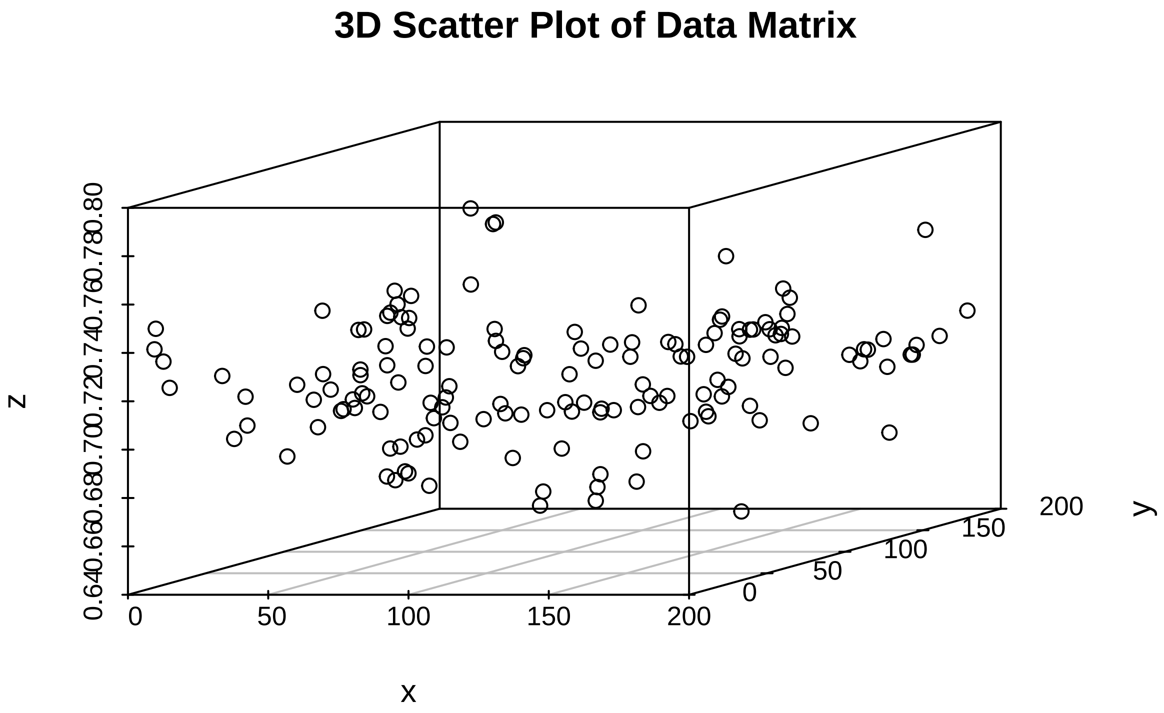
```
# Create a 2D scatter plot of z vs. x and tz vs. y
plot(data$x, tz, xlab = "x", ylab = "tz", main = "Scatter Plot of z vs. x")
```



```
plot(data$y, tz, xlab = "y", ylab = "tz", main = "Scatter Plot of z vs. y")
```



```
# Load the scatterplot3d library
library(scatterplot3d)

# Create a 3D scatter plot of the data matrix
scatterplot3d(data$x, data$y, tz,
              xlab = "x", ylab = "y", zlab = "z",
              main = "3D Scatter Plot of Data Matrix")
```



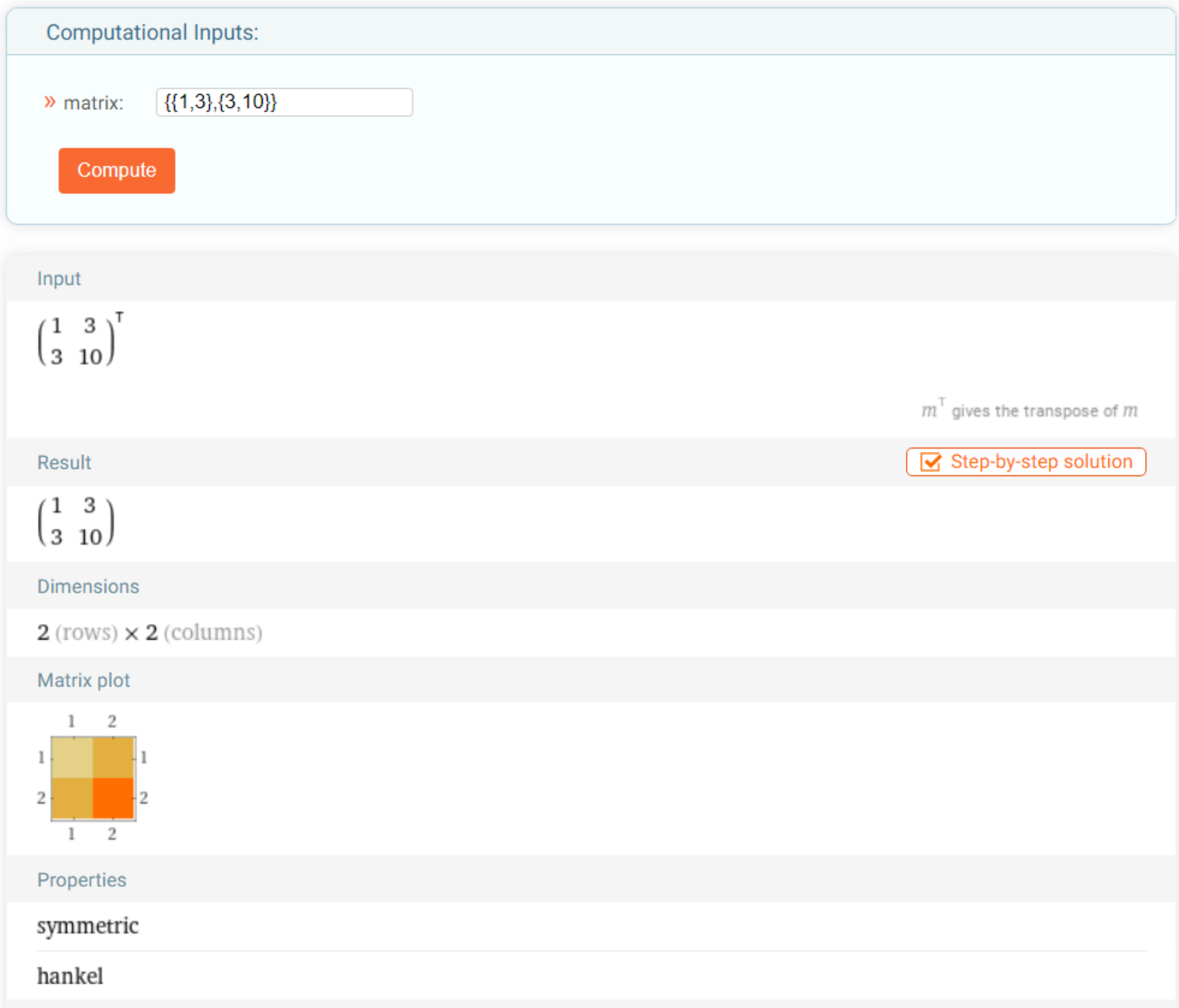
The 2D scatter plots of ‘tz’ against ‘x’ and ‘tz’ against ‘y’ continue to show a lack of noticeable patterns, indicating that there is no clear linear relationship or correlation between these variables. This reinforces the notion that ‘tz’, ‘x’, and ‘y’ may not have a straightforward joint distribution that can be explained solely by linear relationships. The data points appear to be randomly scattered, suggesting that other factors or complex interactions might be influencing their distribution.

Question 3

Answer 3.a There are two conditions that Σ needs to pass in order to be a positive definite matrix

1. Symmetry: A matrix must be symmetric to be considered positive definite. That means Σ must be equal to its transpose, Σ^T.

2. Both Eigenvalues should be positive



Computational Inputs:

» matrix: {{1,3},{3,10}}

Input

$$\begin{pmatrix}1 & 3\\ 3 & 10\end{pmatrix}^{T}$$

m^{T} gives the transpose of m

Result

$$\begin{pmatrix}1 & 3\\ 3 & 10\end{pmatrix}$$

Dimensions

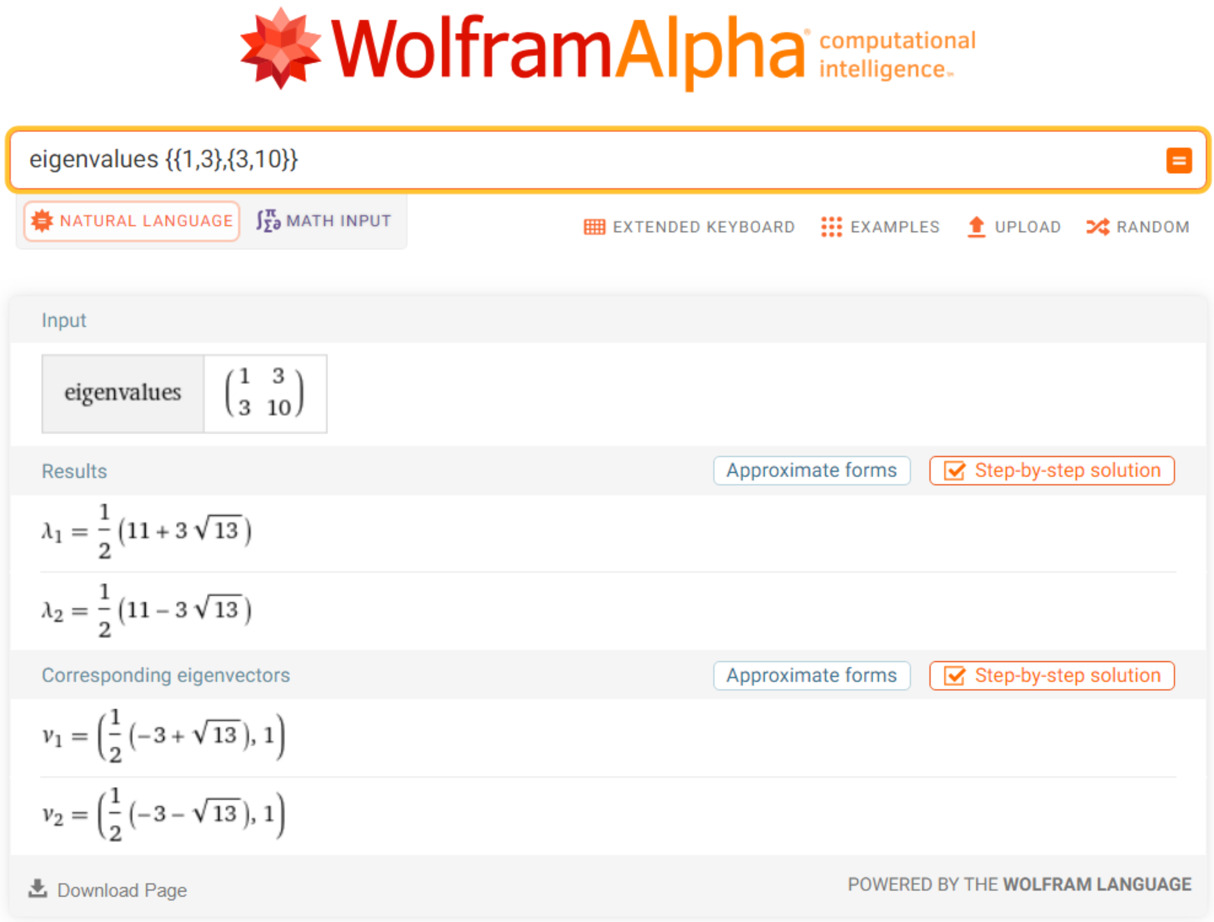
2 (rows) × 2 (columns)

Matrix plot

Properties

symmetric

hankel



eigenvalues {{1,3},{3,10}}

Input

eigenvalues $\begin{pmatrix} 1 & 3 \\ 3 & 10 \end{pmatrix}$

Results

$$\lambda_1 = \frac{1}{2}\left(11 + 3\sqrt{13}\right)$$

$$\lambda_2 = \frac{1}{2}\left(11 - 3\sqrt{13}\right)$$

Corresponding eigenvectors

$$v_1 = \left(\frac{1}{2}\left(-3 + \sqrt{13}\right), 1\right)$$

$$v_2 = \left(\frac{1}{2}\left(-3 - \sqrt{13}\right), 1\right)$$

Both eigenvalues are positive, which satisfies the requirement for positive definiteness. Therefore, Σ is a positive definite matrix.

```
covariance_matrix <- c(1,3,3,10)

dim(covariance_matrix) <- c(2,2)
covariance_matrix
```

Answer 3.b

```
##      [,1] [,2]
## [1,]    1    3
## [2,]    3   10
```

```
round(sqrtm(covariance_matrix))
```

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    3
```

Above is the calculation of square root of Σ^{-1}