**Bangabandhu Sheikh Mujibur Rahman Agricultural University**
**EDGE_Batch-11**
**Quiz Exam**
**Marks: 20 Time: 90 minutes**
**Name: Ummay Habiba**
**Reg. No:18-05-4603 Dept. Agronomy**

**Note**: Submit the completed file to rabiulauwul@bsmrau.edu.bd with subject ***EDGE11_Quiz_Your registration number_ Dept.***

1. **Short Questions** **(6*1=06)**
   a) In R, you can use …**install packages**…… () to install a package from CRAN.
   b) To check the structure of an object in R, the function ……**str**…… () is used.
   c) To subset a data frame by selecting specific rows and columns, the **……[, ]**…… operator is used.
   d) In R, the ……**summary**…… () function provides a summary of key descriptive statistics
   e) In R, the ……**na.rm**…… () function can be used to remove missing values (NA) from a vector x.
   f) The residuals of a regression model are the differences between the observed values and the……**fitted…**……. values predicted by the model.

2. For the ***iris*** data**:** **(7)**
   a) Calculate descriptive statistics ($\boldsymbol{median \pm SD, mean, CV}$) for each numeric variable in a single table.

Answer: iris

```
#For median_sd

Median_SD <- sapply(iris[, 1:4], function(column) {

 median_val <- median(column)

 sd_val <- sd(column)

 paste0(round(median_val, 2), " ± ", round(sd_val, 2))

})


#For Mean
```

```
Mean <- sapply(iris[, 1:4], mean)


#For CV
cv<- function(column) {
  (sd(column) / mean(column)) * 100
}
CV <- sapply(iris[, 1:4], cv)


#For descriptive statistics
descriptive_stats <- data.frame(
  Median_SD, Mean, CV)


print(descriptive_stats)
write.csv(descriptive_stats,"iris_descriptive-dataset.csv")
```

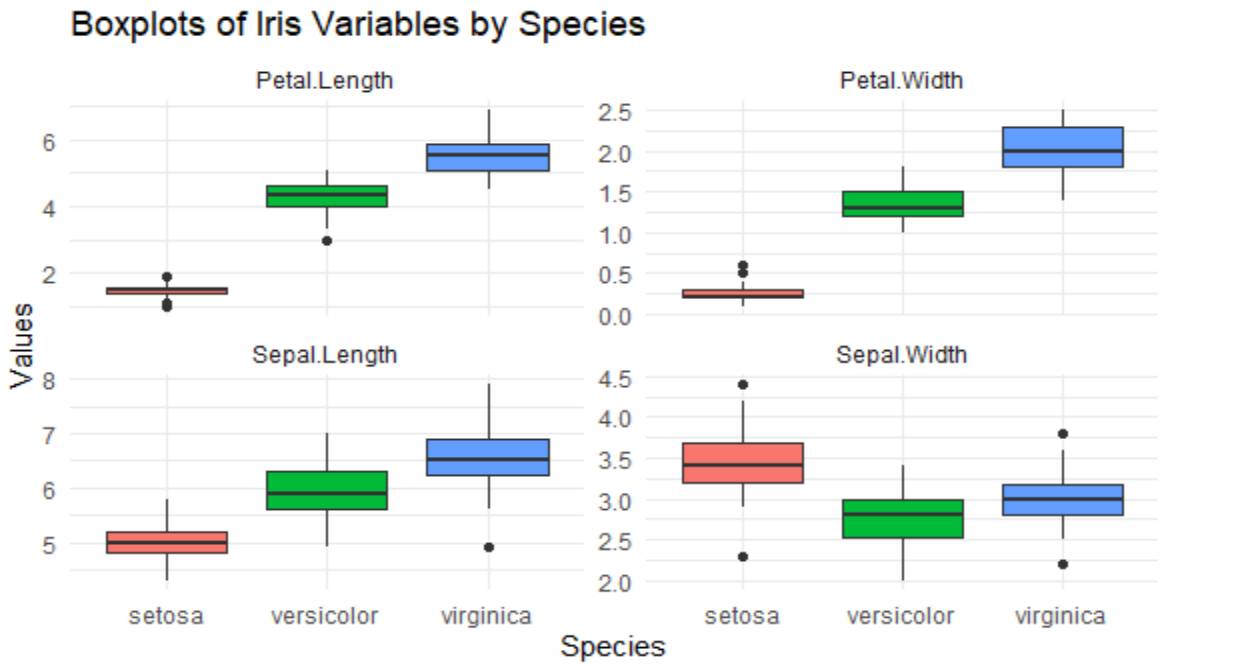
| | Median_SD | Mean | CV |
|---|---|---|---|
| Sepal.Length | 5.8 ± 0.83 | 5.843333 | 14.17113 |
| Sepal.Width | 3 ± 0.44 | 3.057333 | 14.25642 |
| Petal.Length | 4.35 ± 1.77 | 3.758000 | 46.97441 |
| Petal.Width | 1.3 ± 0.76 | 1.199333 | 63.55511 |

b) Construct boxplots with ggplot2 package for each variable by ***Species*** categories with color aesthetic and interpret your results.

Answer: # Load the iris dataset

```
data(iris)
```

```
# Reshape the data into a long format for easier plotting
iris_long <- iris %>%
  pivot_longer(cols = -Species, names_to = "Variable", values_to = "Value")

# Create boxplots
ggplot(iris_long, aes(x = Species, y = Value, fill = Species)) +
  geom_boxplot() +
  facet_wrap(~Variable, scales = "free_y") +
  theme_minimal() +
  labs(
    title = "Boxplots of Iris Variables by Species",
    x = "Species",
    y = "Values"
  ) +
  theme(legend.position = "none")
```



There are four variables of iris such as petal length, petal width, sepal length and sepal width. In case of petal length, virginica has the larger petal length and petal width. In versicolor and Virginica have similar sepal length, and setosa has larger sepal width.

**3.** For the provided dataset of "***vegitables"***, answer the following questions: **(7)**

a) Identify missing values in each variable and impute them using the mean values of the corresponding variables.

Answer:

```
##missing value

data<-read.csv("1734953626384_vegitables.csv")

str(data)

data[data==""]<-NA

data[data=="" | data ==" "] <- NA

summary(data)

is.na(data)

table(is.na(data))

sum(is.na(data))

which(is.na(data))

data1<-na.omit(data)

mean(data, na.rm=TRUE)

is.na(data1)

table(is.na(data1))
```

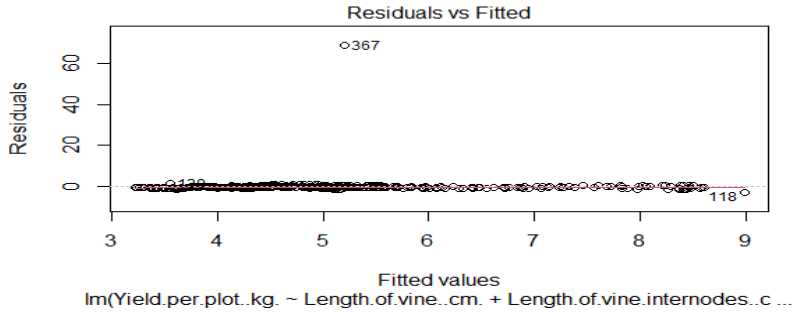
there are six not available data in this dataset. 39 466 933 1184 1688 2137

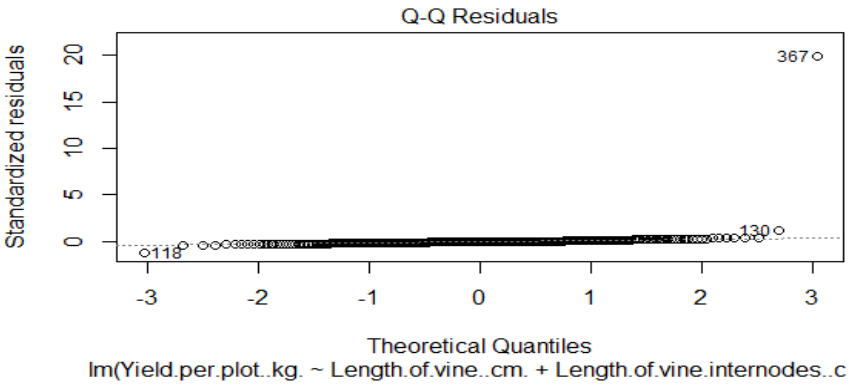
b) Fit a suitable multiple linear regression model for the dataset and interpret your findings.

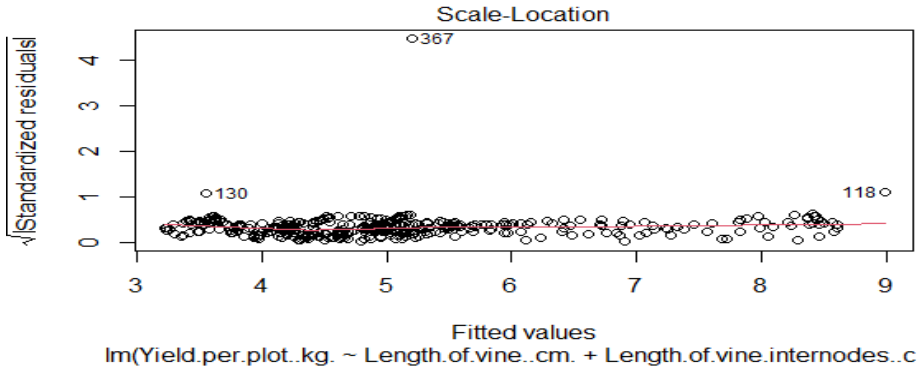
Answer:

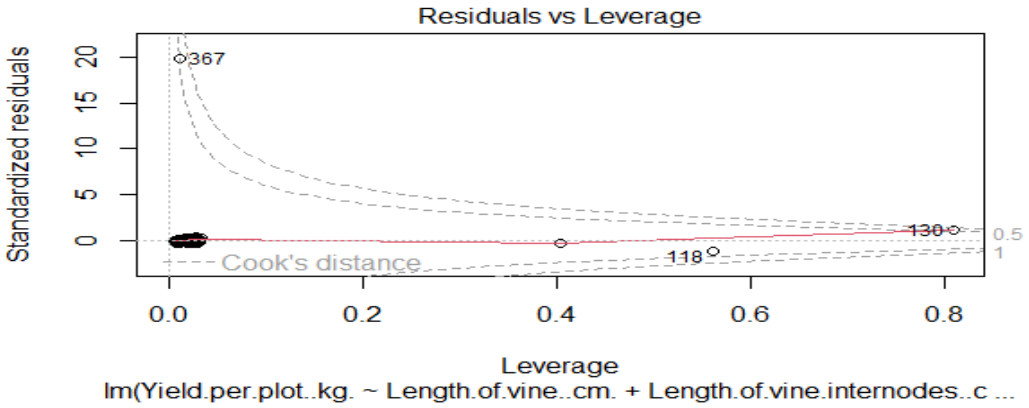
##regression model

```
dat<-read.csv('1734953626384_vegitables.csv',header=T)

head(dat)

model <- lm(Yield.per.plot..kg. ~ Length.of.vine..cm. +Length.of.vine.internodes..cm. +Petiole.length..cm. + Number.of.leaves.per.plant+Number.of.branches..main.+Number.of.days.required.for.maturity+Number.of.tubers.per.plant, data = dat)

summary(model)

AIC(model)

abline(lm(Yield.per.plot..kg. ~ Length.of.vine..cm. +Length.of.vine.internodes..cm. +Petiole.length..cm. + Number.of.leaves.per.plant+Number.of.branches..main.+Number.of.days.required.for.maturity+Number.of.tubers.per.plant, data = dat)))

plot(model)
```




Q-Q Residuals
Standardized residuals
0
5
10
15
20
367
130
118
-3
-2
-1
0
1
2
3
Theoretical Quantiles
lm(Yield.per.plot..kg. ~ Length.of.vine..cm. + Length.of.vine.internodes..c ...


Scale-Location
√|Standardized residuals|
0
1
2
3
4
367
130
118
3
4
5
6
7
8
9
Fitted values
lm(Yield.per.plot..kg. ~ Length.of.vine..cm. + Length.of.vine.internodes..c ...


Residuals vs Leverage
Standardized residuals
0
5
10
15
20
367
130
118
Cook's distance
0.5
1
0.0
0.2
0.4
0.6
0.8
Leverage
lm(Yield.per.plot..kg. ~ Length.of.vine..cm. + Length.of.vine.internodes..c ...