Statistical data analysis, Assignment 6

Name: 徐竣霆
ID: 110060012

Problem 1

(a)

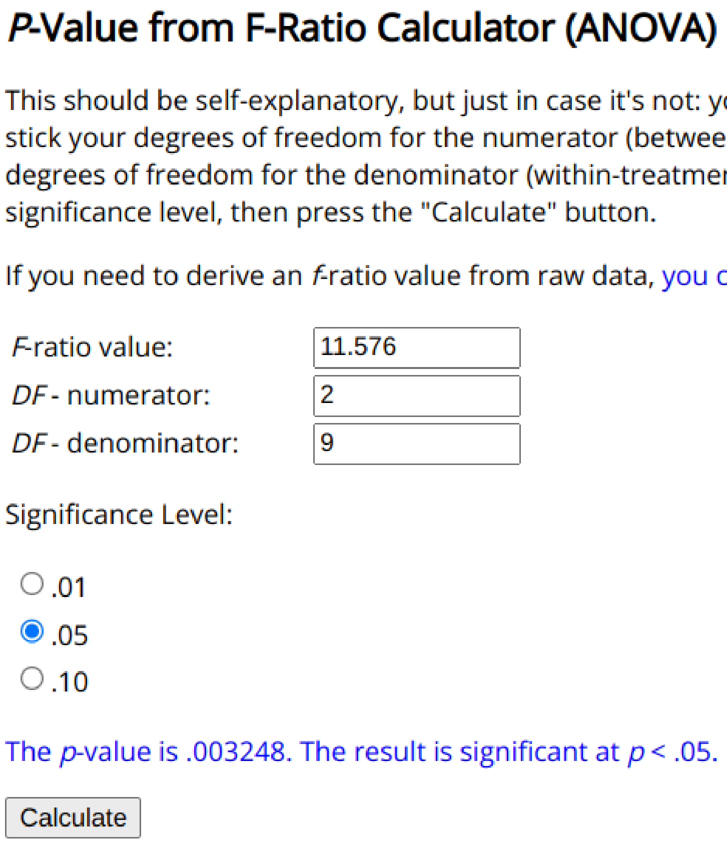
$\bar{Y} = 55.2$
$SSB = (4 \times (59.5 - 55.2)^2) + (4 \times (55.4 - 55.2)^2) + (4 \times (50.7 - 55.2)^2) = 155.12$

$SSE = (6.7 + 7.1 + 6.3) \times 3 = 60.3$
$SST = SSB + SSE = 155.12 + 60.3 = 215.42$

$MSB = 155.12/2 = 77.56$
$MSE = 60.3/9 = 6.7$

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
(Between) EC levels	2	155.12	77.56	11.58	0.003248
(Within Error) Residuals	9	60.3	6.7		
Total	11	215.42			

(b)



P-Value from F-Ratio Calculator (ANOVA)

This should be self-explanatory, but just in case it's not: your *F*-ratio value goes in the *F*-ratio value box, you stick your degrees of freedom for the numerator (between-treatments) in the *DF* - numerator box, your degrees of freedom for the denominator (within-treatments) in the *DF* - denominator box, select your significance level, then press the "Calculate" button.

If you need to derive an *F*-ratio value from raw data, you can find an ANOVA calculator here.

F-ratio value: 11.576

DF - numerator: 2

DF - denominator: 9

Significance Level:

○ .01
● .05
○ .10

The *p*-value is .003248. The result is significant at *p* < .05.

reject H_0, the null hypothesis that here are no effects due to the EC of the soil.

Problem 2

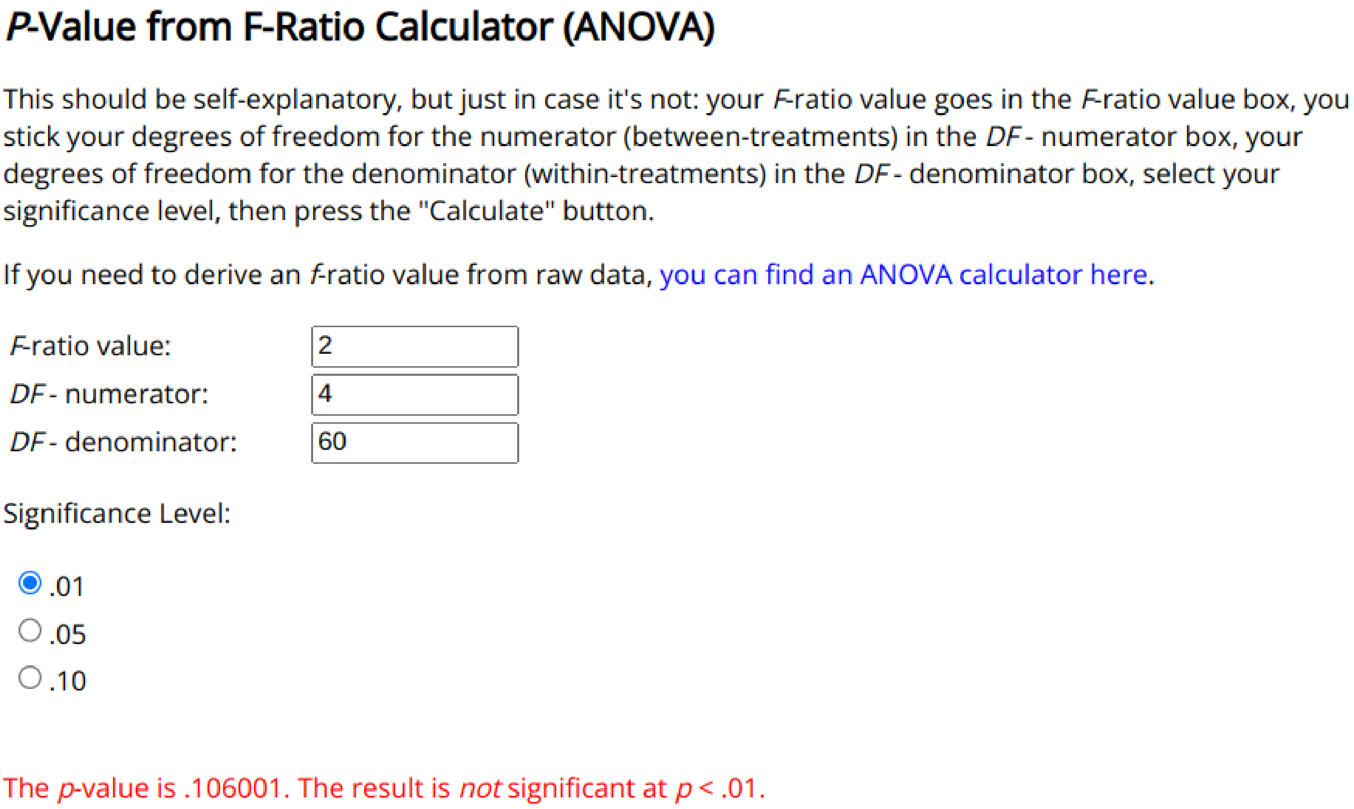
(a)

(a) = 4(five groups)
(b) = 10.0(40 / 4)
(c) = 2.00(10 / 5)
(d) = 300(60 * 5)

(b)

Number of all observations = Df_between + DF_within + 1 = $4 + 60 + 1 = 65$
Value of total sum of square = SSB + SSE = $40 + 300 = 340$

(c)



P-Value from F-Ratio Calculator (ANOVA)

This should be self-explanatory, but just in case it's not: your *F*-ratio value goes in the *F*-ratio value box, you stick your degrees of freedom for the numerator (between-treatments) in the *DF* - numerator box, your degrees of freedom for the denominator (within-treatments) in the *DF* - denominator box, select your significance level, then press the "Calculate" button.

If you need to derive an *F*-ratio value from raw data, you can find an ANOVA calculator here.

F-ratio value:	2
DF - numerator:	4
DF - denominator:	60

Significance Level:

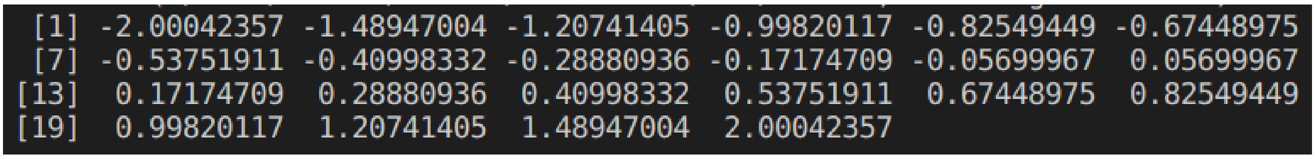
● .01
○ .05
○ .10

The *p*-value is .106001. The result is *not* significant at *p* < .01.

No, there is no significant evidence shows there is a difference in the treatment mean.

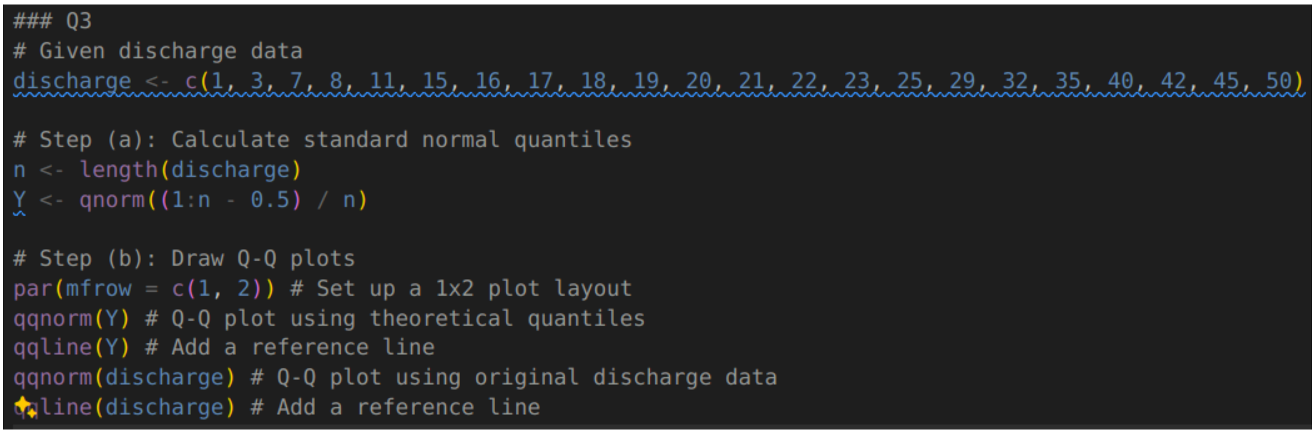
Problem 3

- (a) Y:



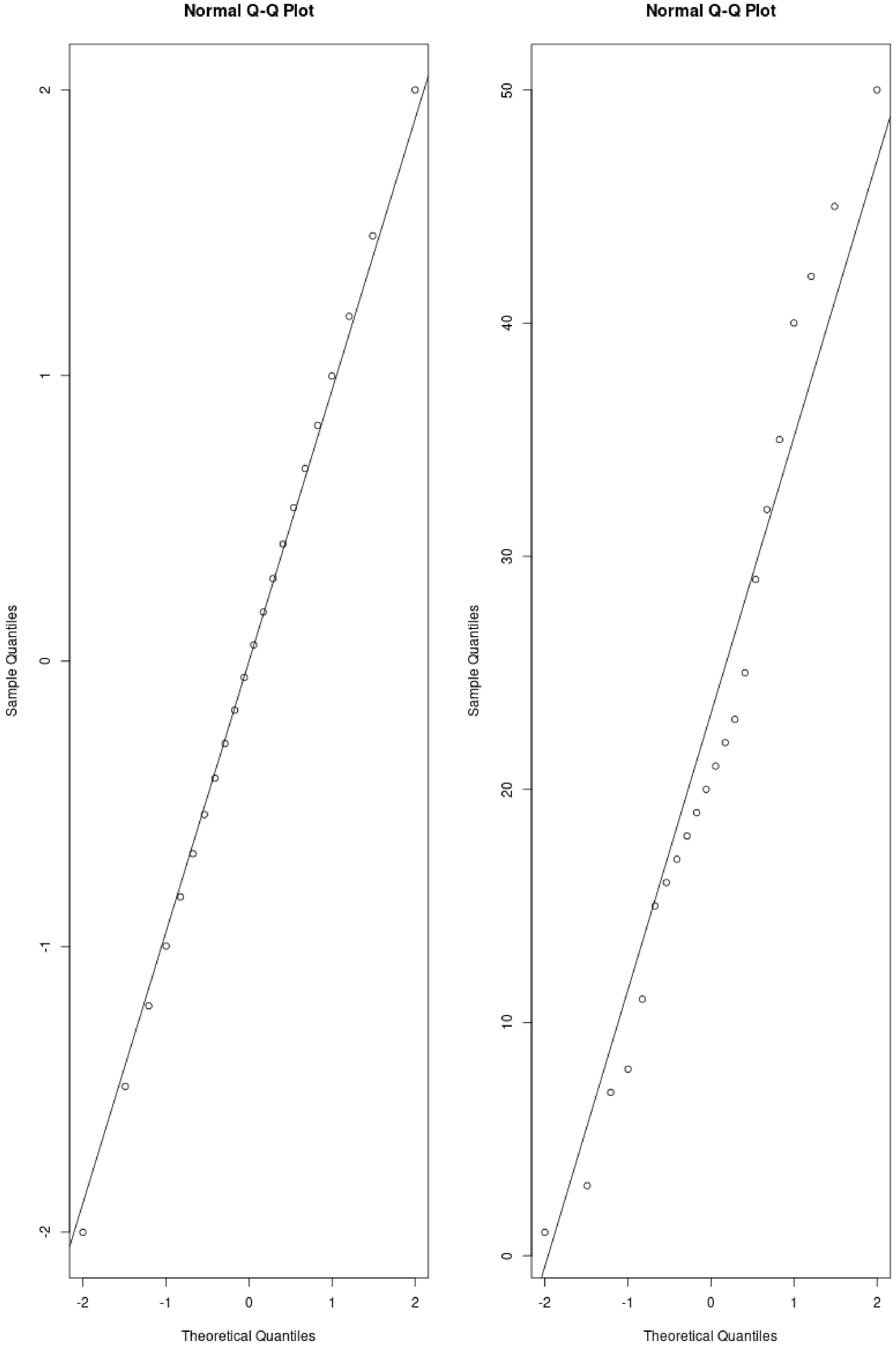
```
 [1] -2.00042357 -1.48947004 -1.20741405 -0.99820117 -0.82549449 -0.67448975
 [7] -0.53751911 -0.40998332 -0.28880936 -0.17174709 -0.05699967  0.05699967
[13]  0.17174709  0.28880936  0.40998332  0.53751911  0.67448975  0.82549449
[19]  0.99820117  1.20741405  1.48947004  2.00042357
```

- code:



```
### Q3
# Given discharge data
discharge <- c(1, 3, 7, 8, 11, 15, 16, 17, 18, 19, 20, 21, 22, 23, 25, 29, 32, 35, 40, 42, 45, 50)

# Step (a): Calculate standard normal quantiles
n <- length(discharge)
Y <- qnorm((1:n - 0.5) / n)

# Step (b): Draw Q-Q plots
par(mfrow = c(1, 2)) # Set up a 1x2 plot layout
qqnorm(Y) # Q-Q plot using theoretical quantiles
qqline(Y) # Add a reference line
qqnorm(discharge) # Q-Q plot using original discharge data
qqline(discharge) # Add a reference line
```

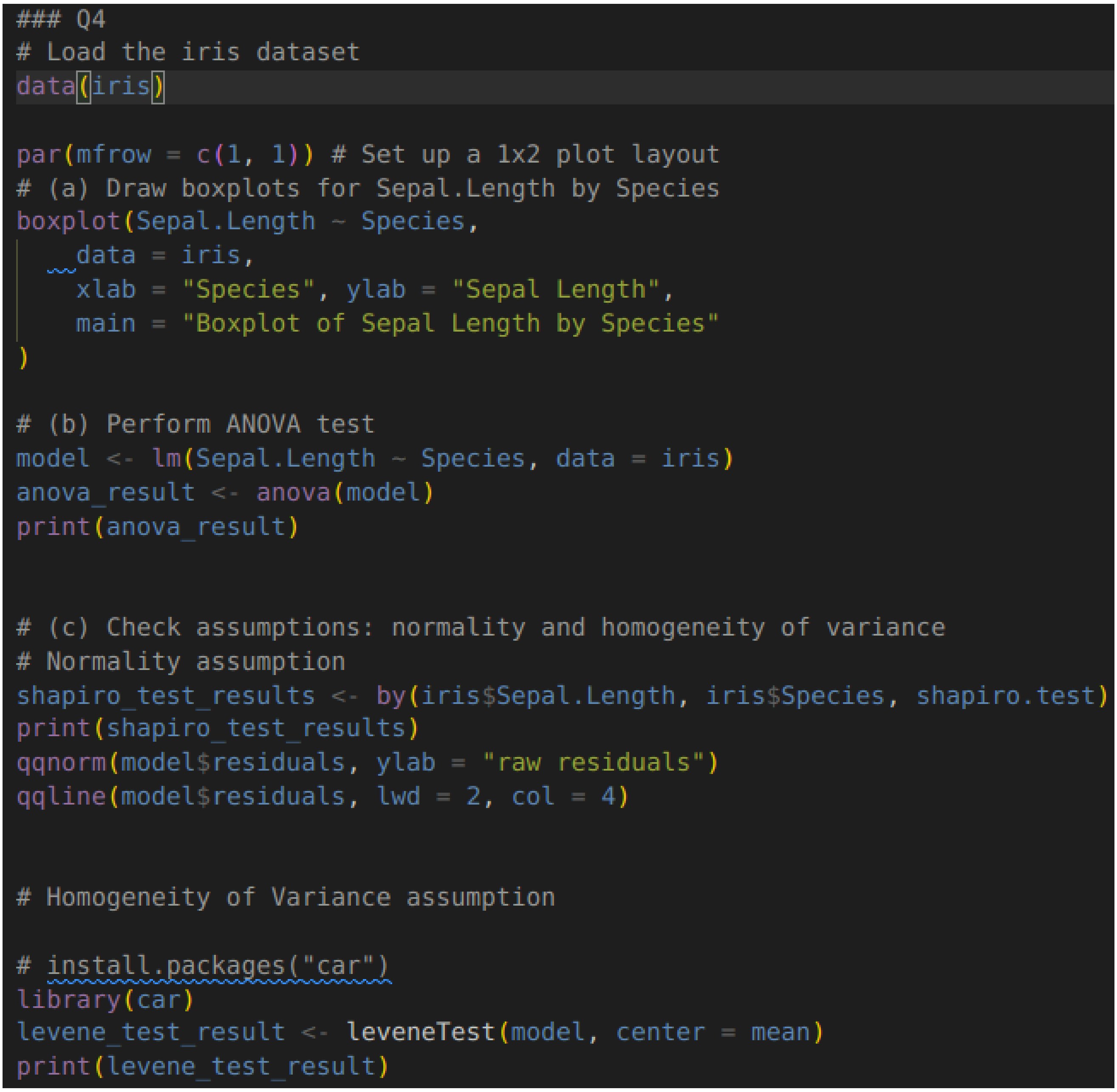
- (b) Q-Q plot:



nearly identical => normal

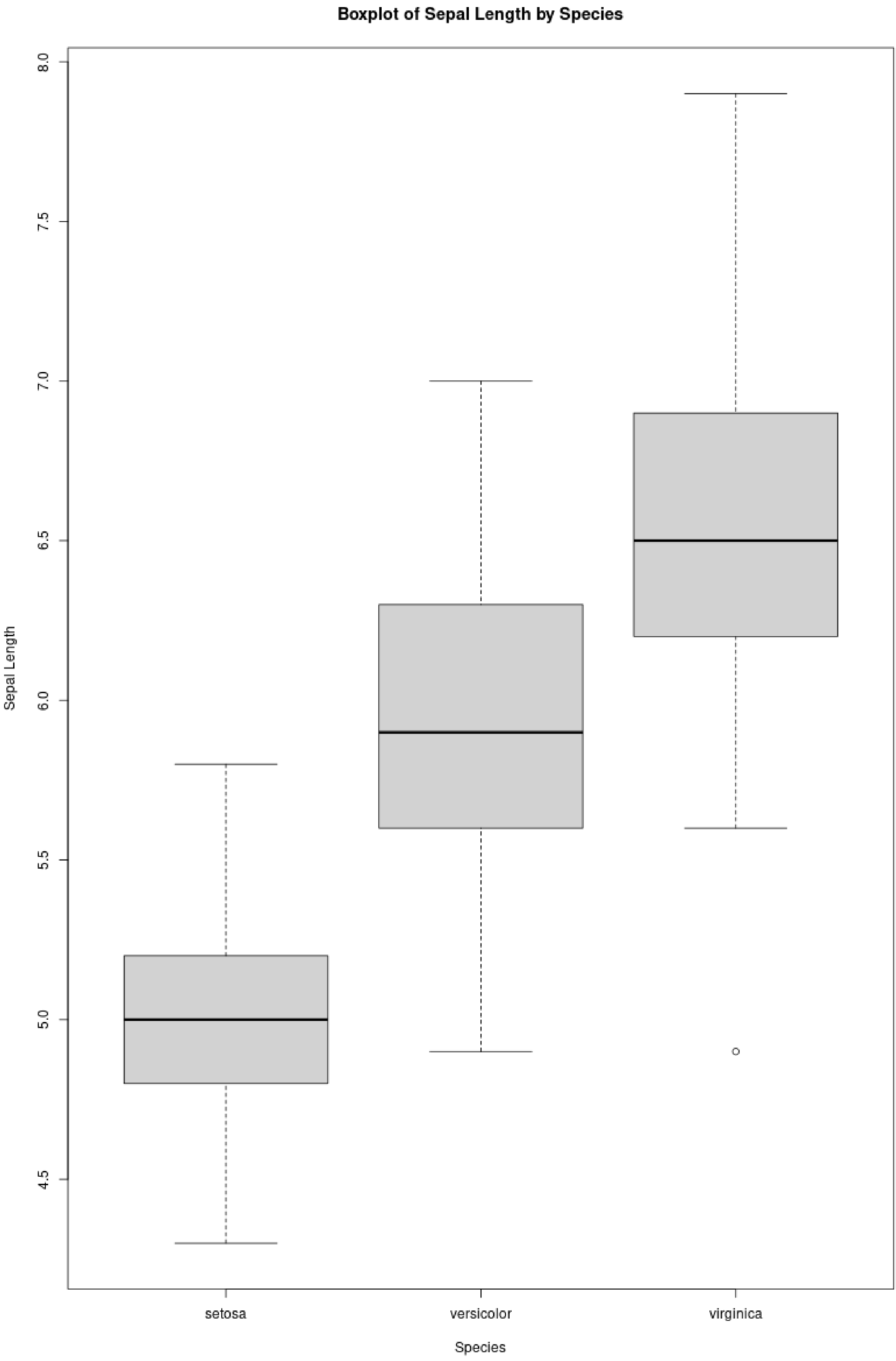
Problem 4

- code:



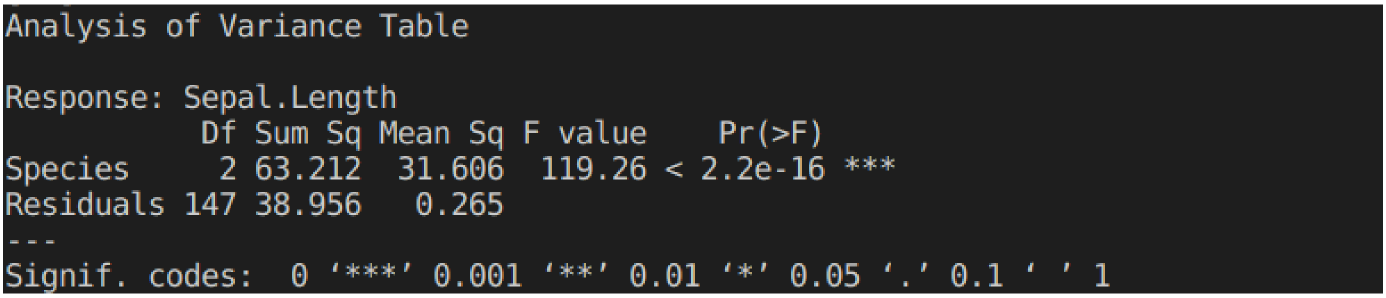
```
### Q4
# Load the iris dataset
data(iris)

par(mfrow = c(1, 1)) # Set up a 1x2 plot layout
# (a) Draw boxplots for Sepal.Length by Species
boxplot(Sepal.Length ~ Species,
        data = iris,
        xlab = "Species", ylab = "Sepal Length",
        main = "Boxplot of Sepal Length by Species"
)

# (b) Perform ANOVA test
model <- lm(Sepal.Length ~ Species, data = iris)
anova_result <- anova(model)
print(anova_result)


# (c) Check assumptions: normality and homogeneity of variance
# Normality assumption
shapiro_test_results <- by(iris$Sepal.Length, iris$Species, shapiro.test)
print(shapiro_test_results)
qqnorm(model$residuals, ylab = "raw residuals")
qqline(model$residuals, lwd = 2, col = 4)


# Homogeneity of Variance assumption

# install.packages("car")
library(car)
levene_test_result <- leveneTest(model, center = mean)
print(levene_test_result)
```

(a)



We can see that the length of sepals are the longest among virginica, with mean located at 6.5. Following is the versicolor, the mean of sepal length is arouud 5.8. There are significant difference between setosa and above two, the mean of sepal length located at 5.0.

(b)

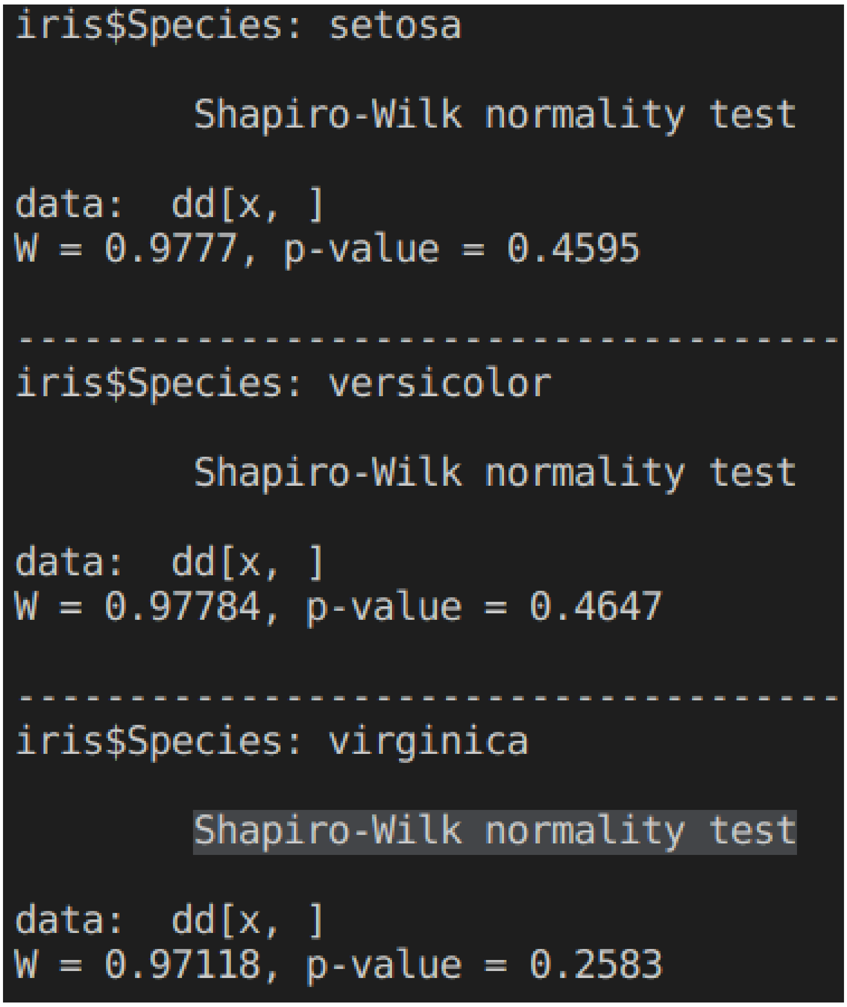


```
Analysis of Variance Table

Response: Sepal.Length
           Df Sum Sq Mean Sq F value    Pr(>F)    
Species     2 63.212  31.606  119.26 < 2.2e-16 ***
Residuals 147 38.956   0.265                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

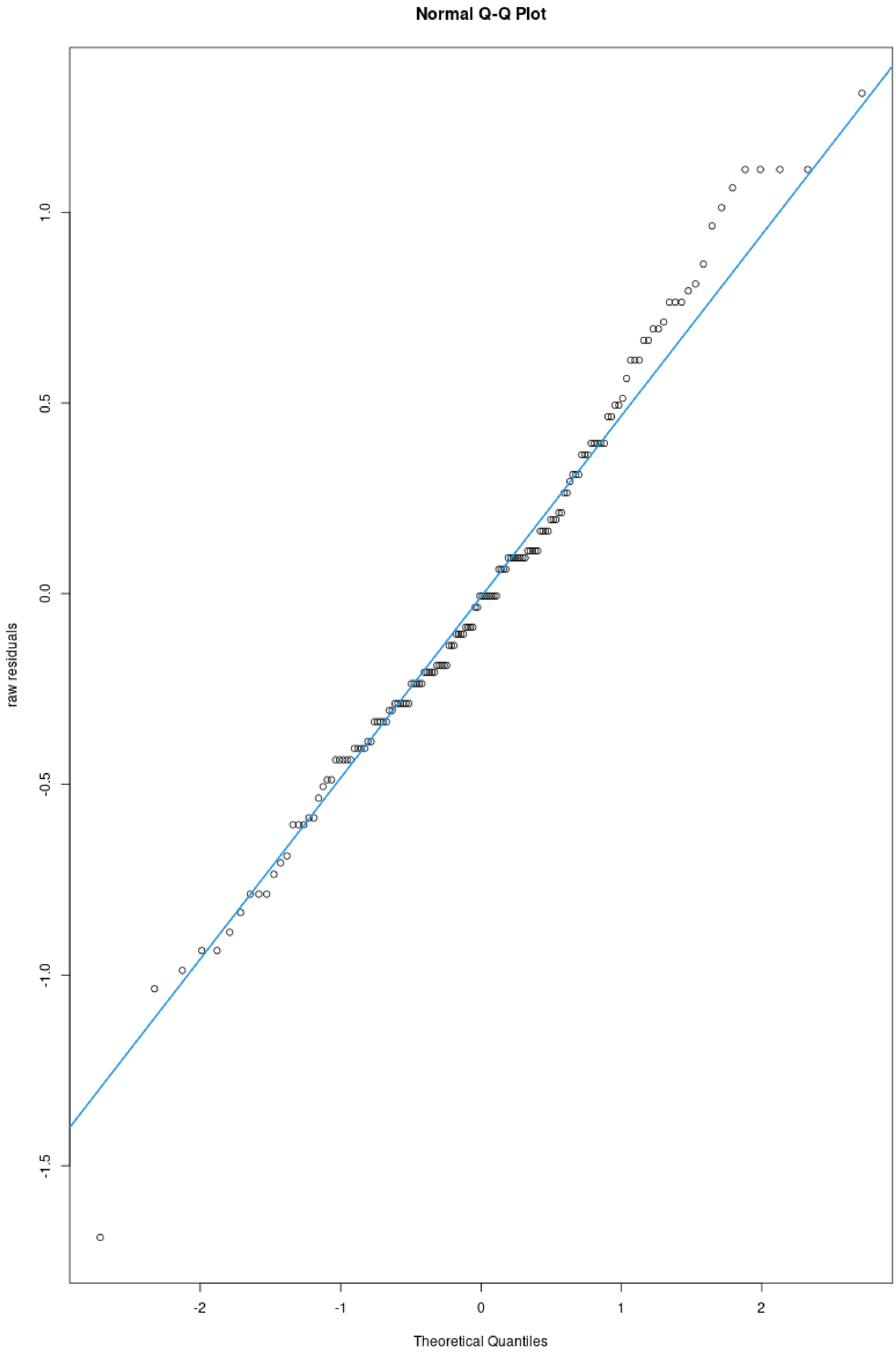
Yes, there is differences in the species group mean. I use the F value and Pr(>F), the F value is 119.26, the corresponding p-value is 2.2×10^{-16}, which is significant smaller than 0.05. Therefore, we reject the null hypothesis.

(c)

- normality



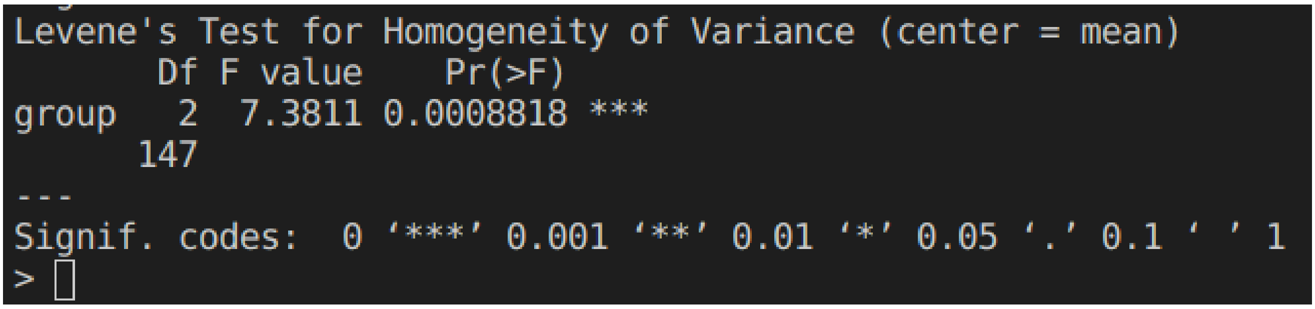
```
iris$Species: setosa

	Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.9777, p-value = 0.4595

-------------------------------------
iris$Species: versicolor

	Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97784, p-value = 0.4647

-------------------------------------
iris$Species: virginica

	Shapiro-Wilk normality test

data:  dd[x, ]
W = 0.97118, p-value = 0.2583
```



Based on the Shapiro–Wilk test, we see that the p-value for each species is greater than the chosen alpha(0.05). Therefore, we fail to reject the null hypothesis that the data is from a normally distributed population.

=> It's normal(checked)

- homogeneity



```
Levene's Test for Homogeneity of Variance (center = mean)
       Df F value    Pr(>F)    
group   2  7.3811 0.0008818 ***
      147                      
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> 
```

the p-value = 0.0008818, significantly smaller then the chosen alpha(0.05), thus, we reject the null hypothesis that the population variances are equal. We can conclude that there is a difference between the variances in the population.