Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Project Report Marks: 25
Name: ....Rabbani Rushsa..........................
Reg. No:...2018-05-4804.....................Dept.......Agricultural Economics...................

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.

Ans to the ques no: 1

(a)

#Loading the data

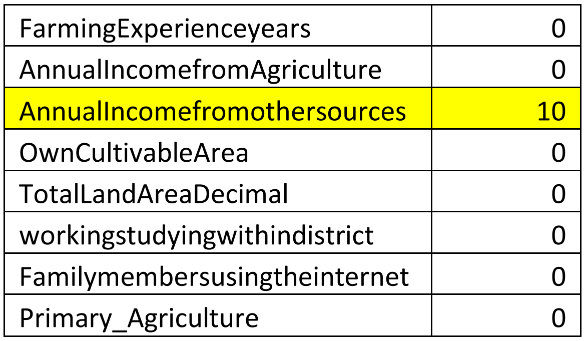
My_PCA_Data<- read.csv("Project Assignment_Rushsa.csv") [1:160,1:12]

#Missing value

colSums(is.na(My_PCA_Data))

Result:

Variable	Missing Values
Distancekm	0
Ageyears	0
Sex	0
FamilySize	0
Yearofschooling	0



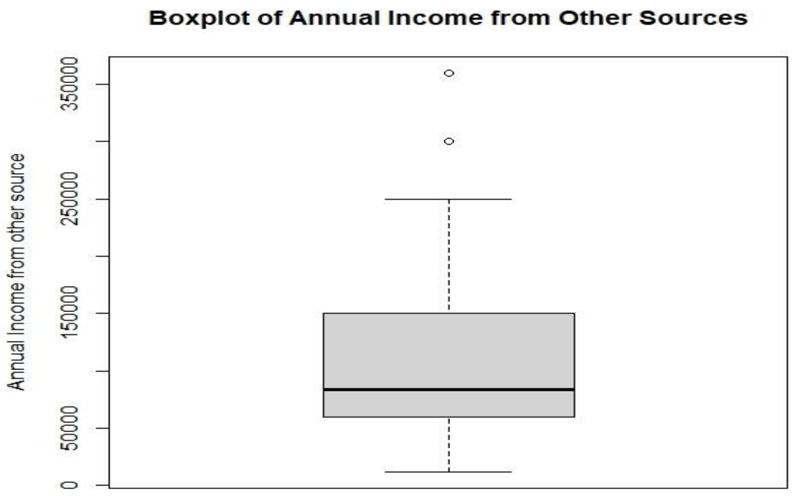
FarmingExperienceyears	0
AnnualIncomefromAgriculture	0
AnnualIncomefromothersources	10
OwnCultivableArea	0
TotalLandAreaDecimal	0
workingstudyingwithindistrict	0
Familymembersusingtheinternet	0
Primary_Agriculture	0

#Impute missing value by mean

```
My_PCA_Data$AnnualIncomefromothersources[is.na(My_PCA_Data$AnnualIncomefromother
sources)] <-
 median(My_PCA_Data$AnnualIncomefromothersources, na.rm = TRUE)
```

#Check outlier

```
boxplot(My_PCA_Data$AnnualIncomefromothersources,
    main = "Boxplot of Annual Income from Other Sources",
    ylab = "Annual Income from other source")
```



# Calculate lower and upper bounds using MAD

```
lower_bound <- median(My_PCA_Data$AnnualIncomefromothersources, na.rm = TRUE) -
  3 * mad(My_PCA_Data$AnnualIncomefromothersources, na.rm = TRUE)
```

Result: lower_bound : -67225.2

```
upper_bound <- median(My_PCA_Data$AnnualIncomefromothersources, na.rm = TRUE) +
  3 * mad(My_PCA_Data$AnnualIncomefromothersources, na.rm = TRUE)
```

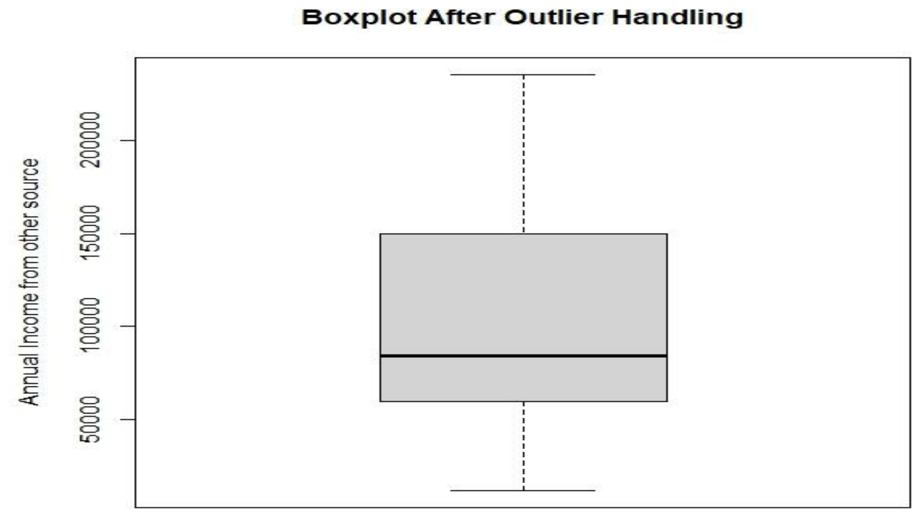
Result: upper bound: 235225.2

# Identify indices of outliers

```
outliers <- which(My_PCA_Data$AnnualIncomefromothersources < lower_bound |
                  My_PCA_Data$AnnualIncomefromothersources > upper_bound)
outliers
 [1]   7  11  25  31  35  37  39  58  70  82  85  88  89  90  92  93  98 102 104 106 122 124
```

# Replace outliers with the calculated bounds

```
My_PCA_Data$AnnualIncomefromothersources[My_PCA_Data$AnnualIncomefromothersources < lower_bound] <- lower_bound

My_PCA_Data$AnnualIncomefromothersources[My_PCA_Data$AnnualIncomefromothersources > upper_bound] <- upper_bound

boxplot(My_PCA_Data$AnnualIncomefromothersources,
    main = "Boxplot After Outlier Handling",
    ylab = "Annual Income")
```



(b)

```
# Perform PCA

correlation<- cor(My_PCA_Data)

mean(correlation)

eigen(correlation)

PCA_result <- prcomp(My_PCA_Data, scale. = TRUE)

summary(PCA_result)

install.packages("devtools")

library(devtools)

install_github("vqv/ggbiplot")

ggscreeplot(PCA_result)+aes(color="red")
```

Component	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9	PC10	PC11	PC12
Standard	1.9264	1.4397	1.1245	1.05554	0.9619	0.88201	0.83578	0.78742	0.64949	0.44738	0.4214	0.12707

Deviation												
Proportion of Variance	0.3093	0.1727	0.1054	0.09285	0.0771	0.06483	0.05821	0.05167	0.03515	0.01668	0.0148	0.00135
Cumulative Proportion	0.3093	0.482	0.5874	0.68022	0.7573	0.82215	0.88036	0.93203	0.96718	0.98386	0.9987	1

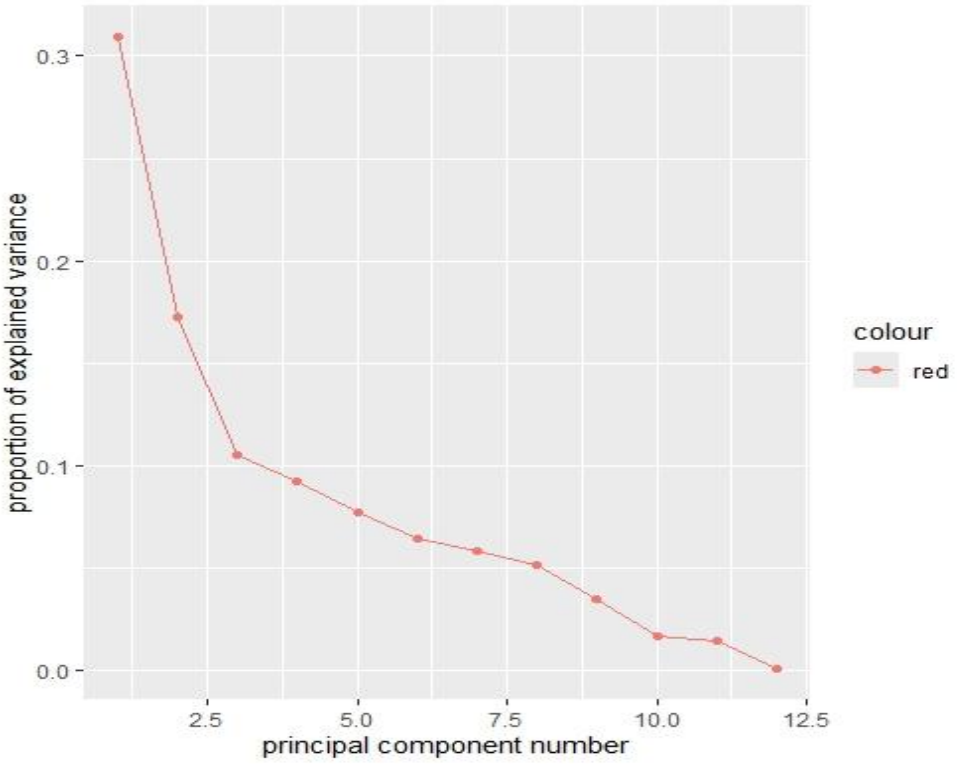


Figure 1 Screeplot

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 four components account for roughly 68.02% of the total variance (PC1: 30.93%, PC2: 17.27%, PC3: 10.54%, PC4: 9.29%). 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 the fourth 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)
```

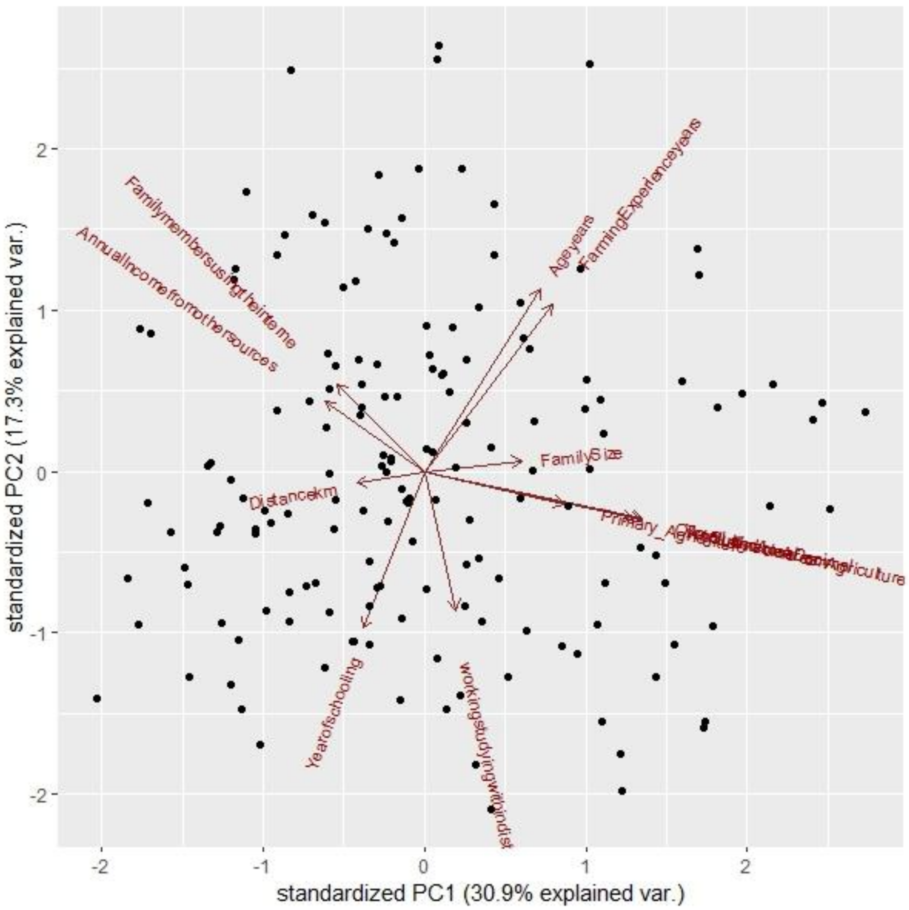


Figure 2 Biplot

The biplot illustrates the relationship between the first two principal components (PC1 and PC2), which together explain 48.2% of the total variance in the data.

Key Features:

- Axes (PC1 and PC2): PC1 accounts for 30.9% of the variance, while PC2 explains 17.3%. 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 "Ageyears" and "FarmingExperienceyears" are positively associated and strongly influence PC1.
- "AnnualIncomefromothersources" and "Familymembersusingtheinternet" align with PC2, representing a separate dimension.
- Negative correlations are evident between "Primary_Agriculture" and variables like "TotalLandAreaDecimal" and "OwnCultivableArea."

Insights:

- Observations on the positive side of PC1 are associated with older individuals with more farming experience.
- Points aligning with PC2 suggest individuals with higher income diversity and internet usage.
- The biplot simplifies complex relationships, showing how variables and observations interact in a reduced-dimension space.

(d)

```
library(psych)

# KMO Test

KMO(My_PCA_Data)

# Bartlett’s Test

bartlett.test(My_PCA_Data)
```

KMO	0.73
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 = 0.73**, 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

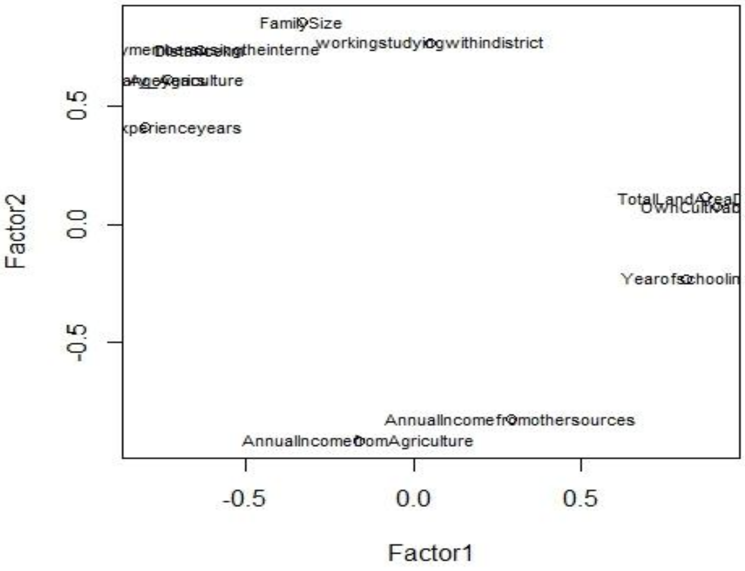
```
fact_result<-factanal(factors=2, covmat = cov(My_PCA_Data))

Rotation<-factanal(factors=2, covmat = cov(My_PCA_Data), rotation = "varimax")

print(fact_result)

plot(load)loads<-fact_result$loadings

fa.diagram(loads)
```

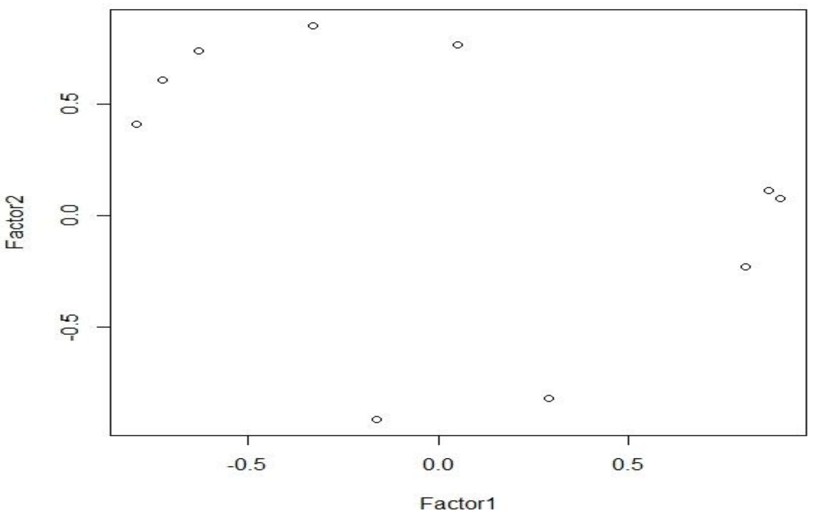
#Plot

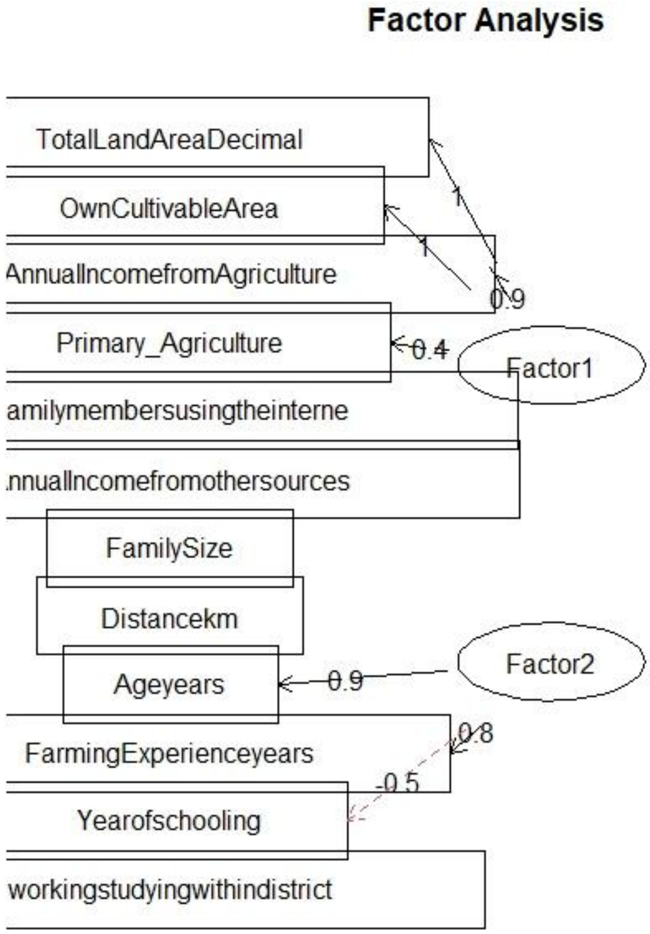
```
plot(load,type="n")

text(load,labels=names(My_PCA_Data), cex= .7)

plot(load)
```

Variable	Factor 1	Factor 2	Uniqueness	SS Loadings	Proportion of Variance	Cumulative Variance
Distancekm	-0.174	-0.121	0.955	3.091	25.80%	25.80%

Ageyears	0.142	0.924	0.125			
FamilySize	0.208	0.167	0.929			
Yearofschooling	-0.49		0.756			
FarmingExperienceyears	0.191	0.82	0.291			
AnnualIncomefromAgriculture	0.855	0.112	0.256			
AnnualIncomefromothersources	-0.226		0.948			
OwnCultivableArea	0.986		0.022			
TotalLandAreaDecimal	0.991		0.012			
workingstudyingwithindistrict	0.166	-0.224	0.922			
Familymembersusingtheinterne	-0.231		0.943			
Primary_Agriculture	0.374	0.125	0.844			




Factor2
0.5
0.0
-0.5
-0.5
0.0
0.5
Factor1


Factor Analysis
TotalLandAreaDecimal
OwnCultivableArea
AnnualIncomefromAgriculture
Primary_Agriculture
amilymembersusingtheinterne
nnualIncomefromothersources
FamilySize
Distancekm
Ageyears
FarmingExperienceyears
Yearofschooling
workingstudyingwithindistrict
1
1
0.9
0.4
Factor1
Factor2
0.9
0.8
-0.5

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.

Ans to the ques no: 2

(a)

# Loading the data

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

# Defining 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:** $H_0: \mu_{Block1}=\mu_{Block2}=\mu_{Block3}$

  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:** $H_0: \mu_{Variety1}=\mu_{Variety2}=\mu_{Variety3}$

  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:**

  $H_0: \mu_{Nitrogen1}=\mu_{Nitrogen2}=\mu_{Nitrogen3}=\mu_{Nitrogen4}=\mu_{Nitrogen5}$

  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):**

  $H_0:(\mu_{Variety\times Nitrogen})_{ij}= \mu_{Variety\ i}+\mu_{Nitrogen\ j}$

<u>Interpretation:</u> 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 = "purple")

text(Barplot.SE, 0,Nitro.Mean$groups , cex = 2, pos = 3, col = "white")
```

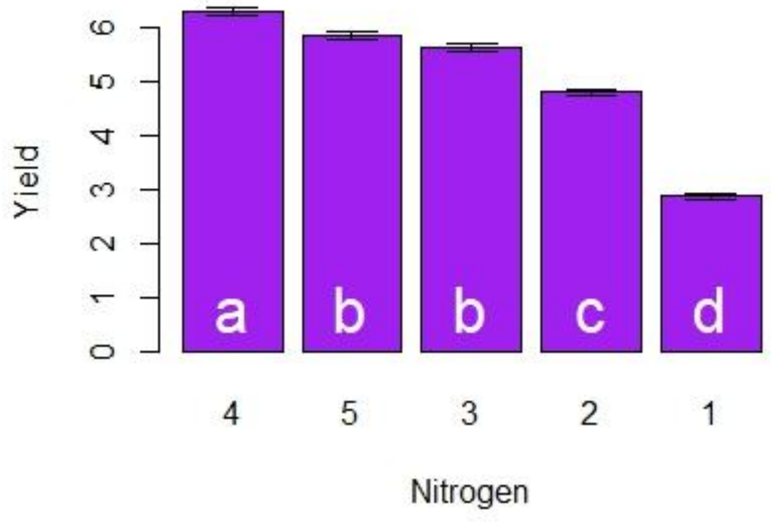


Figure 1 Barplot Nitrogen