Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Project Report Marks: 25
Name: Deepak Thakur
Reg. No:2019-05-5046 Dept: Entomology

Note: Submit the completed file as pdf to nazmol.stat.bioin@bsmrau.edu.bd and rabiulauwul@bsmrau.edu.bd with subject: *EDGE_11_Project_Your registration number_ Department by 13th of January, 2025.*

Problem# 1: Choose a multivariate dataset (with at least 10 variables) in your subject area and solve the following issue. (***Attach your dataset in csv file to the email***)

a) Pre-process your dataset with imputing outliers and missing values.

Ans:

```
library(dplyr)
data <- read.csv("ent.data.csv")
colSums(is.na(data))
numeric_columns <- sapply(data, is.numeric)
data[numeric_columns] <- lapply(data[numeric_columns], function(col) {
  col[is.na(col)] <- median(col, na.rm = TRUE)
  return(col)})
# for outliers
find_outliers <- function(col) {
  Q1 <- quantile(col, 0.25, na.rm = TRUE)
  Q3 <- quantile(col, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
```

```
  upper_bound <- Q3 + 1.5 * IQR
  return(which(col < lower_bound | col > upper_bound)) }
outlier_indices <- lapply(data[numeric_columns], find_outliers)
sapply(outlier_indices, length)
data[numeric_columns] <- lapply(data[numeric_columns], function(col) {
  Q1 <- quantile(col, 0.25, na.rm = TRUE)
  Q3 <- quantile(col, 0.75, na.rm = TRUE)
  IQR <- Q3 - Q1
  lower_bound <- Q1 - 1.5 * IQR
  upper_bound <- Q3 + 1.5 * IQR
  col[col < lower_bound | col > upper_bound] <- NA
  col[is.na(col)] <- median(col, na.rm = TRUE)
  return(col) })
colSums(is.na(data))
```

b) Interpret how many principle components should be retained for your data with justification.

Ans:

```
data<-read.csv("ent.data.csv")
data<-na.omit(data[,-5])
co<-cor(data)
mean(co)
dim(co)
eigen(co)
```

```
prcomp(data,scale. = TRUE)

library("devtools")

install_github("vqv/ggbiplot")

pca<-prcomp(data, scale.= TRUE)

pca$x

ggscreeplot(pca)+aes(color="green")
```

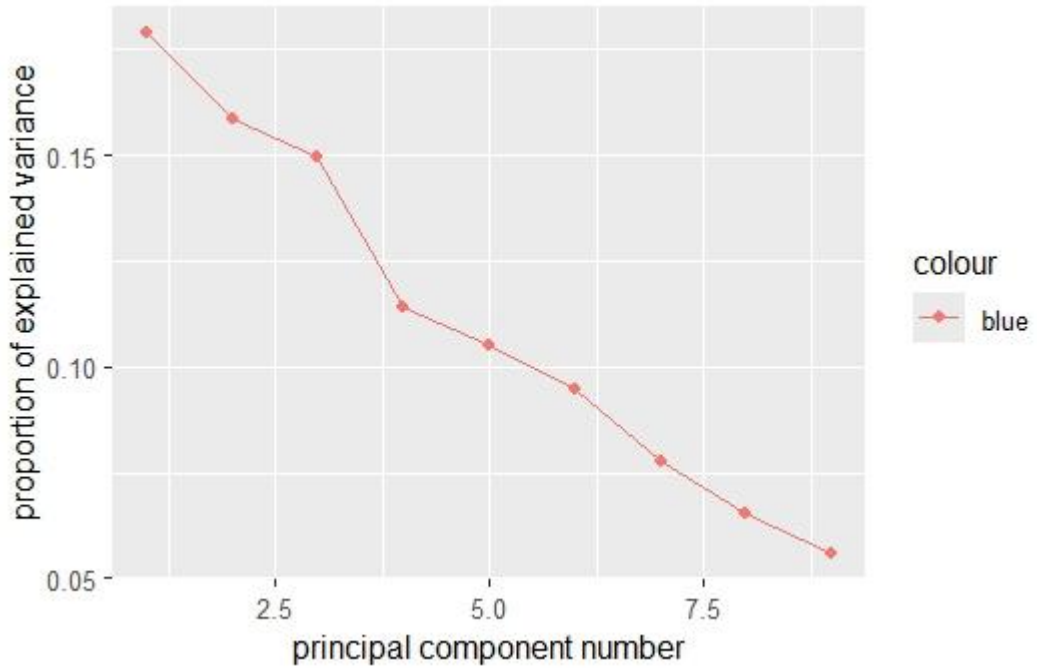
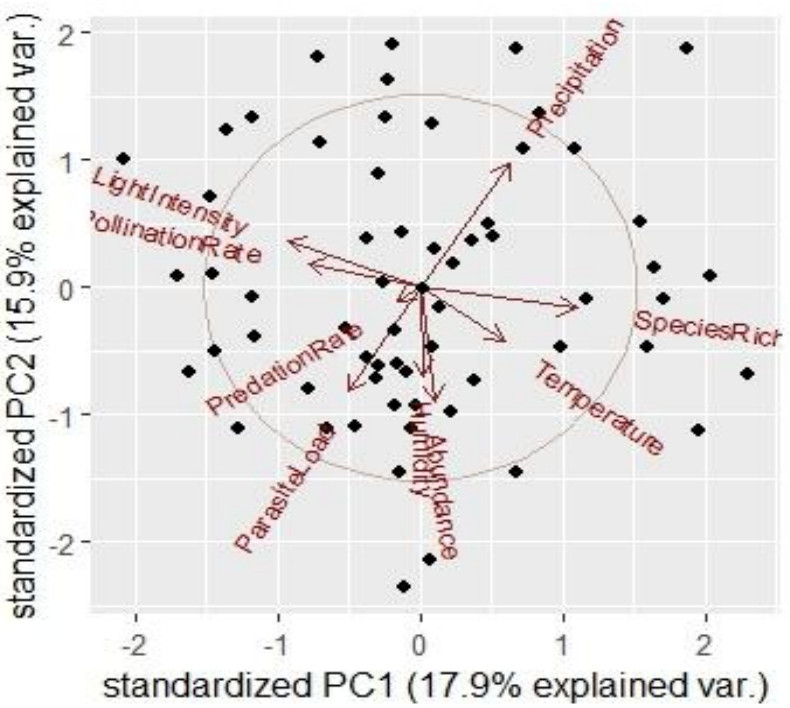


Fig: screeplot

c) Construct a bi-plot with ggplot2 package for the selected principle components and describe the plots.

```
library(ggbiplot)

ggscreeplot(pca)+

  aes(colour = "blue")

pca$sdev

summary(pca)
```

```
ggbiplot(pca,circle = TRUE) #graph

fa_result<-factanal(factors = 2,covmat = cov(data))

biplot(pca)
```



Interpretation of above biplot

This biplot visualizes the relationships between the variables and the first two principal components (PC1 and PC2), which explain 17.9% and 15.9% of the variance, respectively. The red arrows represent the variables, with their directions and lengths indicating their contributions to the components. Variables pointing in similar directions are positively correlated, while those pointing in opposite directions are negatively correlated. For example, Precipitation and SpeciesRichness are strongly associated with PC1, while PollinationRate and LightIntensity align more with PC2. Shorter arrows indicate weaker contributions to the variance.

d) Test whether your data is suitable for factor analysis or not.

```
library(psych)
library(corrplot)
data <- read.csv("ent.data.csv")
numeric_data <- data[sapply(data, is.numeric)]
numeric_data <- lapply(numeric_data, function(col) {
  col[is.na(col)] <- median(col, na.rm = TRUE)
  return(col) })
numeric_data <- as.data.frame(numeric_data)
cor_matrix <- cor(numeric_data, use = "pairwise.complete.obs")
bartlett_test <- cortest.bartlett(cor_matrix, n = nrow(numeric_data))
print(bartlett_test)
kmo_result <- KMO(cor_matrix)
print(kmo_result)
```

Interpretation of KMO result

```
Kaiser-Meyer-Olkin factor adequacy
Call: KMO(r = cor_matrix)
Overall MSA =  0.52
MSA for each item =
   Temperature       Humidity  Precipitation SpeciesRichness      Abundance
          0.51           0.43           0.49           0.54           0.54
 ParasiteLoad PollinationRate  PredationRate  LightIntensity
          0.53           0.50           0.52           0.56
```

The overall MSA (Measure of Sampling Adequacy) is 0.52, which is below the commonly accepted threshold of 0.6 for factor analysis suitability.

A value of 0.5–0.6 is considered mediocre, and anything below 0.5 indicates that the dataset is likely unsuitable for factor analysis without adjustments.

Each variable has its own MSA score:

Temperature (0.51), Humidity (0.43), Precipitation (0.49), and other variables all show mediocre or low MSA values.

Light Intensity (0.56) has the highest score but is still within the mediocre range.

Variables like Humidity (0.43) and Precipitation (0.49) fall below 0.5, indicating they contribute poorly to the overall adequacy of the dataset.

Conclusion: The dataset overall has limited suitability for factor analysis.

```
corrplot(cor_matrix, method = "circle")
```

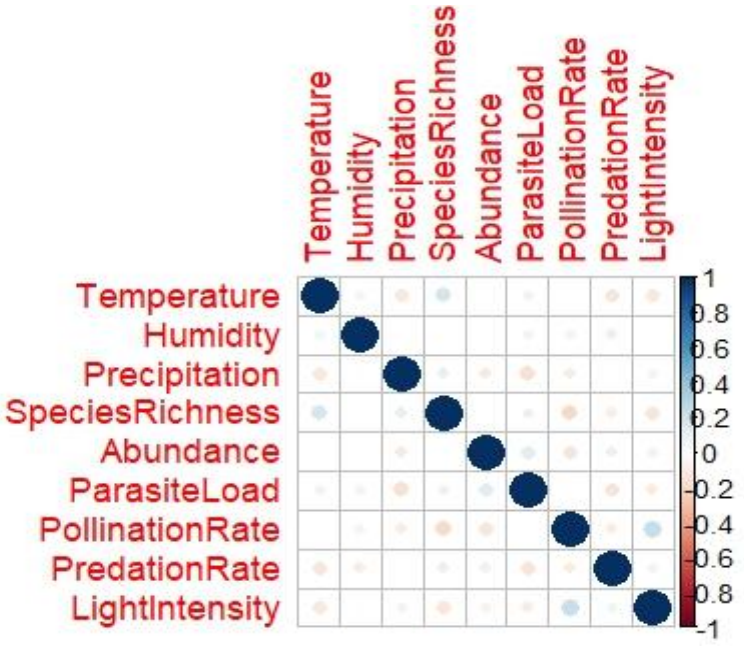


Fig: corrpot

e) Construct a suitable plot to visualize the factors with their loadings with factor analysis.

Ans:

```
library(psych)
library(ggplot2)
library(reshape2)
data <- read.csv("ent.data.csv")
numeric_data <- data[sapply(data, is.numeric)]
numeric_data <- lapply(numeric_data, function(col) {
  col[is.na(col)] <- median(col, na.rm = TRUE)
  return(col) })
numeric_data <- as.data.frame(numeric_data)
fa_result <- fa(numeric_data, nfactors = 2, rotate = "varimax")
print(fa_result$loadings)
loadings <- as.data.frame(as.table(fa_result$loadings))
colnames(loadings) <- c("Variable", "Factor", "Loading")
ggplot(loadings, aes(x = Factor, y = Variable, fill = Loading)) +
  geom_tile(color = "white") +
  scale_fill_gradient2(low = "blue", high = "red", mid = "white", midpoint = 0) +
  theme_minimal() +
  labs(title = "Factor Loadings Heatmap", fill = "Loading") +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```

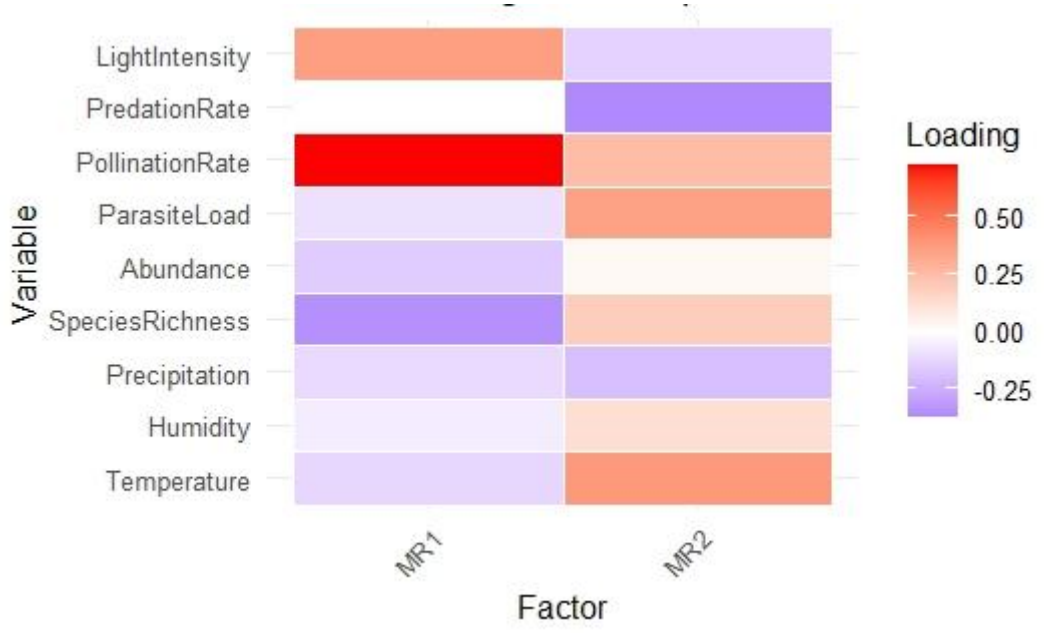


Fig: factors loading heat map

Problem # 2: A two-factor factorial design was conducted considering tree blocks, three levels/treatments of variety, and five levels/treatments of nitrogen. Afterward, the yield of certain plant characteristics was observed. The data regarding this experiment were given in the file "Data_Factorial_Design". Answer the following question using this data.

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

```
library(stats)

data <- read.csv("Data_Factorial_Design.csv")

data$VARIETY <- as.factor(data$VARIETY)

data$NITROGEN <- as.factor(data$NITROGEN)

data$REPLICAT <- as.factor(data$REPLICAT)

aov_result <- aov(YIELD ~ VARIETY * NITROGEN + Error(REPLICAT/(VARIETY * NITROGEN)), data = data)

summary(aov_result)
```

Source	Sum of squares	df	F- value	P-value
variety	1.925613	2	22.091	1.75×10−6
Nitrogen	66.02670	4	378.730	6.20×10−24
replicat	1.254773	2	14.395	5.02×10−5
Variety: Nitrogen	6.101631	8	17.500	5.23×10−9
Residual	1.220360	28	-	-

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1]

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

ANS:

Null Hypotheses for All Possible Effects:

In the context of a two-way ANOVA with replication, the null hypotheses (H_0) represent the assumption that there are no significant effects of each factor or their interactions on the response variable (in this case, YIELD). We will consider the following effects:

• **Main Effect of Block (replicat)**: H0: μBlock1=μBlock2=μBlock3

Interpretation: Since p<0.05, we can reject the null hypothesis by concluding that there are significant differences in all block levels.

• **Main Effect of Variety:** H0: μVariety1=μVariety2=μVariety3

Interpretation: Since p<0.05, we can reject the null hypothesis by concluding that there are significant differences in all variety levels.

• **Main Effect of Nitrogen:**

H0: μNitrogen1=μNitrogen2=μNitrogen3=μNitrogen4=μNitrogen5

Interpretation: Since p<0.05, we can reject the null hypothesis by concluding that there are significant differences in all Nitrogen levels.

• Interaction Effect (Variety × Nitrogen):

H0 :(μVariety×Nitrogen) ij= variety i+μNitrogen j

Interpretation: Since p<0.05, we can reject the null hypothesis by concluding that there is a significant interaction effect between variety and nitrogen.

c) Perform a post-hoc test for the levels/treatments of nitrogen and draw a bar diagram with lettering.

ANS:

Nitrogen	Yield	Groups
4	6.302222	a
5	5.858889	a
3	5.628889	a
2	4.804444	b
1	2.875556	c

From the post hoc test we can conclude that

Group a: Nitrogen level 4.5,3, highest yield, most distinct. •

Group b: Nitrogen levels 2, moderate yields.

Group c: Nitrogen level 1, lowest yield

IN R

```
tukey_result <- HSD.test(aov(YIELD ~ NITROGEN, data = data), "NITROGEN", group = TRUE)
print(tukey_result)
bar_data <- tukey_result$groups
bar_data$NITROGEN <- rownames(bar_data)
colnames(bar_data) <- c("mean", "groups", "NITROGEN") # Adjust based on `print(bar_data)`
```

```
bar_plot <- ggplot(bar_data, aes(x = NITROGEN, y = mean, fill = NITROGEN)) +
  geom_bar(stat = "identity", color = "black") +
  geom_text(aes(label = groups), vjust = -0.5, size = 5) +
  labs(title = "Nitrogen Levels with Post-Hoc Grouping", x = "Nitrogen Levels", y = "Mean Yield")
+  theme_minimal()
```

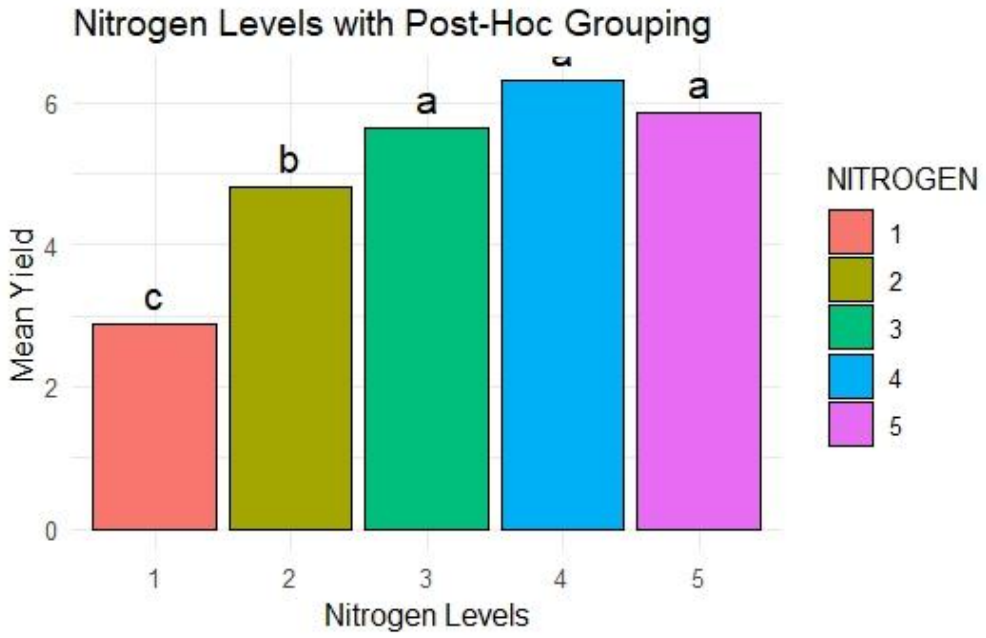


Fig: Barplot nitrogen