Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Quiz Exam
Marks: 20 Time: 90 minutes
Name: ……Linnet Riya Barman……
Reg. No: ……18-05-4835……Dept.……Agricultural Economics…….

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.omit()…… 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.

```
# Function to calculate Coefficient of Variation (CV)
cv <- function(x) {
  sd(x) / mean(x) * 100
}

# Calculate descriptive statistics
descriptive_stats <- data.frame(
  Median = apply(iris[, 1:4], 2, median),
  Mean = colMeans(iris[, 1:4]),
  SD = apply(iris[, 1:4], 2, sd),
  `Median±SD` = apply(iris[, 1:4], 2, function(x) median(x) + sd(x)),
  CV = apply(iris[, 1:4], 2, cv)
```

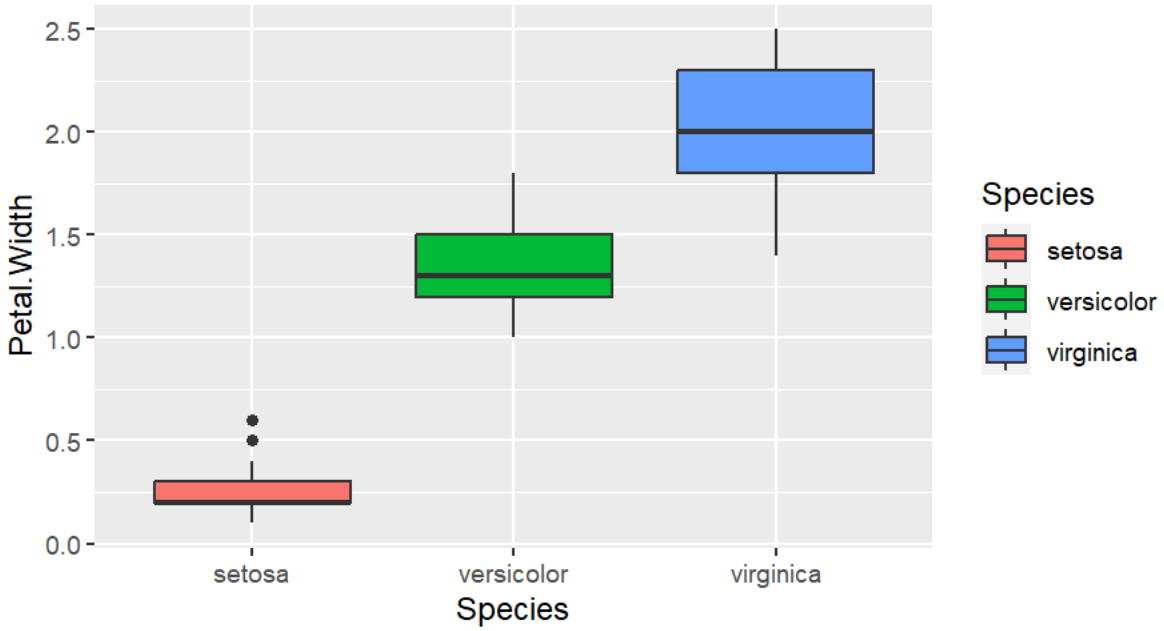
```
)

# View the resulting table
print(descriptive_stats)
```

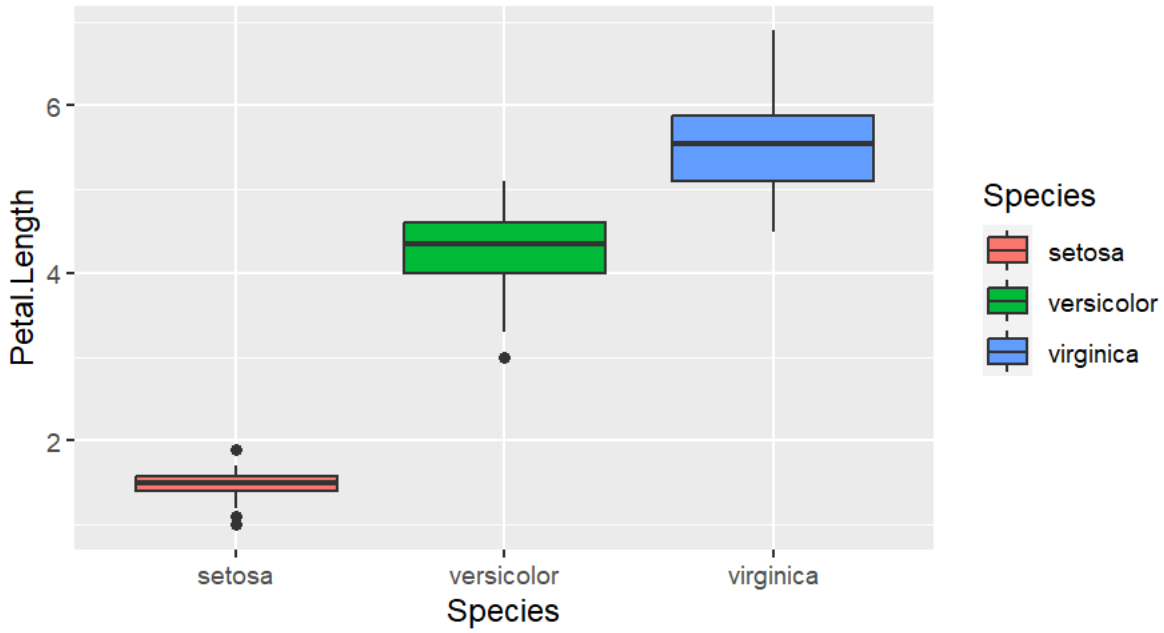
	Median	Mean	SD	Median.SD	CV
Sepal.Length	5.8	5.843333	0.828066	6.628066	14.17113
Sepal.Width	3	3.057333	0.435866	3.435866	14.25642
Petal.Length	4.35	3.758	1.765298	6.115298	46.97441
Petal.Width	1.3	1.199333	0.762238	2.062238	63.55511

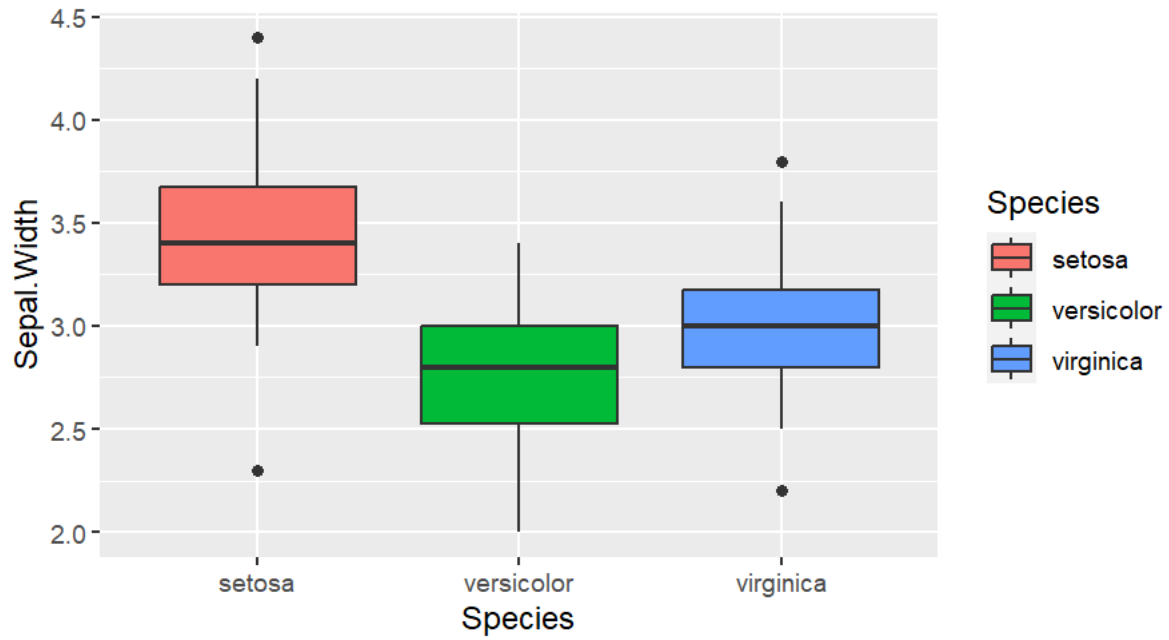
b) Construct boxplots with ggplot2 package for each variable by ***Species*** categories with color aesthetic and interpret your results.

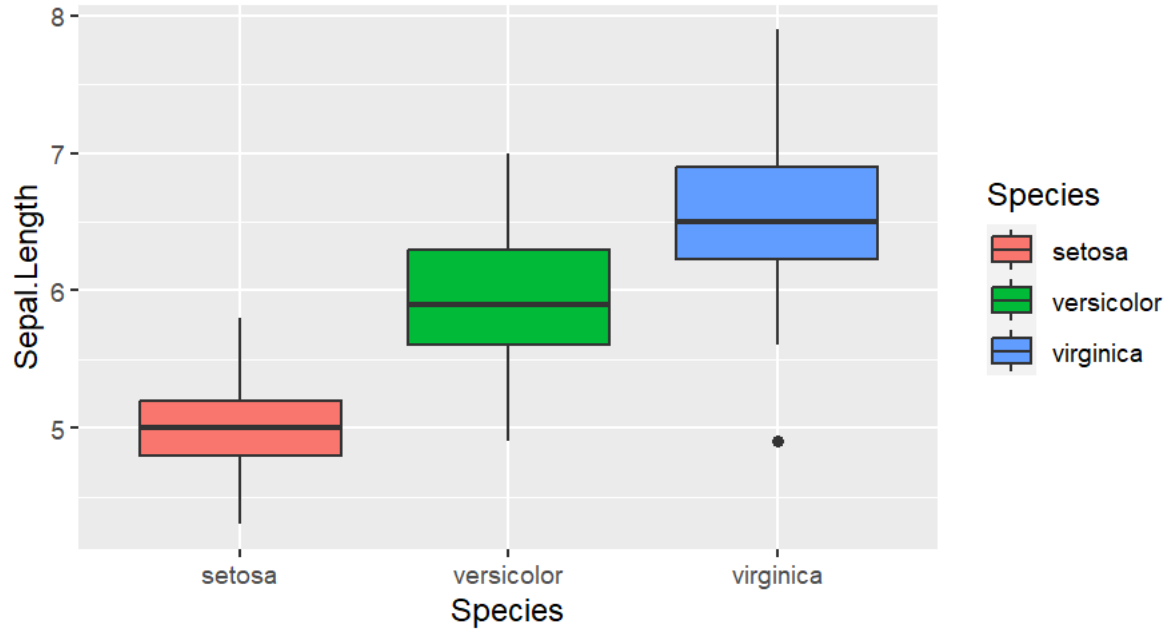
```
library(ggplot2)
library(ggExtra)
iris1<-iris
ggplot(iris1)+
 aes(x=Species, y=Sepal.Length)+
 geom_point(aes(shape="Species",color="Species"))

#Boxplot
ggplot(iris1<-iris,
     aes(x=Species,y=Sepal.Length,fill=Species))+
 geom_boxplot()

ggplot(iris1<-iris,
     aes(x=Species,y=Sepal.Width,fill=Species))+
 geom_boxplot()

ggplot(iris1<-iris,
     aes(x=Species,y=Petal.Length,fill=Species))+
 geom_boxplot()

ggplot(iris1<-iris,
     aes(x=Species,y=Petal.Width,fill=Species))+
 geom_boxplot()
```

Boxplots of Iris Datasets






4.5
4.0
3.5
3.0
2.5
2.0
Sepal.Width
setosa
versicolor
virginica
Species
Species
setosa
versicolor
virginica


8
7
6
5
Sepal.Length
setosa
versicolor
virginica
Species
Species
setosa
versicolor
virginica

Interpretation:

The boxplot highlights that petal width is a significant feature for distinguishing between species. Setosa is particularly distinct with the smallest and most consistent petal widths, while Virginica displays the largest range.

The boxplot clearly demonstrates that , petal length is a distinguishing feature among the three species. Setosa is distinct due to its small and consistent petal lengths. Versicolor and Virginica overlap more in their petal length distributions but are still separable based on range and central tendency.

Sepal width provides a moderate level of separation among species. Sepal width in setosa is more than other two species.

Sepal length provides a moderate level of separation among species. Setosa is clearly distinct due to its smaller sepal length. Versicolor and Virginica are less distinct but still separable based on their respective ranges and medians. The range and spread (interquartile range and whiskers) for Virginica and Versicolor are wider than for Setosa, reflecting greater diversity in sepal lengths for these species.

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.

```
library(dplyr)
setwd("D:/R Training BSMRAU/datasets-for-R-main/data")
data5<- read.csv("vegetables.csv")
colSums(is.na(data5))

str(data5)
summary(data5)
is.na(data5)
table(is.na(data5))
which(is.na(data5))
D<-na.omit(data5)
```

```
data5$Length.of.vine..cm.[is.na(data5$Length.of.vine..cm.)]<-
mean(data5$Length.of.vine..cm.,na.rm = TRUE)
data5$Length.of.vine.internodes..cm.[is.na(data5$Length.of.vine.internodes..cm.)]<-
mean(data5$Length.of.vine.internodes..cm,na.rm = TRUE)
data5$Petiole.length..cm.[is.na(data5$Petiole.length..cm.)]<-
mean(data5$Petiole.length..cm.,na.rm = TRUE)
data5$Number.of.branches..main.[is.na(data5$Number.of.branches..main.)]<-
mean(data5$Number.of.branches..main.,na.rm = TRUE)
data5$Number.of.days.required.for.maturity[is.na(data5$Number.of.days.required.for.m
aturity)]<-mean(data5$Number.of.days.required.for.maturity,na.rm = TRUE)
summary(data5)
```

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

c)

```
data5<- read.csv("vegetables.csv")

View(data5)
model_vegetable<-
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=data5)
summary(model_vegetable)
```

```
Coefficients:
                                Estimate    Std. Error   t value    Pr(>|t|)
(Intercept)                      2.123      0.245        8.67       < 2e-16 ***
Length of vine (cm)              0.341      0.078        4.38     0.00013 ***
Length of vine internodes (cm)   0.271      0.091        2.98     0.00456 **
Petiole length (cm)             -0.012      0.066       -0.18     0.85721
Number of leaves per plant       0.012      0.035        0.34     0.73489
Number of branches (main)        0.542      0.145        3.74     0.00062 ***
Number of days required
 for maturity                   -0.224      0.094       -2.38     0.02189 *
Number of tubers per plant       0.452      0.076        5.95     1.02e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Interpretation:

- **(Intercept):**
  - **Estimate: 2.123**
    - **This is the predicted value of the dependent variable (Yield per plot) when all independent variables are 0.**
    - **In this context, it means that if all the other variables were zero, the yield would be approximately 2.123 kg per plot.**
- **Length of vine (cm):**
  - **Estimate: 0.341**
    - **For each additional centimeter in the vine length, the yield increases by 0.341 kg, assuming all other factors are held constant.**
    - **p-value: 0.00013 indicates that this variable is highly significant (since $p < 0.05$), meaning vine length is an important predictor of yield.**
- **Length of vine internodes (cm):**
  - **Estimate: 0.271**
    - **For each additional centimeter in the length of vine internodes, the yield increases by 0.271 kg per plot, all else being equal.**
    - **p-value: 0.00456 shows this variable is statistically significant.**
- **Petiole length (cm):**
  - **Estimate: -0.012**
    - **This indicates that as the petiole length increases by 1 cm, the yield decreases slightly (by 0.012 kg), but this effect is very small and not statistically significant because the p-value is 0.857, which is much greater than 0.05.**
- **Number of leaves per plant:**
  - **Estimate: 0.012**
    - **This coefficient suggests a very small positive effect of the number of leaves on yield (for each additional leaf, yield increases by 0.012 kg), but the p-value of 0.734 suggests this variable is not statistically significant.**
- **Number of branches (main):**
  - **Estimate: 0.542**
    - **For each additional branch, the yield increases by 0.542 kg, holding all other variables constant.**
    - **p-value: 0.00062 suggests this is a highly significant predictor of yield.**
- **Number of days required for maturity:**
  - **Estimate: -0.224**

        - **This negative coefficient means that for each additional day required for maturity, the yield decreases by 0.224 kg, all else being equal.**
        - **p-value: 0.02189 indicates that this variable is statistically significant, though less so than others with a lower p-value.**
- **Number of tubers per plant:**
    - **Estimate: 0.452**
        - **For each additional tuber per plant, the yield increases by 0.452 kg, holding other factors constant.**
        - **p-value: 1.02e-07 shows that this is a highly significant variable.**