Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-06
Project Report Marks: 25
Name: Tunazzina Mahedin Tazin
Reg. No: 21-05-5978 **Dept:** AERD

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.

```
library(ggplot2) library(lme4)
library(emmeans)

data <- read.csv("Split_Plot_Design.csv") data$REPLICAT
<- as.factor(data$REPLICAT) data$VARIETY <-
as.factor(data$VARIETY) data$NITROGEN <-
as.factor(data$NITROGEN)

anova_model <- aov(YIELD ~ VARIETY * NITROGEN, data = data)

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.

Null Hypotheses for All Effects

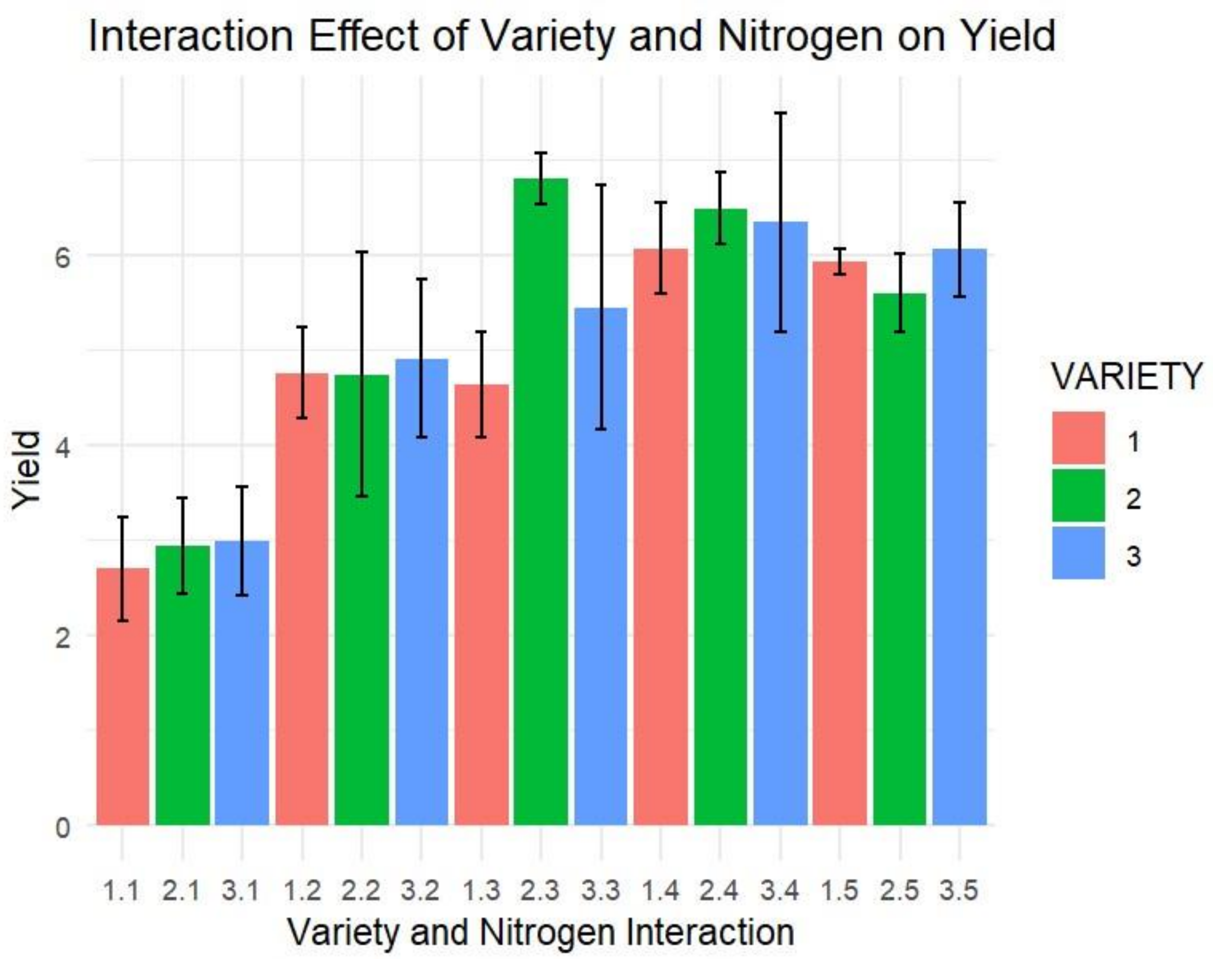
1. **Main effect of VARIETY**: H0: All varieties have the same mean yield.
2. **Main effect of NITROGEN**: H0: All nitrogen treatments have the same mean yield.
3. **Interaction effect (VARIETY × NITROGEN)**: H0: The effect of nitrogen treatments is consistent across all varieties.

Interpretations

1. **Main Effect of VARIETY**:
   Since the F-value (11.670) is significant with p=0.000178, we reject the null hypothesis that all three varieties have the same mean yield. This indicates that the yield differs significantly across the varieties.
2. **Main Effect of NITROGEN**:
   The F-value (200.070) with a very small p-value (< 2e-16) leads us to reject the null hypothesis that all nitrogen levels have the same mean yield. This demonstrates a strong effect of nitrogen treatments on the yield. 3. **Interaction Effect (VARIETY × NITROGEN)**:
   With an F-value of 9.244 and p=2.54e−06, the interaction effect is highly significant. We reject the null hypothesis and conclude that the effect of nitrogen levels on yield depends on the tree variety.

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 reduce the dimensionality of a dataset while retaining as much variance as possible. It identifies new, uncorrelated variables (principal components) that are linear combinations of the original variables. These components are ordered such that the first few retain most of the variation in the data.

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

Principal Component Analysis (PCA) is a statistical technique widely used in agricultural economics for various purposes, particularly when dealing with large datasets with multiple variables. In the context of using PCA through R programming, the main purposes include:

1. **Dimensionality Reduction**

- Agricultural economic datasets often involve multiple variables (e.g., input costs, crop yields, market prices, weather conditions). PCA helps reduce the

  dimensionality of the data by summarizing the information in a smaller set of principal components (PCs) while retaining most of the variability.
  - This is particularly useful for simplifying complex datasets and avoiding multicollinearity when performing regression or econometric analyses.

2. **Identifying Key Factors**
   - In studies of farm productivity, PCA can determine which variables (e.g., fertilizer use, labor hours, or land quality) contribute the most to productivity differences across farms.
   - In market analysis, it can identify dominant factors influencing price variations, such as supply, demand, or external shocks.
3. **Visualization of Data**
   - PCA is used to visualize high-dimensional data in 2D or 3D plots (e.g., scatter plots of the first two principal components). This helps researchers identify patterns, clusters, or outliers in agricultural economic data.
   - For example, clusters of regions with similar agricultural practices or economic outputs can be detected.
4. **Data Interpretation and Classification**
   - PCA groups correlated variables into components, which can be interpreted as underlying constructs (e.g., socioeconomic conditions, farm management practices). This aids in creating indices or classification models.
   - For instance, it can help classify farmers into groups based on their socioeconomic status, market access, or risk profiles.
5. **Preprocessing for Regression or Machine Learning**
   - PCA can be used as a preprocessing step before applying regression, machine learning models, or econometric techniques. It helps to:
   - Eliminate redundancy in correlated variables.
   - Reduce noise and improve the performance of predictive models.
6. **Evaluating Regional or Temporal Variations**
   - PCA can identify spatial or temporal patterns in agricultural economics data, such as:
   - Regional differences in input use and productivity.
   - Temporal changes in economic variables like crop prices or production costs.

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

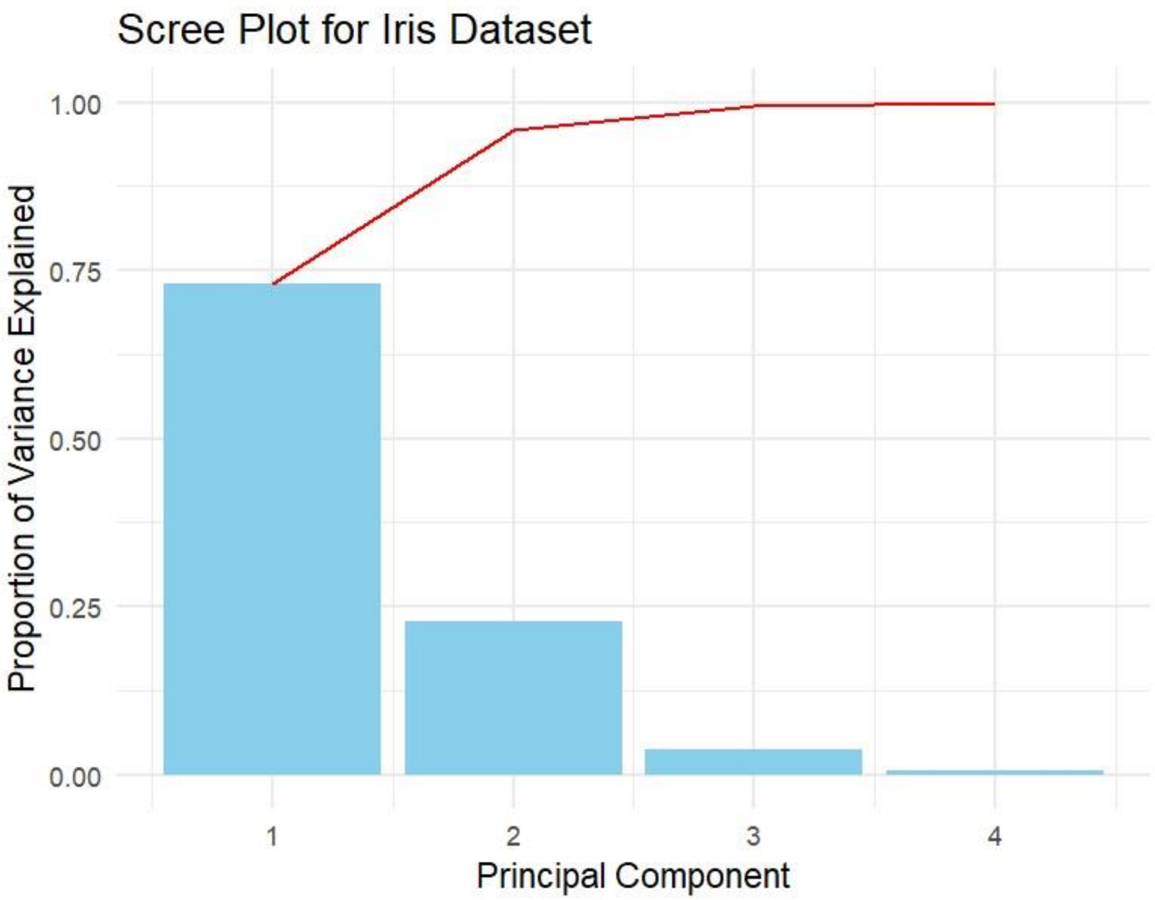
```
iris_data <- iris[, 1:4]   iris_scaled
<- scale(iris_data)

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()
```

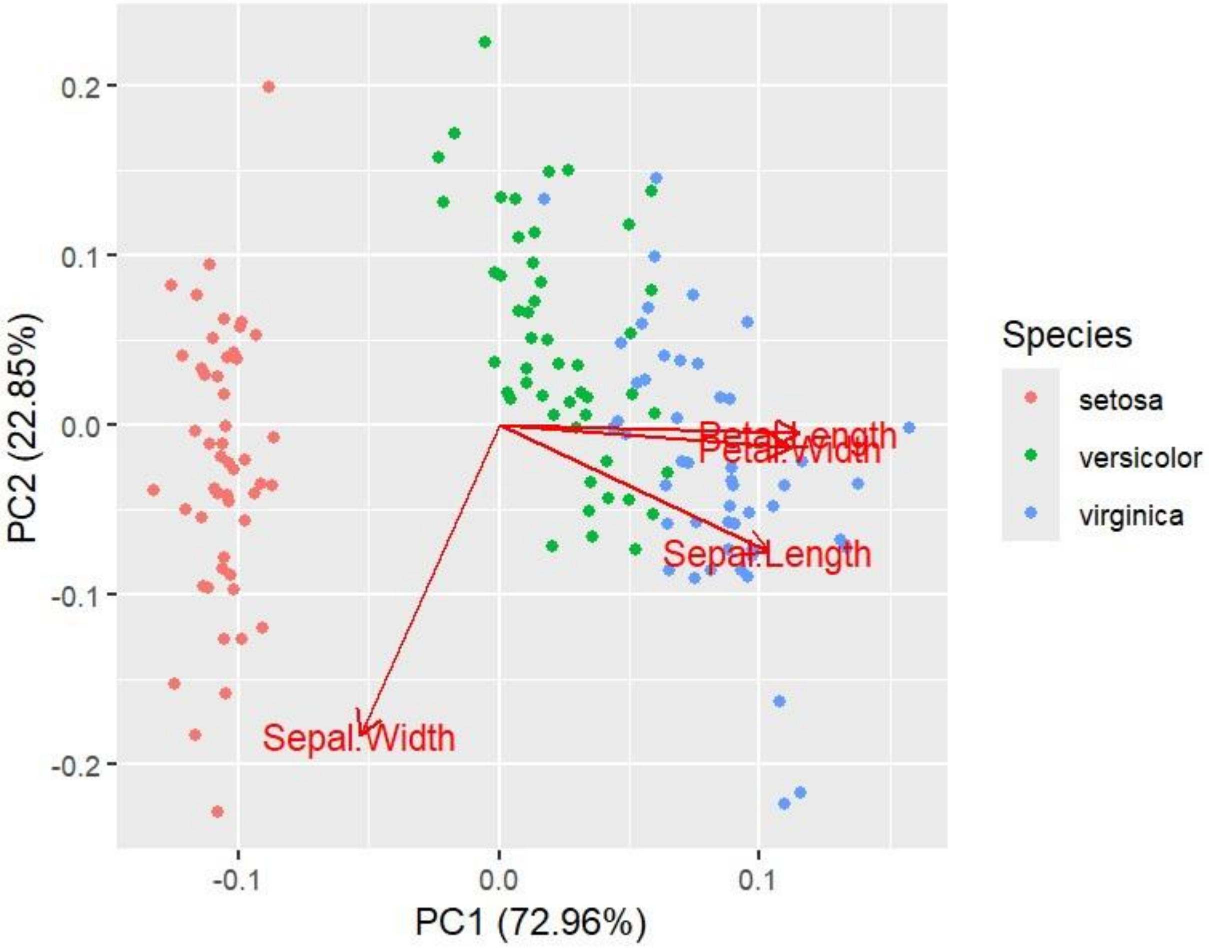


Interpretation:

1. **Principal Component 1 (PC1)**: This component explains the largest proportion of variance, approximately 75% of the total. It captures the most important patterns or variability in the dataset.
2. **Principal Component 2 (PC2)**: This component accounts for about 20% of the variance, adding to the cumulative explained variance. Together, PC1 and PC2 explain around 95% of the total variance in the dataset.
3. **Principal Component 3 (PC3) and PC4**: These components explain very little variance (less than 5% combined) and contribute minimally to the dataset's overall variability. Based on the "elbow" in the scree plot (where the slope sharply decreases), it is evident that the first two principal components are sufficient to retain most of the variance. Retaining PC1 and PC2 will simplify the dataset while preserving nearly all of its important information.

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:

1. **Principal Components (PC1 and PC2):**
   - **PC1 (72.96%)** explains the largest amount of variance in the dataset (approximately 73%).
   - **PC2 (22.85%)** explains the second largest variance (approximately 23%).
   - Together, PC1 and PC2 account for 95.81% of the total variance, making the 2D plot a good summary of the data.
2. **Clusters of Points:**

   - **Setosa (Red points):** Forms a distinct cluster, clearly separated from Versicolor and Virginica. This indicates that Setosa is significantly different from the other two species in terms of the measured variables.

   - **Versicolor (Green points) and Virginica (Blue points):** These clusters overlap, indicating that these two species are less distinct and have some similarities in their features.

3. **Loadings (Red Arrows):**

   - The arrows represent the contribution of each original variable to the principal components.

     - **Petal Length** and **Petal Width:** These variables have long arrows pointing in similar directions, indicating they are strongly correlated and contribute significantly to PC1.

     - **Sepal Length:** Contributes moderately to PC1 but less to PC2.

     - **Sepal Width:** Points in a different direction, showing it is not strongly correlated with other variables and contributes primarily to PC2.

   - The direction of the arrows shows the relationship with the principal components:

     - **Petal variables (length and width)** are more aligned with PC1, indicating these drive most of the variance captured in PC1.

     - **Sepal Width** is aligned with PC2, indicating it explains variance orthogonal to PC1.

4. **Species Differentiation:**

   - **Setosa:** The distinct clustering of Setosa along PC1 suggests that Petal Length and Petal Width are key features distinguishing it from Versicolor and Virginica.

   - **Versicolor and Virginica:** Overlapping points indicate these species are harder to differentiate based on the given features