Peter Mamaev

MGMT-4600 Lab 2 Part 1

* Generate Central Tendency values for AIR\_E variable
* Generate Central Tendency values for WATER\_E variable

> class(modernepi\_data$AIR\_E)

[1] "numeric"

> class(modernepi\_data$WATER\_E)

[1] "numeric"

> mean(modernepi\_data$AIR\_E, na.rm = TRUE)

[1] 49.46074

> median(modernepi\_data$WATER\_E, na.rm = TRUE)

[1] 71.17

> mode = mfv(modernepi\_data$AIR\_E, na\_rm = TRUE)

> print(mode)

[1] 44.69 51.96 53.74

> mode = mfv(modernepi\_data$WATER\_E, na\_rm = TRUE)

> print(mode)

[1] 0.00 71.40 73.08

> class(modernepi\_data$NOX\_pt)

[1] "numeric"

> class(modernepi\_data$SO2\_pt)

[1] "numeric"

> mean(modernepi\_data$NOX\_pt, na.rm = TRUE)

[1] 47.50777

> mean(modernepi\_data$SO2\_pt, na.rm = TRUE)

[1] 53.05288

mode = mfv(modernepi\_data$AIR\_E, na\_rm = TRUE)

> print(mode)

[1] 44.69 51.96 53.74

> class(modernepi\_data$CLIMATE)

[1] "numeric"

> class(modernepi\_data$AGRICULTURE)

[1] "numeric"

> mean(modernepi\_data$CLIMATE, na.rm = TRUE)

[1] 55.33374

> mean(modernepi\_data$AGRICULTURE, na.rm = TRUE)

[1] 70.85822

> median(modernepi\_data$CLIMATE, na.rm = TRUE)

[1] 55.43

> median(modernepi\_data$AGRICULTURE, na.rm = TRUE)

[1] 75.29

> class(modernepi\_data$CLIMATE)

[1] "numeric"

> class(modernepi\_data$AGRICULTURE)

[1] "numeric"

> mean(modernepi\_data$CLIMATE, na.rm = TRUE)

[1] 55.33374

> mean(modernepi\_data$AGRICULTURE, na.rm = TRUE)

[1] 70.85822

> median(modernepi\_data$CLIMATE, na.rm = TRUE)

[1] 55.43

> median(modernepi\_data$AGRICULTURE, na.rm = TRUE)

[1] 75.29

> mode = mfv(modernepi\_data$CLIMATE, na\_rm = TRUE)

> print(mode)

[1] 35.05 49.85 60.74 64.45 75.40

> mode = mfv(modernepi\_data$AGRICULTURE, na\_rm = TRUE)

> print(mode)

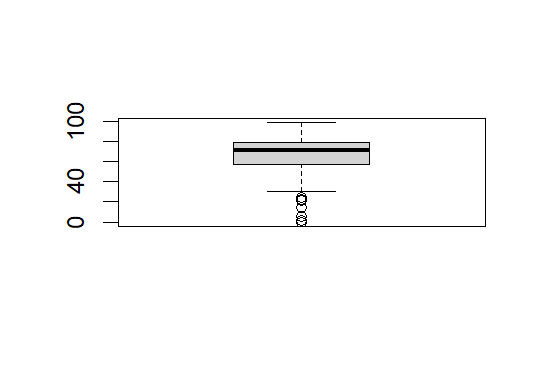
[1] 54.55

* Generate the Boxplots for AIR\_E variable

A line drawing of a rectangular object

Description automatically generated

* Generate the Boxplots for WATER\_E variable



* Generate Central Tendency values for NOX\_pt variable
* Generate Central Tendency values for SO2\_pt variable

> class(modernepi\_data$NOX\_pt)

[1] "numeric"

> class(modernepi\_data$SO2\_pt)

[1] "numeric"

> mean(modernepi\_data$NOX\_pt, na.rm = TRUE)

[1] 47.50777

> mean(modernepi\_data$SO2\_pt, na.rm = TRUE)

[1] 53.05288

> median(modernepi\_data$NOX\_pt, na.rm = TRUE)

[1] 48.03795

> median(modernepi\_data$SO2\_pt, na.rm = TRUE)

[1] 53.7276

> hist(modernepi\_data$NOX\_pt)

> hist(modernepi\_data$SO2\_pt)

> mode = mfv(modernepi\_data$AIR\_E, na\_rm = TRUE)

> print(mode)

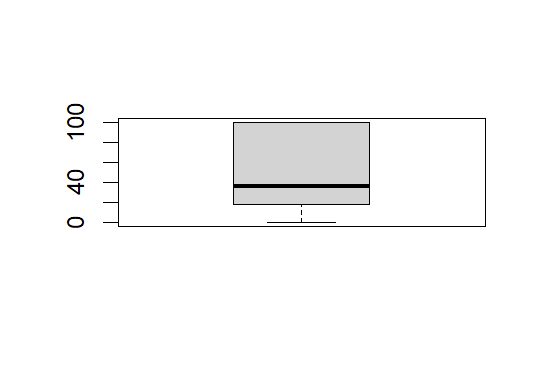
[1] 44.69 51.96 53.74

> mode = mfv(modernepi\_data$WATER\_E, na\_rm = TRUE)

> print(mode)

[1] 0.00 71.40 73.08

* Generate the Boxplots for OZONE\_pt variable



* Generate the Boxplots for WQI\_pt variable

A black and white rectangular object with lines

Description automatically generated with medium confidence

* Generate Central Tendency values for CLIMATE variable
* Generate Central Tendency values for AGRICULTURE variable

> # This is done to check the numerical nature of the values

> class(modernepi\_data$CLIMATE)

[1] "numeric"

> class(modernepi\_data$AGRICULTURE)

[1] "numeric"

> mean(modernepi\_data$CLIMATE, na.rm = TRUE)

[1] 55.33374

> mean(modernepi\_data$AGRICULTURE, na.rm = TRUE)

[1] 70.85822

> median(modernepi\_data$CLIMATE, na.rm = TRUE)

[1] 55.43

> median(modernepi\_data$AGRICULTURE, na.rm = TRUE)

[1] 75.29

> hist(modernepi\_data$CLIMATE)

> hist(modernepi\_data$AGRICULTURE)

> # Multiple modes because a single number can appear multiple times as the mode

> mode = mfv(modernepi\_data$CLIMATE, na\_rm = TRUE)

> print(mode)

[1] 35.05 49.85 60.74 64.45 75.40

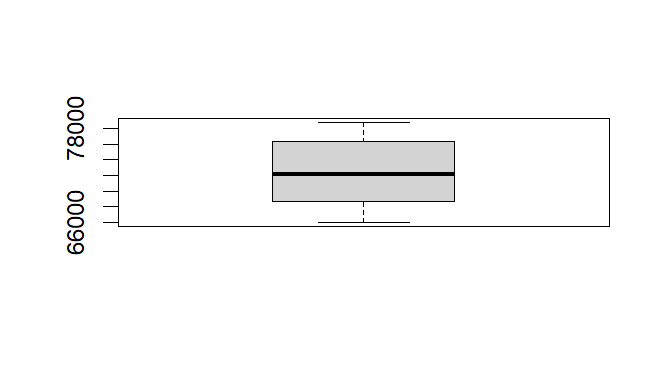
> # Here, on the other hand, using the same method, there is only one most frequent number.

> mode = mfv(modernepi\_data$AGRICULTURE, na\_rm = TRUE)

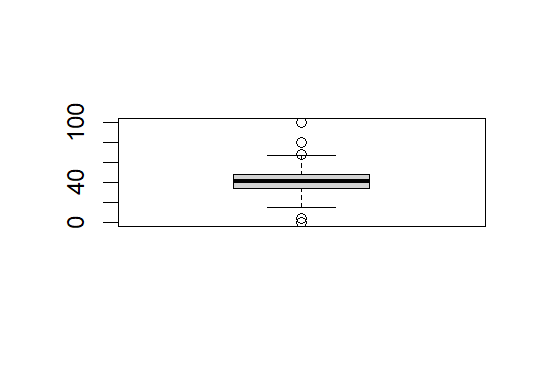
> print(mode)

[1] 54.55

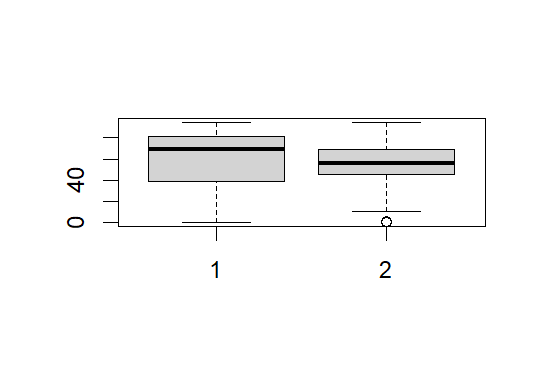
* Generate the Boxplots for FISHERIES\_pt variable



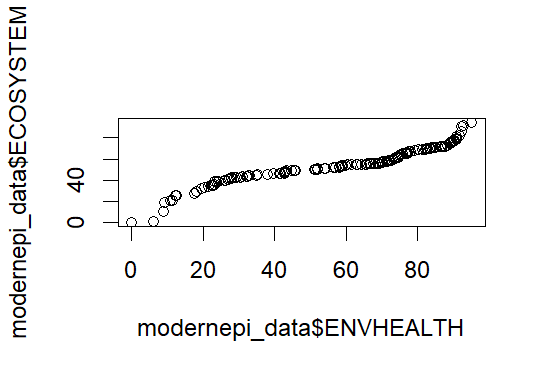
* Generate the Boxplots for NMVOC\_pt variable



boxplot(ENVHEALTH,ECOSYSTEM)

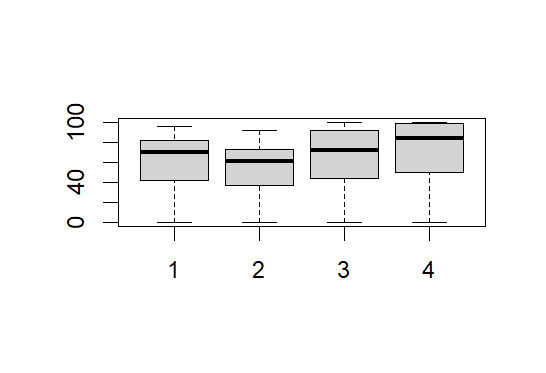


qqplot(ENVHEALTH,ECOSYSTEM



> #SLIDE 12

> boxplot(epi\_data$ENVHEALTH, epi\_data$DALY, epi\_data$AIR\_H, epi\_data$WATER\_H)



> lmENVH <- lm(epi\_data$ENVHEALTH ~ epi\_data$DALY + epi\_data$AIR\_H + epi\_data$WATER\_H)

> lmENVH

Call:

lm(formula = epi\_data$ENVHEALTH ~ epi\_data$DALY + epi\_data$AIR\_H +

epi\_data$WATER\_H)

Coefficients:

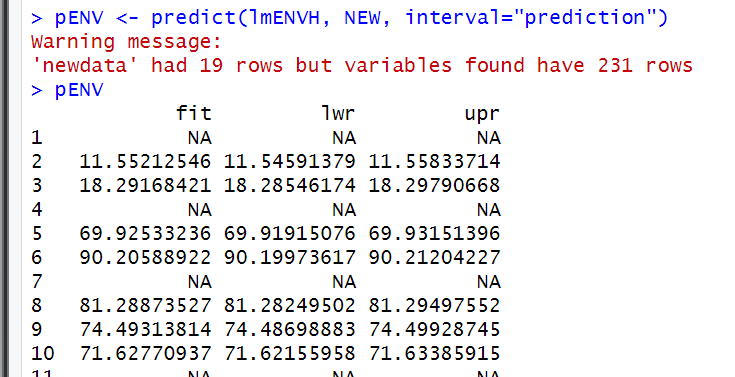
(Intercept) epi\_data$DALY epi\_data$AIR\_H

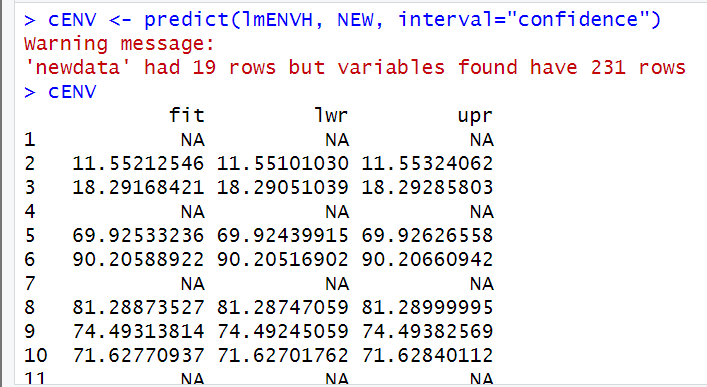
-2.673e-05 5.000e-01 2.500e-01

epi\_data$WATER\_H

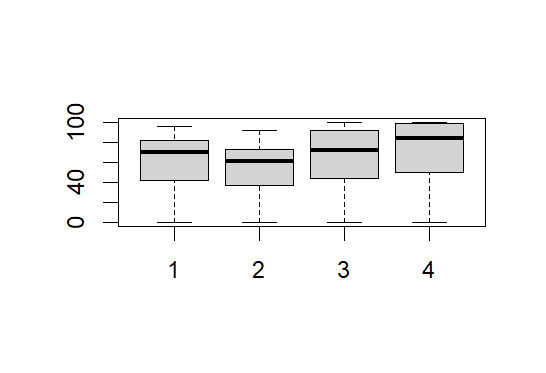
2.500e-01

> #SLIDE 13





* Boxplot ENVHEALTH, DALY, AIR\_H, WATER\_H



> lmENVH <- lm(epi\_data$ENVHEALTH ~ epi\_data$DALY + epi\_data$AIR\_H + epi\_data$WATER\_H)

> lmENVH

Call:

lm(formula = epi\_data$ENVHEALTH ~ epi\_data$DALY + epi\_data$AIR\_H +

epi\_data$WATER\_H)

Coefficients:

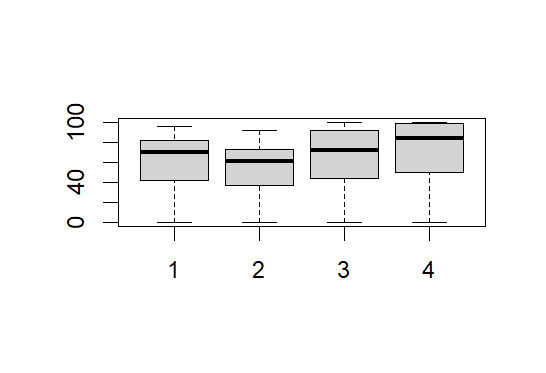
(Intercept) epi\_data$DALY epi\_data$AIR\_H

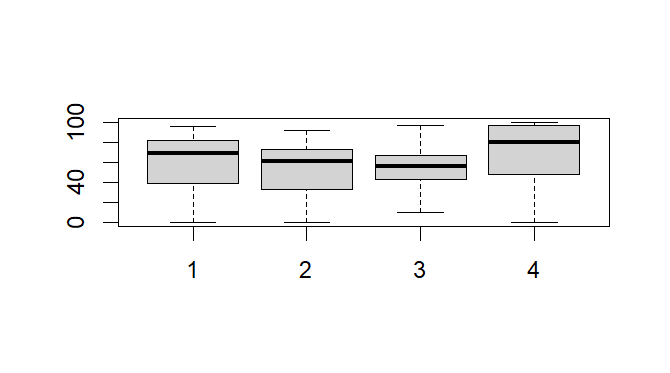
-2.673e-05 5.000e-01 2.500e-01

epi\_data$WATER\_H

2.500e-01

* Using the response variable as: AIR\_E and CLIMATE\_E
* > boxplot(ENVHEALTH,DALY,AIR\_E,WATER\_H)
* > boxplot(ENVHEALTH,DALY,CLIMATE,WATER\_H)





> Model1

Call:

lm(formula = epi\_data$AIR\_E ~ epi\_data$DALY + epi\_data$AIR\_H +

epi\_data$WATER\_H)

Coefficients:

(Intercept) epi\_data$DALY epi\_data$AIR\_H

59.2903 -0.1248 0.1686

epi\_data$WATER\_H

-0.1798

> summary(Model1)

Call:

lm(formula = epi\_data$AIR\_E ~ epi\_data$DALY + epi\_data$AIR\_H +

epi\_data$WATER\_H)

Residuals:

Min 1Q Median 3Q Max

-32.708 -7.328 -1.739 8.117 38.182

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 59.29025 2.55759 23.182 < 2e-16 \*\*\*

epi\_data$DALY -0.12482 0.07707 -1.620 0.10710

epi\_data$AIR\_H 0.16863 0.05104 3.304 0.00115 \*\*

epi\_data$WATER\_H -0.17982 0.07021 -2.561 0.01126 \*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.42 on 178 degrees of freedom

(49 observations deleted due to missingness)

Multiple R-squared: 0.1803, Adjusted R-squared: 0.1664

F-statistic: 13.05 on 3 and 178 DF, p-value: 9.654e-08

> cModel1 <- coef(Model1)

> AIR\_ENEW <- c(seq(5, 95, 5))

> help(seq)

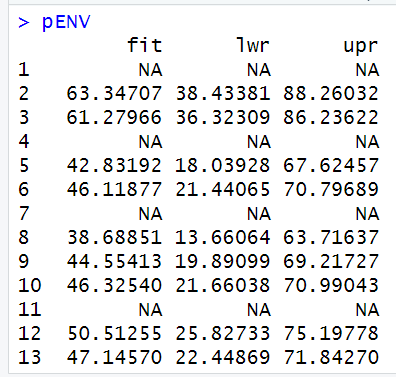
> NEW\_AIRE <- data.frame(DALYNEW, AIR\_ENEW, WATER\_HNEW)

> pENV <- predict(Model1, NEW\_AIRE, interval = "prediction")

Warning message:

'newdata' had 19 rows but variables found have 231 rows

> pENV

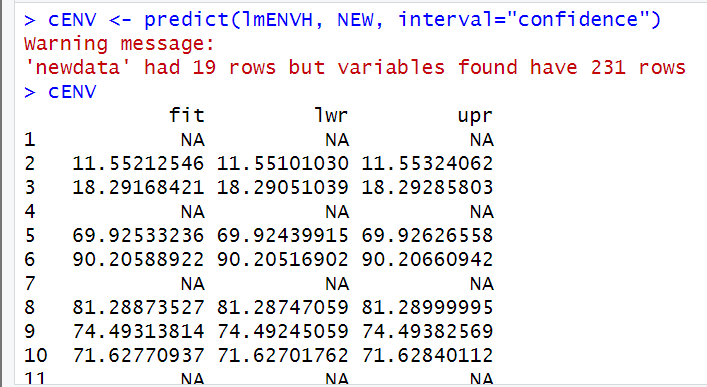


> cENV <- predict(Model1, NEW\_AIRE, interval = "confidence")

Warning message:

'newdata' had 19 rows but variables found have 231 rows

> cENV



> Model2

> Model2 <- lm(epi\_data$CLIMATE ~ epi\_data$DALY + epi\_data$AIR\_H + epi\_data$WATER\_H)

Call:

lm(formula = epi\_data$CLIMATE ~ epi\_data$DALY + epi\_data$AIR\_H +

epi\_data$WATER\_H)

Coefficients:

(Intercept) epi\_data$DALY epi\_data$AIR\_H

75.3487 -0.1732 0.0181

epi\_data$WATER\_H

-0.1538

> summary(Model2)

Call:

lm(formula = epi\_data$CLIMATE ~ epi\_data$DALY + epi\_data$AIR\_H +

epi\_data$WATER\_H)

Residuals:

Min 1Q Median 3Q Max

-37.578 -9.768 1.165 9.164 44.434

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 75.34874 3.01412 24.999 <2e-16 \*\*\*

epi\_data$DALY -0.17323 0.09050 -1.914 0.0573 .

epi\_data$AIR\_H 0.01810 0.05919 0.306 0.7602

epi\_data$WATER\_H -0.15385 0.08161 -1.885 0.0611 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.15 on 168 degrees of freedom

(59 observations deleted due to missingness)

Multiple R-squared: 0.255, Adjusted R-squared: 0.2417

F-statistic: 19.17 on 3 and 168 DF, p-value: 9.704e-11

> cModel2 <- coef(Model2)

> CLIMATE\_NEW <- c(seq(5, 95, 5))

> cModel2 <- coef(Model2)

> CLIMATE\_NEW <- c(seq(5, 95, 5))

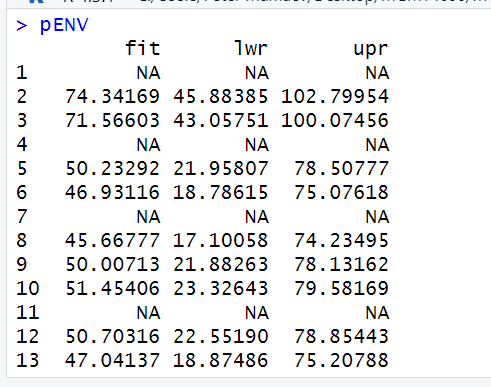
> NEW\_CLIMATE <- data.frame(DALYNEW, CLIMATE\_NEW, WATER\_HNEW)

> pENV <- predict(Model2, NEW\_CLIMATE, interval = "prediction")

Warning message:

'newdata' had 19 rows but variables found have 231 rows

> pENV

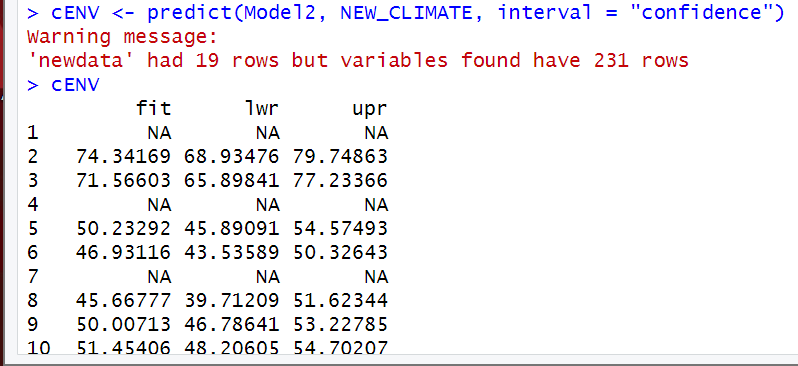


> cENV <- predict(Model2, NEW\_CLIMATE, interval = "confidence")

Warning message:

'newdata' had 19 rows but variables found have 231 rows

> cENV



On EPI dataset, use the: ENVHEALTH and ECOSYSTEM Explain the distributions using shapiro.test() function.

Run the Shapiro.Wilk normality test for the following variables in 2010EPI.csv using shapiro.test() function. Interpret your results for normality.

> # While the Shapiro-Wilks test is possible, the number of rows for

> # 2010 EPI data (referred to here as modernepi\_data) is too high.

> # As such we wouldn't receive a good assessment for its values.

> # The dataset for epi\_data is within acceptable size to do Shapiro-Wilks test on.

> nrow(epi\_data)

[1] 231

> library(dplyr)

> # data\_sl <- sample(modernepi\_data))

> data\_sl <- na.omit(modernepi\_data$ENVHEALTH)

> shapiro.test(data\_sl)

Shapiro-Wilk normality test

data: data\_sl

W = 0.92019, p-value = 8.178e-08

* From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution.

> # Ommitting the NA values from the 2010 EPI data to compress it to a size where the Shapiro-Wilks can be performed.

> data\_sl2 <- na.omit(modernepi\_data$ECOSYSTEM)

> shapiro.test(data\_sl2)

Shapiro-Wilk normality test

data: data\_sl2

W = 0.98129, p-value = 0.02654

* From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution.

> shapiro.test(epi\_data$ENVHEALTH)

Shapiro-Wilk normality test

data: epi\_data$ENVHEALTH

W = 0.91613, p-value = 1.083e-08

* From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution.

> shapiro.test(epi\_data$ECOSYSTEM)

Shapiro-Wilk normality test

data: epi\_data$ECOSYSTEM

W = 0.98129, p-value = 0.02654

* From the output, the p-value < 0.05 implying that the distribution of the data are significantly different from normal distribution.