Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Project Report Marks: 25
Name: …Ummay…Habiba……
Reg. No:…18-05-4603……Dept.…Agronomy……..

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.
b) Interpret how many principle components should be retained for your data with justification.
c) Construct a bi-plot with ggplot2 package for the selected principle components and describe the plots.
d) Test whether your data is suitable for factor analysis or not.
e) Construct a suitable plot to visualize the factors with their loadings with factor analysis.

Answer to the ques no: 1

(a)

#Loading the data

```
PCA_Data<- read.csv("Project_Habiba.csv") [-c(1:4)]
```

#Missing value

```
colSums(is.na(PCA_Data))
```

Result:

Variable	Missing Values
Emergence days	0

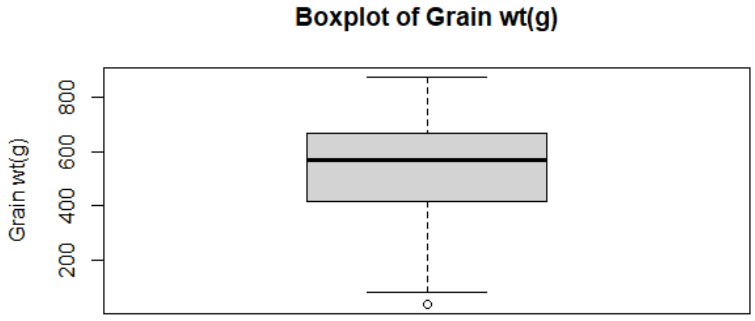
Plant height	**0**
Pod number	**0**
Seed number	**0**
100 Seed wt.	**0**
Pod wall wt.	**0**
Stem wt.	**0**
Grain wt.	**5**
Straw wt.	**0**
Total seed wt.	**0**

#Impute missing value by mean

```
PCA_Data$Grain.wt.g.[is.na(PCA_Data$Grain.wt.g.)] <-
    median(PCA_Data$Grain.wt.g., na.rm = TRUE)
```

#Check outlier

```
boxplot(PCA_Data$Grain.wt.g.,
    main = "Boxplot of Grain wt(g)",
    ylab = "Grain wt(g)")
```



 # Calculate lower and upper bounds using MAD

```
lower_bound <- median(PCA_Data$Grain.wt.g,na.rm = TRUE) -
  3 * mad(PCA_Data$Grain.wt.g.,na.rm = TRUE)
```

Result: lower_bound : 89.8615

```
upper_bound <- median(PCA_Data$Grain.wt.g.,na.rm = TRUE) +
  3 * mad(PCA_Data$Grain.wt.g.,na.rm = TRUE)
```

Result: upper bound: 1046.139

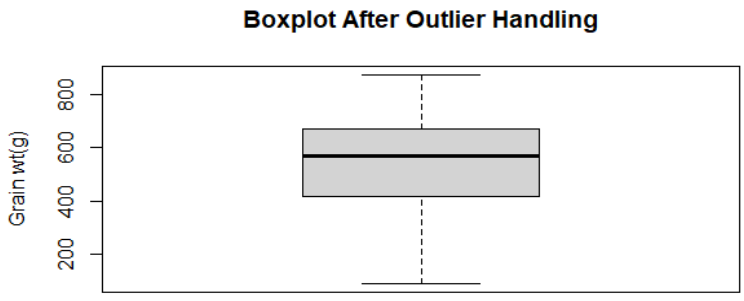
 # Identify indices of outliers

```
outliers <- which(PCA_Data$Grain.wt.g. < lower_bound |
                    PCA_Data$Grain.wt.g. > upper_bound)
```

Outliers [1] 21 30

 # Replace outliers with the calculated bounds

```
PCA_Data$Grain.wt.g.[PCA_Data$Grain.wt.g.< lower_bound] <- lower_bound
PCA_Data$Grain.wt.g.[PCA_Data$Grain.wt.g. > upper_bound] <- upper_bound
boxplot(PCA_Data$Grain.wt.g.,
        main = "Boxplot After Outlier Handling",
        ylab = "Grain wt(g)")
```



(b)

```
# Perform PCA
correlation<- cor(PCA_Data)
mean(correlation)
eigen(correlation)
PCA_result <- prcomp(PCA_Data, scale. = TRUE)
summary(PCA_result)
install.packages("devtools")
library(devtools)
install_github("vqv/ggbiplot",force=TRUE)
library(ggbiplot)
ggscreeplot(PCA_result)+
  aes(colour = "red")
```

Component	**PC1**	**PC2**	**PC3**	**PC4**	**PC5**	**PC6**	**PC7**	**PC8**	**PC9**	**PC10**
Standard deviation	2.3057	1.3881	0.9921	0.7691	0.68960	0.64175	0.33870	0.28580	0.2303	0.21082
Proportion of Variance	0.5316	0.1927	0.09843	0.05916	0.04755	0.04118	0.01147	0.00817	0.0053	0.00444
Proportion of Variance	0.5316	0.7243	0.82272	0.88187	0.92943	0.97061	0.98209	0.99025	0.9956	1.00000

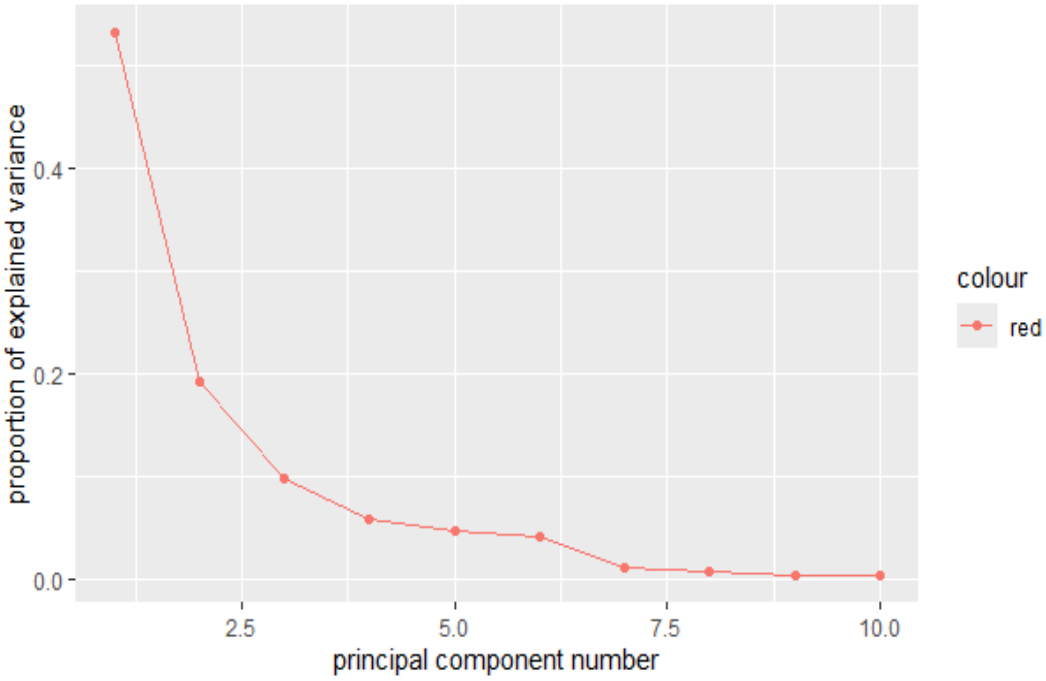


Figure 1 Screeplot

To determine the number of principal components to retain, we evaluate the cumulative proportion of variance explained and examine the scree plot:

Cumulative Proportion:

The first three components account for roughly 82.23% of the total variance (PC1: 53.16%, PC2: 19.27%, PC3: 9.8%). This level of variance is typically sufficient for retaining components, as it represents a significant portion of the dataset's variability.

ScreePlot:

The scree plot reveals an "elbow" after three component, indicating a slower rate of decline in explained variance beyond this point. This supports the decision to retain four components.

(c)

#To draw bi-plot

install.packages("devtools")

library(devtools)

install_github("vqv/ggbiplot")

library(ggbiplot)

ggbiplot(PCA_result

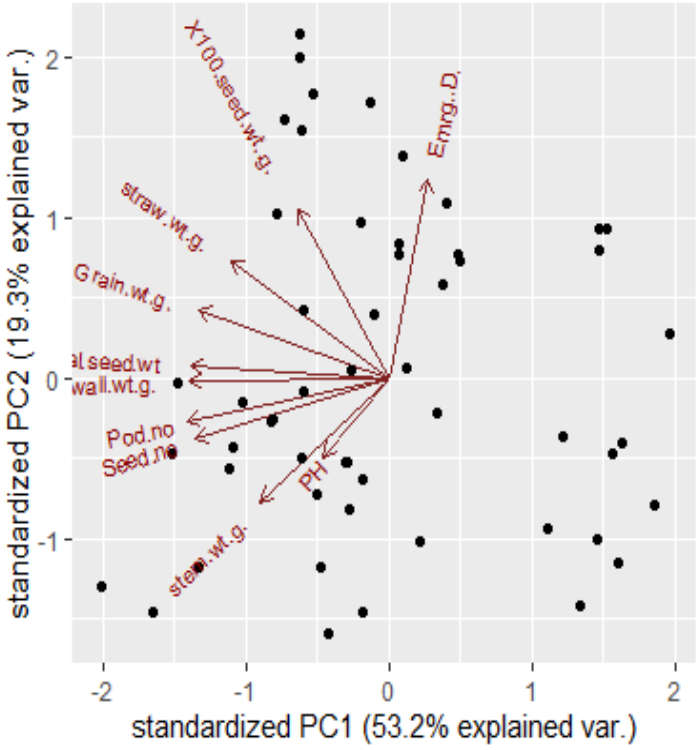


Figure 2 Biplot

The biplot illustrates the relationship between the first two principal components (PC1 and PC2), which together explain 72.5% of the total variance in the data.

Key Features:

Axes (PC1 and PC2): PC1 accounts for 53.2% of the variance, while PC2 explains 19.3%.

These components capture the most significant patterns in the dataset.

Points (Observations): Each point represents an observation, with closer points indicating greater similarity. The spread shows how the data varies along the principal components.

Arrows (Variables): The arrows indicate the contributions of the original variables:

- Longer arrows signify variables with a stronger influence on the components.
- Arrows pointing in similar directions suggest a positive correlation, while perpendicular arrows indicate little or no correlation.
- Opposing arrows reveal negative correlations.

Observations:

- Variables like 100.Seed.wt, Grain.wt.g, and straw.wt.g have long arrows, indicating that they are significant contributors to the variation captured by PC1 and PC2.
- Variables like Emrg D(Emergence day) are also important contributors, but they point in a different direction, suggesting that they capture a distinct source of variation.
- Observations near 100.Seed.wt are likely associated with higher values of seed weight.
- Observations near Emrg D(Emergence day) may have higher values related to that variable.

The most significant contributors to variability are variables like 100.Seed.wt, straw.wt.g, Grain.wt.g, and wall.wt.g. These variables are positively correlated. The second major source of variation is linked to Emergence day and Plant height. These variables point in a different direction than others, indicating they are capturing distinct patterns or sources of variability. Again, variables such as 100.Seed.wt, Grain.wt.g, straw.wt.g, and wall.wt.g are strongly positively correlated. This indicates that plants with higher seed weight are likely to have higher grain and straw weight, suggesting a potential productivity or yield relationship.

(d)

```
library(psych)

# KMO Test
 KMO(PCA_Data)

# Bartlett's Test

bartlett.test(PCA_Data)
```

KMO	0.79
Bartlett's Test	p-value < 2.2e-16

For Kaiser-Meyer-Olkin (KMO) test

• KMO > 0.9: Marvelous – Excellent suitability for factor analysis.

• KMO between 0.8 and 0.9: Great – Very good suitability.

• KMO between 0.7 and 0.8: Good – Adequate, acceptable for factor analysis.

• KMO between 0.6 and 0.7: Mediocre – Marginally acceptable, might need further

checks.

• KMO < 0.6: Not suitable – Factor analysis may not be appropriate for this data.

Here, Overall KMO = 0.79, which falls in the "Good" range (between 0.7 and 0.8). This value suggests that the data is adequate for factor analysis, as the KMO value is above 0.7, indicating that there is sufficient common variance between the variables.

Bartlett's Test

If the p-value is less than 0.05, we can conclude that the data is suitable for factor analysis. Here, the p-value is very small (< 0.05), which indicates that the correlation matrix is significantly different from an identity matrix. This suggests that the variables in the data are correlated enough to justify the use of factor analysis. In other words, Bartlett's test indicates that factor analysis is

appropriate for the data.

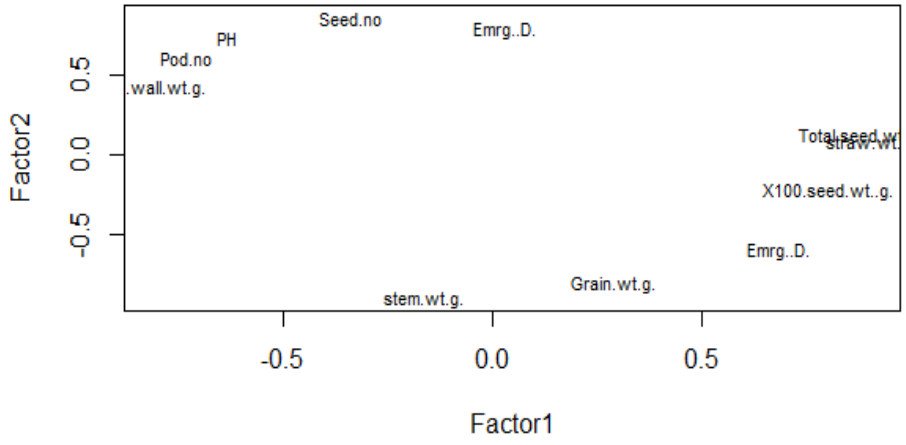
(e)

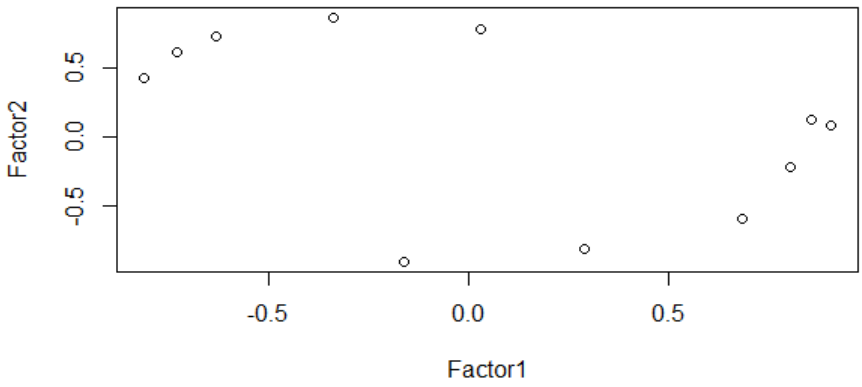
 # Perform factor analysis

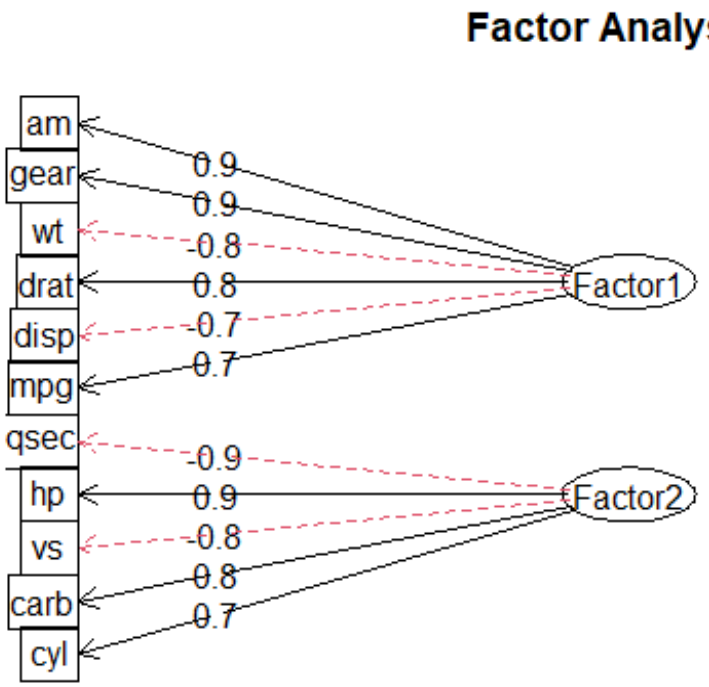
```
fact_result<-factanal(factors=2, covmat = cov(PCA_Data))
Rotation<-factanal(factors=2, covmat = cov(PCA_Data), rotation = "varimax")
print(fact_result)
plot(load)loads<-fact_result$loadings
fa.diagram(loads)
#Plot
plot(load,type="n")
text(load,labels=names(PCA_Data), cex= .7)
plot(load)
```

Variable	**Factor1**	**Factor2**	**Uniqueness**	**SS Loadings**	**Proportion of Variance**	**Cumulative Variance**
Emergence days	-0.397	0.347	0.722	4.067 2.350	0.407 0.235	0.407 0.642
Plant height	0.200	0.133	0.942			
Pod number	0.943	0.282	0.031			
Seed number	0.922	0.192	0.114			

100 Seed wt.		0.601	0.636			
Pod wall wt.	0.843	0.387	0.140			
Stem wt.	0.602		0.637			
Grain wt.	0.539	0.789	0.087			
Straw wt.	0.318	0.908	0.074			
Total seed wt.	0.815	0.370	0.200			







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.
b) Write down the null hypothesis of all possible effects and interpret the results based on the ANOVA table.
c) Perform a post-hoc test for the levels/treatments of nitrogen and draw a bar diagram with lettering.

Answer to the ques no: 2

(a)

 # Loading the data

```
Data.factorial <- read.csv("Data_Factorial_Design.csv")
```

 # factors

```
block <- c("Block1", "Block2", "Block3")

variety <- c("Variety1", "Variety2", "Variety3")

nitrogen <- c("Nitrogen1", "Nitrogen2", "Nitrogen3", "Nitrogen4", "Nitrogen5")
```

 # Determining the total number of blocks, varieties, and nitrogen levels

```
b <- length(block)

v <- length(variety)

n <- length(nitrogen)
```

 # Generating factorial combinations

```
Block <- gl(b, v * n, b * v * n, factor(block))

Varfact <- gl(v, n, b * v * n, factor(variety))

NitroFact <- gl(n, 1, b * v * n, factor(nitrogen))
```

 # Performing ANOVA for Randomized Complete Block Design (RCBD)

```
ANOVA.twoFact.Factorial.RCBD <- aov(data = Data.factorial, YIELD ~ Varfact + Block + NitroFact + Varfact * NitroFact)

summary(ANOVA.twoFact.Factorial.RCBD)
```

Result:

Table 1: ANOVA.twoFact.Factorial.RCBD

Sources	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
Varfact	2	1.93	0.963	22.09	1.75E-06	***
Block	2	1.25	0.627	14.39	5.02E-05	***
NitroFact	4	66.03	16.507	378.73	<2.00E-16	***
Varfact:NitroFact	8	6.1	0.763	17.5	5.23E-09	***
Residuals	28	1.22	0.044			

[Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1]

(b)

The null hypotheses are:

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

Interpretation: Since p<0.05 (table 2), 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 (table 1), 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 (table 1), 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= μVariety i+μNitrogen j

Interpretation: Since p<0.05 (table 1), we can reject the null hypothesis by concluding that there is a significant interaction effect between variety and nitrogen.

(c)

library(agricolae)

 # Post-hoc test for Nitrogen levels

PostHoc.Test.nitrogen<-with(Data.factorial,HSD.test(YIELD,NITROGEN,DFerror = 28,MSerror = 0.044))

NITROGEN	YIELD	groups
4	6.302222	a
5	5.858889	b
3	5.628889	b
2	4.804444	c
1	2.875556	d

From PostHoc test we can conclude that,

• Group a: Nitrogen level 4, highest yield, most distinct.

• Group b: Nitrogen levels 3 and 5, moderate yields.

• Group c: Nitrogen level 2, moderate-low yields

• Group d: Nitrogen level 1, lowest yield.

#Barplot

```
Mutplcom.NitroFact<-with(Data.factorial,HSD.test
 (YIELD,NITROGEN,DFerror=28,MSerror=0.044))
Nitro.Mean <- Mutplcom.NitroFact$groups
Nitro.SE.Mat <- Mutplcom.NitroFact$means
Nitro.SE.Mat <- Mutplcom.NitroFact$means[, "se"]
Mean.Mat <- Mutplcom.NitroFact$means
Mean.Mat <- Mean.Mat[order(-Mean.Mat$YIELD), ]
Nitro.Nitro.Mean <- Nitro.Mean$YIELD
Nitro.SE <- Mean.Mat[, "se"]
Nitro.SE.Mat <- Mutplcom.NitroFact$means[order(Mutplcom.NitroFact$means[,”se”])]
library(gplots)
Barplot.SE <- barplot2(Nitro.Nitro.Mean, names.arg = rownames(Nitro.Mean), xlab =
"Nitrogen",
 ylab = "Yield", horiz = F, plot.ci = T, ci.l = Nitro.Nitro.Mean - Nitro.SE,
 ci.u = Nitro.Nitro.Mean + Nitro.SE, col = "lightpink")
text(Barplot.SE, 0,Nitro.Mean$groups , cex = 2, pos = 3, col = "black")
```

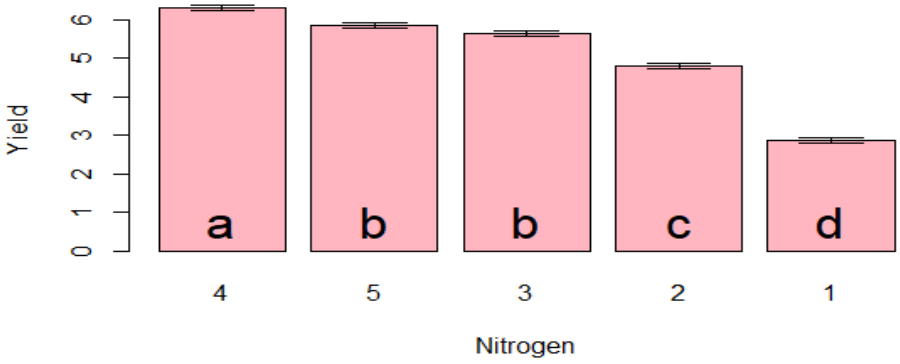


Figure 1 Barplot Nitrogen