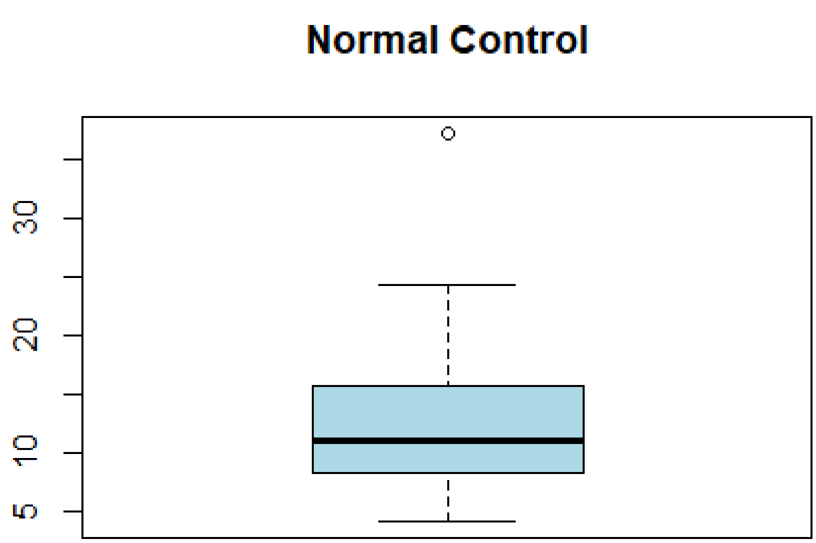
Problem Set 9

2023-11-02

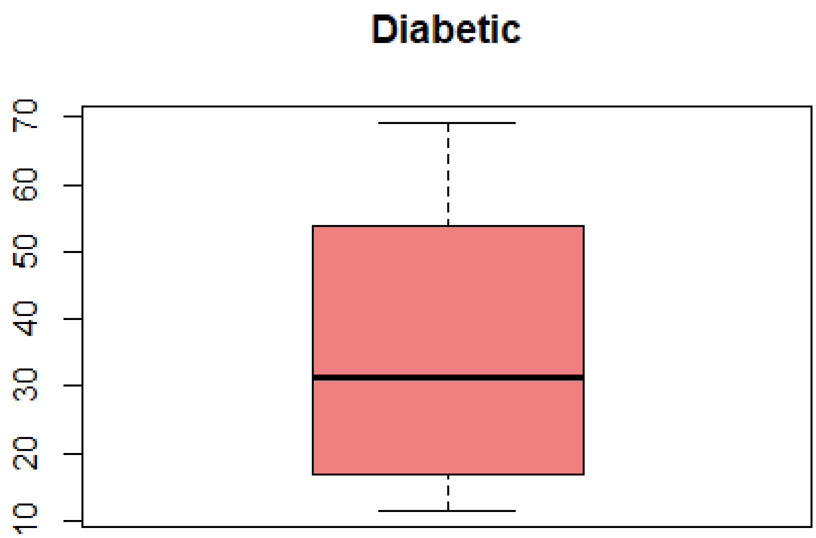
Question - 1

(a) To determine if the measurements appear to be samples from symmetric distributions.

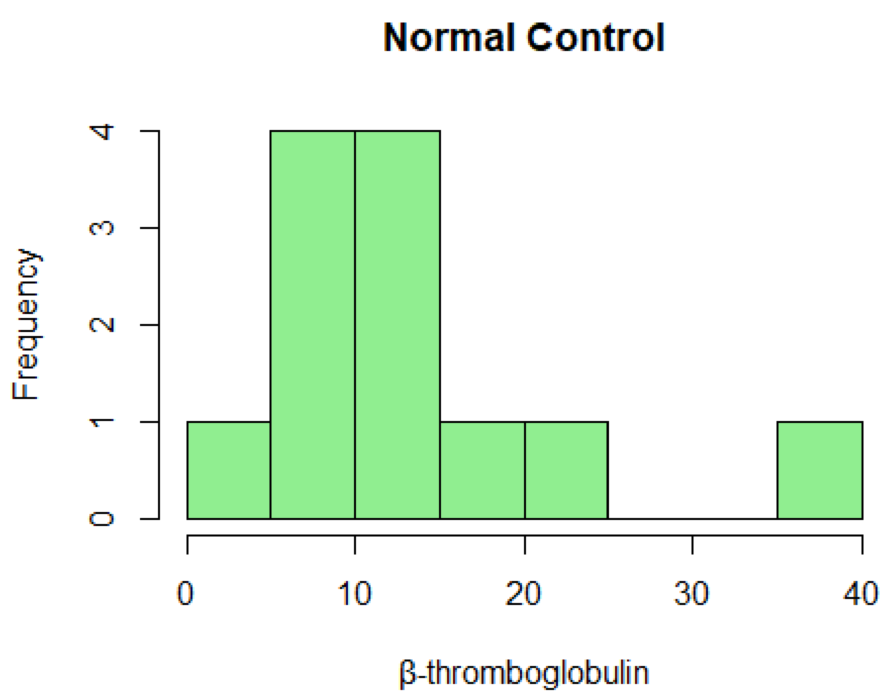
```
normal <- c(4.1, 6.3, 7.8, 8.5, 8.9, 10.4, 11.5, 12.0, 13.8, 17.6, 24.3,
37.2)
diabetic <- c(11.5, 12.1, 16.1, 17.8, 24.0, 28.8, 33.9, 40.7, 51.3, 56.2,
61.7, 69.2)

summary(normal)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   4.100   8.325  10.950  13.533  14.750  37.200

summary(diabetic)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   11.50   17.38   31.35   35.27   52.52   69.20

boxplot(normal, main = "Normal Control", col = "lightblue")
```



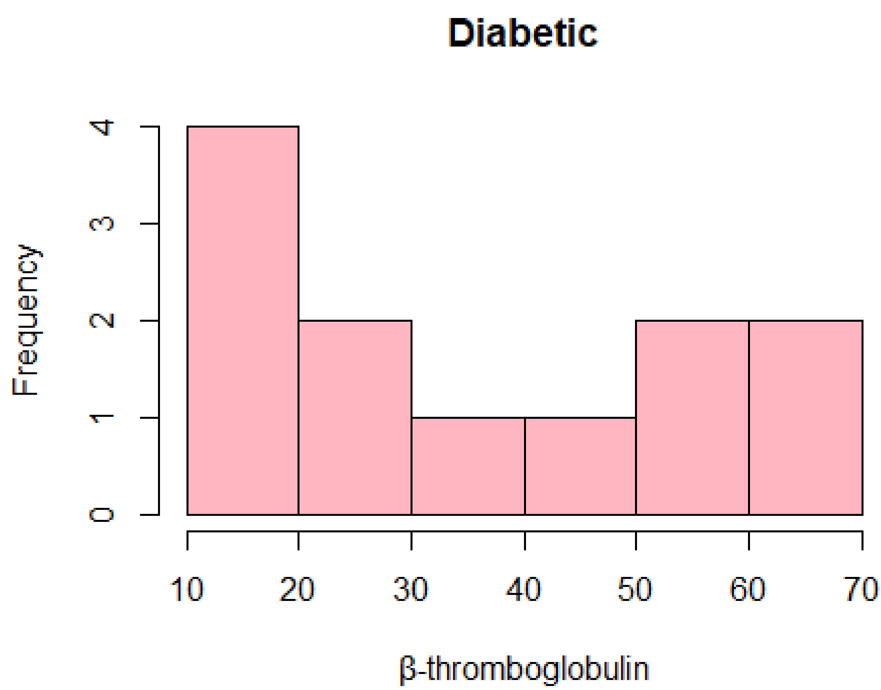

```
boxplot(diabetic, main = "Diabetic", col = "lightcoral")
```



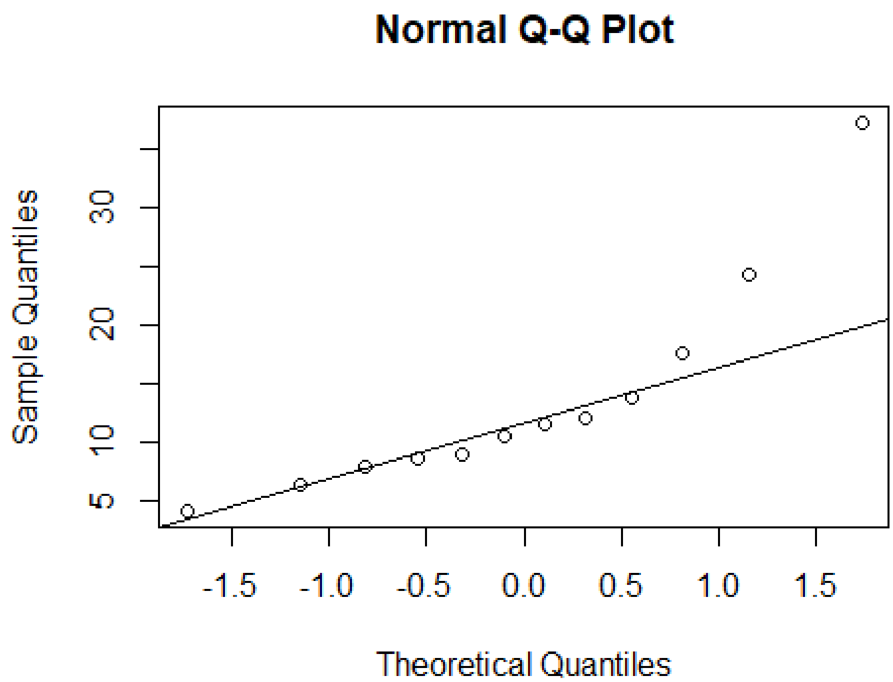

```
hist(normal, main="Normal Control", xlab="β-thromboglobulin",
col="lightgreen")
```



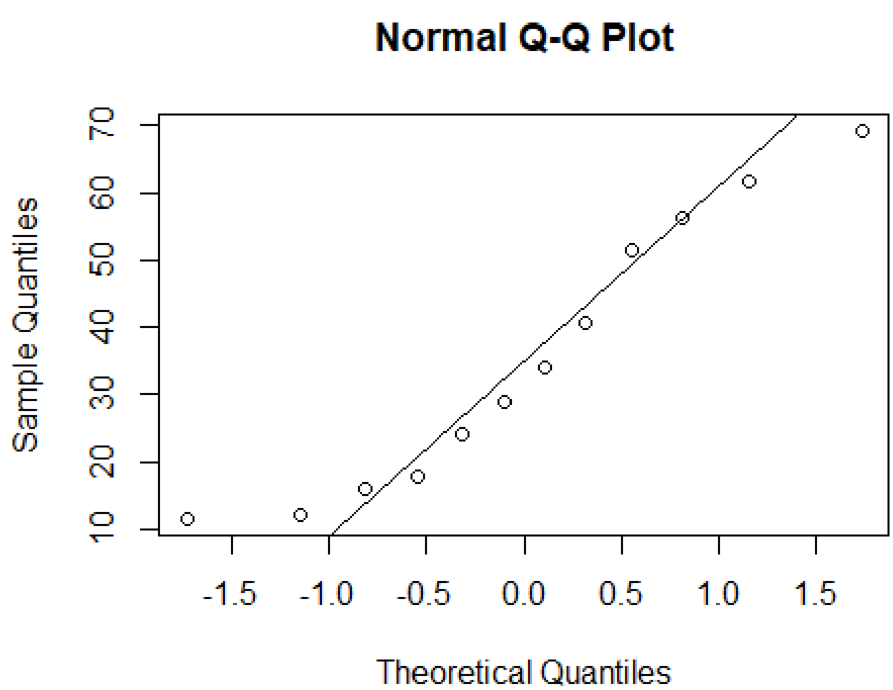
```r
hist(diabetic, main="Diabetic", xlab="β-thromboglobulin", col="lightpink")
```




```r
qqnorm(normal)
qqline(normal)
```



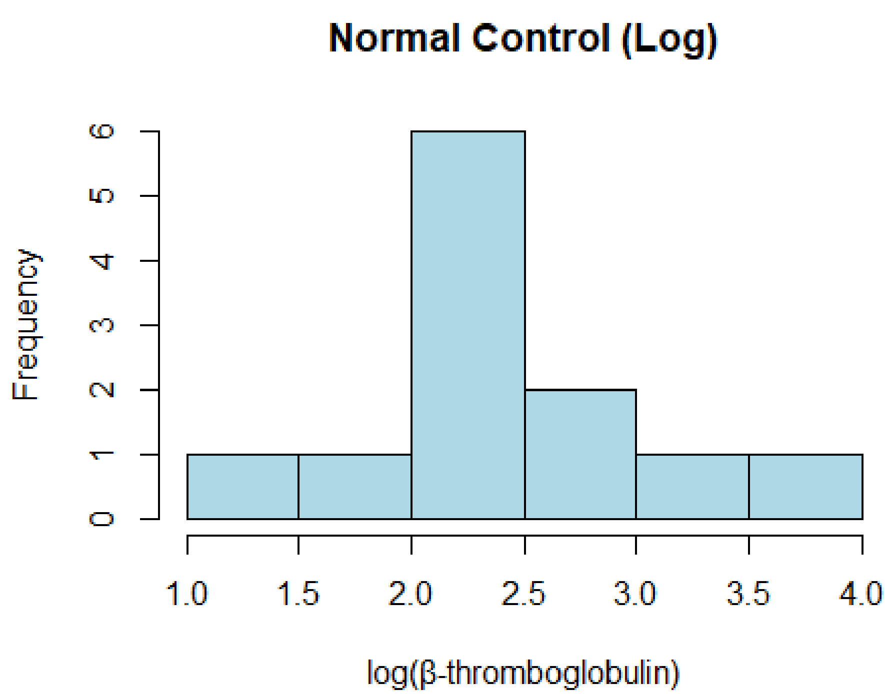

```r
qqnorm(diabetic)
qqline(diabetic)
```



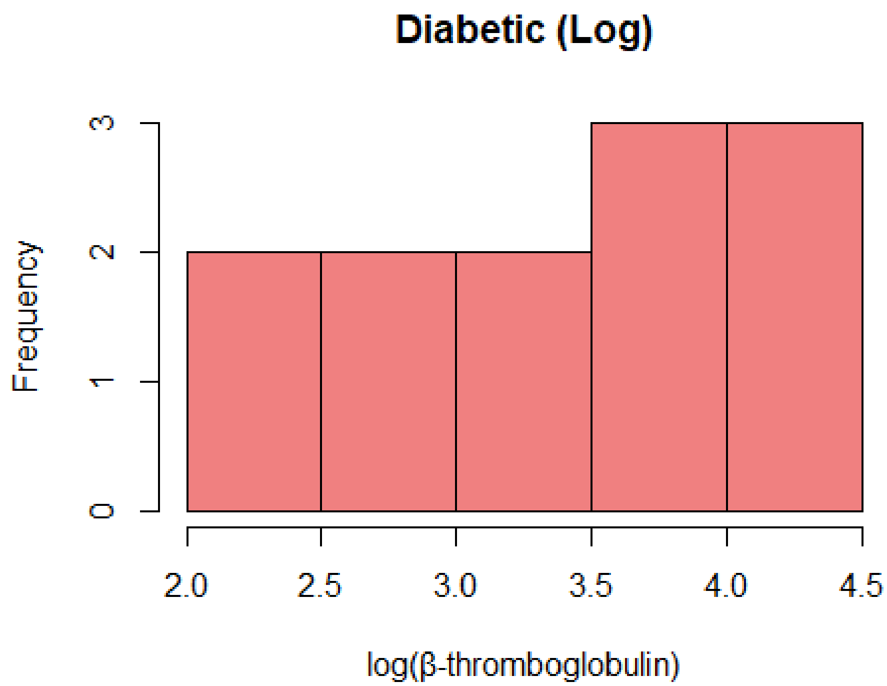
In conclusion, based on the graphs shown above and summary statistics, it appears that the measurements of urinary β-thromboglobulin excretion in both the diabetic and normal control groups are not from symmetric distributions. Both groups exhibit right-skewed distributions, with a majority of values concentrated on the left side and a few high values on the right side and the upper tail of the distribution is longer than the lower tail of the distribution.

(b) Data Transformations on the above measurements

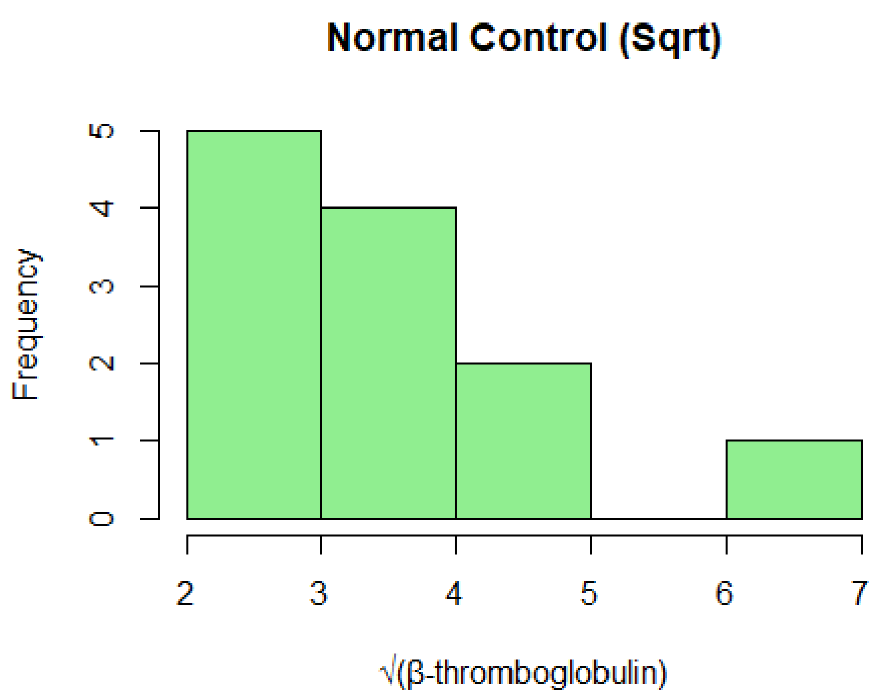
```
#Natural logarithm transformation
normal_log <- log(normal)
diabetic_log <- log(diabetic)

#Square root transformation
normal_sqrt <- sqrt(normal)
diabetic_sqrt <- sqrt(diabetic)

hist(normal_log, main="Normal Control (Log)", xlab="log(β-thromboglobulin)",
col="lightblue")
```



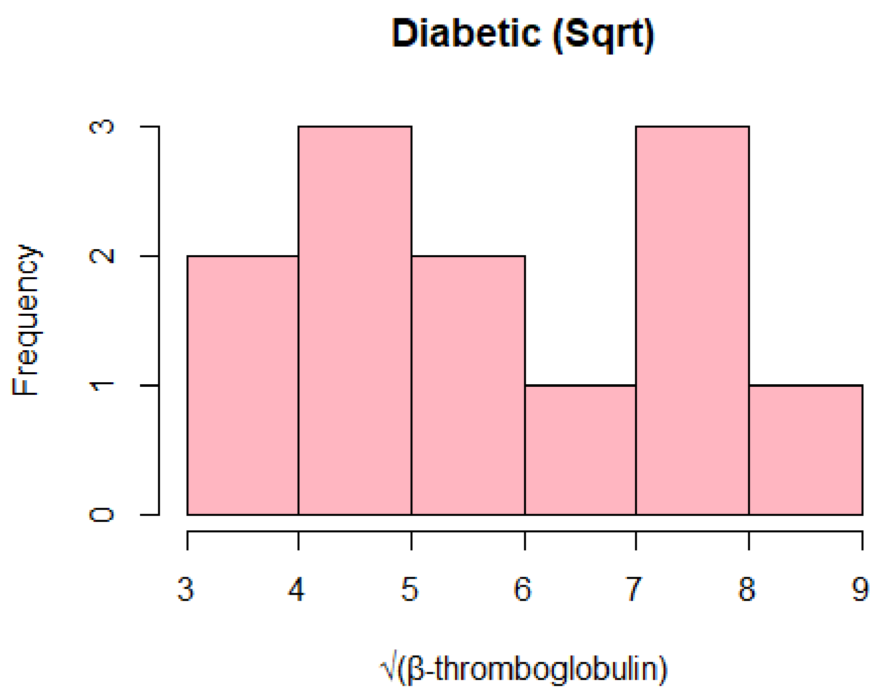

```
hist(diabetic_log, main="Diabetic (Log)", xlab="log(β-thromboglobulin)",
col="lightcoral")
```



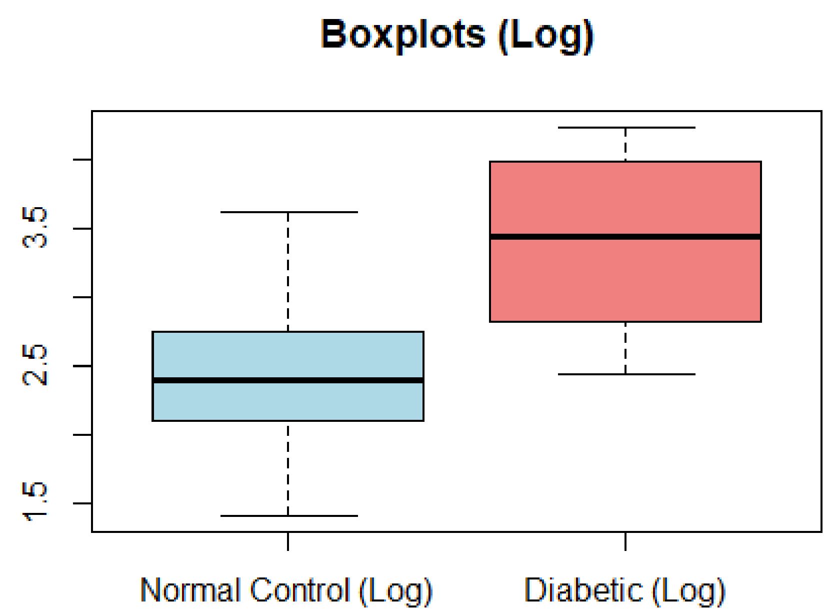

```
hist(normal_sqrt, main="Normal Control (Sqrt)", xlab="√(β-thromboglobulin)", col="lightgreen")
```



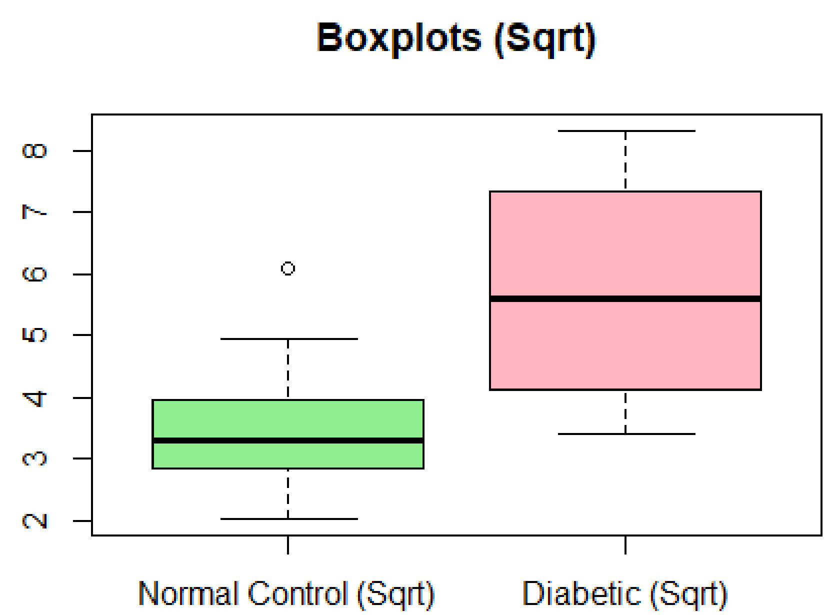
```r
hist(diabetic_sqrt, main="Diabetic (Sqrt)", xlab="√(β-thromboglobulin)",
col="lightpink")
```




```r
boxplot(normal_log, diabetic_log, names=c("Normal Control (Log)", "Diabetic
(Log)"), col=c("lightblue", "lightcoral"), main="Boxplots (Log)")
```




```r
boxplot(normal_sqrt, diabetic_sqrt, names=c("Normal Control (Sqrt)",
"Diabetic (Sqrt)"), col=c("lightgreen", "lightpink"), main="Boxplots (Sqrt)")
```

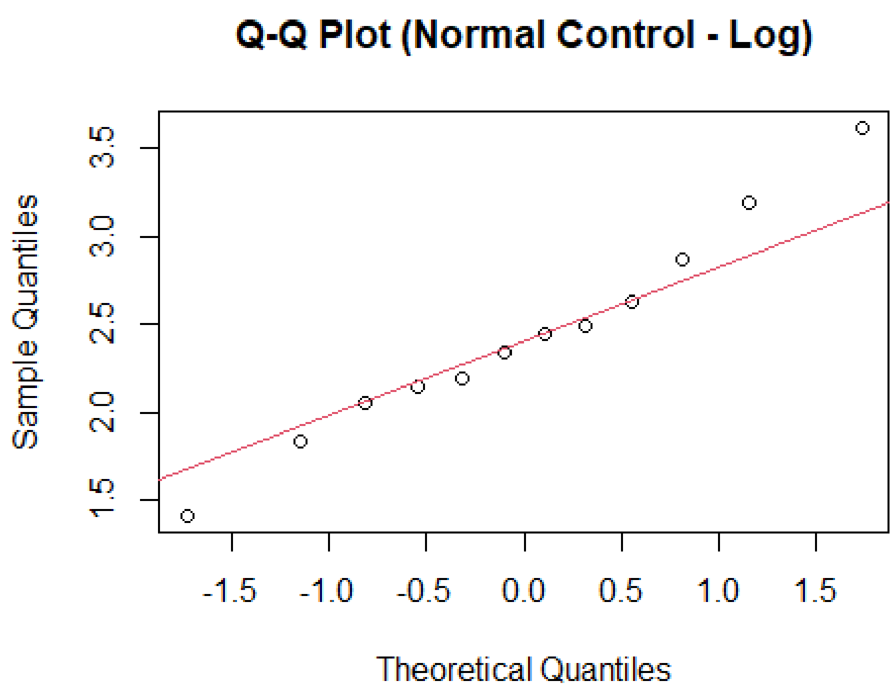


As evident from the above plots, after applying the natural logarithm transformation (log), the distributions appear more symmetric and closer to a normal distribution and after applying the square root transformation (sqrt), the distributions also become more symmetric but with some right skew still visible.

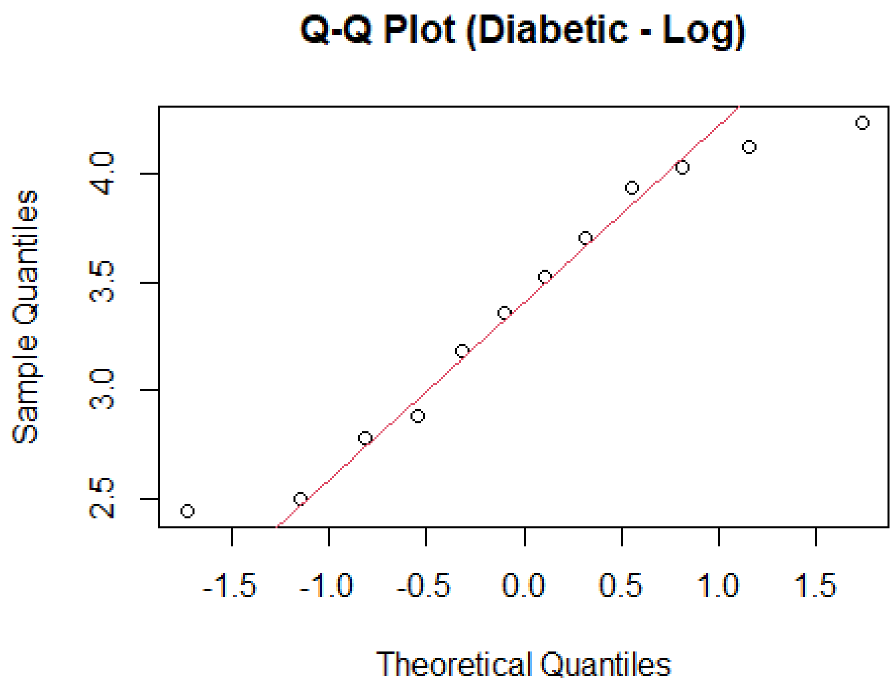
Therefore, I would prefer natural logarithm transformation (log)

(c)To determine whether the transformed measurements appear to be samples from normal distributions, we can plot QQ-plot for the transformations as shown below

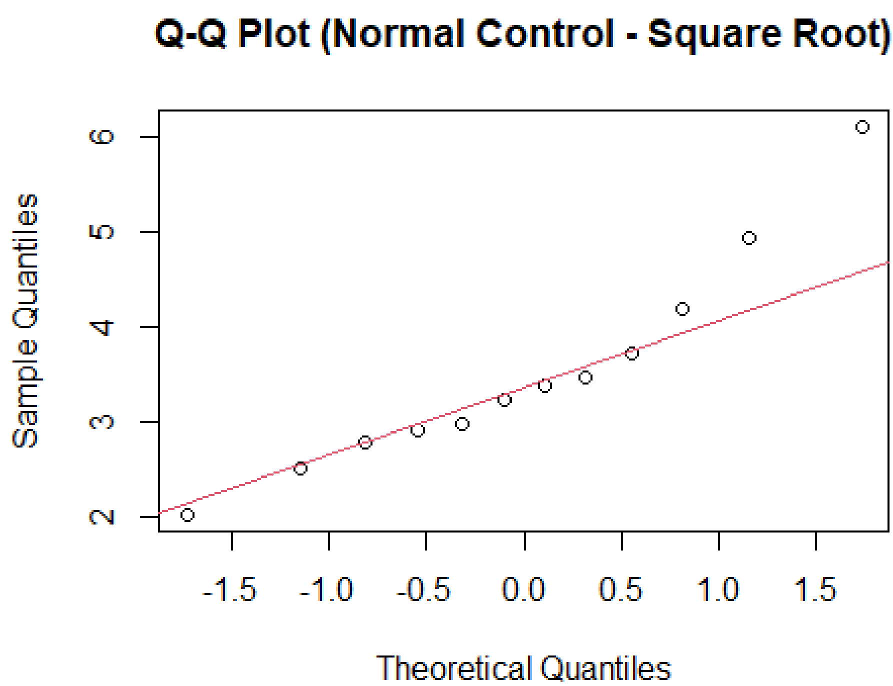
```
# Normal logarithm
qqnorm(normal_log, main="Q-Q Plot (Normal Control - Log)")
qqline(normal_log, col=2)
```



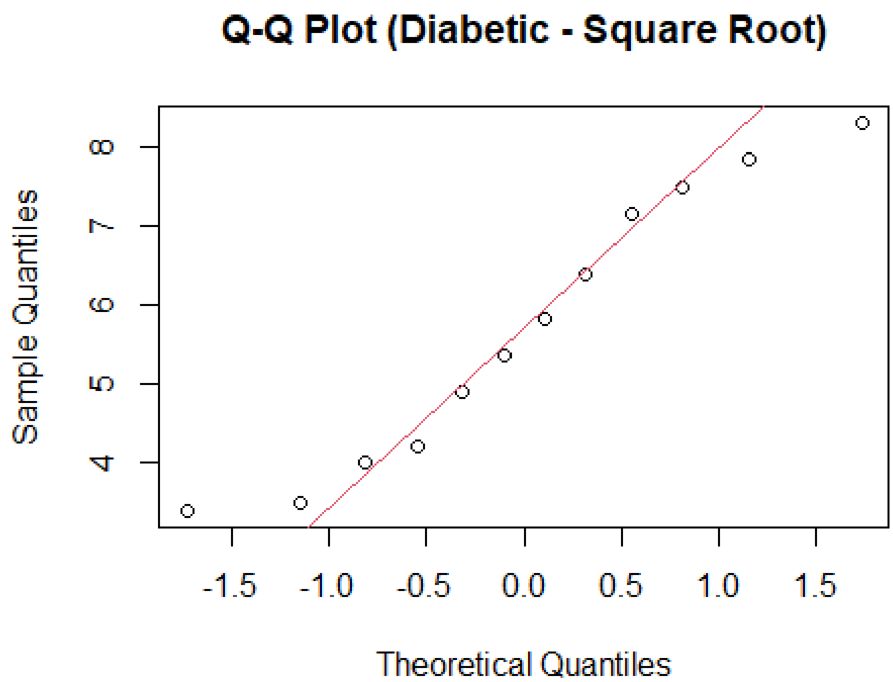

```
# Diabetic logarithm
qqnorm(diabetic_log, main="Q-Q Plot (Diabetic - Log)")
qqline(diabetic_log, col=2)
```




```r
# Normal square root
qqnorm(normal_sqrt, main="Q-Q Plot (Normal Control - Square Root)")
qqline(normal_sqrt, col=2)
```



```r
# Diabetic square root
qqnorm(diabetic_sqrt, main="Q-Q Plot (Diabetic - Square Root)")
qqline(diabetic_sqrt, col=2)
```



For the normal control group, the Q-Q plot after the logarithm transformation shows a fairly straight line, indicating that the data approximates a normal distribution. For the diabetic group, the Q-Q plot after the logarithm transformation is less linear but still shows a moderate resemblance to a normal distribution.

The Q-Q plots for both groups after the square root transformation show slight deviations from a normal distribution, but they still demonstrate a relatively good fit.

In conclusion, both transformations (logarithm and square root) tend to improve the normality of the data. The logarithm transformation appears to be slightly better in making the data approximate a normal distribution, especially for the normal control group

(d) We can perform Welch Two sampled test to compare the means of the transformed data between the diabetic and normal control groups which helps us to find the evidence for the researchers claim that diabetic patients have increased urinary β-thromboglobulin excretion.

we assume that the transformed measurements are samples from normal distributions

```r
t_test_log <- t.test(diabetic_log, normal_log, alternative = "greater")
t_test_log

##
##  Welch Two Sample t-test
```

```
##
## data:  diabetic_log and normal_log
## t = 3.8041, df = 21.9, p-value = 0.0004888
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  0.5250797       Inf
## sample estimates:
## mean of x mean of y
##  3.390628  2.433349
```

The p-value using welch two sampled test is 0.0004888 which is very small and therefore supports the researchers claim.

Question - 2

The heights of men and women are approximately normally distributed.A random sample of seven men has average height 68.5 inches with standard deviation 3.0 inches, while a random sample of seven women has average height 65.5 inches with standard deviation 2.5 inches

```
n_men <- 7
n_women <- 7

mu_men <- 68.5
mu_women <- 65.5

sd_men <- 3.0
sd_women <- 2.5
```

(a) The reason for using the t-distribution and not the standard normal distribution to calculate P-values and confidence intervals in this case is because we are dealing with small sample sizes and the heights are randomly sampled from two populations: Men and Women. The t-distribution allows for a more accurate representation of the population's standard error and accounts for the additional variability introduced by small samples

(b) According to Welch's approximation, the number of degrees of freedom is 11.6. Using the fact that the R code qt(.975,df=11.62) gives the value 2.187, find an approximate 95% confidence interval for the average difference in heights between men and women at the university.

```
# average difference in heights between men and women at the university
delta.hat <- mu_men - mu_women
q <- 2.187

# approximate 95% confidence interval
std.error <- sqrt(sd_men^2 / n_men + sd_women^2 / n_women)

delta.hat - q * std.error

## [1] -0.2280086

delta.hat + q * std.error
```

```
## [1] 6.228009
```

(c) The fact that the confidence interval in part (b) contains the value 0 suggests that there could be no significant difference in average heights between men and women at the university. A confidence interval provides a range of plausible values for the population parameter (in this case, the average height difference) based on the sample data. Since the interval includes values both below and above 0, it suggests that there is uncertainty about the true population difference. The data do not provide enough evidence to conclude that the average heights are definitively equal or not equal. Instead, it indicates that the population difference could be anywhere within the confidence interval.

Question - 3

(From the Fall 2016 final.) In a randomized experiment in Georgia, a treatment group of 592 convicts received cash payments upon being released from prison, while a control group of 154 convicts received no money upon release. In the first year after release, the members of the treatment group averaged 16.8 weeks of paid work, with a standard deviation of 15.9 weeks. The members of the control group averaged 24.3 weeks of paid work, with a standard deviation of 17.3 weeks. The samples were large and right-skewed.

```
n_treat_group <- 592
n_control_group <- 154

mu_treat <- 16.8
mu_control <- 24.3

sd_treat <- 15.9
sd_control <- 17.3
```

(a)We may do Welch's t-test even though the samples are right-skewed because the samples are described as large, which means that the Central Limit Theorem starts to apply, and the sampling distribution of the means becomes approximately normal even if the underlying data is skewed. The researcher prefers not to transform the data, possibly because the interpretation of the results would be easier with the original data.

In addition, Welch's t-test is robust to deviations from normality, specifically adapted for scenarios where the variances between groups are not equal, especially when dealing with large sample sizes. It can provide valid results even with non-normally distributed data.

(b)

Welch's two-sample test:

```
# average difference in heights between men and women at the university
delta_hat <- mu_control - mu_treat
```

Find the standard error of Delta hat:

```r
# standard error
std_error <- sqrt(sd_treat^2 / n_treat_group + sd_control^2 /
n_control_group)
```

Find Welch's t-statistic:

```r
t_welch <- delta_hat / std_error
t_welch

## [1] 4.871275
```

Find the DF:

```r
DF <- (sd_treat^2/n_treat_group + sd_control^2 / n_control_group)^2 /
  ((sd_treat^2/n_treat_group)^2/(n_treat_group - 1) + (sd_control^2 /
n_control_group)^2/(n_control_group - 1))
DF

## [1] 224.8164
```

Find the P-value:

```r
2 * (1 - pt(abs(t_welch), df = DF))

## [1] 2.091618e-06
```

Since the P-value is very small, we reject the null hypothesis that the treatment and control will, on average, result in the same number of weeks worked.

(c) To compute 95% confidence interval for the average difference in weeks worked between the treatment and control

```r
q <- qt(.975, df = DF)

delta_hat - qt(.975, df = DF) * std_error

## [1] 4.466032

delta_hat + qt(.975, df = DF) * std_error

## [1] 10.53397
```

We are 95% confident that the difference between average weeks worked by treatment group and average weeks worked by control group is 4.4 to 10.5 weeks.

Question - 4

(a) Given: Group A has an average score of 61 with a sample standard deviation of 10. Group B has an average score of 59 with a sample standard deviation of 13. Both sets of scores are approximately normal

To determine if there is a significant difference between the averages of the two groups (Group A and Group B), you can perform a two-sample t-test. The null hypothesis(H0) is that there is no significant difference, while the alternative hypothesis (H1) is that there is a significant difference between the groups.

```
n_group_A <- 100
n_group_B <- 100

mu_group_A <- 61
mu_group_B <- 59

sd_group_A <- 10
sd_group_B <- 13
```

Welch's two-sample test:

```
# average difference in heights between men and women at the university
delta.group <- mu_group_A - mu_group_B
```

Find the standard error of Delta hat:

```
# standard error
std_error.group <- sqrt(sd_group_A^2 / n_group_A + sd_group_B^2 / n_group_B)
```

Find Welch's t-statistic:

```
t_welch.group <- delta.group / std_error.group
t_welch.group

## [1] 1.219422
```

Find the DF:

```
DF_group <- (sd_group_A^2/n_group_A + sd_group_B^2 / n_group_B)^2 /
  ((sd_group_A^2/n_group_A)^2/(n_group_A - 1) + (sd_group_B^2 /
n_group_B)^2/(n_group_B - 1))
DF_group

## [1] 185.7768
```

Find the P-value:

```
2 * (1 - pt(abs(t_welch.group), df = DF_group))
```

```
## [1] 0.2242302
```

(b) To compute 90% confidence interval for the difference in averages between the two groups

```
# 90% confidence interval = 0.95
q.group <- qt(.95, df = DF_group)

delta.group - qt(.95, df = DF_group) * std_error.group

## [1] -0.7112808

delta.group + qt(.95, df = DF_group) * std_error.group

## [1] 4.711281
```

(c) The P-value as calculated above is 0.2242302 which is greater than the standard significance level 5%(0.05) and therefore is still compatible with null hypothesis and fails to prove the alternative hypothesis that there is a significant difference between the groups

The 90% confidence interval calculated in part (b) provides a range of values for the difference in average scores between two groups (Group A and Group B) which is -0.7 and 4.7.

Since this interval includes zero, it suggests that there may not be a significant difference.Since zero is included within this range, we cannot rule out the possibility of no difference, but at the same time, we cannot confirm it either; there is simply not enough evidence to definitively state there is no difference based on this interval alone. Further investigation with a larger sample size could provide a more definitive conclusion.

Question - 5

In this case, we would use a two-sample t-test for the difference in means. Specifically, we are interested in whether the aerobic exercise had a significant effect on VO2 levels, so we want to compare the changes in VO2 between the treatment group and the control group. The assumptions for this test include:

The data in each group (treatment and control) are approximately normally distributed.

The samples are independent.

To test the hypothesis that the VO2 change is positive under the assumption that the sample is random, Let's consider the following hypothesis

Step 1: Hypotheses

1) Null Hypothesis (H0): The population mean VO2 change is not positive or is less than or equal to 0.

H0:delta <= 0

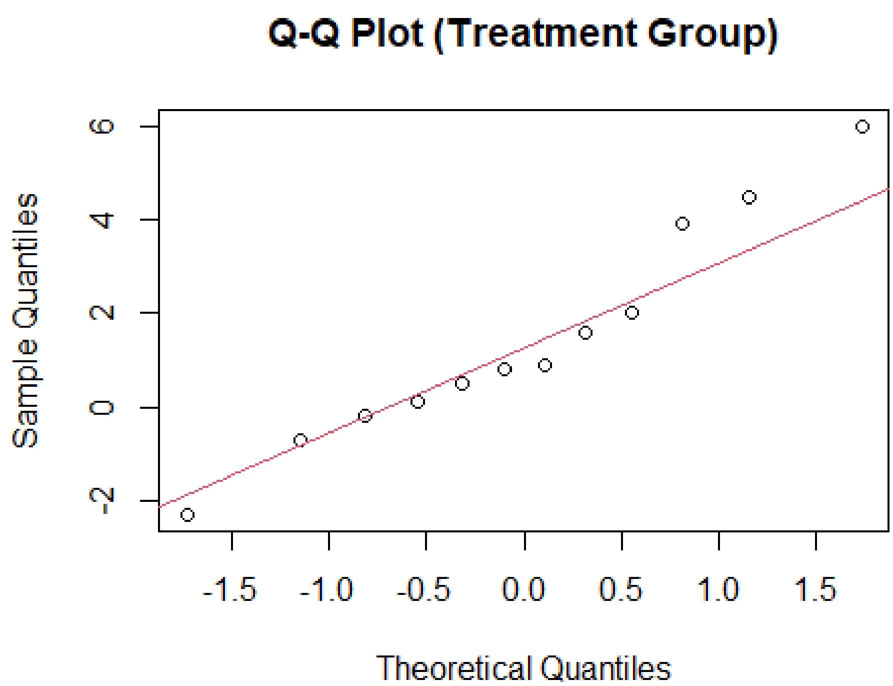
2) Alternative Hypothesis (H1): The population mean VO2 change is positive.

H1:delta > 0

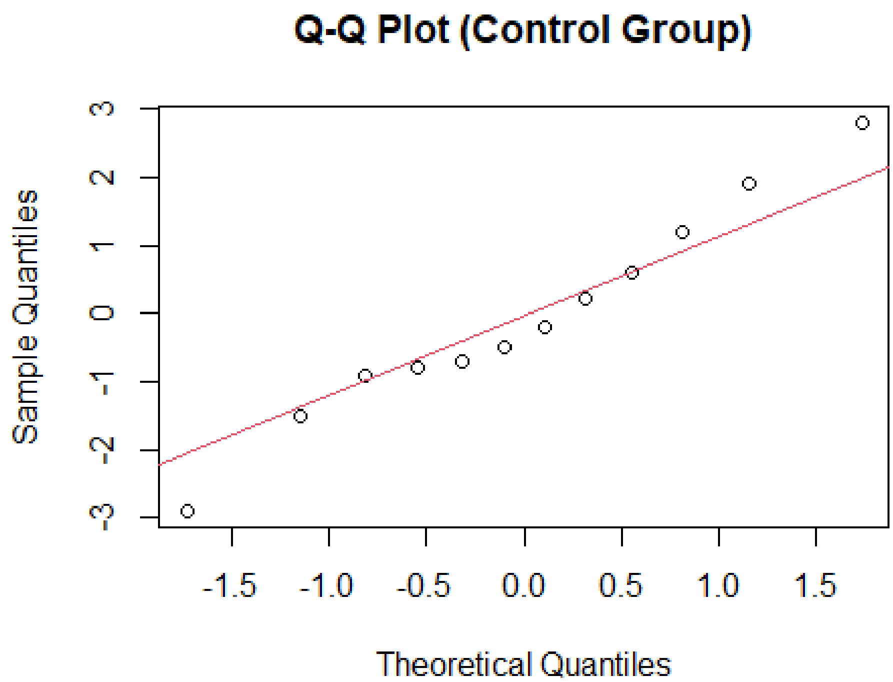
where delta = µ1 - µ2, µ1: population mean change in VO2 for the treatment group and µ2: population mean change in VO2 for the control group

```
group_treatment <- c(-2.3,-0.7,-0.2, 0.1, 0.5, 0.8, 0.9, 1.6, 2.0, 3.9, 4.5,
6.0)
group_control <- c(-2.9,-1.5,-0.9,-0.8,-0.7,-0.5,-0.2, 0.2, 0.6, 1.2, 1.9,
2.8)


# Normal square root
```

```r
qqnorm(group_treatment, main="Q-Q Plot (Treatment Group)")
qqline(group_treatment, col=2)
```




```r
# Diabetic square root
qqnorm(group_control, main="Q-Q Plot (Control Group)")
qqline(group_control, col=2)
```



The qqnorm plots indicate that the data does not exactly fit with normal distribution. But as the sample size is small we can take a leap of faith and assume that the data follows normal distributions.

(b) To find the P-value for a one-tailed test, we can perform a one-tailed two-sample t-test.

```
t_test <- t.test(group_treatment, group_control, alternative = "greater")
t_test$p.value

## [1] 0.04119593

#alternatively

delta <- mean(group_treatment) - mean(group_control)
Standard_error <- sqrt(sd(group_treatment)^2/12 + sd(group_control)^2/12)

t.stat <- delta/Standard_error
df <- ((var(group_treatment)/12 +
var(group_control)/12)^2)/((var(group_treatment)/12)^2/(12-1) +
(var(group_control)/12)^2/(12-1))

#one tailed P-value (Right tailed)

(1 - pt(abs(t.stat), df=df))

## [1] 0.04119593
```

(c) To find a 95% confidence interval for the difference in population means, we can use the two-sample t-test with a confidence level of 95%

```
# 95% confidence interval
delta - qt(.975, df = df)*Standard_error

## [1] -0.2108894

delta + qt(.975, df = df)*Standard_error

## [1] 3.194223
```

(d) The interval as shown above suggests that there is a range of possible values for the difference in VO2 changes between the treatment group and the control group which is -0.21 to 3.19. Based on the data we have, we are quite certain that the actual difference in VO2 changes falls within this range. The comparison between the average change in VO2 levels between the group that engaged in aerobic exercise and the one that didn't tells us that the true impact of the exercise could vary. This says that sometimes aerobic exercise may result in a slight reduction of VO2 by as much as 0.21 mL/kg/min or an increase of up to 3.2 mL/kg/min. As we have both positive and negative results by doing aerobic exercise, we cannot come to any conclusion from this study alone. There can be a possibility of having no impact at all.

Question - 6

(a) The experimental unit in this study is the individual with type 2 diabetes. The measurements taken on the experimental units are their glycemic index values. This study involves one independent sample (before and after consuming dates with and without coffee) from the same group of individuals, so it's a problem with one independent sample.

(b) The null and alternative hypotheses for an appropriate two-tailed t-test can be formulated as follows:

Null Hypothesis: The mean glycemic index for dates is the same with or without coffee for individuals with type 2 diabetes. In mathematical notation, H0: µ(with coffee) - µ1(without coffee) = 0

Alternative Hypothesis: The mean glycemic index for dates is different with coffee compared to without coffee for individuals with type 2 diabetes. In mathematical notation, H1: µ(with coffee) - µ1(without coffee) ≠ 0

Calculating t-statistic:

```
#Given: differences between the measurements (“without coffee” minus “with
coffee”) had mean 11.5 with standard deviation 21
xbar <- 11.5
s <- 21
n <- 10
c(xbar, s, n)
```

```
## [1] 11.5 21.0 10.0
t.stat <- (xbar - 0) / (s / sqrt(n))
t.stat
## [1] 1.731723
```

Calculating P-value for two-tailed test

```
# left-tail P-value
lt_p <- pt(t.stat, df = n - 1)

# right-tail P-value
rt_p <- 1-pt(t.stat, df = n - 1)

#Two tailed P-value
min(lt_p, rt_p)*2
## [1] 0.1173671
```

(c) The P-value of 0.12 is greater than the standard significance level of 0.05 (for a two-tailed test). This means that the null hypothesis is not rejected based on the observed data. However, this does not mean we can conclude that we are sure that, on average, dates have the same glycemic index with or without coffee.

A high P-value (greater than 0.05) suggests that there isn't strong evidence to conclude that the mean glycemic index with and without coffee is different, but it doesn't prove that they are the same. The result indicates that, based on the sample data, we do not have sufficient evidence to reject the null hypothesis.

In our case, more research with a larger sample size or different study designs may be needed to draw more definitive conclusions about the impact of coffee on the glycemic index of dates for individuals with type 2 diabetes.