Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-06
Project Report Marks: 25
Name: Nahrin Jannat Mim
Reg. No: 2023-11-6928 **Dept:** Plant Pathology

Note: Submit the completed file as pdf to nazmol.stat.bioin@bsmrau.edu.bd and rabiulauwul@bsmrau.edu.bd with subject: *EDGE_06_Project_Your registration number_ Department by 26th of December, 2024.*

Problem# 1:

A split-plot design was conducted considering tree blocks, three levels/treatments of variety in the main plot, and five levels/treatments of nitrogen in the split-plot. Afterward, the yield of certain plant characteristics was observed. The data regarding this experiment were given in the file "Split_Plot_Design". Answer the following question using this data.

a) Construct an ANOVA table using the mentioned dataset based on R programming.

```
# Load necessary libraries
library(ggplot2)
library(lme4)
library(emmeans)
library(datasets)

# Load the dataset
data <- read.csv("Split_Plot_Design.csv")

# Convert factors to categorical variables
data$REPLICAT <- as.factor(data$REPLICAT)
data$VARIETY <- as.factor(data$VARIETY)
data$NITROGEN <- as.factor(data$NITROGEN)

# Fit the split-plot model
anova_model <- aov(YIELD ~ VARIETY + Error(REPLICAT/VARIETY) + NITROGEN
+ VARIETY:NITROGEN, data = data)

# Display the ANOVA table
summary(anova_model)
```

```
                  Df Sum Sq Mean Sq F value   Pr(>F)
VARIETY            2   1.93   0.963  11.670 0.000178 ***
NITROGEN           4  66.03  16.507 200.070  < 2e-16 ***
VARIETY:NITROGEN   8   6.10   0.763   9.244 2.54e-06 ***
```

Residuals 30 2.48 0.083

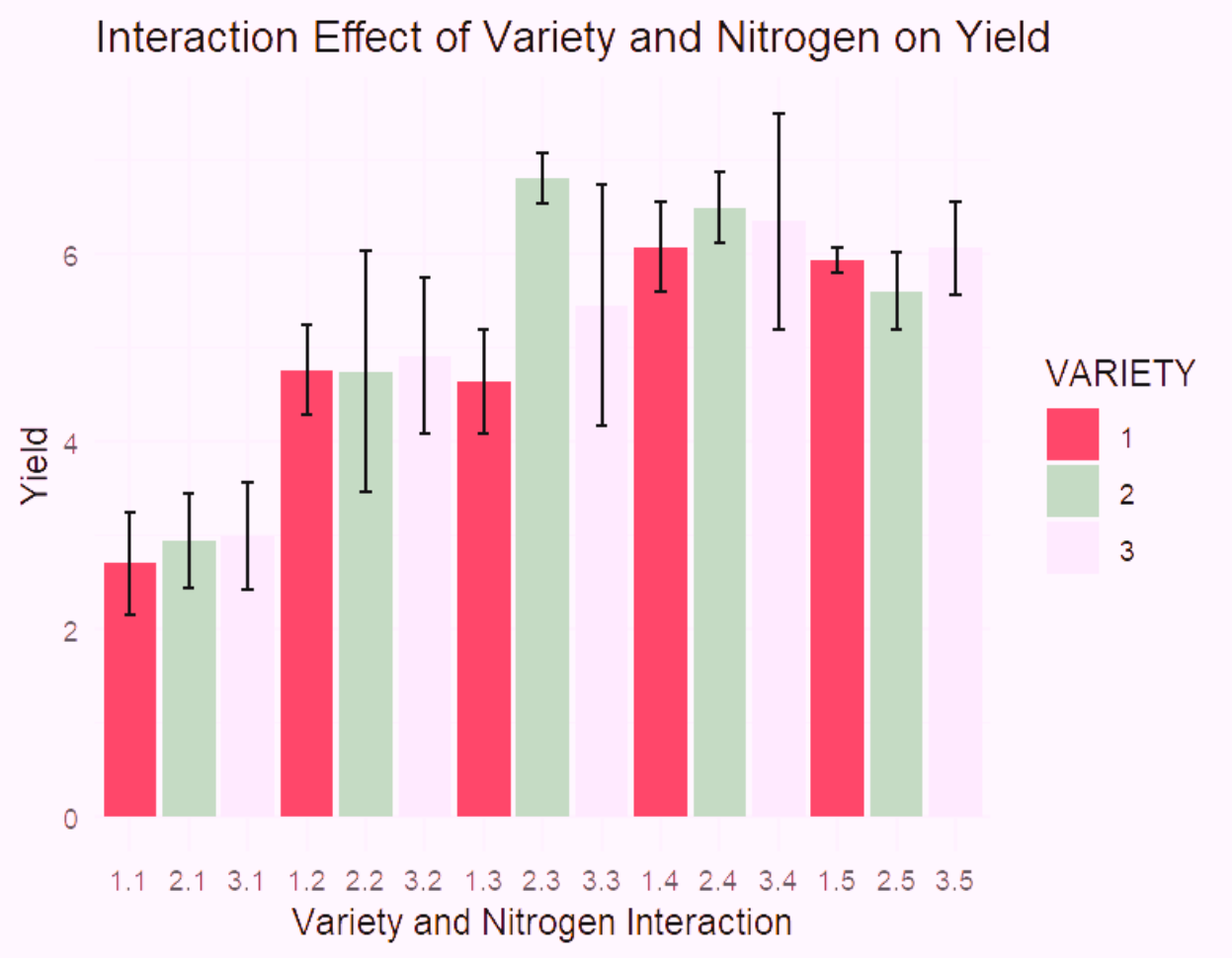
b) Write down the null hypothesis of all possible effects and interpret the results based on the ANOVA table.

Main Effects and Interaction Analysis

1. **Main Effect of VARIETY**:
   - Null Hypothesis (H0H_0H0): All tree varieties have the same mean yield.
   - FFF-value: 11.670, ppp-value: 0.000178 (highly significant at α=0.05\alpha = 0.05α=0.05).
   - **Interpretation**: The tree variety significantly affects yield, as the null hypothesis is rejected. Different varieties have different mean yields.
2. **Main Effect of NITROGEN**:
   - Null Hypothesis (H0H_0H0): All nitrogen treatments have the same mean yield.
   - FFF-value: 200.070, ppp-value: < 2e-16 (extremely significant).
   - **Interpretation**: The nitrogen treatment significantly affects yield. Different nitrogen levels result in significantly different yields.
3. **Interaction Effect (VARIETY × NITROGEN)**:
   - Null Hypothesis (H0H_0H0): The effect of nitrogen treatments is consistent across all varieties.
   - FFF-value: 9.244, ppp-value: 2.54e-06 (highly significant).
   - **Interpretation**: There is a significant interaction between tree variety and nitrogen treatments. This means the impact of nitrogen levels on yield varies depending on the variety.
4. **Residuals**:
   - Residual variance indicates unexplained variability in the model. The Mean Square Error (MSE) for residuals is 0.083.

c) Perform a post-hoc test for the interaction effect (variety × nitrogen) and draw a bar diagram with lettering.

```
emmeans_interaction <- emmeans(anova_model, ~ VARIETY * NITROGEN)
pairs(emmeans_interaction)

ggplot(data, aes(x = interaction(VARIETY, NITROGEN), y = YIELD, fill = VARIETY)) +
  stat_summary(fun = mean, geom = "bar", position = position_dodge()) +
  stat_summary(fun.data = mean_cl_normal, geom = "errorbar", position = position_dodge(0.9), width = 0.2) +
  labs(x = "Variety and Nitrogen Interaction", y = "Yield", title = "Interaction Effect of Variety and Nitrogen on Yield") +
  theme_minimal()
```



Problem# 2:

a) What is principal component analysis?

Principal Component Analysis (PCA) is a statistical technique used to simplify complex datasets by reducing their dimensions while retaining most of the original information. It transforms correlated variables into a smaller number of uncorrelated variables called **principal components**, which capture the maximum variance in the data.

b) What are the main purposes of principle component analysis in your study area?

In **Plant Pathology**, **Principal Component Analysis (PCA)** is used for:

1. **Dimensionality Reduction**: Simplifies complex datasets by reducing variables while retaining key information.
2. **Data Visualization**: Helps visualize high-dimensional data, revealing trends and patterns in plant diseases or pathogens.
3. **Pattern Recognition**: Identifies underlying patterns for disease classification and pathogen identification.
4. **Identifying Key Variables**: Highlights critical factors influencing plant disease, guiding research and interventions.
5. **Noise Reduction**: Reduces the impact of irrelevant data and measurement errors.

6. **Exploratory Data Analysis**: Discovers hidden relationships and guides further research. PCA aids in understanding complex disease data and improving disease management strategies.

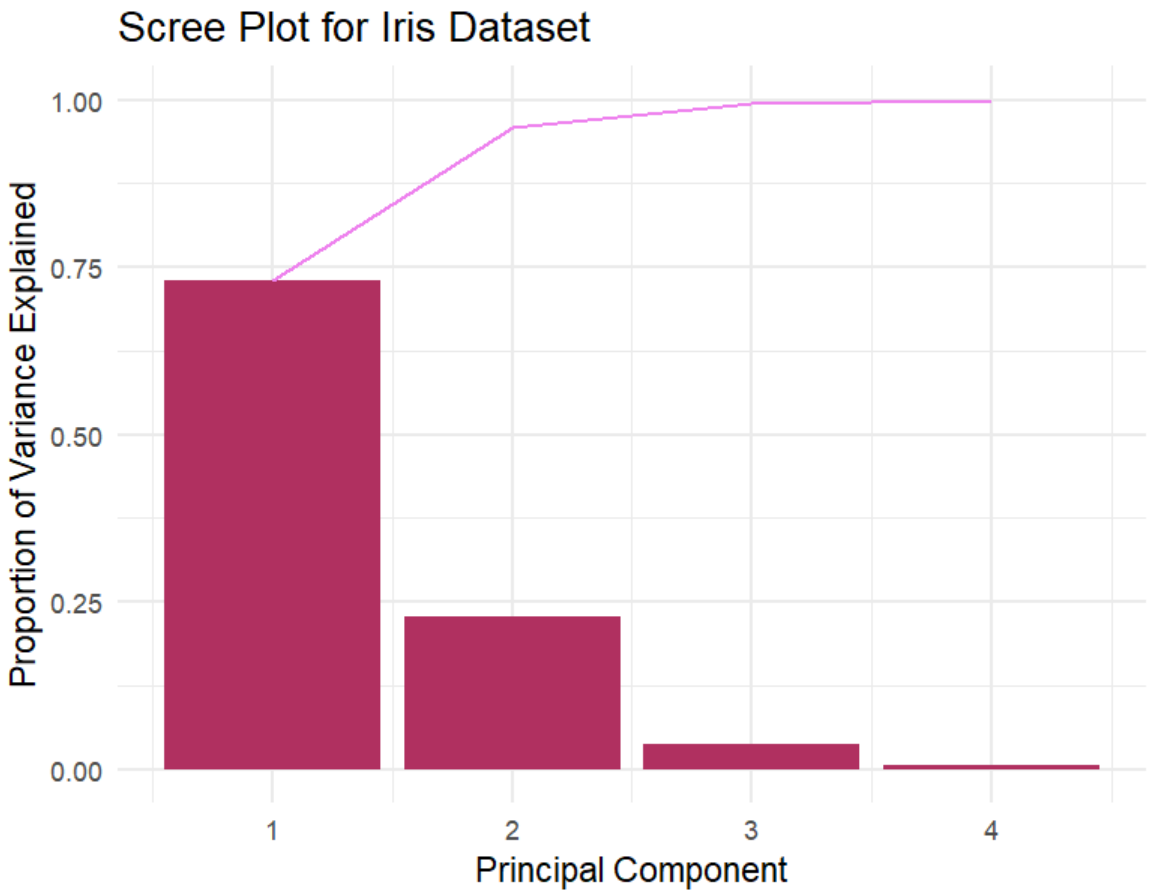
c) Compute the eigenvalue and eigenvector using the iris data based on R programming.

```
iris_data <- iris[, 1:4]
iris_scaled <- scale(iris_data)

iris_cov <- cov(iris_scaled)
iris_eigen <- eigen(iris_cov)

iris_eigen_values <- iris_eigen$values
iris_eigen_vectors <- iris_eigen$vectors
print("Eigenvalues:")
print(iris_eigen_values)
print("Eigenvectors:")
print(iris_eigen_vectors)
```

```
"Eigenvalues:"
[1] 2.91849782 0.91403047 0.14675688 0.02071484
[1] "Eigenvectors:"
           [,1]       [,2]       [,3]       [,4]
[1,]  0.5210659 -0.37741762  0.7195664  0.2612863
[2,] -0.2693474 -0.92329566 -0.2443818 -0.1235096
[3,]  0.5804131 -0.02449161 -0.1421264 -0.8014492
[4,]  0.5648565 -0.06694199 -0.6342727  0.5235971
```

d) Construct a scree plot and interpret how many principal components should be retained to interpret the iris dataset.

```
scree_plot <- data.frame(Principal_Component = 1:length(iris_eigen_values),
                         Variance_Explained = iris_eigen_values / sum(iris_eigen_values))
ggplot(scree_plot, aes(x = Principal_Component, y = Variance_Explained)) +
  geom_bar(stat = "identity", fill = "skyblue") +
  geom_line(aes(y = cumsum(Variance_Explained)), color = "red", group = 1) +
  labs(title = "Scree Plot for Iris Dataset", x = "Principal Component", y = "Proportion of
Variance Explained") +
  theme_minimal()
```

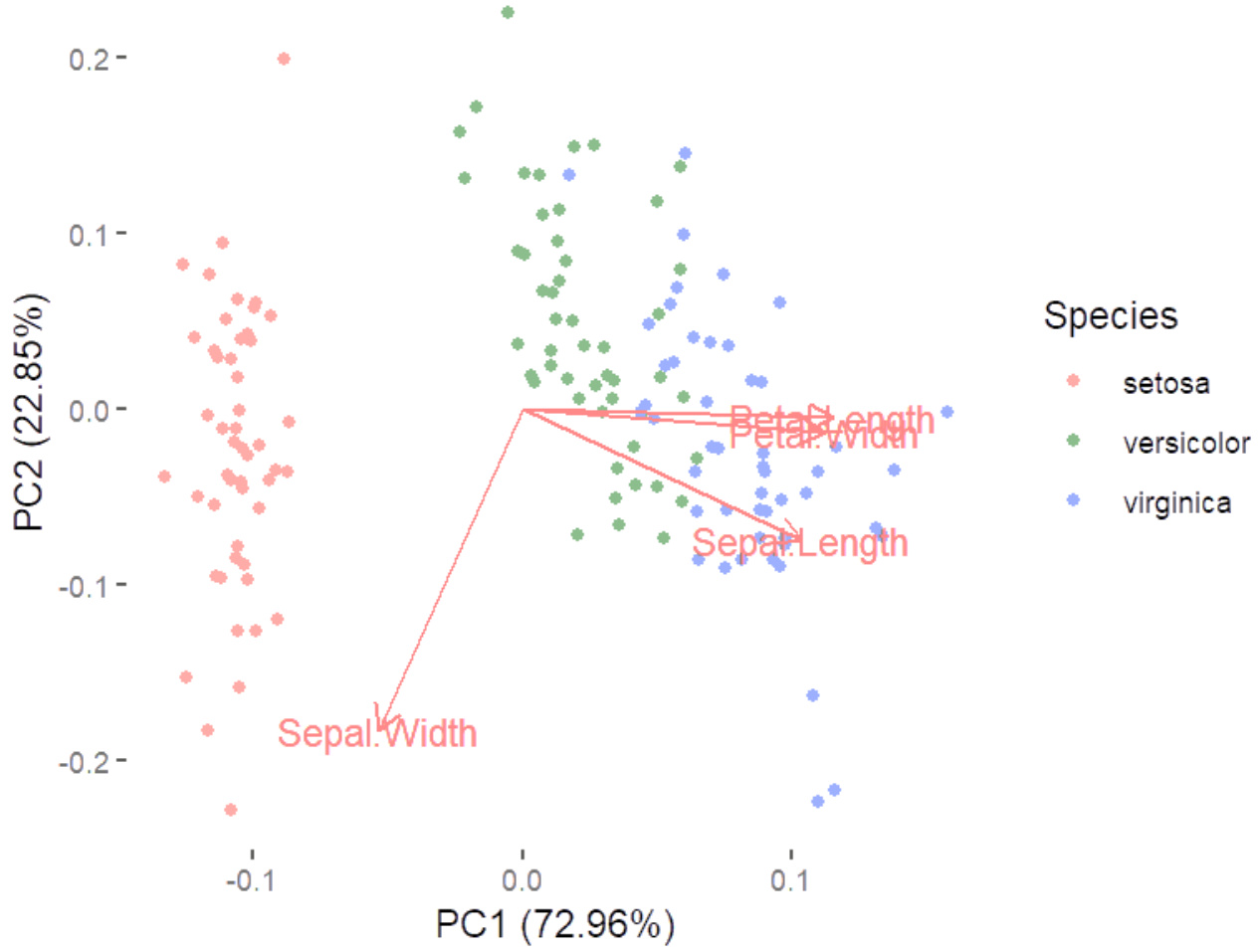


Interpretation:

- Observe clusters of points to identify natural groupings (e.g., species in the Iris dataset).
- Examine the direction and magnitude of arrows to determine which variables influence specific principal components. Variables pointing in the same direction are positively correlated, while those in opposite directions are negatively correlated.
- Check how species or other categorical factors are separated along the principal components to understand major sources of variation.

e) Construct a bi-plot for the iris data based on R programming and interpret the results.

```
library(ggplot2)
library(ggfortify)

data(iris)
iris_pca <- prcomp(iris[, 1:4], center = TRUE, scale. = TRUE)
summary(iris_pca)

autoplot(iris_pca, data = iris, colour = 'Species', loadings = TRUE, loadings.label = TRUE)
```



Interpretation:

- **Observe clusters of points to identify natural groupings (e.g., species in the Iris dataset).**
- **Examine the direction and magnitude of arrows to determine which variables influence specific principal components. Variables pointing in the same direction are positively correlated, while those in opposite directions are negatively correlated.**
- **Check how species or other categorical factors are separated along the principal components to understand major sources of variation.**