Bangabandhu Sheikh Mujibur Rahman Agricultural University
EDGE_Batch-11
Project Report Marks: 25
Name: ……Nusrat Jahan Naba ……
Reg. No:…2018-05-4821……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.

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.

Solution 1

a)

Code:

```
data<-read.csv("Project Data_Naba.csv")
colSums(is.na(data))

data$`Income_from _secondary_sources`[is.na(data$`Income_from _secondary_sources`)] <-
  median(data$`Income_from _secondary_sources`, na.rm = TRUE)

boxplot(data$`Income_from _secondary_sources`)

# Calculate lower and upper bounds using MAD

lower_bound <- median(data$`Income_from _secondary_sources`, na.rm = TRUE) -
  3 * mad(data$`Income_from _secondary_sources`, na.rm = TRUE)
```

Result: -38956

```
upper_bound <- median(data$`Income_from _secondary_sources`, na.rm = TRUE) +
  3 * mad(data$`Income_from _secondary_sources`, na.rm = TRUE)
```

Result: 138956

```
# Identify indices of outliers

outliers <- which(data$`Income_from _secondary_sources` < lower_bound |
                    data$`Income_from _secondary_sources` > upper_bound)


# Replace outliers with the calculated bounds

data$`Income_from _secondary_sources`[data$`Income_from _secondary_sources` < lower_bound] <-
lower_bound

data$`Income_from _secondary_sources`[data$`Income_from _secondary_sources` > upper_bound] <-
upper_bound

boxplot(data$`Income_from _secondary_sources`,
        main = "Boxplot After Outlier Handling",
  ylab = "Annual Income")
```

Result:

Here is the table formatted with the data of missing value:

Attribute	Missing Values
Age	0
Education Level	0
Primary Occupation	0
Income from Primary Sources	0
Secondary Occupation	0
Income from Secondary Sources	4
No. of Family Members	0
Years of Experience in Agriculture	0
Years of Experience in Banana Farming	0
Total Land	0
Total Banana Land	0
Received Agricultural Training	0
Labor Cost for Fertilizer Application	0
Total Output	0
Output Price (in Tk)	0

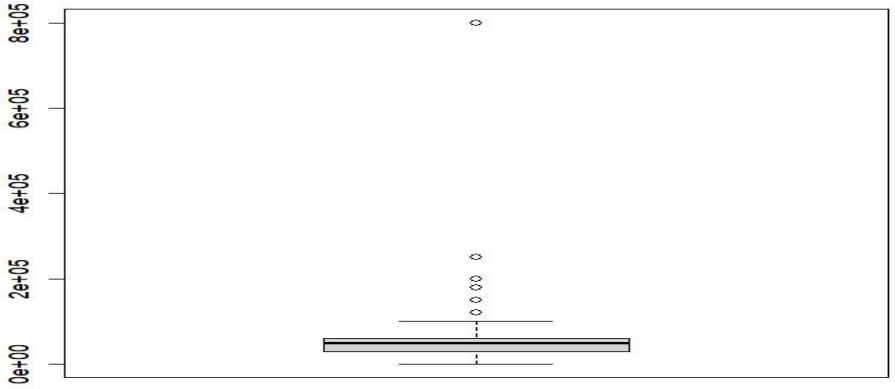

Figure: Boxplot of outlier

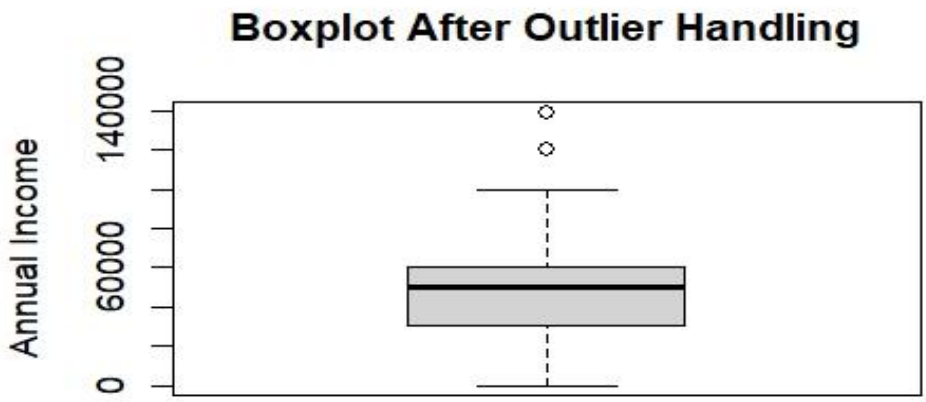

Figure: Boxplot after outlier handling

b)

Code:

```
co<-cor(data)
View(co)
mean(co)
dim(co)
eigen(co)
PCA <- prcomp(data, scale. = TRUE)
summary(PCA)
install.packages("devtools")
library(devtools)
install_github("vqv/ggbiplot")
library(ggbiplot)
ggscreeplot(PCA)
```

Result:

Here is the table formatted with the Summary of PCA:

Component	Standard Deviation	Proportion of Variance	Cumulative Proportion
PC1	2.0136	0.2703	0.2703
PC2	1.5912	0.1688	0.4391
PC3	1.3638	0.124	0.5631
PC4	1.05416	0.07408	0.63718
PC5	1.01724	0.06899	0.70617
PC6	0.9389	0.05877	0.76494
PC7	0.89291	0.05315	0.81809
PC8	0.83051	0.04598	0.86407
PC9	0.7744	0.03998	0.90405
PC10	0.70845	0.03346	0.93751
PC11	0.6606	0.0291	0.9666
PC12	0.51135	0.01743	0.98404
PC13	0.38298	0.00978	0.99382
PC14	0.25052	0.00418	0.99801
PC15	0.17297	0.00199	1

Interpretation:

PC1 accounts for 27.03% of the variance, while the first 4 PCs together explain 63.72%. Expanding to the first 8 PCs captures 86.41% of the variance, and by PC11, 96.66% of the variance is explained.

The variance contribution decreases steadily across the components, with PC1 having the highest contribution (27.03%) and PC15 contributing the least (0.20%). Retaining the first 8 PCs is recommended, as they explain 86.41% of the variance, which is a significant portion of the dataset's variability.

Adding PCs beyond PC8 contributes only a small amount of additional variance (e.g., PC9 adds just 3.99%). By focusing on the first 8 PCs, the majority of the dataset's variability is preserved, dimensionality is reduced, and the model remains simpler and more interpretable. Components beyond PC8 contribute minimal information and are less relevant for analysis.

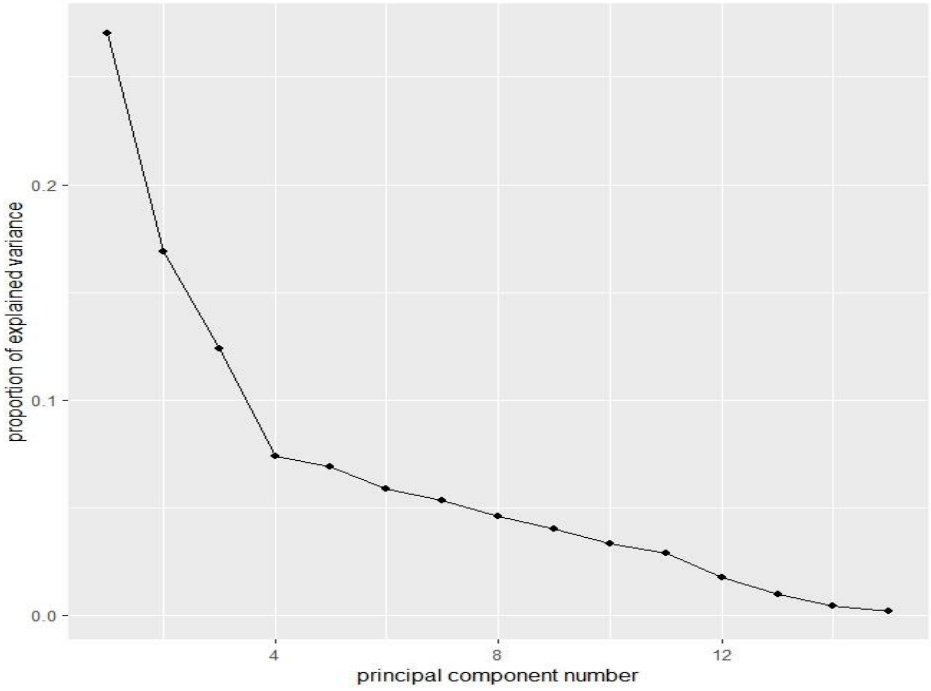

Figure: Screeplot

Interpretation

The X-axis represents the principal components (PC1 to PC15), while the Y-axis shows the proportion of variance explained by each component.

The plot reveals a sharp decline in the variance explained after the first few components (PC1 to PC4). Beyond PC4, the proportion of variance decreases steadily, creating an "elbow" shape in the curve. The initial components, particularly PC1 to PC4, account for the majority of the dataset's variance. The "elbow" around PC4 suggests that retaining the first four components may be sufficient, as subsequent components contribute progressively less variance.

The biplot illustrates the relationships between variables and observations in terms of the first two principal components, which explain a substantial portion of the variance. The scree plot indicates that retaining 4 to 8 principal components may be optimal, depending on the desired level of explained variance (e.g., 70–90%).

c)Code:

```
ggbiplot(PCA)
```

Result:

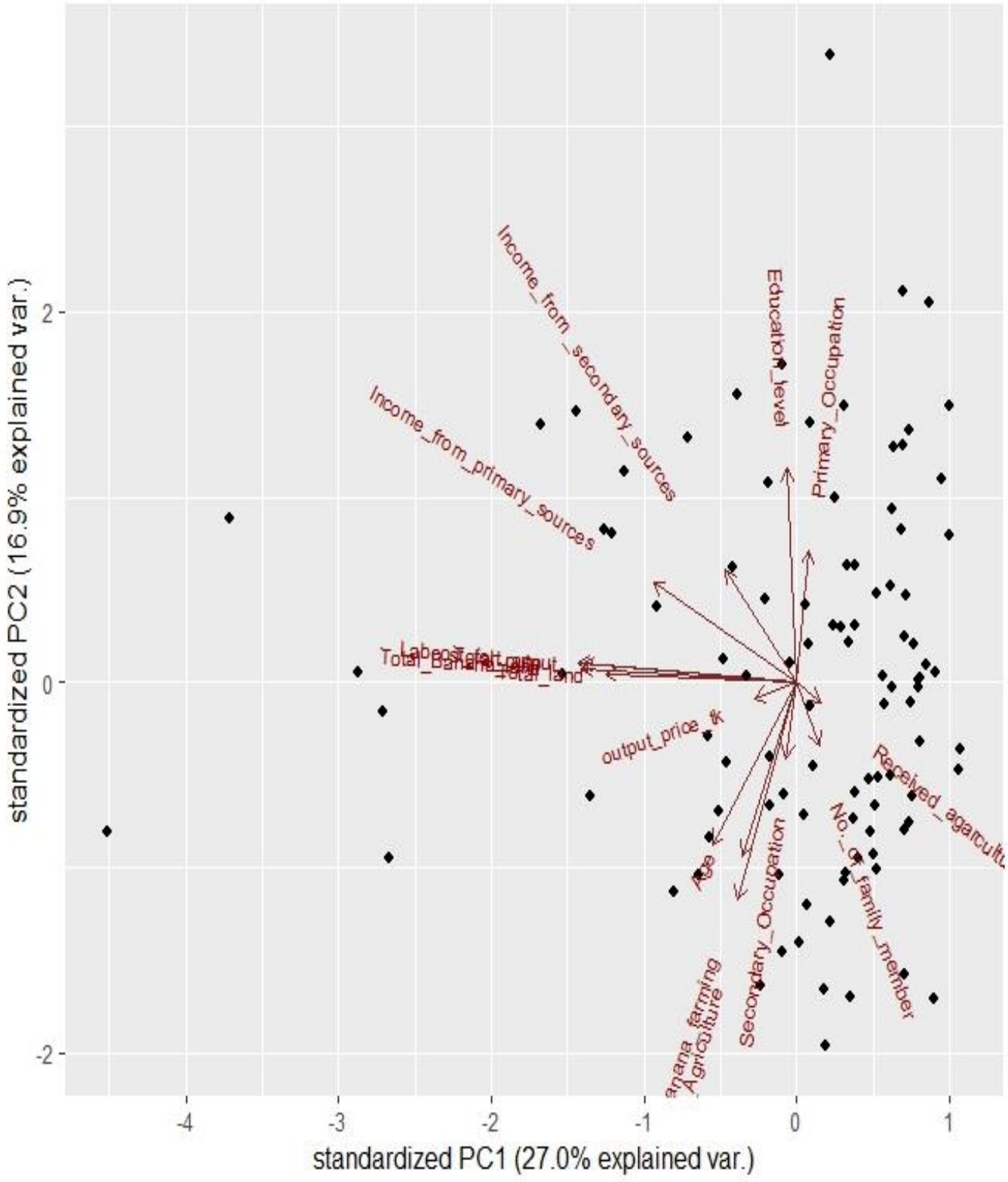

Figure: Biplot

Interpretation:

The X-axis represents the first principal component (PC1), which accounts for 27.0% of the variance, while the Y-axis shows the second principal component (PC2), explaining 16.9% of the variance. The red arrows represent the contribution of the original variables to the principal components. Longer arrows indicate that the corresponding variables have a stronger influence on the principal components.

Variables such as "Total Banana Land" and "Income from Primary Sources" have a substantial impact on PC1, while "Education Level" and "Primary Occupation" are more closely associated with PC2.

The observations scattered across the plot demonstrate the variability in the data explained by the first two principal components.

d)

Code:

```
library(psych)

# KMO Test

KMO(data)

# Bartlett's Test

bartlett.test(data)
```

Result:

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)

Code:

```
# Plot loadings

loads<-fact_result$loadings

fa.diagram(loads)

PCA$x

plot(load,type="n")

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

plot(load)# Add labels
```

Result:

Here is the table representing the loadings for the variables across Factor1 and Factor2:

Variable	Factor1	Factor2
Age	0.159	0.509
Education_level	0.134	-0.467
Primary_Occupation		
Income_from_primary_sources	0.461	
Secondary_Occupation		
Income_from_secondary_sources	0.186	
No._of_family_member		
Years_of_Experience_in_Agriculture		0.997
Years_of_Experience_in_Banana_farming		0.8
Total_land	0.717	0.146
Total_Banana_land	0.956	0.153
Received_agarcultural_training		
Labcost_fert_app	0.985	0.139
Total_output	0.942	0.16
output_price_tk		0.173

Interpretation

- Blank cells indicate no significant loading for the respective factor.
- The loadings represent the contribution of each variable to the factors (Factor1 and Factor2).

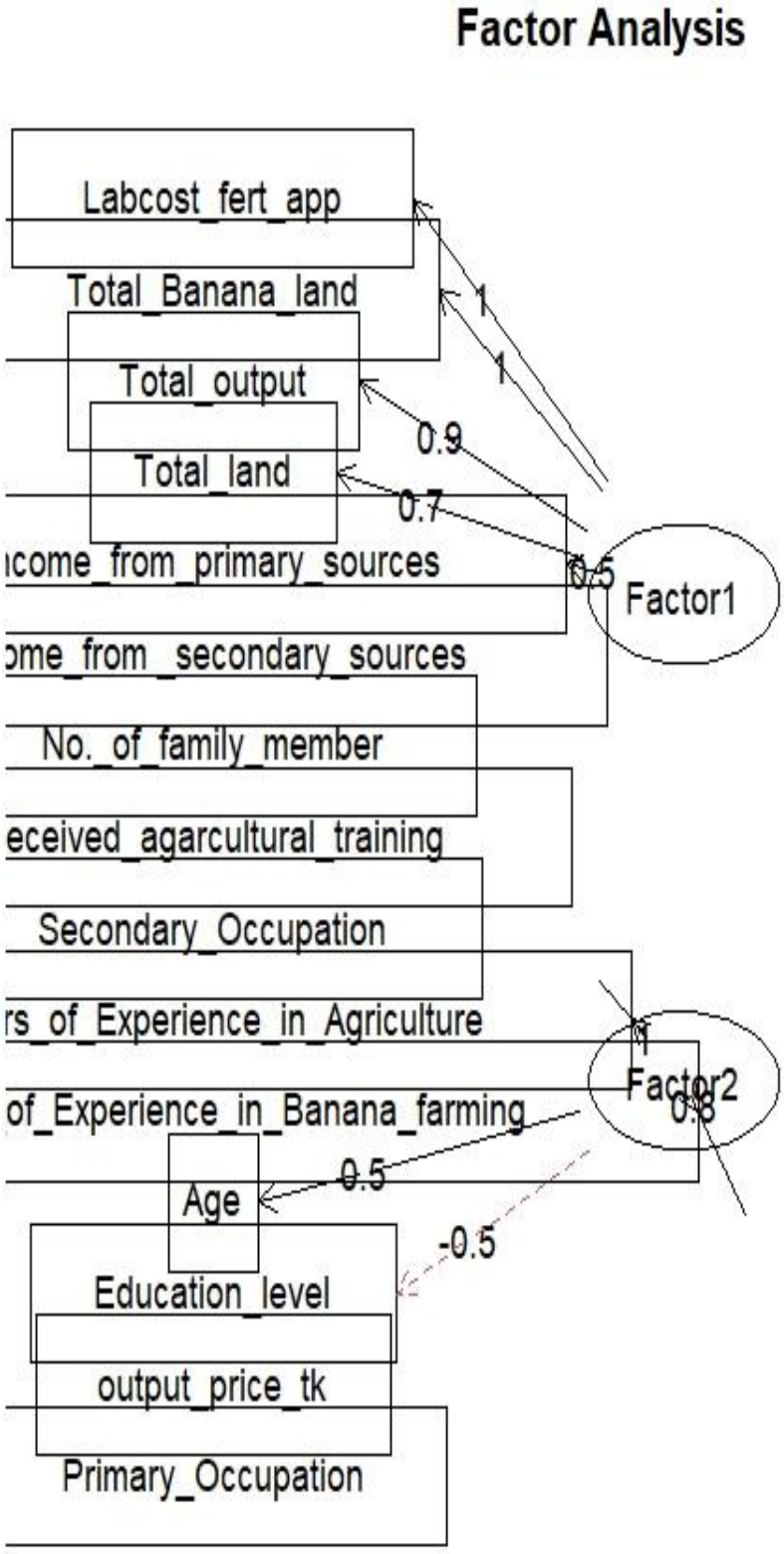

Figure: Factor Analysis

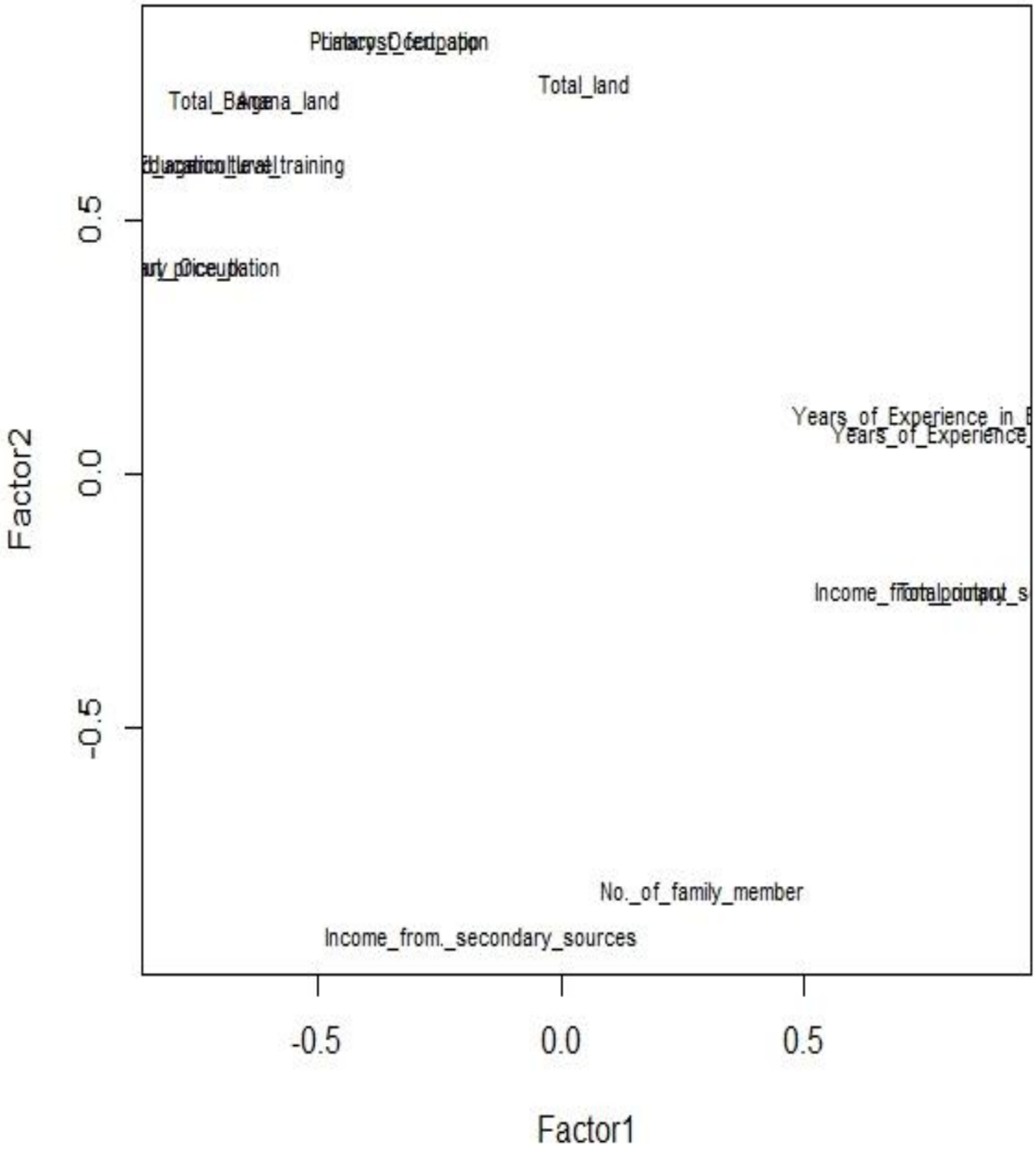

Figure: Factor loadings

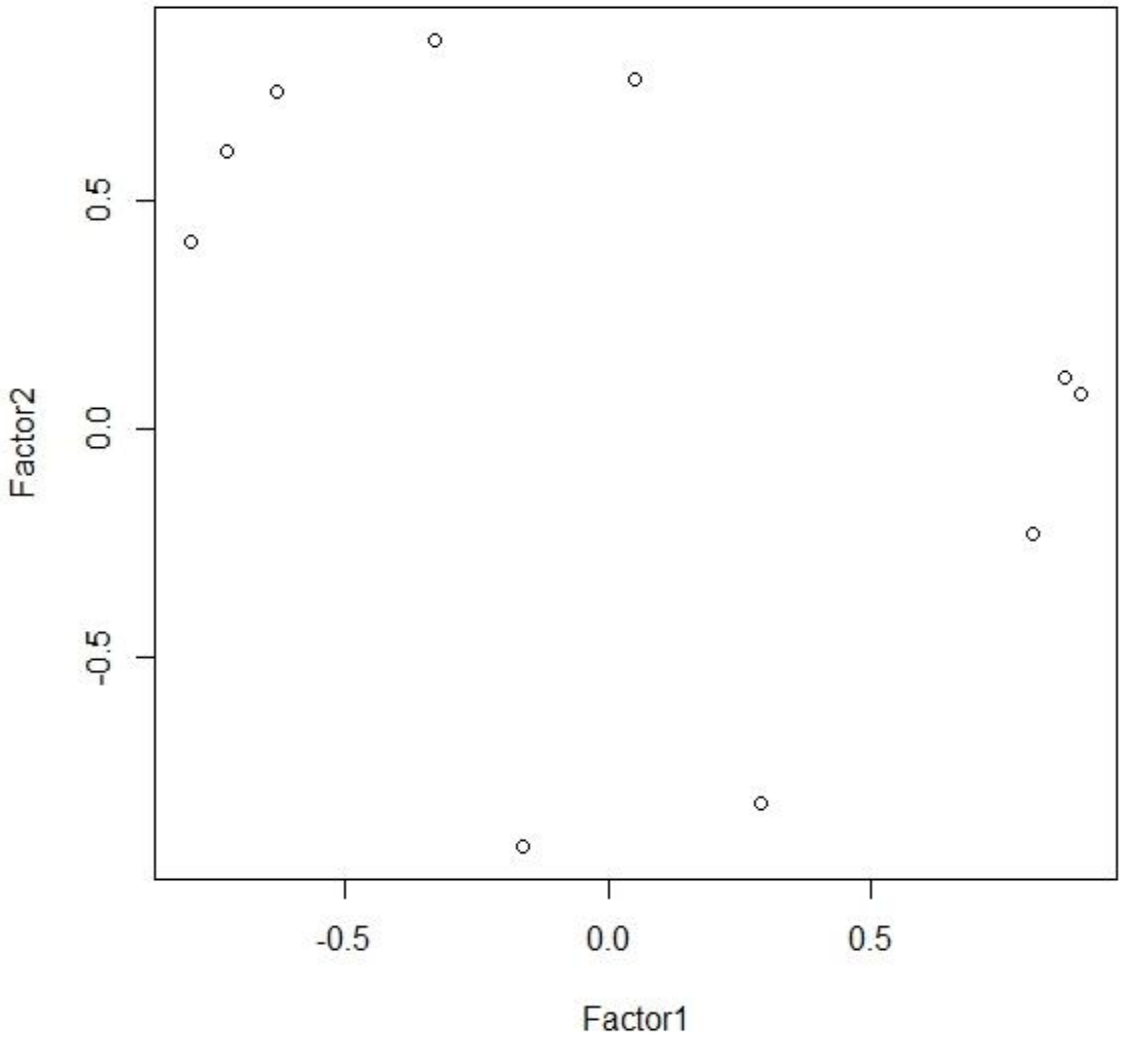

Figure: Factor loadings

Solution 2

(a)

Code

```
# 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 : ANOVA.twoFact.Factorial.RCBD

	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: 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: H_0: 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):

H_0:(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)

Code

```
library(agricolae)

# Post-hoc test for Nitrogen levels

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

Result:

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

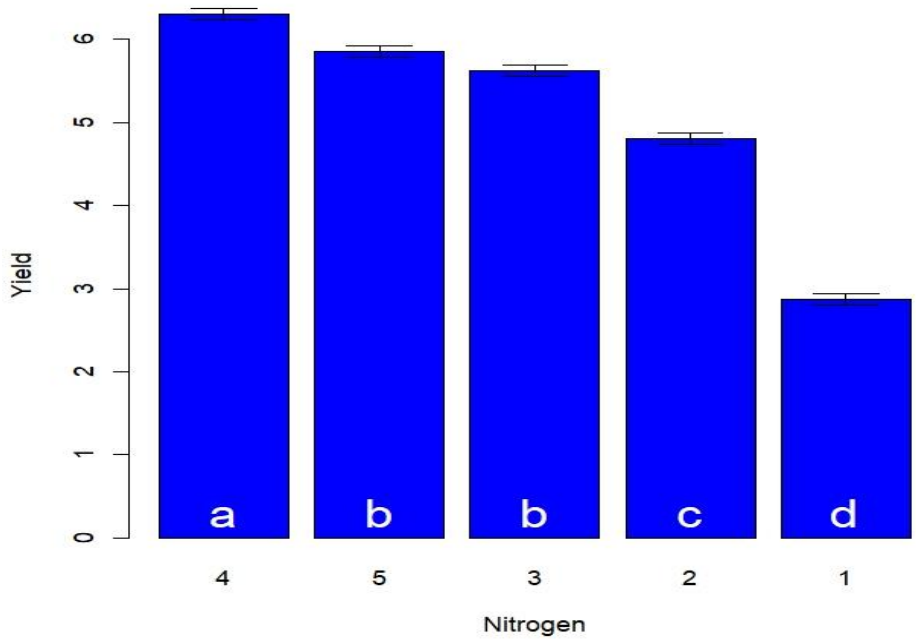

Figure: Burplot