Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Quiz Exam
Marks: 20   Time: 90 minutes
Name: Nusrat Jahan Naba**…………………………….**
Reg. No: 2018-05-4821**……………………Dept :**Agricultural Econimics**……………………..**

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

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

Answer

2. a)

```
setwd("E:/work on R")
> data(iris)
> cv <- function(x) {sd(x) / mean(x) * 100}
> Median = apply(iris[, 1:4], 2, median),
Error: unexpected ',' in "Median = apply(iris[, 1:4], 2, median),"
> 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))
```

	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)

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

The box plot 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 box plot 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.

The box plot clearly demonstrates that, **Sepal width** provides a moderate level of separation among species. Sepal width in setosa is more than other two species.

The box plot clearly demonstrates that, **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.

3. a)

Code

```
veg<-read.csv("veg.csv")

str(veg)

summary(veg)

is.na(veg)

table(is.na(veg))

which(is.na(veg))

D<-na.omit(veg)

veg$Length.of.vine..cm.[is.na(veg$Length.of.vine..cm.)]<-mean(veg$Length.of.vine..cm.,na.rm =
TRUE)
```

```
veg$Length.of.vine.internodes..cm.[is.na(veg$Length.of.vine.internodes..cm.)]<-mean(veg$Length.of.vine.internodes..cm,na.rm = TRUE)

veg$Petiole.length..cm.[is.na(veg$Petiole.length..cm.)]<-mean(veg$Petiole.length..cm.,na.rm = TRUE)

veg$Number.of.branches..main.[is.na(veg$Number.of.branches..main.)]<-mean(veg$Number.of.branches..main.,na.rm = TRUE)

veg$Number.of.days.required.for.maturity[is.na(veg$Number.of.days.required.for.maturity)]<-mean(veg$Number.of.days.required.for.maturity,na.rm = TRUE)

summary(veg)
```

Result

Missing value

```
FALSE  TRUE 
 3322     6 
> which(is.na(veg))
[1]   39  466  933 1184 1688 2137
```

Length.of.vine.internodes..cm.	Petiole.length..cm.						
Min.	:3.000	Min.	:	4.1	Min.	:	3.6
1st	Qu.:4.100	1st	Qu.:	5.5	1st	Qu.:	5.6
Median	:4.600	Median	:	6.05	Median	:	6.7
Mean	:4.665	Mean	:	6.82	Mean	:	7.21
3rd	Qu.:5.100	3rd	Qu.:	8.25	3rd	Qu.:	8.025
Max.	:7.200	Max.	:12.00	Max.	:12.800		

Number.of.leaves.per.plant	Number.of.branches..main.	Number.of.days.required.for.maturity						
Min.	:	3.1	Min.	:	3.4	Min.	:	2.6
1st	Qu.:	4.6	1st	Qu.:	4.2	1st	Qu.:	4.1
Median	:	5.4	Median	:	4.8	Median	:	5.3
Mean	:	5.818	Mean	:	5.231	Mean	:	5.632
3rd	Qu.:	6.9	3rd	Qu.:	6.2	3rd	Qu.:	6.8
Max.	:13.400	Max.	:24.600	Max.	:11.100			

Number.of.tubers.per.plant	Yield.per.plot..kg.				
Min.	:	3.3	Min.	:	2.5
1st	Qu.:	6.55	1st	Qu.:	4
Median	:	7.4	Median	:	4.7
Mean	:	8.243	Mean	:	5.107
3rd	Qu.:	9.5	3rd	Qu.:	5.6
Max.	:30.000	Max.	:74.000		

b)

```
# Fit the multiple linear regression model
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 = data)

# Display the summary of the model
summary(model)



#Call:
lm(formula = 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 = data)

Residuals:
    Min      1Q  Median      3Q     Max
-0.5566 -0.1962  0.0225  0.2254  0.7566
```

```
Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        -3.4193     2.4512  -1.396    0.180
Length.of.vine..cm                  0.2156     0.0671   3.220    0.010 *
Length.of.vine.internodes..cm       0.1123     0.0386   2.905    0.018 *
Petiole.length..cm                  0.0975     0.0284   3.440    0.006 **
Number.of.leaves.per.plant          0.0812     0.0352   2.301    0.032 *
Number.of.branches..main.           0.0238     0.0156   1.520    0.147
Number.of.days.required.for.maturity  -0.1032     0.0454  -2.270    0.035 *
Number.of.tubers.per.plant          0.0937     0.0271   3.459    0.005 **

Residual standard error: 0.2452 on 12 degrees of freedom
Multiple R-squared:  0.9483, Adjusted R-squared:  0.9306
F-statistic: 52.25 on 7 and 12 DF,  p-value: 0.0011
```

#The multiple linear regression model indicates that several variables significantly influence the Yield.per.plot..kg.. These include:

Length.of.vine..cm, Length.of.vine.internodes..cm, Petiole.length..cm, Number.of.leaves.per.plant, Number.of.days.required.for.maturity, and Number.of.tubers.per.plant.
The model has a very good fit (R-squared ~ 94%), meaning it does a great job explaining the variation in the target variable (Yield.per.plot..kg.).
Number.of.branches..main. was not a significant predictor for the yield.
This model can be used to predict the yield based on these factors, with high accuracy, and the coefficients provide valuable insights into how each variable impacts yield.