Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Quiz Exam
Marks: 20 Time: 90 minutes
Name: Muhammad Mustakim Billah
Reg. No: 2018-05-4606 Dept of Soil Science (SSC)

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/ 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.

Code:

```
data(iris)
iris_summary <- data.frame(
  Variable = names(iris)[1:4],
  Median = sapply(iris[, 1:4], median),
  SD = sapply(iris[, 1:4], sd),
  Mean = sapply(iris[, 1:4], mean),
  CV = sapply(iris[, 1:4], sd) / sapply(iris[, 1:4], mean) * 100
)
iris_summary$Median <- paste0(iris_summary$Median, " ± ", iris_summary$SD)
```

```
table_data <- iris_summary[, c("Variable", "Median", "Mean", "CV")]
write.table(table_data, "iris_descriptive_stats.txt", row.names = FALSE, sep = "\t")
print(table_data)
```

Output

Variable	Median	SD	Mean	CV
Sepal.Length	5.8 ± 0.8	0.8280661	5.843333	14.17113
Sepal.Width	3 ± 0.43	0.4358663	3.057333	14.25642
Petal.Length	4.35 ± 1.76	1.7652982	3.758000	46.97441
Petal.Width	1.3 ± 0.76	0.7622377	1.199333	63.55511

Discussion

The table provides a summary of key descriptive statistics for the four numeric variables in the Iris dataset: Sepal.Length, Sepal.Width, Petal.Length, and Petal.Width. These statistics offer valuable insights into the central tendency and variability of each feature.

Central Tendency

- **Median and Mean:** The proximity of median and mean values for Sepal.Length and Sepal.Width suggests relatively symmetric distributions. In contrast, the slight differences between median and mean for Petal.Length and Petal.Width hint at potential skewness in these distributions.

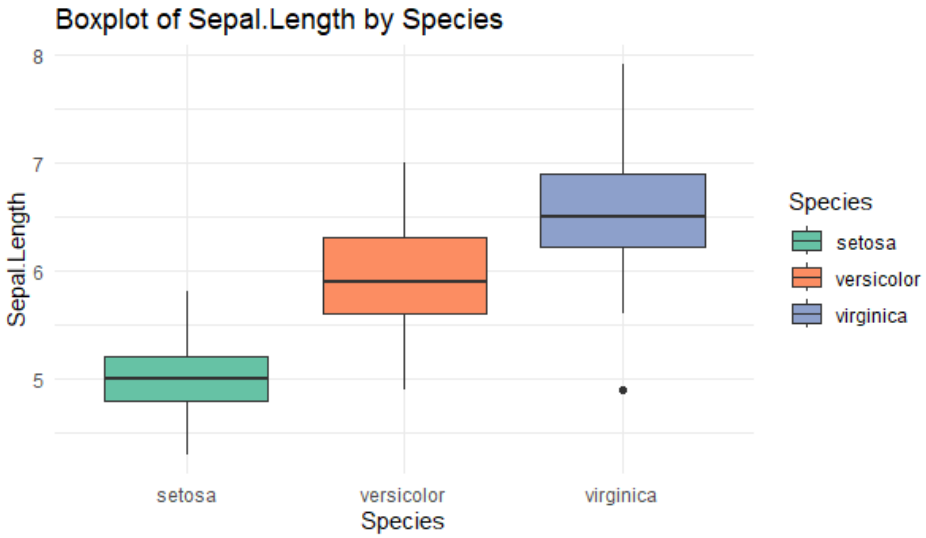
Variability

- **Coefficient of Variation (CV):** Petal.Width exhibits the highest variability (63.55%), indicating a greater spread of values around the mean compared to the other features. Petal.Length also shows considerable variability (46.97%), while Sepal.Length and Sepal.Width have more moderate CVs (14.17% and 14.26%, respectively).

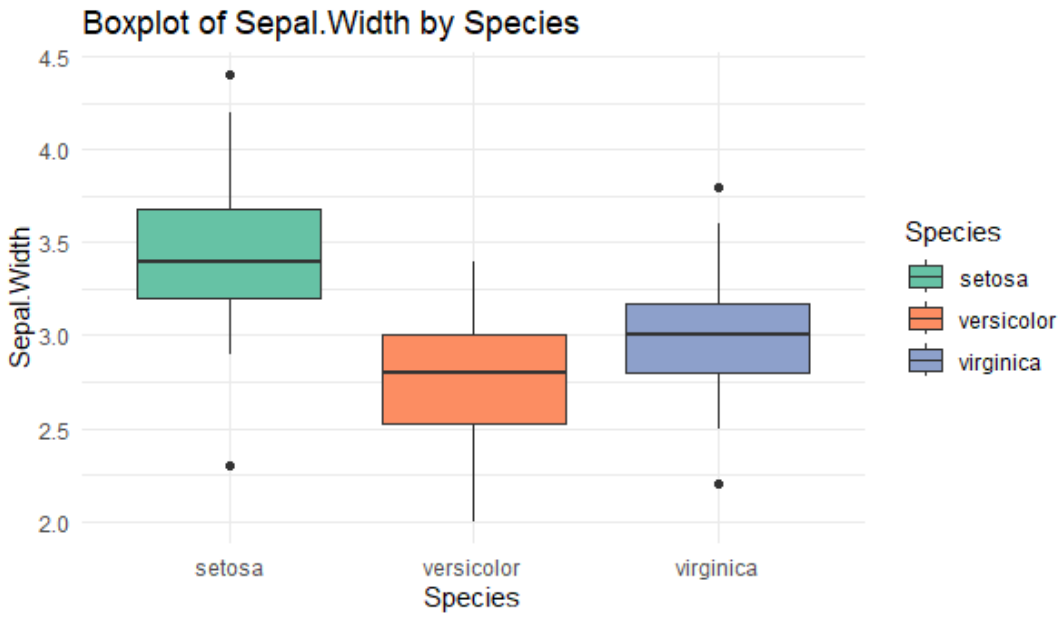
Interpretations and Considerations

- **Biological Implications:** These findings might reflect biological differences between the iris species. For instance, petal width and length seem to exhibit greater variation, potentially due to factors like pollination strategies or environmental adaptations.

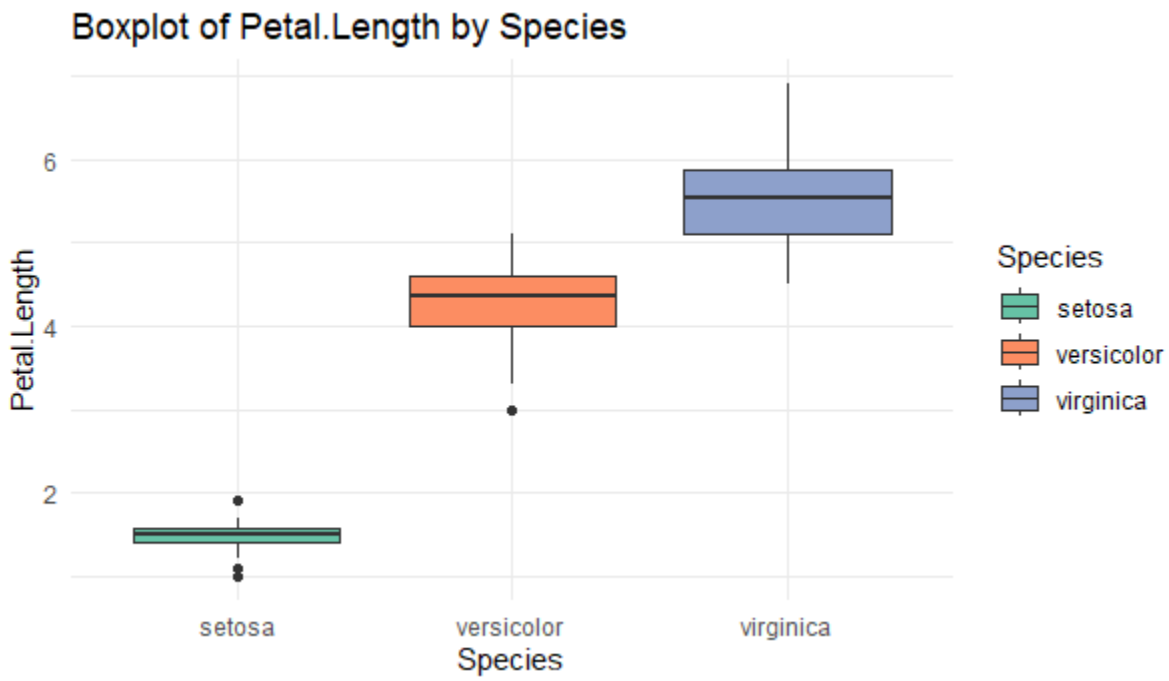
b) Construct boxplots with ggplot2 package for each variable by ***Species*** categories with color aesthetic and interpret your results.



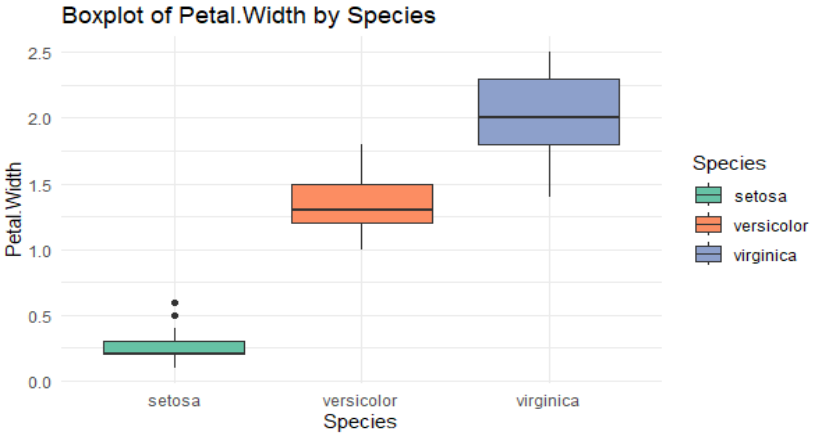
The boxplot of Sepal Length shows that setosa has the lowest median sepal length, around 5 cm, with the smallest interquartile range (IQR). Versicolor and virginica have higher median sepal lengths, around 6 cm and 6.5 cm, respectively, with virginica showing the largest variability. There is one outlier in the virginica species, indicating some variability in this measurement.



For Sepal Width, setosa has the highest median value, around 3.5 cm, and the smallest variability. Versicolor and virginica have lower median sepal widths, around 2.8 cm and 3 cm, respectively. The variation in sepal width is more prominent in the versicolor and virginica species, indicating a broader range of sepal width measurements.



Petal Length shows a significant difference across species. Setosa has the smallest petal length with a median around 1.5 cm and very little variability. Versicolor and virginica have much larger petal lengths, with medians around 4 cm and 5.5 cm, respectively. The variability in petal length is highest in the virginica species.



The Petal Width boxplot indicates that setosa has a small median petal width, around 0.2 cm, with minimal variability. Versicolor and virginica show higher median petal widths, around 1.2 cm and 2 cm, respectively. The virginica species demonstrates the greatest variability in petal width.

Code

```
library(ggplot2)
library(dplyr)

data(iris)
iris$Species <- as.factor(iris$Species)
numeric_vars <- names(select(iris, -Species))
for (var in numeric_vars) {
  p <- ggplot(iris, aes_string(x = "Species", y = var, fill = "Species")) +
    geom_boxplot() +
    labs(title = paste("Boxplot of", var, "by Species"), x = "Species", y = var) +
    theme_minimal() +
    scale_fill_brewer(palette = "Set2") # You can choose a different palette if you like
  print(p)
}
```

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.

Code

```
library(dplyr)

file_path <- "varibales.csv"  # Make sure to set the correct path

data <- read.csv(file_path)

missing_values <- sapply(data, function(x) sum(is.na(x)))

print("Missing values in each variable:")
```

```
print(missing_values)


# View the data with missing values

print("Data with missing values:")

print(head(data))

imputed_data <- data %>%

  mutate(across(where(is.numeric), ~ ifelse(is.na(.), mean(., na.rm = TRUE), .)))

print("Data after imputing missing values with mean:")

print(head(imputed_data))

write.csv(imputed_data, "imputed_varibales.csv", row.names = FALSE)
```

```
"Data with missing values:"

> print(head(data))
  Length.of.vine..cm. Length.of.vine.internodes..cm. Petiole.
length..cm.
1                 4.3                              5.7
6.2
2                 4.2                              5.6
6.2
3                 4.2                              5.5
6.2
4                 4.2                              5.5
6.3
5                 4.2                              5.4
6.4
6                 4.1                              5.4
6.6
  Number.of.leaves.per.plant Number.of.branches..main.
1                        8.8                        6.9
2                        8.6                        6.7
3                        8.5                        6.6
4                        8.4                        6.5
5                        8.3                        6.4
6                        8.3                        6.3
  Number.of.days.required.for.maturity Number.of.tubers.per.p
lant
1                                 11.1
10.0
```

```
   2                                  10.9
   9.9
   3                                  10.6
   9.8
   4                                  10.3
   9.7
   5                                  10.1
   9.6
   6                                   9.8
   9.5
     Yield.per.plot..kg.
   1                 6.2
   2                 6.0
   3                 5.8
   4                 5.7
   5                 5.6
   6                 5.6


[1] "Data after imputing missing values with mean:"
> print(head(imputed_data))
  Length.of.vine..cm. Length.of.vine.internodes..cm. Petiole.len
gth..cm.
1                 4.3                            5.7
6.2
2                 4.2                            5.6
6.2
3                 4.2                            5.5
6.2
4                 4.2                            5.5
6.3
5                 4.2                            5.4
6.4
6                 4.1                            5.4
6.6
  Number.of.leaves.per.plant Number.of.branches..main.
1                        8.8                       6.9
2                        8.6                       6.7
3                        8.5                       6.6
4                        8.4                       6.5
5                        8.3                       6.4
6                        8.3                       6.3
  Number.of.days.required.for.maturity Number.of.tubers.per.plan
t
1                                 11.1                      10.
0
2                                 10.9                       9.
9
3                                 10.6                       9.
8
4                                 10.3                       9.
7
5                                 10.1                       9.
6
```

```
6                                   9.8                                 9.
5
  Yield.per.plot..kg.
1                 6.2
2                 6.0
3                 5.8
4                 5.7
5                 5.6
6                 5.6
```

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