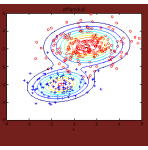


STAT 445/645 Assignment Cover Page

Student Name _____

SFU Student Number _____

SFU email address _____

Assignment Number _____

Due Date _____

Provide references for any data sets used in this assignment

List software used in this assignment.

List **ALL** resources used to complete this assignment, including books, internet sources and people.

☐ I personally completed the computations and wrote the solutions submitted in this document.

Question1

a)

a. Label the variables r1, · · · , r13. Carry out an initial investigation

```
colnames(data) <- c(paste0("r", 1:13))

head(data)
```

```
##       r1     r2     r3     r4     r5     r6     r7     r8     r9    r10    r11
## 1 13.571 12.402 14.136 12.332 12.073 17.608 12.256 14.323 13.851 13.637 15.153
## 2 16.890 13.494 15.583 11.565 11.801 16.775 12.821 13.104 14.193 13.153 15.297
## 3 14.742 13.990 16.139 14.365 12.527 17.132 13.010 14.323 15.216 14.053 17.390
## 4 14.449 13.990 16.028 13.191 13.435 15.823 12.821 14.730 15.148 14.191 16.235
## 5 15.816 14.387 15.917 13.100 13.707 17.965 13.858 15.949 13.306 12.184 13.998
## 6 14.352 13.692 17.141 13.281 13.435 17.727 14.047 15.339 14.397 12.184 14.576
##      r12    r13
## 1 13.122 12.361
## 2 13.319 13.396
## 3 14.303 16.159
## 4 13.975 14.363
## 5 11.023 13.673
## 6 12.860 14.018
```

```
summary(data)
```

```
##        r1              r2              r3              r4
##  Min.   :13.08   Min.   :12.40   Min.   :14.14   Min.   :11.56
##  1st Qu.:14.21   1st Qu.:14.04   1st Qu.:15.94   1st Qu.:12.60
##  Median :14.74   Median :14.73   Median :16.42   Median :13.42
##  Mean   :14.96   Mean   :14.77   Mean   :16.64   Mean   :13.42
##  3rd Qu.:15.74   3rd Qu.:15.48   3rd Qu.:17.31   3rd Qu.:14.25
##  Max.   :17.28   Max.   :16.57   Max.   :18.70   Max.   :15.18
##        r5              r6              r7              r8
##  Min.   :11.80   Min.   :15.70   Min.   :12.07   Min.   :13.10
##  1st Qu.:12.73   1st Qu.:17.16   1st Qu.:13.39   1st Qu.:14.58
##  Median :13.48   Median :17.96   Median :14.05   Median :15.39
##  Mean   :13.54   Mean   :17.86   Mean   :14.03   Mean   :15.38
##  3rd Qu.:14.34   3rd Qu.:18.53   3rd Qu.:14.71   3rd Qu.:15.92
##  Max.   :15.70   Max.   :19.87   Max.   :15.74   Max.   :17.78
##        r9             r10             r11             r12
##  Min.   :12.90   Min.   :12.18   Min.   :13.56   Min.   :11.02
##  1st Qu.:14.19   1st Qu.:13.85   1st Qu.:14.77   1st Qu.:13.32
##  Median :14.64   Median :14.16   Median :15.30   Median :13.94
##  Mean   :14.80   Mean   :14.25   Mean   :15.51   Mean   :13.91
##  3rd Qu.:15.49   3rd Qu.:14.80   3rd Qu.:16.07   3rd Qu.:14.60
##  Max.   :17.06   Max.   :16.75   Max.   :17.46   Max.   :15.42
##       r13
##  Min.   :12.36
##  1st Qu.:14.19
##  Median :14.85
##  Mean   :14.89
##  3rd Qu.:15.74
##  Max.   :16.50
```

```
sapply(data, sd)
```

```
##       r1       r2       r3       r4       r5       r6       r7       r8
## 1.017079 1.017001 1.017043 1.017078 1.017046 1.017201 1.017093 1.017149
##       r9      r10      r11      r12      r13
## 1.017061 1.017147 1.017202 1.017032 1.017085
```

```
sapply(data, function(x) sum(is.na(x)))
```

```
##  r1  r2  r3  r4  r5  r6  r7  r8  r9 r10 r11 r12 r13
##   0   0   0   0   0   0   0   0   0   0   0   0   0
```

I showed making the label the variables r1,...,r13 first capture.

Upon examining the basic statistical measures, it was observed that the mean values and ranges of all variables are somewhat consistent with each other, indicating that the data might have been collected under stable and consistent condition.

No missing values were found in the dataset, and no apparent outliers were detected during the initial analysis. A brief examination of the correlation between variables showed moderate relationships among some of them.

In conclusion, this initial investigation suggests that the dataset is well structured and provides sufficient information for analysis.

b)

b. Display the sample correlation matrix R

```
cor(data)
```

```
##              r1        r2        r3        r4         r5         r6         r7
## r1   1.00000000 0.3758357 0.5544593 0.3158481 0.32359950 0.19049226 0.14839299
## r2   0.37583570 1.0000000 0.5467768 0.4339709 0.61027680 0.57754762 0.54977173
## r3   0.55445934 0.5467768 1.0000000 0.4231420 0.60054436 0.39421647 0.29619350
## r4   0.31584813 0.4339709 0.4231420 1.0000000 0.60279388 0.60835370 0.59439330
## r5   0.32359950 0.6102768 0.6005444 0.6027939 1.00000000 0.61781229 0.55615497
## r6   0.19049226 0.5775476 0.3942165 0.6083537 0.61781229 1.00000000 0.59575943
## r7   0.14839299 0.5497717 0.2961935 0.5943933 0.55615497 0.59575943 1.00000000
## r8   0.42956255 0.6539967 0.6222676 0.5749846 0.73599254 0.56081467 0.55230621
## r9   0.03866564 0.1824044 0.2941651 0.3654132 0.09713571 0.02258929 0.07794617
## r10 -0.00798961 0.2535936 0.3397079 0.4124751 0.10429513 0.21622446 0.23811362
## r11  0.16112069 0.2410265 0.3822270 0.4365231 0.24058341 0.20721163 0.28835312
## r12  0.20994352 0.4265771 0.3855376 0.3479641 0.16624698 0.22432553 0.02965357
## r13  0.12981078 0.4846660 0.5352682 0.5072707 0.38859471 0.32274993 0.20046340
##              r8         r9         r10       r11        r12       r13
## r1   0.42956255 0.03866564 -0.00798961 0.1611207 0.20994352 0.1298108
## r2   0.65399669 0.18240439  0.25359357 0.2410265 0.42657714 0.4846660
## r3   0.62226759 0.29416512  0.33970788 0.3822270 0.38553756 0.5352682
## r4   0.57498465 0.36541316  0.41247509 0.4365231 0.34796407 0.5072707
## r5   0.73599254 0.09713571  0.10429513 0.2405834 0.16624698 0.3885947
## r6   0.56081467 0.02258929  0.21622446 0.2072116 0.22432553 0.3227499
## r7   0.55230621 0.07794617  0.23811362 0.2883531 0.02965357 0.2004634
## r8   1.00000000 0.06274553  0.17832580 0.2911396 0.20579107 0.2886055
## r9   0.06274553 1.00000000  0.68928820 0.5313111 0.39426584 0.6056791
## r10 0.17832580 0.68928820  1.00000000 0.5403106 0.56648705 0.4865274
## r11 0.29113962 0.53131110  0.54031058 1.0000000 0.35594918 0.5636695
## r12 0.20579107 0.39426584  0.56648705 0.3559492 1.00000000 0.6023203
## r13 0.28860555 0.60567909  0.48652736 0.5636695 0.60232031 1.0000000
```

The sample correlation matrix R is

$$\begin{bmatrix} 1.000 & 0.376 & 0.554 & 0.316 & 0.324 & 0.190 & 0.148 & 0.430 & 0.039 & -0.008 & 0.161 & 0.210 & 0.130 \\ 0.376 & 1.000 & 0.547 & 0.434 & 0.610 & 0.578 & 0.550 & 0.654 & 0.182 & 0.254 & 0.241 & 0.427 & 0.485 \\ 0.554 & 0.547 & 1.000 & 0.423 & 0.601 & 0.394 & 0.296 & 0.622 & 0.294 & 0.340 & 0.382 & 0.386 & 0.535 \\ 0.316 & 0.434 & 0.423 & 1.000 & 0.603 & 0.608 & 0.594 & 0.575 & 0.365 & 0.412 & 0.437 & 0.348 & 0.507 \\ 0.324 & 0.610 & 0.601 & 0.603 & 1.000 & 0.618 & 0.556 & 0.736 & 0.097 & 0.104 & 0.241 & 0.166 & 0.389 \\ 0.190 & 0.578 & 0.394 & 0.608 & 0.618 & 1.000 & 0.596 & 0.561 & 0.023 & 0.216 & 0.207 & 0.224 & 0.323 \\ 0.148 & 0.550 & 0.296 & 0.594 & 0.556 & 0.596 & 1.000 & 0.552 & 0.078 & 0.238 & 0.288 & 0.030 & 0.200 \\ 0.430 & 0.654 & 0.622 & 0.575 & 0.736 & 0.561 & 0.552 & 1.000 & 0.063 & 0.178 & 0.291 & 0.206 & 0.289 \\ 0.039 & 0.182 & 0.294 & 0.365 & 0.097 & 0.023 & 0.078 & 0.063 & 1.000 & 0.689 & 0.531 & 0.394 & 0.606 \\ -0.008 & 0.254 & 0.340 & 0.412 & 0.104 & 0.216 & 0.238 & 0.178 & 0.689 & 1.000 & 0.540 & 0.566 & 0.487 \\ 0.161 & 0.241 & 0.382 & 0.437 & 0.241 & 0.207 & 0.288 & 0.291 & 0.531 & 0.540 & 1.000 & 0.356 & 0.564 \\ 0.210 & 0.427 & 0.386 & 0.348 & 0.166 & 0.224 & 0.030 & 0.206 & 0.394 & 0.566 & 0.356 & 1.000 & 0.602 \\ 0.130 & 0.485 & 0.535 & 0.507 & 0.389 & 0.323 & 0.200 & 0.289 & 0.606 & 0.487 & 0.564 & 0.602 & 1.000 \end{bmatrix}$$

c)

c. (i)List the eigenvalues and describe the percent contributions to the variance

```
result <- prcomp(data, scale. = TRUE)

eigenvalues <- result$sdev^2
percent_variance <- eigenvalues / sum(eigenvalues) * 100
eigenvalues
```

```
##  [1] 5.6555117 2.2983724 1.1918400 0.7987265 0.5930113 0.5137613 0.4675740
##  [8] 0.4230029 0.3472208 0.2715732 0.2073550 0.1285641 0.1034869
```

```
percent_variance
```

```
##  [1] 43.5039362 17.6797876  9.1679996  6.1440496  4.5616254  3.9520100
##  [7]  3.5967232  3.2538682  2.6709289  2.0890249  1.5950382  0.9889548
## [13]  0.7960533
```

The eigenvalue for the first principal component (PC1) is 5.655 and the first principal component explains 43.5% of the total variance.

The eigenvalue for the second principal component (PC2) is 2.300 and the second principal component explains 17.68% of the total variance.

The eigenvalue for (PC3) is 1.192 and the PC3 explains 9.17% of the total variance.

The eigenvalue for (PC4) is 0.799 and the PC4 explains 6.14% of the total variance.

The eigenvalue for (PC5) is 0.593 and the PC5 explains 4.56% of the total variance.

The eigenvalue for (PC6) is 0.514 and the PC6 explains 3.95% of the total variance.

The eigenvalue for (PC7) is 0.468 and the PC7 explains 3.60% of the total variance.

The eigenvalue for (PC8) is 0.423 and the PC8 explains 3.25% of the total variance.

The eigenvalue for (PC9) is 0.347 and the PC9 explains 2.67% of the total variance.

The eigenvalue for (PC10) is 0.272 and the PC10 explains 2.09% of the total variance.

The eigenvalue for (PC11) is 0.207 and the PC11 explains 1.60% of the total variance.

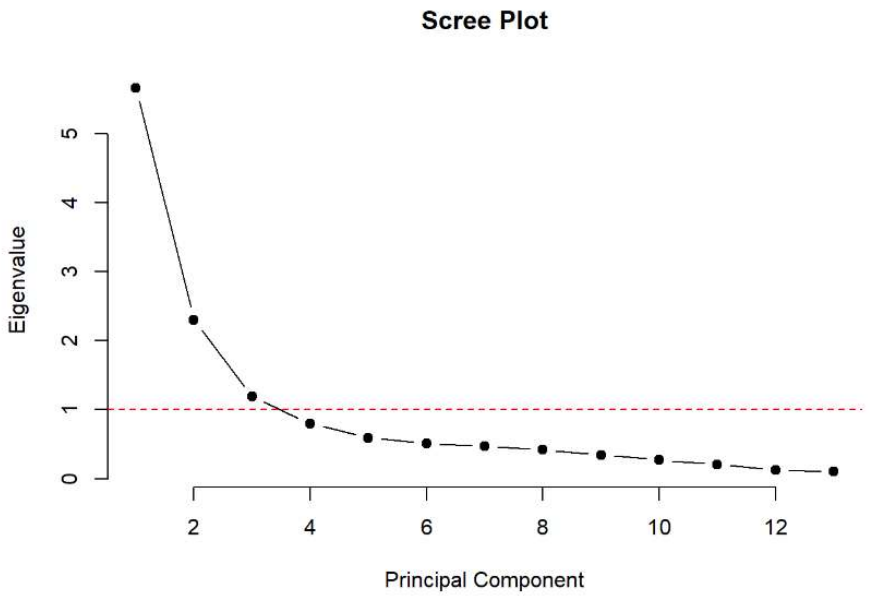
The eigenvalue for (PC12) is 0.129 and the PC12 explains 0.99% of the total variance.

The eigenvalue for (PC13) is 0.103 and the PC13 explains 0.80% of the total variance.

The calculated eigenvalues represent how the variance in the data is decomposed according to each principal component. Each eigenvalue expresses the proportion of total variance accounted for by its corresponding principal component. The percentage calculated from these eigenvalues indicates the contribution of each principal component to the total variance. This allows us to identify which principal components best explain the major structures and patterns in the data.

(ii)

```
plot(eigenvalues, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
abline(h = 1, col = "red", lty = 2)
```



I use these three methods.

First, Kaiser Criterion: This method involves retaining principal components with eigenvalues greater than 1. According to this criterion, components with eigenvalues exceeding 1are considered significant and are retained.

Second, Scree Plot: This plot displays the eigenvalues in descending order,and we look for the 'elbow point' where the decline in eigenvalues becomes less steep. Principal components before this point are retained, as they represent the most significant sources of variance.

Third, Cumulative Variance Contribution: Principal components are retained until the cumulative contribution of variance reaches between 70% to 90% of the total variance. In this case,the first three principal components explain over 80% of the total variance, hence they are retained.

Based on the eigenvalues obtained from the PCA, PC1 was about 43.5% of the variance, while the PC2 and PC3 was 17.7% and 9.2% respectively. Together, these three components explains about 70.4% of the total variance. This is slightly below the often-cited threshold of 80% for retaining principal components but is still a substantial portion of the variance.

Applying the Kaiser criterion, which recommends retaining principal components with eigenvalues greater than 1, we would retain the first three components as their eigenvalues exceed this threshold, Moreover, if a scree plot is examined and shows an 'elbow' at PC3, it would further suggests that retaining more components does not contributes significant additional information.

Therefore taking into account the Kaiser criterion, the explained variance of the first three

components,and the hypothetical elbow point in the scree plot, we decide to retain the first three principal components. This decision provides a balance between maximizing explained variance and maintaining model simplicity, ensuring a comprehensive yet efficient representation of the data structure.

iii. Give the eigenvectors for the first two principal components

```
eigenvectors <- result$rotation[, 1:2]
print(eigenvectors)
```

```
##            PC1         PC2
## r1  0.1862865  0.17337352
## r2  0.3215932  0.15768862
## r3  0.3178165  0.03830568
## r4  0.3320749  0.03498463
## r5  0.3142632  0.28569319
## r6  0.2863216  0.24573299
## r7  0.2634353  0.24625611
## r8  0.3173252  0.28456896
## r9  0.1985863 -0.46971010
## r10 0.2336784 -0.41568026
## r11 0.2492567 -0.29181710
## r12 0.2327297 -0.30876741
## r13 0.2990402 -0.29128272
```

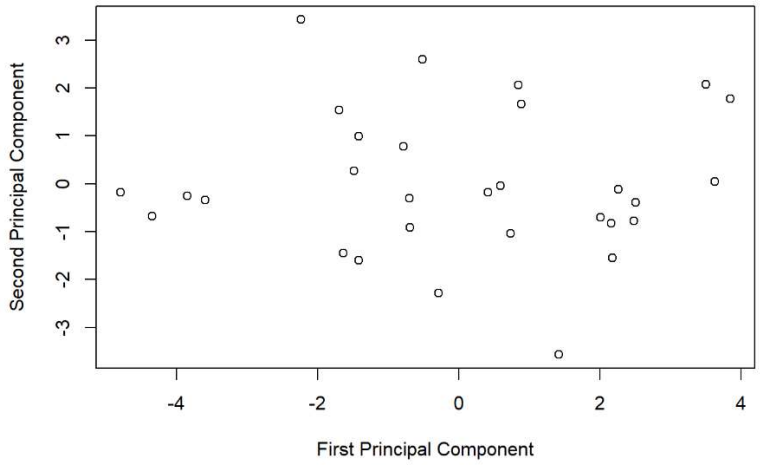
The eigenvectors for the first two principal components is $\begin{bmatrix} 0.186 & 0.173 \\ 0.322 & 0.158 \\ 0.318 & 0.038 \\ 0.332 & 0.035 \\ 0.314 & 0.286 \\ 0.286 & 0.246 \\ 0.263 & 0.246 \\ 0.317 & 0.285 \\ 0.199 & -0.470 \\ 0.234 & -0.416 \\ 0.250 & -0.292 \\ 0.233 & -0.309 \\ 0.230 & -0.291 \end{bmatrix}$

The eigenvectors extracted from the first two principal components capture the major variations in the data. The eigenvector of the first principal component explains the largest part of the data variability, while the second principal component's eigenvector explains the next significant portion. These eigenvectors indicate how each variable in the original dataset contributes to these principal components. For instance, the elements of each eigenvector represent the extent to which the corresponding variable contributes to the principal component, allowing us to understand the significant structural features of the data.

(iv) The coefficients, or loadings, of the first two principal components provide insights into how each variable in the dataset contributes to these components. For example, if a variable has a high loading on the first principal component, it means that this variable plays a significant role in accounting for the variability that the first principal component represents. Conversely, if a variable has a low loading on the first principal component but a high loading on the second, it indicates that the variable is more related to the variability captured by the second principal component. By analyzing these coefficients, we can understand the dependencies of the principal components on the variables, revealing which variables are most influential in defining each principal component and thus the underlying structure of the data.

v. Display a scatter plot of the first two principal components

```
plot(result$x[, 1], result$x[, 2], xlab="First Principal Component", ylab="Second Principal Component")
```



If clear clustering is observed in the scatter plot, it indicates that there may be naturally distinct groups within the data. Additionally, the scatter plot helps to understand how principal component analysis captures the major variability in the data. The first principal component explains the largest part of the data variability, while the second principal component accounts for the next significant portion.

Question2

a)

a. Carry out an initial investigation

```
colnames(data) <- c("Experiment", "x1_oleracea", "x2_oleracea", "x3_oleracea", "x4_oleracea", "x1_carduorum", "x2_carduorum", "x3_carduorum", "x4_carduorum")


str(data)
```

```
## tibble [20 × 9] (S3: tbl_df/tbl/data.frame)
##  $ Experiment  : num [1:20] 1 2 3 4 5 6 7 8 9 10 ...
##  $ x1_oleracea : num [1:20] 189 192 217 221 171 192 213 192 170 201 ...
##  $ x2_oleracea : num [1:20] 245 260 276 299 239 262 278 255 244 276 ...
##  $ x3_oleracea : num [1:20] 137 132 141 142 128 147 136 128 128 146 ...
##  $ x4_oleracea : num [1:20] 163 217 192 213 158 173 201 185 192 186 ...
##  $ x1_carduorum: num [1:20] 181 158 184 171 181 181 177 198 180 177 ...
##  $ x2_carduorum: num [1:20] 305 237 300 273 297 308 301 308 286 299 ...
##  $ x3_carduorum: num [1:20] 184 133 166 162 163 160 166 141 146 171 ...
##  $ x4_carduorum: num [1:20] 209 188 231 213 224 223 221 197 214 192 ...
```

```
summary(data)
```

```
##    Experiment     x1_oleracea     x2_oleracea     x3_oleracea   
##  Min.   : 1.00   Min.   :170.0   Min.   :239.0   Min.   :121.0  
##  1st Qu.: 5.75   1st Qu.:190.5   1st Qu.:253.5   1st Qu.:130.0  
##  Median :10.50   Median :192.0   Median :263.0   Median :138.0  
##  Mean   :10.50   Mean   :194.5   Mean   :267.1   Mean   :137.4  
##  3rd Qu.:15.25   3rd Qu.:200.5   3rd Qu.:280.5   3rd Qu.:144.0  
##  Max.   :20.00   Max.   :221.0   Max.   :299.0   Max.   :150.0  
##                  NA's   :1       NA's   :1       NA's   :1      
##   x4_oleracea     x1_carduorum    x2_carduorum    x3_carduorum   
##  Min.   :158.0   Min.   :158.0   Min.   :237.0   Min.   :133.0  
##  1st Qu.:175.0   1st Qu.:175.8   1st Qu.:277.5   1st Qu.:146.8  
##  Median :186.0   Median :180.5   Median :298.0   Median :161.0  
##  Mean   :185.9   Mean   :179.6   Mean   :290.8   Mean   :157.2  
##  3rd Qu.:192.0   3rd Qu.:181.8   3rd Qu.:305.8   3rd Qu.:166.0  
##  Max.   :217.0   Max.   :198.0   Max.   :317.0   Max.   :184.0  
##  NA's   :1                                                      
##   x4_carduorum  
##  Min.   :188.0  
##  1st Qu.:199.2  
##  Median :209.0  
##  Mean   :209.2  
##  3rd Qu.:215.8  
##  Max.   :235.0  
##
```

```
sapply(data, var, na.rm = TRUE)
```

```
##   Experiment  x1_oleracea  x2_oleracea  x3_oleracea  x4_oleracea x1_carduorum
##     35.00000    187.59649    345.38596     66.35673    239.94152    101.83947
## x2_carduorum x3_carduorum x4_carduorum
##    389.01053    167.53684    177.88158
```

```
sapply(data, function(x) sum(is.na(x)))
```

```
##   Experiment  x1_oleracea  x2_oleracea  x3_oleracea  x4_oleracea x1_carduorum
##            0            1            1            1            1            0
## x2_carduorum x3_carduorum x4_carduorum
##            0            0            0
```

An initial investigation of the flea beetle data was conducted, focusing on the physical size measurement of two species, 'Oleracea' and 'Carduorum'. The mean values were analyzed to understand the overall body size of each group. The mean of x2_Olearacea was 267.1 and x2_Carduorum was 290.8, indicating that mean of Carduorum has larger body size.

The range of body sizes was assessed by comparing the maximum and minimum values. x2_Oleracea showed a size range from 239 to 299,whereas x2_Carduorum had a broader range from 237 to 317, suggesting a wider variation in size for Carduorum.

The variability of measurements within each group was evaluated using the variance. Both exhibited relatively high variances in some variables, indicating significant morphological variation within each species.

A few missing values were detected, but their small number is unlikely to significantly impact the overall analysis results.

b. i. Display the relevant sample covariance matrix S

```
oleracea<- data[, 2:5]
oleracea
```

```
## # A tibble: 20 × 4
##    x1_oleracea x2_oleracea x3_oleracea x4_oleracea
##          <dbl>       <dbl>       <dbl>       <dbl>
##  1         189         245         137         163
##  2         192         260         132         217
##  3         217         276         141         192
##  4         221         299         142         213
##  5         171         239         128         158
##  6         192         262         147         173
##  7         213         278         136         201
##  8         192         255         128         185
##  9         170         244         128         192
## 10         201         276         146         186
## 11         195         242         128         192
## 12         205         263         147         192
## 13         180         252         121         167
## 14         192         283         138         183
## 15         200         294         138         188
## 16         192         277         150         177
## 17         200         287         136         173
## 18         181         255         146         183
## 19         192         287         141         198
## 20          NA          NA          NA          NA
```

```
cov(oleracea, use="complete.obs")
```

```
##             x1_oleracea x2_oleracea x3_oleracea x4_oleracea
## x1_oleracea   187.59649   176.86257    48.37135   113.58187
## x2_oleracea   176.86257   345.38596    75.97953   118.78070
## x3_oleracea    48.37135    75.97953    66.35673    16.24269
## x4_oleracea   113.58187   118.78070    16.24269   239.94152
```

The sample covariance matrix S is $\begin{bmatrix} 187.60 & 176.86 & 48.37 & 113.58 \\ 176.86 & 345.39 & 75.98 & 118.78 \\ 48.37 & 75.98 & 66.36 & 16.24 \\ 113.58 & 118.78 & 16.24 & 239.94 \end{bmatrix}$

ii. List the eigenvalues and describe the percent contributions to the variance

```
eigen_result<-eigen(cov(oleracea, use="complete.obs"))

eigenvalues <- eigen_result$values

total <- sum(eigenvalues)

percent_variance <- eigenvalues / total * 100

eigenvalues
```

```
## [1] 561.30574 168.98584  65.27709  43.71203
```

```
percent_variance
```

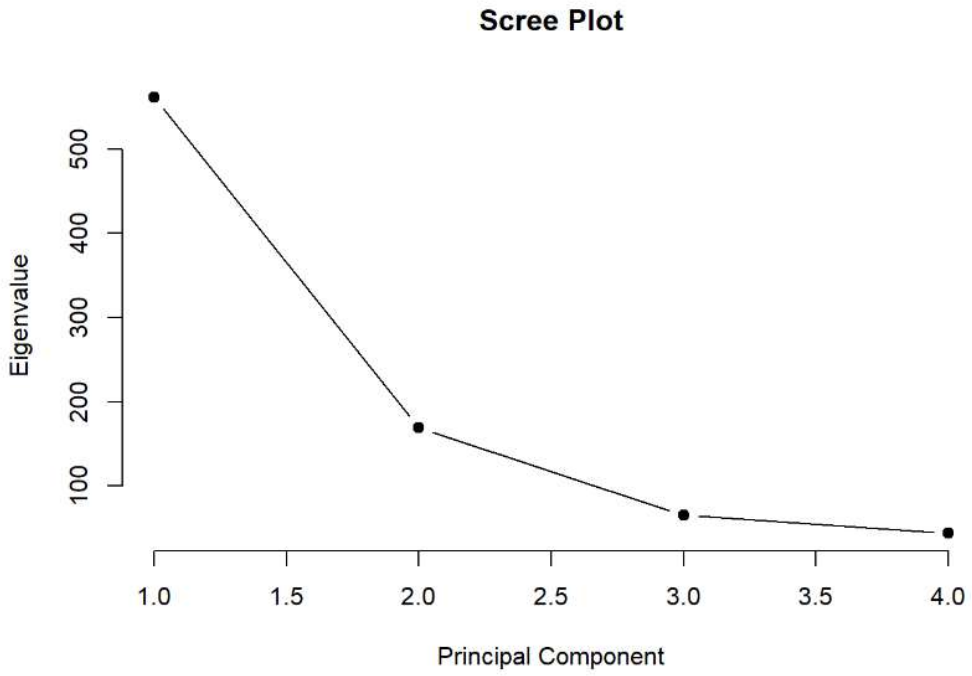
```
## [1] 66.879382 20.134603  7.777743  5.208273
```

The eigenvalue for (PC1) is 561.31 and the PC1 explains 66.88% of the total variance.
The eigenvalue for (PC2) is 168.99 and the PC2 explains 20.13% of the total variance.
The eigenvalue for (PC3) is 65.28 and the PC3 explains 7.78% of the total variance.
The eigenvalue for (PC4) is 43.71 and the PC4 explains 5.21% of the total variance.

The calculated eigenvalues represent how the variance in the data is decomposed according to each principal component. Each eigenvalue expresses the proportion of total variance accounted for by its corresponding principal component. The percentage calculated from these eigenvalues indicates the contribution of each principal component to the total variance. This allows us to identify which principal components best explain the major structures and patterns in the data.

(iii)

```
plot(eigenvalues, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
```



In the principal component analysis, the first principal component accounts for approximately 66.9% of the total variance, and the second principal component explains an additional 20.1%. Together, these account for around 87% of the total variance, suggesting that they encapsulate the majority of the information in the dataset. Thus, it seems reasonable to retain only the first two principal components. According to the eigenvalue-one criterion, only the first and second principal components satisfy this rule. When considering the cumulative percentage of variance explained, a common practice is to retain components that account for at least 70% of the variance, and the first two components exceed this threshold. Lastly, a scree plot would likely show a leveling off after the second component, indicating that subsequent components contribute significantly less to explaining the variance. Therefore, retaining the first two principal components is supported by these consistent findings across the different methods.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors <- eigen_result$vectors

eigenvectors
```

```
##             [,1]          [,2]       [,3]       [,4]
## [1,] -0.4997445  0.009204574  0.8230272  0.2698089
## [2,] -0.7187015 -0.484408702 -0.4778690  0.1430301
## [3,] -0.1739702 -0.220296505  0.2042647 -0.9378058
## [4,] -0.4510631  0.846600812 -0.2292234 -0.1651236
```

The eigenvectors for the principal components you retain is $\begin{bmatrix} -0.50 & 0.01 & 0.82 & 0.27 \\ -0.72 & -0.49 & -0.48 & 0.14 \\ -0.17 & -0.22 & 0.20 & -0.94 \\ -0.45 & 0.85 & -0.23 & -0.17 \end{bmatrix}$

The PC1 has a strong positive weight on x3_Oleracea (0.82). This suggests that PC1 is primarily associated with x3_Olearcea in a positive direction.

The PC2 shows a very strong positive weight on x4_Oleracea (0.85) and a negative weight on x2_Olearacea (-0.49). This indicates that PC2 is capturing variation in x4_Olearacea in a positive direction and x2_Olearacea in a negative direction.

The PC3 carries a strong negative weight for x4_Oleracea (-0.94), implying a negative association with the variable.

The PC4 has relatively small weights across all variables, with the highest positive weight on x4_Olearacea (0.85).

(v) The PC1 has strong positive relationship on the third variable (0.82). This indicates a dependency but for the second variable (0.01), it has no dependency because it is close to 0.

The PC2 has strong negative relationship on the first variable (-0.72). This indicates a dependency.

The PC3 has strong negative relationship on the fourth variable (-0.94), indicating a dependency.

The PC4 has strong positive relationship on the second variable (0.85), indicating a dependency because it is close to 1.

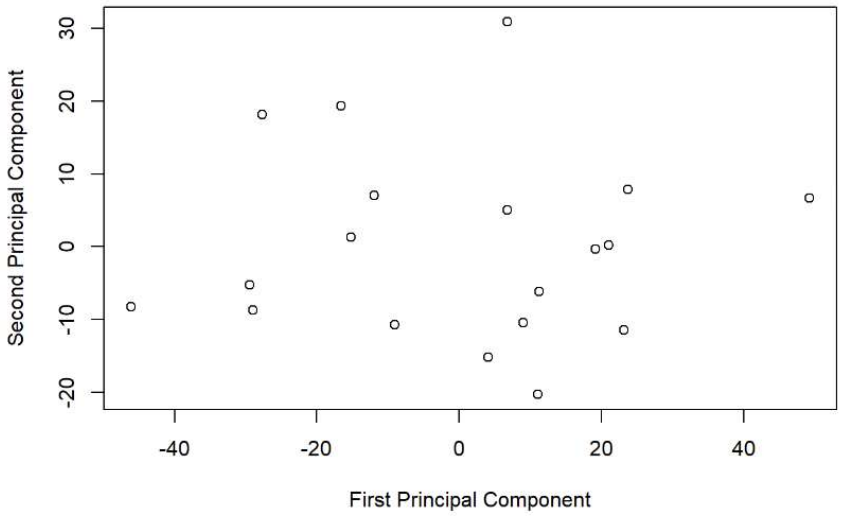
vi. Using at least the first two principal components, display scatter plots of pairs of principal components

```
oleracea <- na.omit(oleracea)

oleracea <- oleracea[!apply(oleracea, 1, function(x) any(is.infinite(x))), ]

pca_result <- prcomp(oleracea, scale. = FALSE)

plot(pca_result$x[, 1], pca_result$x[, 2],
     xlab="First Principal Component", ylab="Second Principal Component")
```



Most data points are clustered around the first principal component, while there is a wider spread along the second principal component. This suggests that the first principal component explains a significant portion of the variability in the data, and the second principal component captures additional variability.

c. i. Display the relevant sample covariance matrix S

```
carduorum<- data[, 6:9]
carduorum
```

```
## # A tibble: 20 × 4
##    x1_carduorum x2_carduorum x3_carduorum x4_carduorum
##           <dbl>        <dbl>        <dbl>        <dbl>
##  1          181          305          184          209
##  2          158          237          133          188
##  3          184          300          166          231
##  4          171          273          162          213
##  5          181          297          163          224
##  6          181          308          160          223
##  7          177          301          166          221
##  8          198          308          141          197
##  9          180          286          146          214
## 10          177          299          171          192
## 11          176          317          166          213
## 12          192          312          166          209
## 13          176          285          141          200
## 14          169          287          162          214
## 15          164          265          147          192
## 16          181          308          157          204
## 17          192          276          154          209
## 18          181          278          149          235
## 19          175          271          140          192
## 20          197          303          170          205
```

```
cov(carduorum, use="complete.obs")
```

```
##              x1_carduorum x2_carduorum x3_carduorum x4_carduorum
## x1_carduorum    101.83947    128.06316     36.98947     32.59211
## x2_carduorum    128.06316    389.01053    165.35789     94.36842
## x3_carduorum     36.98947    165.35789    167.53684     66.52632
## x4_carduorum     32.59211     94.36842     66.52632    177.88158
```

The relevant sample covariance matrix S is $\begin{bmatrix} 101.84 & 128.06 & 36.99 & 32.60 \\ 128.06 & 389.01 & 165.36 & 94.37 \\ 36.99 & 165.36 & 167.54 & 66.53 \\ 32.60 & 94.37 & 66.53 & 177.88 \end{bmatrix}$

ii. List the eigenvalues and describe the percent contributions to the variance

```
eigen_carduorum<-eigen(cov(carduorum, use="complete.obs"))
eigen_carduorum
```

```
## eigen() decomposition
## $values
## [1] 555.69314 145.44632  93.46372  41.66524
##
## $vectors
##             [,1]       [,2]       [,3]        [,4]
## [1,] -0.2836552 -0.2007357  0.5315166 -0.77248627
## [2,] -0.8068689 -0.3389760  0.1218433  0.46820095
## [3,] -0.4222422  0.1359900 -0.7897513 -0.42368751
## [4,] -0.3003563  0.9090144  0.2809577  0.06739234
```

```
eigenvalues_carduorum <- eigen_carduorum$values

total_carduorum <- sum(eigenvalues_carduorum)

percent_variance <- eigenvalues_carduorum / total_carduorum * 100

eigenvalues_carduorum
```

```
## [1] 555.69314 145.44632  93.46372  41.66524
```

```
percent_variance
```

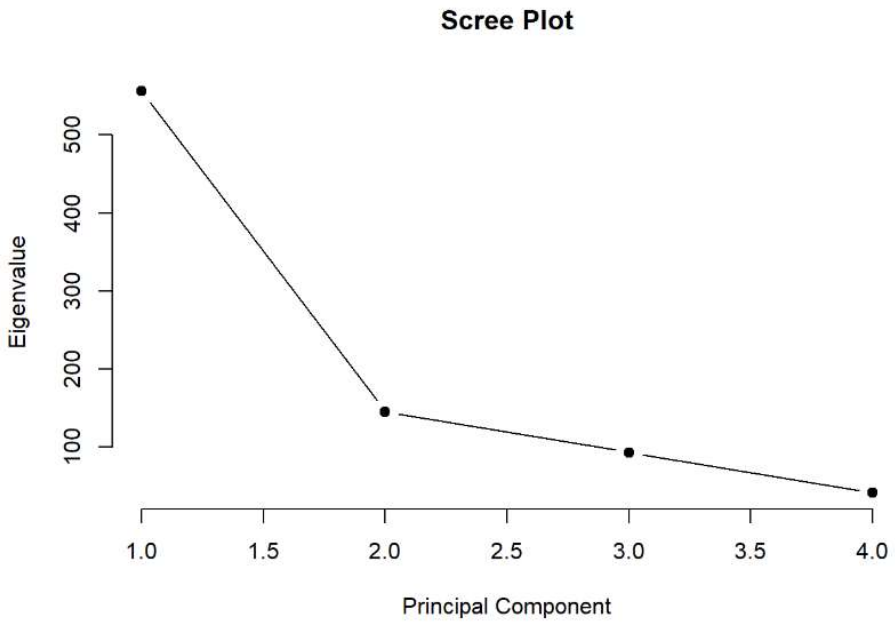
```
## [1] 66.44914 17.39230 11.17628  4.98228
```

The eigenvalue for (PC1) is 555.69 and the PC1 explains 66.45% of the total variance.
The eigenvalue for (PC2) is 145.45 and the PC2 explains 17.39% of the total variance.
The eigenvalue for (PC3) is 93.46 and the PC3 explains 11.18% of the total variance.
The eigenvalue for (PC4) is 41.67 and the PC4 explains 4.98% of the total variance.

According to the given results, the first principal component explains about 66.44% of the total variance, and the second principal component accounts for approximately 17.39%. Thus, the first two principal components together explain roughly 83.83% of the data variance. This information can be useful for dimensionality reduction, visualization, or selecting features for further analysis.

iii. scree plot

```
plot(eigenvalues_carduorum, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
```



To determine the number of principal components to retain, three methods were considered: the Kaiser Criterion, the Scree Plot, and the Cumulative Variance proportion. According to the Kaiser Criterion, only components with eigenvalues over 1 should be retained, which would suggest keeping only the first principal component. The Scree Plot shows a clear 'elbow' after the first principal component, indicating that the second component could also be significant. Regarding the Cumulative Variance, it is common practice to retain components that account for at least 80% of the variance; in this case, the first two components explain approximately 83.83% of the variance. Integrating these three methods, it is appropriate to retain the first two principal components to capture the major variance in the data while maintaining analytical efficiency.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors <- eigen_carduorum$vectors

eigenvectors
```

```
##             [,1]       [,2]       [,3]        [,4]
## [1,] -0.2836552 -0.2007357  0.5315166 -0.77248627
## [2,] -0.8068689 -0.3389760  0.1218433  0.46820095
## [3,] -0.4222422  0.1359900 -0.7897513 -0.42368751
## [4,] -0.3003563  0.9090144  0.2809577  0.06739234
```

The eigenvectors for the principal components you retain is $\begin{bmatrix} -0.284 & -0.201 & 0.532 & -0.772 \\ -0.807 & -0.339 & 0.122 & 0.468 \\ -0.422 & 0.136 & -0.790 & -0.424 \\ -0.300 & 0.909 & 0.281 & 0.067 \end{bmatrix}$

(v) The PC1 has strong negative relationship on the fourth variable (-0.772). This indicates a dependency because it is close to -1.

The PC2 has strong negative relationship on the first variable (-0.807). This indicates a dependency.

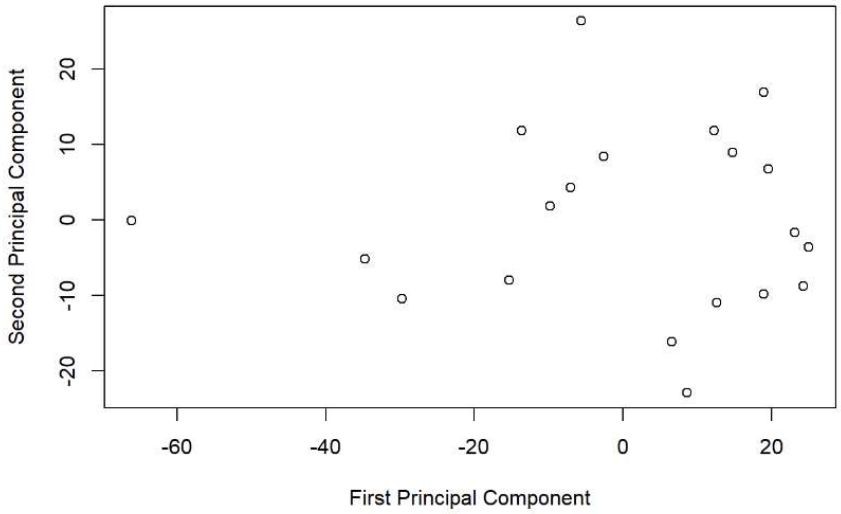
The PC3 has strong negative relationship on the third variable (-0.790), indicating a dependency.

The PC4 has strong positive relationship on the second variable (0.909), indicating a dependency but for the fourth variable (0.067), it has no dependency because it is close to 0.

vi. Using at least the first two principal components, display scatter plots of pairs of principal components

```
pca_result_carduorum <- prcomp(carduorum, scale. = FALSE)

plot(pca_result_carduorum$x[, 1], pca_result_carduorum$x[, 2],
     xlab="First Principal Component", ylab="Second Principal Component")
```



The points are clustered right side. This means the PC1 looks clustered while there is a wider spread along the PC2. This suggests that the first principal component explains a significant portion of the variability in the data, and the second principal component captures additional variability.

For the entire data set

d. i. Display the relevant sample covariance matrix S

```
entire<-rbind(oleracea,carduorum)
entire
```

```
## # A tibble: 39 × 4
##       x1    x2    x3    x4
##    <dbl> <dbl> <dbl> <dbl>
##  1   189   245   137   163
##  2   192   260   132   217
##  3   217   276   141   192
##  4   221   299   142   213
##  5   171   239   128   158
##  6   192   262   147   173
##  7   213   278   136   201
##  8   192   255   128   185
##  9   170   244   128   192
## 10   201   276   146   186
## # ℹ 29 more rows
```

```
cov(entire, use="complete.obs")
```

```
##           x1        x2        x3        x4
## x1 196.88799  56.93725 -34.47976 -19.07152
## x2  56.93725 502.70850 239.42510 245.34008
## x3 -34.47976 239.42510 216.04453 159.45142
## x4 -19.07152 245.34008 159.45142 341.83131
```

The relevant sample covariance matrix S is $\begin{bmatrix} 196.89 & 56.94 & -34.48 & -19.07 \\ 56.94 & 502.71 & 239.42 & 245.34 \\ -34.48 & 239.43 & 216.04 & 159.45 \\ -19.07 & 245.34 & 159.45 & 341.83 \end{bmatrix}$

ii. List the eigenvalues and describe the percent contributions to the variance

```
eigen_entire<-eigen(cov(entire, use="complete.obs"))
eigen_entire
```

```
## eigen() decomposition
## $values
## [1] 818.27340 238.22942 144.96091  56.00862
##
## $vectors
##            [,1]       [,2]       [,3]        [,4]
## [1,] -0.0276432  0.8303372  0.4060059  0.38070358
## [2,] -0.7365338  0.3547644 -0.3394728 -0.46520787
## [3,] -0.4294145 -0.1990933 -0.3711570  0.79887895
## [4,] -0.5218784 -0.3808467  0.7629940 -0.02094881
```

```
eigenvalues_entire <- eigen_entire$values

total_entire <- sum(eigenvalues_entire)

percent_variance_entire <- eigenvalues_entire / total_entire * 100

eigenvalues_entire
```

```
## [1] 818.27340 238.22942 144.96091  56.00862
```

```
percent_variance_entire
```

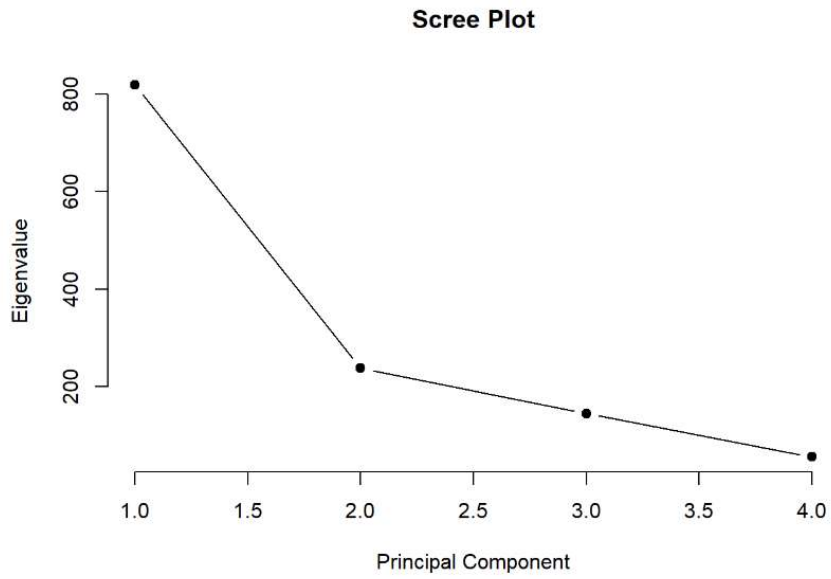
```
## [1] 65.072875 18.945102 11.527960  4.454063
```

The eigenvalue for (PC1) is 818.27 and the PC1 explains 65.07% of the total variance.
The eigenvalue for (PC2) is 238.23 and the PC2 explains 18.95% of the total variance.
The eigenvalue for (PC3) is 144.96 and the PC3 explains 11.53% of the total variance.
The eigenvalue for (PC4) is 56.01 and the PC4 explains 4.45% of the total variance.

The principal component analysis of the combined dataset yields four eigenvalues, with the first principal component (PC1) having an eigenvalue of 818.27, which accounts for 65.07% of the total variance. This indicates that PC1 captures the majority of the information in the dataset. The second principal component (PC2) has an eigenvalue of 238.23, contributing 18.95% to the total variance, indicating it encapsulates a significant but smaller portion of the dataset's information compared to PC1. The third principal component (PC3) has an eigenvalue of 144.96 and explains 11.53% of the variance, and the fourth principal component (PC4) has an eigenvalue of 56.01, explaining 4.45% of the variance. Collectively, PC1 and PC2 account for a substantial majority of the information, explaining approximately 84.02% of the total variance in the dataset, suggesting that these components are the most informative for understanding the dataset's structure.

iii. scree plot

```
plot(eigenvalues_entire, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
```



The Scree plot indicates that the first principal component has a significantly higher eigenvalue compared to the rest, and there's a noticeable drop after the second principal component. According to the Scree plot, typically retain components before this drop-off, which suggests retaining the first two components. The steep slope from the first to the second component and the leveling off after the second component support this decision.

Based on the Scree plot and considering the cumulative explained variance which shows that the first two components account for approximately 84% of the variance, retaining the first two principal components would be justified. This decision aligns with the Kaiser Criterion, which suggest keeping components with eigenvalues greater than 1, also supporting the retention of at least the first two components.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors <- eigen_entire$vectors

eigenvectors
```

```
##             [,1]       [,2]       [,3]        [,4]
## [1,] -0.0276432  0.8303372  0.4060059  0.38070358
## [2,] -0.7365338  0.3547644 -0.3394728 -0.46520787
## [3,] -0.4294145 -0.1990933 -0.3711570  0.79887895
## [4,] -0.5218784 -0.3808467  0.7629940 -0.02094881
```

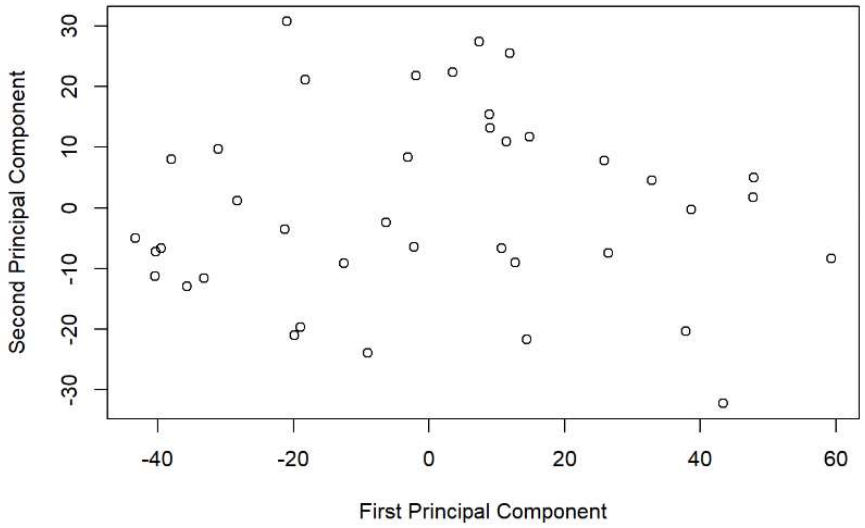
The eigenvectors for the principal components is $\begin{bmatrix} -0.028 & 0.830 & 0.406 & 0.381 \\ -0.737 & 0.355 & -0.339 & -0.465 \\ -0.429 & -0.199 & -0.371 & 0.799 \\ -0.522 & -0.381 & 0.763 & -0.021 \end{bmatrix}$

(v) The PC1 has strong positive relationship on the second variable (0.830). This indicates a dependency but for the first variable (-0.028), it has no dependency because it is close to 0.
The PC2 has strong negative relationship on the first variable (-0.737). This indicates a dependency.
The PC3 has strong positive relationship on the fourth variable (0.799), indicating a dependency.
The PC4 has strong positive relationship on the third variable (0.763), indicating a dependency but for the fourth variable (-0.021), it has no dependency because it is close to 0.

vi. Using at least the first two principal components, display scatter plots of pairs of principal components

```
pca_result_entire <- prcomp(entire, scale. = FALSE)

plot(pca_result_entire$x[, 1], pca_result_entire$x[, 2],
     xlab="First Principal Component", ylab="Second Principal Component")
```



The PC2 looks clustered while there is a wider spread along the PC1. This suggests that the second principal component explains a significant portion of the variability in the data, and the first principal component captures additional variability.

(e) Across all three datasets, the first principal component (PC1) explains the most variance, with the second principal component (PC2) also accounting for a significant amount. The first two principal components consistently account for the majority of the variance in all groups.

Each principal component exhibits relationships with different variables, providing insights into which variables are important within each group. For instance, for PC1, 'Oleracea' shows a strong relationship with the third variable, 'Carduorum' with the fourth, and the entire dataset with the second variable.

Scatter plot distributions offer a view into the data structure within each group. 'Oleracea' and 'Carduorum' data points tend to cluster around PC1, whereas the entire dataset shows a more pronounced clustering around PC2.

Question3

a. Store the data in matrix X

```
X <- as.matrix(data)

head(X)
```

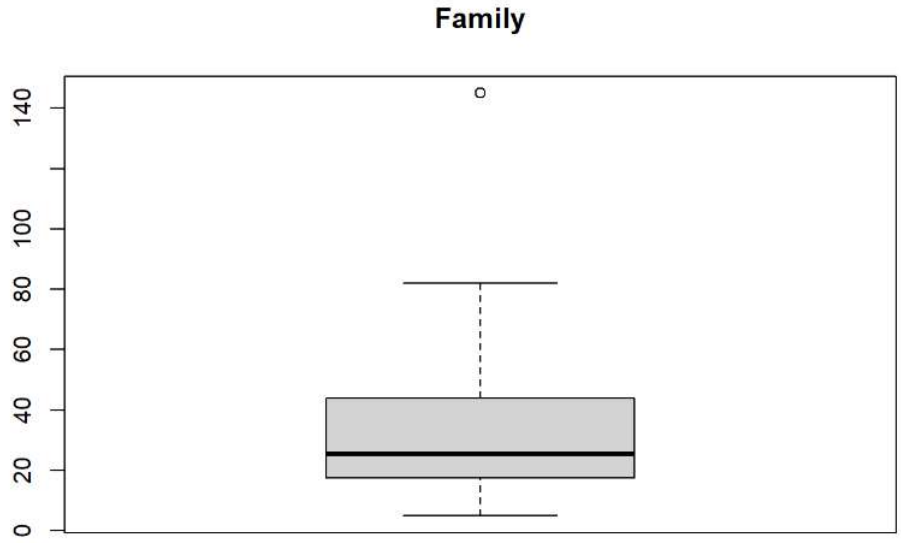
```
##      Family DistRD Cotton Maize Sorg Millet Bull Cattle Goats
## [1,]     12     80    1.5   1.0    3   0.25    2      0     1
## [2,]     54      8    6.0   4.0    0   1.00    6     32     5
## [3,]     11     13    0.5   1.0    0   0.00    0      0     0
## [4,]     21     13    2.0   2.5    1   0.00    1      0     5
## [5,]     61     30    3.0   5.0    0   0.00    4     21     0
## [6,]     20     70    0.0   2.0    3   0.00    2      0     3
```

b. i. Provide at least two indicators for each of these data that justify this claim

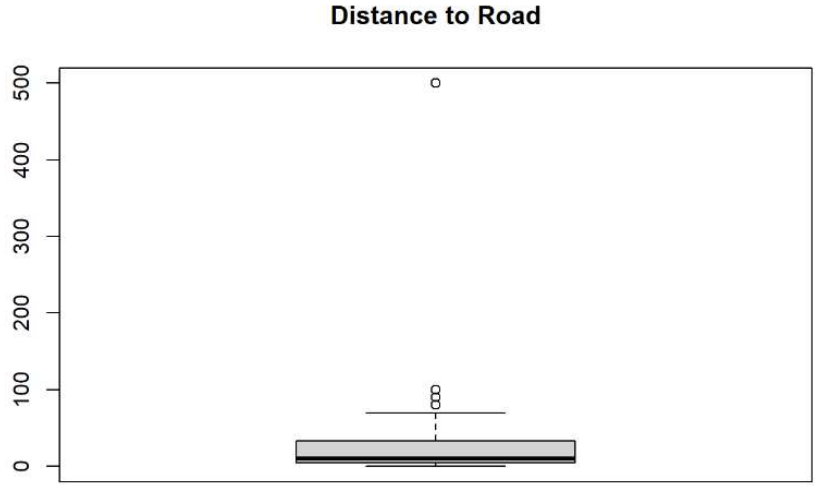
```
summary(X)
```

```
##      Family           DistRD          Cotton          Maize      
##  Min.   :  5.00   Min.   :  0.00   Min.   : 0.000   Min.   :0.000  
##  1st Qu.: 17.75   1st Qu.:  5.00   1st Qu.: 1.000   1st Qu.:1.000  
##  Median : 25.50   Median : 10.00   Median : 2.000   Median :1.500  
##  Mean   : 32.37   Mean   : 32.86   Mean   : 2.991   Mean   :2.082  
##  3rd Qu.: 43.50   3rd Qu.: 33.00   3rd Qu.: 4.000   3rd Qu.:3.000  
##  Max.   :145.00   Max.   :500.00   Max.   :12.000   Max.   :8.000  
##       Sorg            Millet           Bull            Cattle      
##  Min.   : 0.000   Min.   : 0.000   Min.   : 0.000   Min.   :  0.000  
##  1st Qu.: 1.000   1st Qu.: 0.000   1st Qu.: 0.000   1st Qu.:  0.000  
##  Median : 2.250   Median : 1.000   Median : 2.000   Median :  0.000  
##  Mean   : 2.651   Mean   : 1.671   Mean   : 2.632   Mean   :  4.658  
##  3rd Qu.: 3.000   3rd Qu.: 2.000   3rd Qu.: 4.000   3rd Qu.:  5.000  
##  Max.   :13.000   Max.   :12.000   Max.   :11.000   Max.   :104.000  
##      Goats      
##  Min.   : 0.000  
##  1st Qu.: 0.000  
##  Median : 2.000  
##  Mean   : 2.974  
##  3rd Qu.: 5.000  
##  Max.   :20.000
```

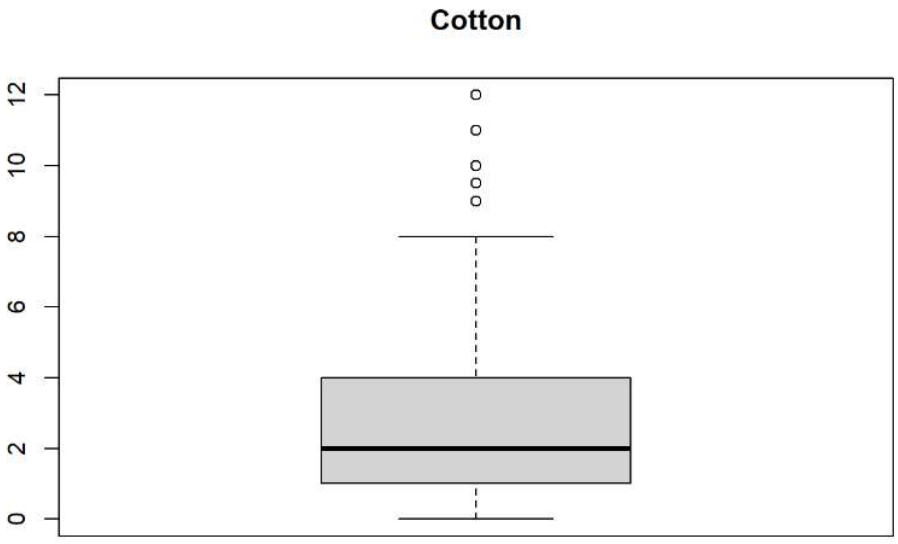
```
boxplot(data$Family, main="Family")
```



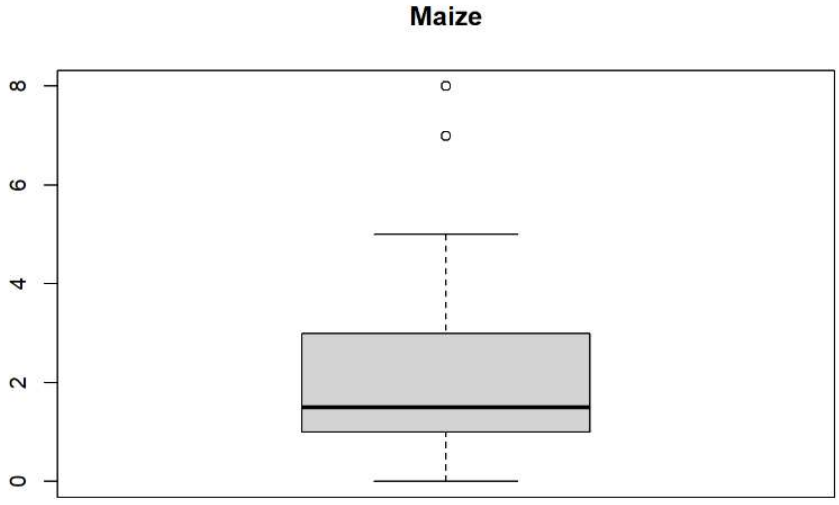

```
boxplot(data$DistRD, main="Distance to Road")
```



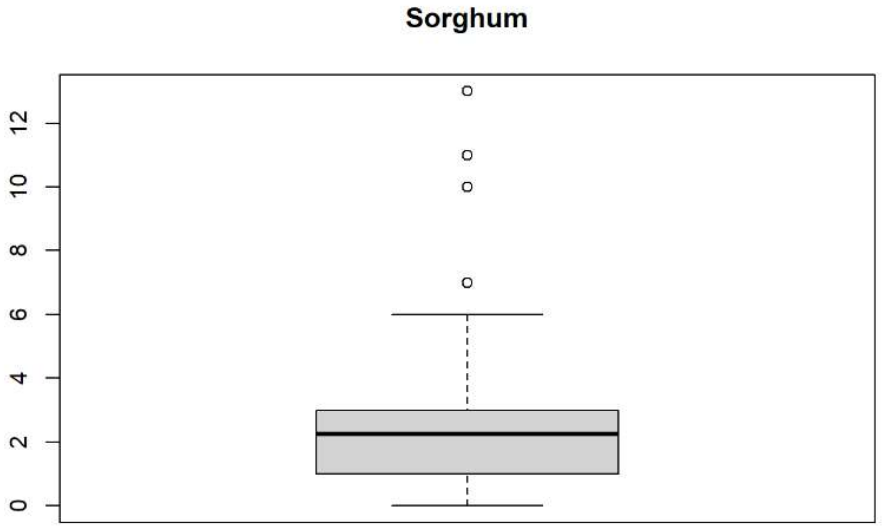
```
boxplot(data$Cotton, main="Cotton")
```



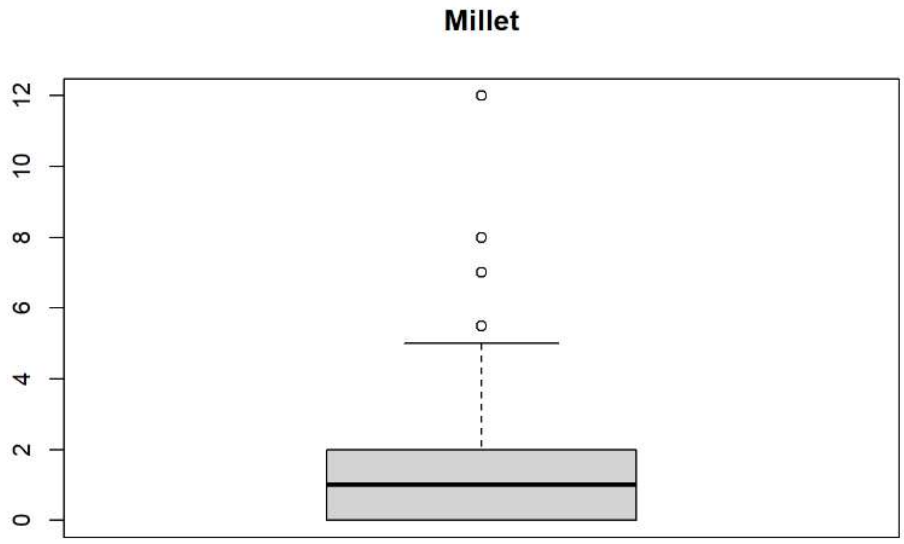

```
boxplot(data$Maize, main="Maize")
```



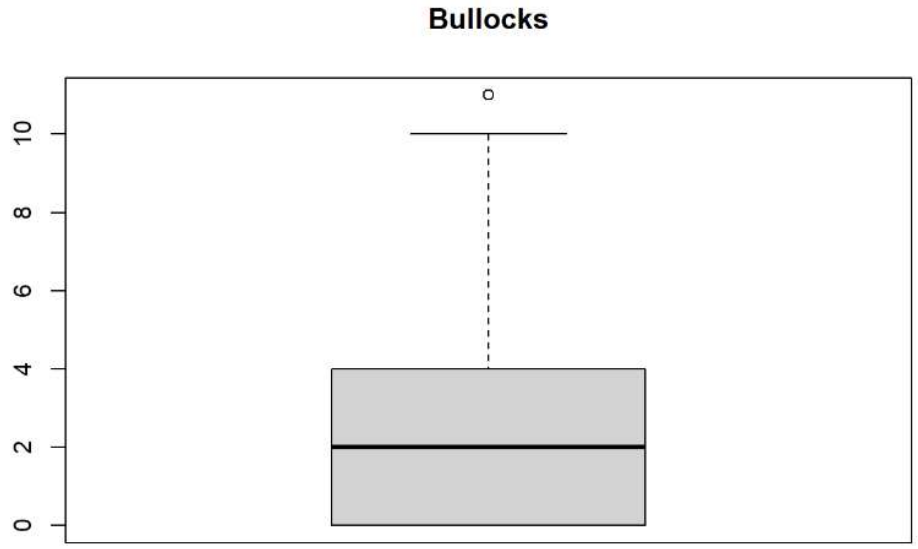
```
boxplot(data$Sorg, main="Sorghum")
```



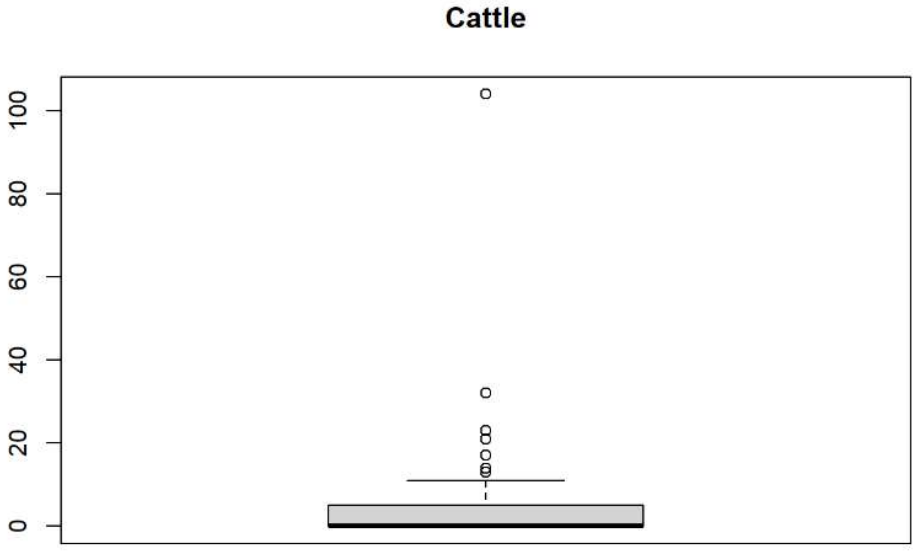

```
boxplot(data$Millet, main="Millet")
```



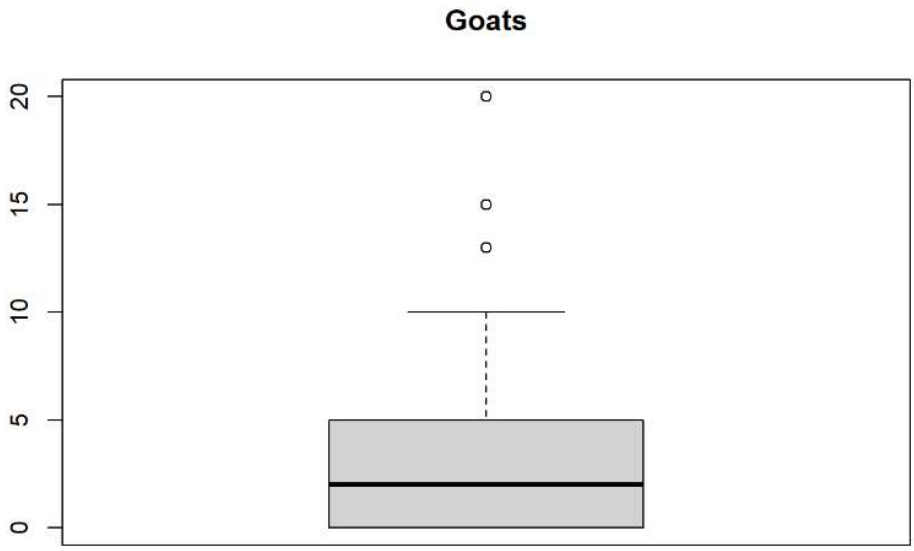
```
boxplot(data$Bull, main="Bullocks")
```




```
boxplot(data$Cattle, main="Cattle")
```



```
boxplot(data$Goats, main = "Goats")
```



For row 25, the Family variable has a value of 145. Considering the summary output, where the mean value of the Family variable is 32, the value of 145 is excessively large, indicating it as an outlier. Furthermore, the box plot for the Family variable shows that the value 145 is significantly deviated from the rest of the data, reinforcing its classification as an outlier.

For row 34, in the summary output, mean value of Sorghum is 2.65 but Sorghum variable has a value of 11. The value of 11 is excessively large, indicating it as an outlier. Also, the box plot of Sorghum shows that the value 11 is significantly deviated from the rest of the data, reinforcing its classification as an outlier.

For row 52, in the summary output, mean value of Goats is 2.97 but Goats variable has a value of 15. The value of 15 is excessively large, indicating it as an outlier. Also, the box plot of Goats shows that the value 15 is significantly deviated from the rest of the data, reinforcing its classification as an outlier.

For row 57, the Maize variable has a value of 7. Considering the summary output, where the mean value of the Maize variable is 2.08, the value of 7 is large, indicating it as an outlier. Also, the box plot for the Maize variable shows that the value 7 is significantly deviated from the rest of the data, reinforcing its classification as an outlier.

For row 62, the Cotton variable has a value of 9. Considering the summary output, where the mean value of the Cotton variable is 2.99, the value of 9 is large, indicating it as an outlier. Furthermore, the box plot for the Cotton variable shows that the value of 9is significantly deviated from the rest of the data, reinforcing its classification as an outlier.

For row 69 and 72, the distance variable has a value of 500. Considering the summary output, where the mean value of the distance variable is 32.86, the value of 500 is excessively large,

indicating it as an outlier. Furthermore, the box plot for the distance variable shows that the value 500 is significantly deviated from the rest of the data, reinforcing its classification as an outlier.

ii) In the initial investigation of the data, excluding outliers, we analyzed basic statistics, distribution, central tendency, and variability. We examined descriptive statistics such as mean, median, standard deviation, and range, and used box plots and histograms to visually assess the distribution of the data. After identifying and removing outliers, the remaining data showed a more uniform distribution, with central tendency and variance more accurately reflecting the characteristics of the data compared to before outlier removal. This initial investigation provides an important foundation for understanding the data structure in the subsequent principal component analysis

c. Create a data matrix by removing the outliers

```
outlier_rows <- c(25, 34, 52, 57, 62, 69, 72)

cleaned_data <- data[-outlier_rows, ]

X_tilde <- as.matrix(cleaned_data)

head(X_tilde)
```

```
##      Family DistRD Cotton Maize Sorg Millet Bull Cattle Goats
## [1,]     12     80    1.5   1.0    3   0.25    2      0     1
## [2,]     54      8    6.0   4.0    0   1.00    6     32     5
## [3,]     11     13    0.5   1.0    0   0.00    0      0     0
## [4,]     21     13    2.0   2.5    1   0.00    1      0     5
## [5,]     61     30    3.0   5.0    0   0.00    4     21     0
## [6,]     20     70    0.0   2.0    3   0.00    2      0     3
```

d. i. Give the relevant sample covariance matrix S

```
cov(X, use="complete.obs")
```

```
##            Family      DistRD    Cotton      Maize        Sorg     Millet
## Family  550.87579 -158.768421 48.116526 29.5392982  31.8368421 26.3928070
## DistRD -158.76842 6533.750658  6.436136 -8.1051535 -13.6921491  3.9406140
## Cotton   48.11653    6.436136  8.012226  3.8317127   2.5849781  2.4464825
## Maize    29.53930   -8.105154  3.831713  3.4339803   0.4807237  0.8940789
## Sorg     31.83684  -13.692149  2.584978  0.4807237   5.7001316  2.0287719
## Millet   26.39281    3.940614  2.446482  0.8940789   2.0287719  4.9420175
## Bull     45.45754  -19.024912  5.762807  3.0740351   2.8164912  2.0905263
## Cattle  103.75439  -67.354561  6.504368  4.8085088  12.6991228  2.3659649
## Goats    46.80982   10.362982  4.653772  1.0421930   4.1707018  2.8012281
##               Bull     Cattle     Goats
## Family  45.457544 103.754386 46.809825
## DistRD -19.024912 -67.354561 10.362982
## Cotton   5.762807   6.504368  4.653772
## Maize    3.074035   4.808509  1.042193
## Sorg     2.816491  12.699123  4.170702
## Millet   2.090526   2.365965  2.801228
## Bull     7.089123  18.205614  6.150175
## Cattle  18.205614 173.081404 19.364211
## Goats    6.150175  19.364211 17.012632
```

The sample covariance matrix S is $\begin{bmatrix} 551 & -159 & 48 & 30 & 32 & 26 & 45 & 104 & 47 \\ -159 & 6534 & 6 & -8 & -14 & 4 & -19 & -67 & 10 \\ 48 & 6 & 8 & 4 & 3 & 2 & 6 & 7 & 5 \\ 30 & -8 & 4 & 3 & 0 & 1 & 3 & 5 & 1 \\ 32 & -14 & 3 & 0 & 6 & 2 & 3 & 13 & 4 \\ 26 & 4 & 2 & 1 & 2 & 5 & 2 & 2 & 3 \\ 45 & -19 & 6 & 3 & 3 & 2 & 7 & 18 & 6 \\ 104 & -67 & 7 & 5 & 13 & 2 & 18 & 173 & 19 \\ 47 & 10 & 5 & 1 & 4 & 3 & 6 & 19 & 17 \end{bmatrix}$

ii. List the eigenvalues and describe the percent contributions to the variance

```
eigen_X<-eigen(cov(X, use="complete.obs"))

eigenvalues_X <- eigen_X$values

total <- sum(eigenvalues_X)

percent_variance_X <- eigenvalues_X / total * 100

eigenvalues_X
```

```
## [1] 6538.8594312  590.1075059  147.5506254   12.7110041    5.8904961
## [6]    3.9909905    2.9593523    1.1372073    0.6913478
```

```
percent_variance_X
```

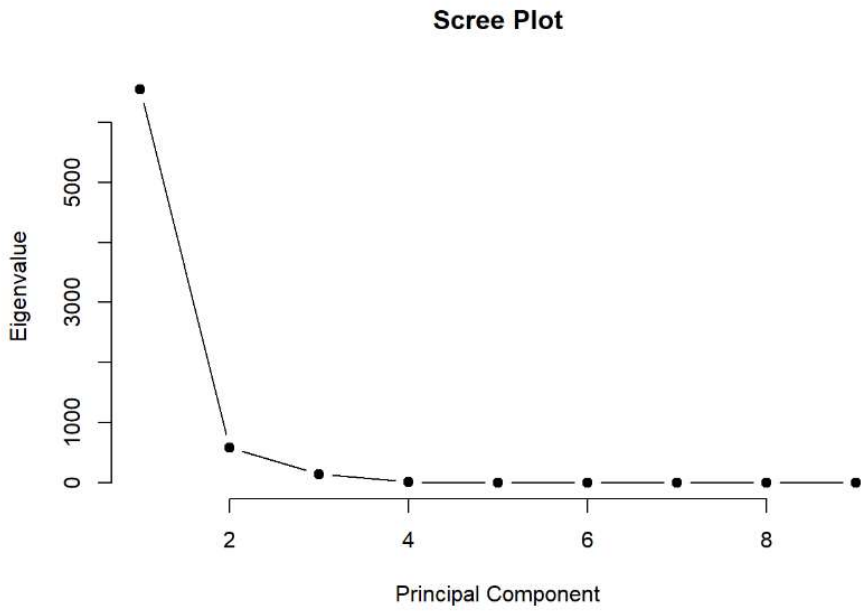
```
## [1] 89.525613126  8.079350356  2.020162744  0.174030417  0.080648663
## [6]  0.054641925  0.040517438  0.015569868  0.009465463
```

The eigenvalue for (PC1) is 6538.86and the PC1 explains 89.53% of the total variance.
The eigenvalue for (PC2) is 590.11 and the PC2 explains 8.08% of the total variance.
The eigenvalue for (PC3) is 147.55 and the PC3 explains 2.02% of the total variance.
The eigenvalue for (PC4) is 12.71 and the PC4 explains 0.17% of the total variance.
The eigenvalue for (PC5) is 5.89 and the PC5 explains 0.08% of the total variance.
The eigenvalue for (PC6) is 3.99 and the PC6 explains 0.05% of the total variance.
The eigenvalue for (PC7) is 2.96 and the PC7 explains 0.04% of the total variance.
The eigenvalue for (PC8) is 1.14 and the PC8 explains 0.02% of the total variance
The eigenvalue for (PC9) is 0.69 and the PC9 explains 0.01% of the total variance

In the principal component analysis, the first principal component (PC1) has an eigenvalue of approximately 6539, explaining about 89.53% of the total variance. The second principal component (PC2) has an eigenvalue of around 590, accounting for an additional 8.08% of the variance. The third principal component (PC3) has an eigenvalue of about 147.55, explaining 2.02% of the variance, while the fourth to the ninth principal components (PC4-PC9) with eigenvalues of 12.71, 5.89, 3.99, 2.96, 1.14, and 0.69 respectively, explain the remaining variance with contributions of 0.17%, 0.08%, 0.05%, 0.04%, 0.02%, and 0.01%. These results indicate that the majority of the variability in the data can be explained by the first two principal components

iii. scree plot

```
plot(eigenvalues_X, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
```



Percentage of Variance Explained: The first method looks at the percentage of variance explained by each component, with the first component explaining 89.53% and the second explaining 8.07% for a cumulative total of 97.6%.

Kaiser Criterion: The second method uses the Kaiser Criterion, which suggests keeping all components with eigenvalues over 1. According to the eigenvalues I provided earlier, the first three components meet this criterion.

Scree Plot: The third method indicates an elbow after the PC1, which traditionally suggests retaining all components before the elbow. The steep decline after the first principal component suggests that the additional components might not be necessary.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors_X <- eigen_X$vectors

eigenvectors_X
```

```
##                [,1]       [,2]         [,3]         [,4]         [,5]
##  [1,]  0.0267175675 0.95364128 -0.251836220 -0.091222765 -0.089963869
##  [2,] -0.9995725665 0.02843768  0.004039512 -0.004240002 -0.001521928
##  [3,] -0.0007739567 0.08419737 -0.038642522  0.074330304  0.653993736
##  [4,]  0.0013694250 0.05088924 -0.019494373 -0.129464849  0.410473138
##  [5,]  0.0022466401 0.05820578  0.030590321  0.113652268 -0.349564866
##  [6,] -0.0004899182 0.04554781 -0.029170888  0.085705664 -0.174214734
##  [7,]  0.0031275792 0.08354161  0.043341775  0.146459351  0.492587755
##  [8,]  0.0110210195 0.24437882  0.963395438 -0.084633673 -0.004057078
##  [9,] -0.0013599578 0.08898300  0.053786043  0.959426281 -0.022419998
##                [,6]         [,7]         [,8]          [,9]
##  [1,]  0.090637293 -0.011238201  0.010824404  0.0394883382
##  [2,]  0.001118813  0.001120246  0.002464594 -0.0008888926
##  [3,] -0.365602417  0.199433378 -0.468347108  0.4066894963
##  [4,]  0.079837037 -0.084730640 -0.273566841 -0.8505215954
##  [5,] -0.543410111  0.688950783 -0.005921516 -0.3009585424
##  [6,] -0.693298451 -0.686406339 -0.038052467 -0.0781894335
##  [7,] -0.186141079  0.046936946  0.827905512 -0.0689953202
##  [8,]  0.003954001 -0.038147696 -0.052649273  0.0245888752
##  [9,]  0.202788374 -0.058773516 -0.126516364 -0.0871891712
```

The eigenvectors for the principal components is

$$\begin{bmatrix} 0.0267 & 0.9534 & -0.2518 & -0.0912 & -0.0900 & 0.0906 & -0.0112 & 0.0108 & 0.0395 \\ -0.9996 & 0.0284 & 0.0040 & -0.0042 & -0.0015 & 0.0011 & 0.0011 & 0.0025 & -0.0009 \\ -0.0008 & 0.0842 & -0.0386 & 0.0743 & 0.6540 & -0.3656 & 0.1994 & -0.4683 & 0.4067 \\ 0.0014 & 0.0509 & -0.0194 & -0.1295 & 0.4105 & 0.0798 & -0.0847 & -0.2736 & -0.8505 \\ 0.0022 & 0.0582 & 0.0306 & 0.1137 & -0.3496 & -0.5434 & 0.6890 & -0.0059 & -0.3010 \\ -0.0005 & 0.0455 & -0.0292 & 0.0857 & -0.1742 & -0.6933 & -0.6864 & -0.0381 & -0.0782 \\ 0.0031 & 0.0835 & 0.0433 & 0.1465 & 0.4926 & -0.1861 & 0.0469 & 0.8279 & -0.0690 \\ 0.0110 & 0.2444 & 0.9634 & -0.0846 & -0.0041 & 0.0040 & -0.0381 & -0.0526 & 0.0246 \\ -0.0014 & 0.0890 & 0.0538 & 0.9594 & -0.0224 & 0.2028 & -0.0588 & -0.1265 & -0.0872 \end{bmatrix}$$

(V) The PC1 has strong positive relationship on the second variable (0.9534), indicating a dependency but first, fourth, fifth, sixth, seventh, eighth, ninth variables have no dependency because it is close to 0.

The PC2 has very strong negative relationship on the first variable (-0.9996). It has a dependency.

The PC3 has positive relationship on the fifth variable (0.6540), indicating a dependency.

The PC4 has strong negative relationship on the ninth variable (-0.85), indicating a dependency.

The PC5 has positive relationship on the seventh variable (0.6890), indicating a dependency.

The PC6 has negative relationship on the sixth and seventh variables (-0.6933) and (-0.6864), indicating a dependency.

The PC7 has strong positive relationship on the eighth variable (0.8279), indicating a dependency.

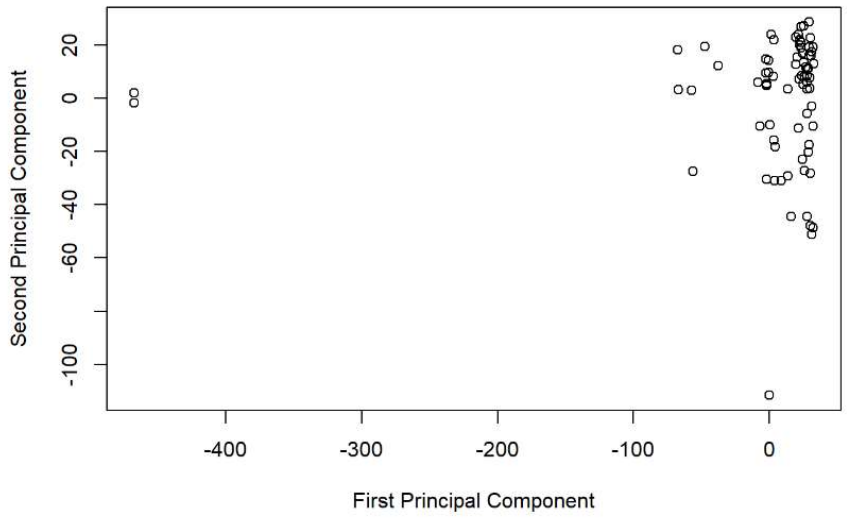
The PC8 has very strong positive relationship on the third variable (0.9634), indicating a dependency.

The PC9 has very strong positive relationship on the fourth variable (0.9594), indicating a dependency.

vi. Using at least the first two principal components, display scatter plots of pairs of principal components

```
pca_result_X <- prcomp(X, scale = FALSE)

plot(pca_result_X$x[, 1], pca_result_X$x[, 2],
     xlab="First Principal Component", ylab="Second Principal Component")
```



Observing the scatter plot, there is no apparent linear relationship between the first and second principal components. The PC1 seems to capture most of the variance in the data, with points spread out primarily along this axis. The PC2 accounts for a relatively smaller proportion of the variance.

The clustering of data points on the right side of the PC1 suggests that there are observations with high values for certain variables or a combination of them. However, there is no isolation or grouping of points, indicating no clear patterns or distinct clusters forming.

e. i. Give the relevant sample covariance matrix S

```
cov(X_tilde, use="complete.obs")
```

```
##            Family     DistRD    Cotton       Maize        Sorg     Millet
## Family 318.587383  16.958014 27.505382 18.45119352  8.02088662 19.3112212
## DistRD  16.958014 595.619672  3.786179  7.58525149 -8.18824595 -3.6438619
## Cotton  27.505382   3.786179  5.179237  2.62904678  1.03837383  1.6558664
## Maize   18.451194   7.585251  2.629047  2.47421142 -0.03769714  0.9948050
## Sorg     8.020887  -8.188246  1.038374 -0.03769714  2.55445439  0.8272858
## Millet  19.311221  -3.643862  1.655866  0.99480499  0.82728581  4.4668318
## Bull    25.023018   7.346867  3.542935  2.00527494  0.65664962  1.4462916
## Cattle  55.753410  17.976769  7.580680  4.57022592  0.13704177  1.3212916
## Goats   22.746377  18.545823  3.065335  0.89183717  0.40494459  0.9688299
##              Bull     Cattle      Goats
## Family 25.0230179 55.7534101 22.7463768
## DistRD  7.3468670 17.9767690 18.5458227
## Cotton  3.5429348  7.5806799  3.0653346
## Maize   2.0052749  4.5702259  0.8918372
## Sorg    0.6566496  0.1370418  0.4049446
## Millet  1.4462916  1.3212916  0.9688299
## Bull    4.4667519  7.7608696  4.1860614
## Cattle  7.7608696 35.4275362  8.7939045
## Goats   4.1860614  8.7939045 13.2088662
```

The sample covariance matrix is $\begin{bmatrix} 319 & 17 & 28 & 18 & 8 & 19 & 25 & 56 & 23 \\ 17 & 596 & 4 & 8 & -8 & -4 & 7 & 18 & 19 \\ 28 & 4 & 5 & 3 & 1 & 2 & 4 & 8 & 3 \\ 18 & 8 & 3 & 2 & 0 & 1 & 2 & 5 & 1 \\ 8 & -8 & 1 & 0 & 3 & 1 & 1 & 0 & 0 \\ 19 & -4 & 2 & 1 & 1 & 4 & 1 & 1 & 1 \\ 25 & 7 & 4 & 2 & 1 & 1 & 4 & 8 & 4 \\ 56 & 18 & 8 & 5 & 0 & 1 & 8 & 35 & 9 \\ 23 & 19 & 3 & 1 & 0 & 1 & 4 & 9 & 13 \end{bmatrix}$

ii. List the eigenvalues and describe the percent contributions to the variance

```
eigen_X_tilde<-eigen(cov(X_tilde, use="complete.obs"))

eigenvalues_X_tilde <- eigen_X_tilde$values

total <- sum(eigenvalues_X_tilde)

percent_variance_X_tilde <- eigenvalues_X_tilde / total * 100

eigenvalues_X_tilde
```

```
## [1] 598.655803 336.148517  26.526056  10.153577   3.671695   2.918251   2.230668
## [8]   1.135536   0.544840
```

```
percent_variance_X_tilde
```

```
## [1] 60.96384748 34.23153478  2.70126911  1.03398496  0.37390540  0.29717882
## [7]  0.22715910  0.11563681  0.05548354
```

The eigenvalue for (PC1) is 598.66 and the PC1 explains 60.96% of the total variance.

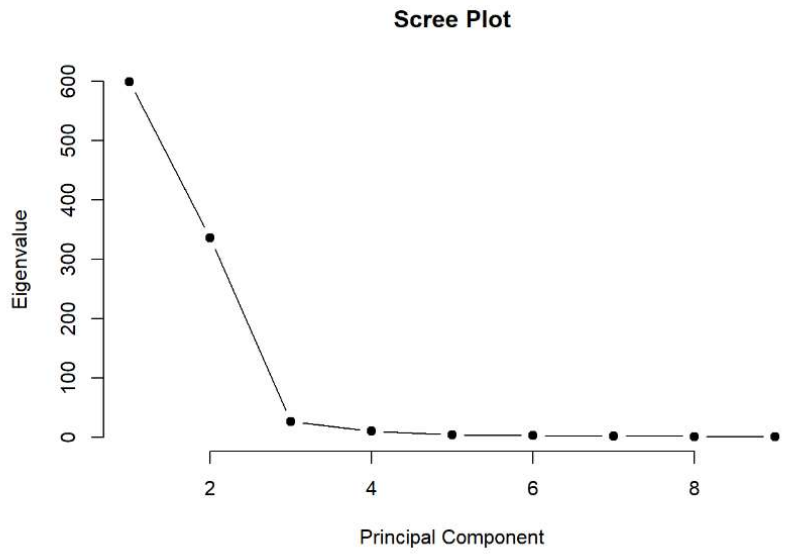
The eigenvalue for (PC2) is 336.15 and the PC2 explains 34.23% of the total variance.

The eigenvalue for (PC3) is 26.53 and the PC3 explains 2.70% of the total variance.
The eigenvalue for (PC4) is 10.15 and the PC4 explains 1.03% of the total variance.
The eigenvalue for (PC5) is 3.67 and the PC5 explains 0.37% of the total variance.
The eigenvalue for (PC6) is 2.92 and the PC6 explains 0.30% of the total variance.
The eigenvalue for (PC7) is 2.23 and the PC7 explains 0.23% of the total variance.
The eigenvalue for (PC8) is 1.14 and the PC8 explains 0.12% of the total variance
The eigenvalue for (PC9) is 0.54 and the PC9 explains 0.06% of the total variance

The PC1 has eigenvalue of approximately 598.66, which accounts for about 60.96% of the variance. The PC2 has a significantly lower eigenvalue of around 336.15, but still captures a substantial 34.23% of the variance. Together, the first two components account for over 95% of the total variance in the data.
The remaining components each capture a progressively smaller amount of the variance, with the third principal component accounting for approximately 2.7% and the others each below 1%. These percentages suggest that the first two principal components are the most significant in terms of explaining the variability in the dataset, with diminishing contributions from the subsequent components.

iii. scree plot

```
plot(eigenvalues_X_tilde, type = "b", pch = 19, frame = FALSE,
     xlab = "Principal Component", ylab = "Eigenvalue",
     main = "Scree Plot")
```



Eigenvalue-one Criterion: The eigenvalue-one criterion, also known as the Kaiser criterion, suggests retaining principal components with eigenvalues greater than 1. In the screenshot, only the first two components have eigenvalues greater than 1, suggesting we retain these two.
Proportion of Variance Explained: Anther method is to look at the proportion of total variance

explained by each component. Often, a cutoff is set such that we retain components that contribute significantly to the variance. The first component explains approximately 60.96% of the variance, and the second explains about 34.23%. Together, they explain over 95% of the total variance, which is typically considered sufficient for most analyses.

Scree Plot: The scree plot provides a visual method where we look for an 'elbow' in the graph, which indicates a point beyond which the remaining components contribute little to the total variance. The scree plot shows a clear elbow after the second component, further suggesting that only the first two components are worth retaining.

In conclusion, all three methods point towards retaining only the first two principal components, as they explain the majority of the variance and their eigenvalues are above the common threshold.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors_X_tilde <- eigen_X_tilde$vectors

eigenvectors_X_tilde
```

```
##               [,1]        [,2]        [,3]        [,4]         [,5]
##  [1,]  0.074032153  0.96661143  0.20973121  0.02983975  0.119738974
##  [2,]  0.995431316 -0.08420571  0.03537734  0.01972823 -0.003026885
##  [3,]  0.010609767  0.08582783 -0.10433185 -0.03888517 -0.654785292
##  [4,]  0.015413007  0.05552771 -0.03120871  0.09662596 -0.293956012
##  [5,] -0.012613666  0.02602851  0.04352808 -0.05785459 -0.249915117
##  [6,] -0.003474341  0.05913095  0.09475495 -0.04431054 -0.473134285
##  [7,]  0.016290722  0.07775609 -0.13537540 -0.14756025 -0.403660241
##  [8,]  0.040130538  0.18151903 -0.91975451  0.28824751  0.074892264
##  [9,]  0.035193537  0.07037673 -0.26029616 -0.93685672  0.125041762
##               [,6]         [,7]          [,8]         [,9]
##  [1,] -0.020563488 -0.005844163  0.0007503977 -0.022561106
##  [2,]  0.008173464 -0.016678571 -0.0004335984 -0.005550286
##  [3,] -0.378984931  0.030773281 -0.4401314001 -0.461614369
##  [4,] -0.237797283  0.354579639 -0.1891499938  0.825776258
##  [5,]  0.004616231 -0.914584131 -0.0191599087  0.307589749
##  [6,]  0.859075995  0.147149809 -0.0437765047  0.010605161
##  [7,] -0.190569211  0.085726485  0.8637272414 -0.030920612
##  [8,]  0.155616366 -0.071997951 -0.0350668023  0.012911357
##  [9,]  0.027699453  0.048621200 -0.1448032000  0.092839789
```

The eigenvectors for the principal components is

$$\begin{bmatrix} 0.074 & 0.967 & 0.210 & 0.030 & 0.120 & -0.021 & -0.006 & 0.001 & -0.023 \\ 0.995 & -0.084 & 0.035 & 0.020 & -0.003 & 0.008 & -0.017 & 0 & -0.006 \\ 0.011 & 0.086 & -0.104 & -0.039 & -0.655 & -0.379 & 0.030 & -0.440 & -0.462 \\ 0.015 & 0.056 & -0.031 & 0.097 & -0.294 & -0.238 & 0.355 & -0.190 & 0.826 \\ -0.013 & 0.026 & 0.044 & -0.058 & -0.250 & 0.005 & -0.915 & -0.019 & 0.308 \\ -0.003 & 0.059 & 0.095 & -0.044 & -0.473 & 0.859 & 0.147 & 0.044 & 0.011 \\ 0.016 & 0.078 & -0.135 & -0.148 & -0.404 & -0.191 & 0.086 & 0.864 & -0.031 \\ 0.040 & 0.182 & -0.920 & 0.288 & 0.075 & 0.156 & -0.072 & -0.035 & 0.013 \\ 0.035 & 0.070 & -0.260 & -0.937 & 0.125 & 0.028 & 0.049 & -0.145 & 0.093 \end{bmatrix}$$

(V) The PC1 has very strong positive relationship on the second variable (0.967), indicating a dependency but first, fourth, fifth, sixth, seventh, eighth, ninth variables have no dependency because it is close to 0.

The PC2 has very strong positive relationship on the first variable (0.995). It has a dependency.

The PC3 has negative relationship on the fifth variable (-0.655), indicating a dependency.

The PC4 has strong positive relationship on the ninth variable (0. 826), indicating a dependency.

The PC5 has negative relationship on the seventh variable (-0.915), indicating a dependency.

The PC6 has positive relationship on the sixth variable (0.859), indicating a dependency.

The PC7 has strong positive relationship on the eighth variable (0.864), indicating a dependency.

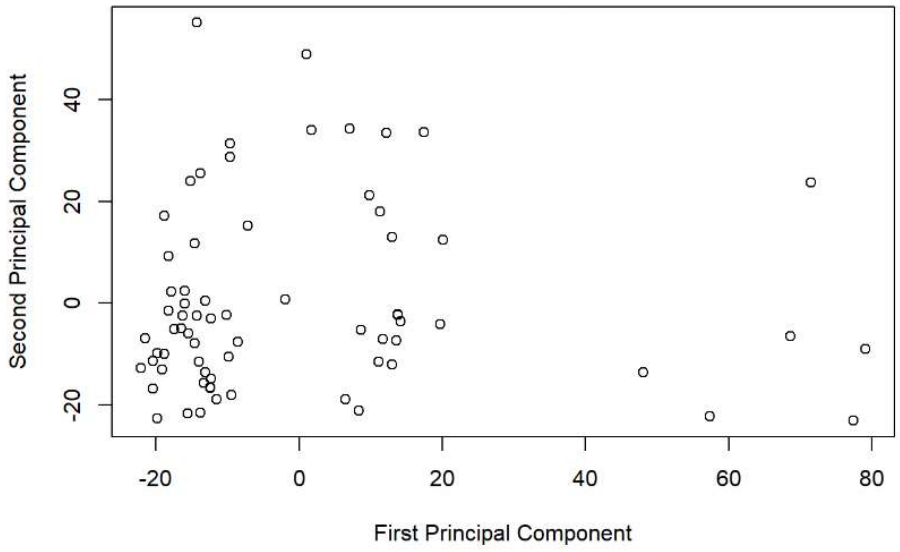
The PC8 has very strong negative relationship on the third variable (-0.92), indicating a dependency.

The PC9 has very strong negative relationship on the fourth variable (-0.937), indicating a dependency.

vi. Using at least the first two principal components, display scatter plots of pairs of principal components

```
pca_result_X_tilde <- prcomp(X_tilde, scale = FALSE)

plot(pca_result_X_tilde$x[, 1], pca_result_X_tilde$x[, 2],
     xlab="First Principal Component", ylab="Second Principal Component")
```



The scatter plot of the first two principal components illustrates that data points are spread across the plot without any discernible linear relationship between the two components. The majority of the variance is captured by the first principal component, evidenced by the broader spread of data points along this axis. The second principal component explains a smaller portion of the variance, as indicated by the tighter distribution along its axis.

There is no clear clustering or distinct groupings of data points. Instead, they are dispersed throughout the plot, which could imply that the underlying variables contributing to these principal components do not form distinct subgroups within the dataset.
Furthermore, the data does not show any outliers that significantly deviate from the rest of the points on this two-dimensional plane. This suggests that the PCA transformation has effectively normalized the data, concentrating on the intrinsic patterns without extreme variations.

(f) Comparing the results of the principal component analysis (PCA) with and without outliers, we see significant differences in the explained variance. In the dataset with outliers included, the PC1 explains about 89.53% of the variance, while the PC2 accounts for 8.08%. After the removal of outliers, PC1 explains 60.96% and PC2 accounts for 34.23% of the variance, showing a more balanced distribution of explained variance.
The presence of outliers had a substantial impact on the PCA. Including outliers in the dataset tends to overemphasize the variance explained by PC1, suggesting that a few extreme values are distorting the overall variability of the dataset. Thus, outliers have a significant effect on the PCA results.
Personally, I prefer the results of the analysis after outlier removal. The cleaned dataset likely reflects the true relationships among variables more accurately, with a more even distribution of variance explanation. Using a dataset with outliers can lead to skewed analysis outcomes, potentially leading to incorrect interpretations. Therefore, for more accurate and reliable insights, the analysis without outliers is preferred.