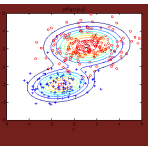


STAT 445/645 Assignment Cover Page

Student Name	JooHyeok Seo
SFU Student Number	301258279
SFU email address	jsa184@sfu.ca
Assignment Number	Final Project
Due Date	2024. 04. 18

Provide references for any data sets used in this assignment

adapted from R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, Pearson, New Jersey, 2019 by Don Estep in 2023

List software used in this assignment.

I solved the problems and visualize them using R

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

Independency
- Refer to Assignment 1

Scale and Outliers
- Refer to Assignment 3

PCA Analysis
- Refer to Assignment 4

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

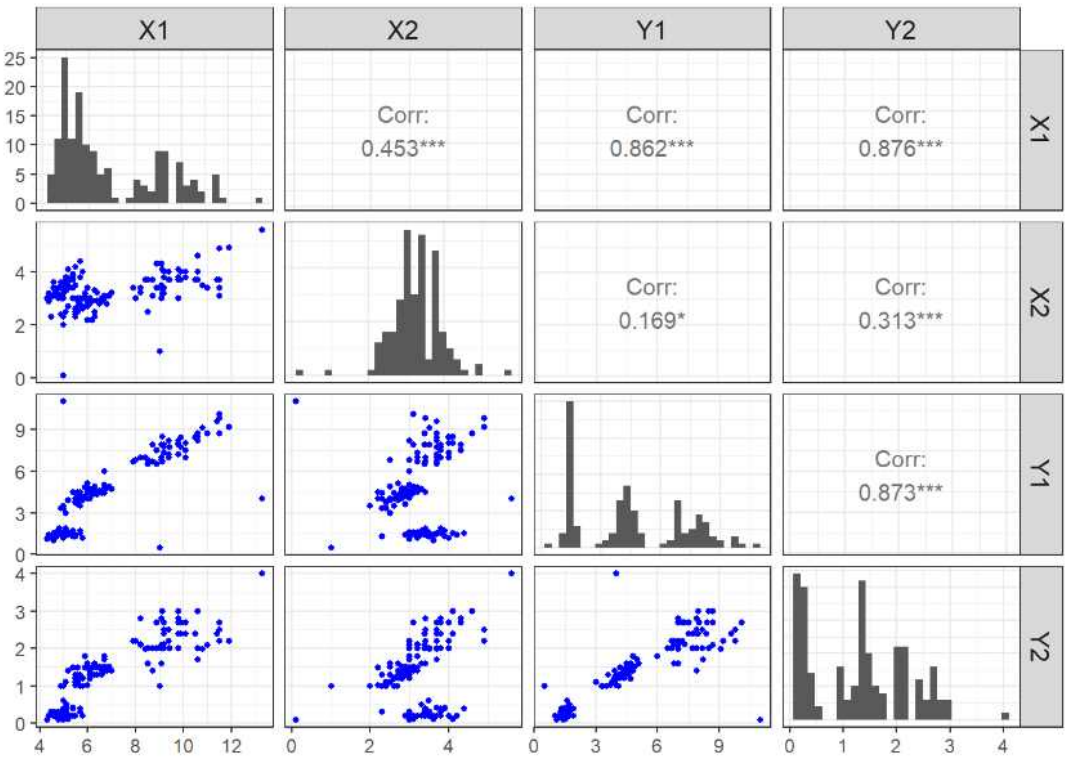
a. Initial investigation

```
summary(data)
```

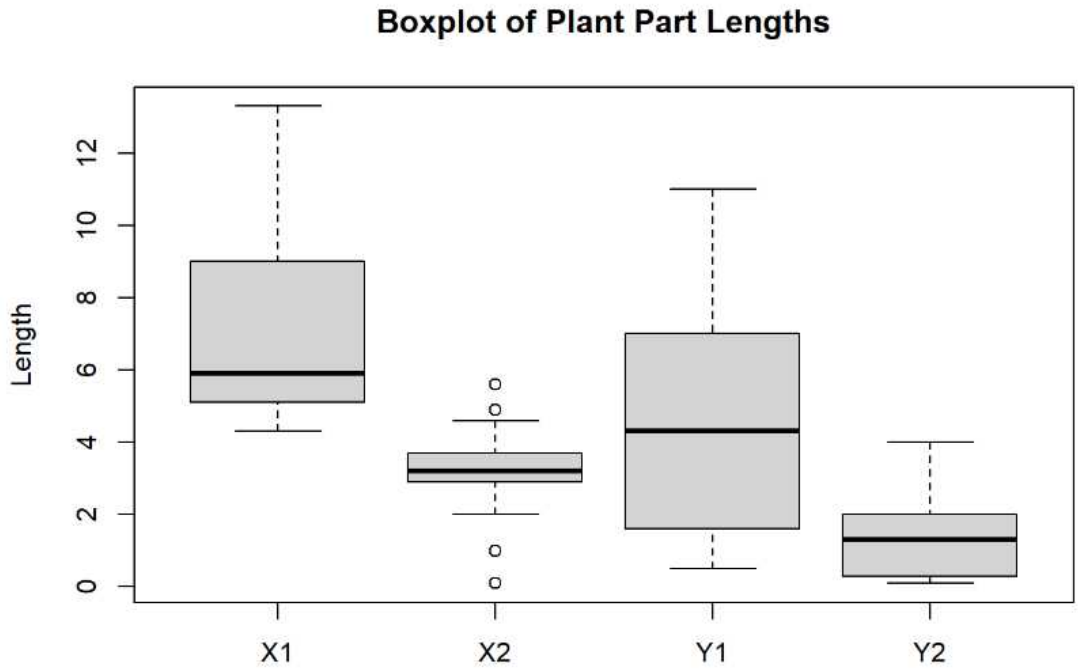
```
##        X1               X2               Y1               Y2
##  Min.   : 4.303   Min.   :0.101   Min.   : 0.503   Min.   :0.096
##  1st Qu.: 5.104   1st Qu.:2.904   1st Qu.: 1.595   1st Qu.:0.300
##  Median : 5.903   Median :3.204   Median : 4.298   Median :1.303
##  Mean   : 6.881   Mean   :3.262   Mean   : 4.514   Mean   :1.297
##  3rd Qu.: 9.004   3rd Qu.:3.696   3rd Qu.: 7.000   3rd Qu.:2.001
##  Max.   :13.297   Max.   :5.596   Max.   :10.999   Max.   :4.000
```

```
ggpairs(data,lower=list(continuous = wrap("points", color = "blue", size = 1)),
        diag = list(continuous = "barDiag"))+
  theme_bw()+
  theme(strip.text=element_text(size=12))
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



```
boxplot(data,
        main = "Boxplot of Plant Part Lengths",
        names = c("X1", "X2", "Y1", "Y2"),
        ylab = "Length")
```



For variable X1, the mean value is 6.881, with a minimum and maximum of about 4.303 and 13.297, respectively, indicating a considerable range and variability in the measurements. On the other hand, X2 shows a mean value of 3.262, with a minimum of 0.101 and a maximum of 5.596, suggesting lower variability in comparison to X1.

The variable Y1 has the mean of 4.514, spanning from a minimum of 0.503 to a maximum of 10.999, which points to a wide measurement range as well. Lastly, the mean for Y2 is roughly 1.297, with the smallest range noted, from 0.096 to 4.000, making it the most narrowly distributed variable among the others.

The medians and quartiles further reflect these tendencies. The median values for each variables are about 5.903 for X1, 3.204 for X2, 4.298 for Y1 and 1.303 for Y2. The interquartile range for X1 and Y1 are also broader compared to those of X2 and Y2, accentuating differences between the two groups.

There is potential for scale and outlier issue because the range of variables X1 and Y1 are much larger than variables X2 and Y2. Also, there are visible outliers in X2, indicating outlier issues. This means that the data point is significantly different from other observations.

The correlation between X1 and Y1, X1 and Y2, and Y1 and Y2 is 0.862, 0.876 and 0.873. It indicates a strong positive linear relationship which is seem in third row first column plot, fourth row first column plot and fourth row third column plot. Since these values are close to 1, there is a strong dependency.

From the box plots, we can infer potential deviations from normal distribution for X1, X2 and Y2, as indicated by the position of the median and the presence of outliers. Y1 is the variable that appears most symmetrically distributed, with the median line close to the center of the box, suggesting a closer alignment with a normal distribution. However, X1 and X2 show their medians closer to the bottom of the box, and Y2's median is positioned slightly above the midpoint. These observations hint at skewness in the distributions, with X1 and X2 potentially skewed towards higher values, and Y2 skewed towards lower values.

b) For the X Group

i. Display the relevant sample covariance matrix S

```
xvalue<- data[, 1:2]
xvalue
```

```
## # A tibble: 153 × 2
##       X1    X2
##    <dbl> <dbl>
##  1  5.70 3.00
##  2  9.1  3.00
##  3  6.80 2.80
##  4  6.30 2.30
##  5  5.40 3.70
##  6  6.30 3.30
##  7  5.80 2.70
##  8  5.00 0.101
##  9 10.5  3.7
## 10  5.40 3.00
## # ℹ 143 more rows
```

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

```
##           X1        X2
## X1 4.6723916 0.6381401
## X2 0.6381401 0.4254285
```

The sample covariance matrix is $\begin{bmatrix} 4.67 & 0.64 \\ 0.64 & 0.43 \end{bmatrix}$

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

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

eigenvalues <- eigen_result$values

total <- sum(eigenvalues)

percent_variance <- eigenvalues / total * 100

eigenvalues
```

```
## [1] 4.7662050 0.3316151
```

```
percent_variance
```

```
## [1] 93.494962  6.505038
```

The eigenvalue for (PC1) is 4.766 and the PC1 explains 93.49% of the total variance.
The eigenvalue for (PC2) is 0.331 and the PC2 explains 6.51% of the total variance

The calculated eigenvalues reflect the decomposition of variance within the data, with each principal component expressing the amount of the total variance it accounts for. These percentages of variance highlight which components best describe the underlying structure and patterns within the dataset, with PC1 being particularly prominent in this case.

iii) Eigenvalue Criterion: A common rule of thumb is to keep principal components with eigenvalues greater than 1. In this case, the first eigenvalue is about 4.77, which exceeds this threshold, suggesting we should retain the first principal component. The second eigenvalue is about 0.33, which is less than 1, so typically we would not retain this component based on this criterion alone.
Percentage of Variance Explained: The PC1 explains about 93.5% of the variance, and the PC2 explains about 6.5%. Since the PC1 accounts for the majority of the information, retaining only the PC1 may be justified.
Combining these approaches, we can conclude that the PC1 significantly captures the structure within the data, and additional components contribute relatively little. Therefore, retaining only one principal component seems to be a reasonable approach for this dataset.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors <- eigen_result$vectors

eigenvectors
```

```
##             [,1]       [,2]
## [1,] -0.9893660  0.1454473
## [2,] -0.1454473 -0.9893660
```

The eigenvectors for the principal components is $\begin{bmatrix} -0.9894 & 0.1454 \\ -0.1454 & -0.9894 \end{bmatrix}$

For the PC1, the eigenvector is [-0.9894, 0.1454]. This indicates that PC1 is primarily associated with the first variable with a strong negative weight and a much smaller positive weight on the second variable, suggesting that as the first variable increases, the value on PC1 decreases, and as the second variable increases, the value on PC1 slightly increases.

For the PC2, the eigenvector is [-0.1454, -0.9894]. This demonstrates that PC2 has a small negative weight on the first variable and a strong negative weight on the second variable, indicating that PC2 captures the variation in both variables in a negative direction.

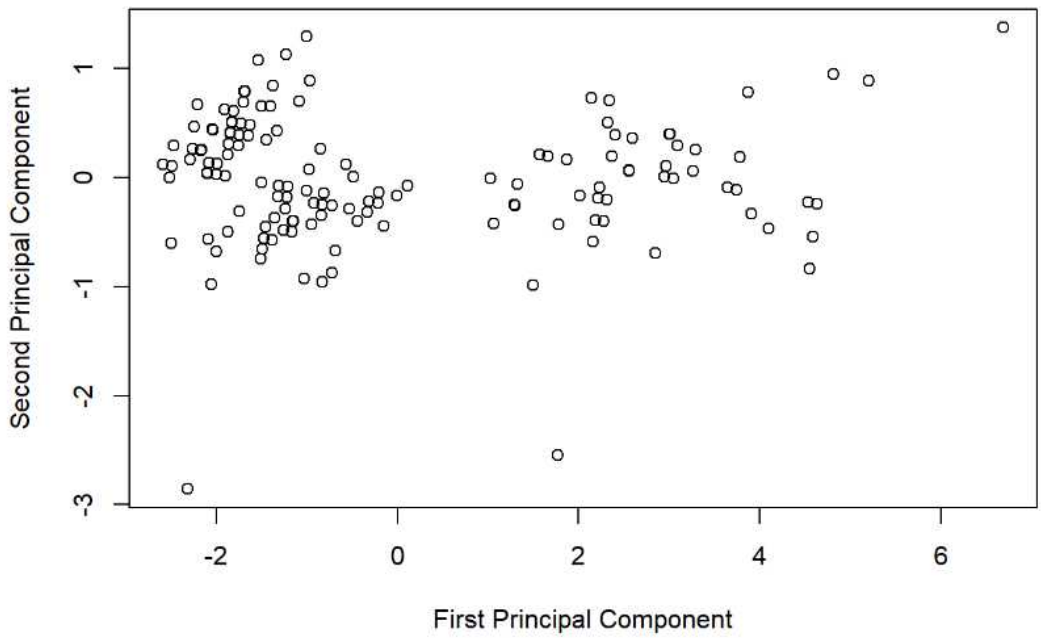
v) The PC1 has strong negative relationship on the first variable (-0.9894). This indicates a dependency but for the second variable (0.1454), it has no dependency because it is close to 0.

The PC2 has strong negative relationship on the second variable (-0.9894), indicating a dependency because it is close to 1.

vi. Display scatter plots of pairs of principal components

```
first_pca_result <- prcomp(xvalue, scale. = FALSE)

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



The scatter plot of the principal components presents a visual depiction of how the data is distributed with respect to the two main axes of variation. Most data points are clustered around the first principal component, with the cluster center near the origin, indicating that the first principal component has a significant role in explaining the variability in the data. There is also a wide spread along the second principal component, although it is less pronounced than the first, suggesting that while the first principal component captures the majority of the variation, the second principal component accounts for additional variability that is not explained by the first. This spread indicates the existence of another dimension of variability which the first principal component does not capture.

c) For the Y Group

i. Display the relevant sample covariance matrix S

```
Yvalue<- data[, 3:4]
Yvalue
```

```
## # A tibble: 153 × 2
##       Y1    Y2
##    <dbl> <dbl>
##  1  4.20 1.20
##  2  6.80 2.10
##  3  4.80 1.40
##  4  4.40 1.30
##  5  1.50 0.204
##  6  4.70 1.60
##  7  3.90 1.20
##  8 11.0  0.096
##  9  8.40 2.40
## 10  4.50 1.50
## # ℹ 143 more rows
```

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

```
##          Y1        Y2
## Y1 7.352679 2.1342820
## Y2 2.134282 0.8128447
```

The sample covariance matrix is $\begin{bmatrix} 7.35 & 2.13 \\ 2.13 & 0.81 \end{bmatrix}$

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

```
eigen_Yvalue<-eigen(cov(Yvalue, use="complete.obs"))
eigenvalues_Yvalue <- eigen_Yvalue$values

total_Y <- sum(eigenvalues_Yvalue)

percent_variance_Y <- eigenvalues_Yvalue / total_Y * 100

eigenvalues_Yvalue
```

```
## [1] 7.9875690 0.1779549
```

```
percent_variance_Y
```

```
## [1] 97.820656  2.179344
```

The eigenvalue for (PC1) is 7.988 and the PC1 explains 97.82% of the total variance.
The eigenvalue for (PC2) is 0.178 and the PC2 explains 2.18% of the total variance

The eigenvalues derived from the analysis suggest that the variance within the dataset is predominantly accounted for by the first principal component. These values help in understanding the extent to which each principal component captures the variability in the data, illustrating the underlying structure and patterns that exist within the dataset, especially highlighted by the dominance of PC1.

iii) Eigenvalue Criterion: A common rule of thumb is to keep principal components with eigenvalues greater than 1. In this case, the first eigenvalue is approximately 7.99, which exceeds this threshold, suggesting we should retain the first principal component. The second eigenvalue is about 0.18, which is less than 1, so typically we would not retain this component based on this criterion alone. Percentage of Variance Explained: The first principal component (PC1) explains about 97.82% of the variance, and the second principal component (PC2) explains about 2.18%. Since the PC1 accounts for the vast majority of the information, retaining only the PC1 may be justified. Combining these approaches, we can conclude that the PC1 significantly captures the structure within the data, and additional components contribute relatively little. Therefore, retaining only one principal component seems to be a reasonable approach for this dataset.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors_Y <- eigen_Yvalue$vectors

eigenvectors_Y
```

```
##            [,1]       [,2]
## [1,] -0.9584905  0.2851244
## [2,] -0.2851244 -0.9584905
```

The eigenvectors for the principal components is $\begin{bmatrix} -0.9585 & 0.2851 \\ -0.2851 & -0.9585 \end{bmatrix}$

For the PC1, the eigenvector is [-0.9585, 0.2851]. This indicates that PC1 is primarily associated with the first variable with a strong negative weight and a much smaller positive weight on the second variable, suggesting that as the first variable increases, the value on PC1 decreases, and as the second variable increases, the value on PC1 slightly increases.

For the PC2, the eigenvector is [-0.2851, -0.9585]. This demonstrates that PC2 has a small negative weight on the first variable and a strong negative weight on the second variable, indicating that PC2 captures the variation in both variables in a negative direction.

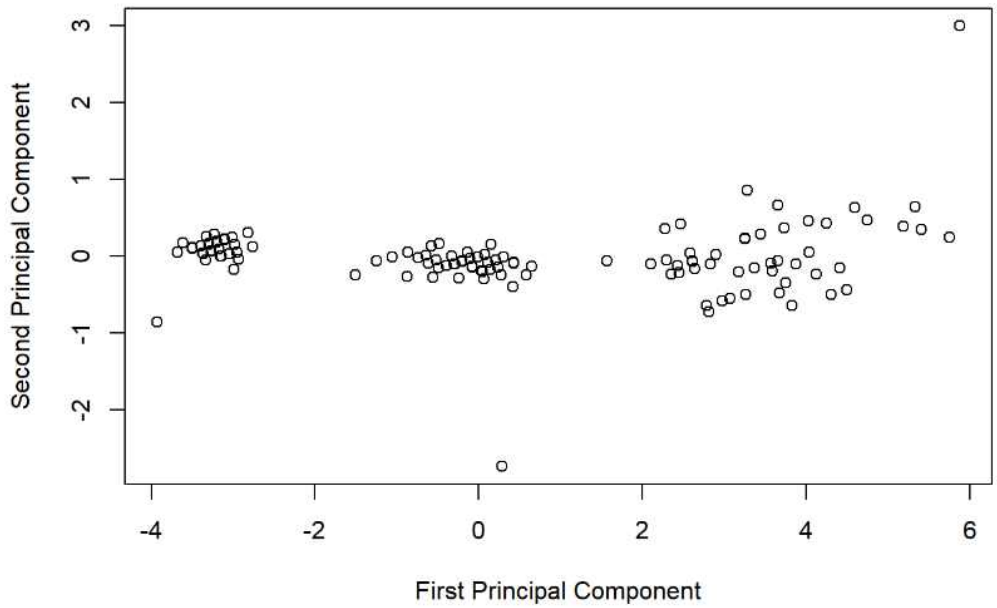
v) The PC1 has strong negative relationship on the first variable (-0.9585). This indicates a dependency but for the second variable (0.2851), it has no dependency because it is close to 0.

The PC2 has strong negative relationship on the second variable (-0.9585), indicating a dependency because it is close to 1.

vi. Display scatter plots of pairs of principal components

```
pca_result_Y <- prcomp(Yvalue, scale. = FALSE)

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



The scatter plot displayed indicates the distribution of data points across the first and second principal components. Observing the plot, it appears that a substantial number of data points are concentrated around the origin on the first principal component, with a spread out along the second principal component. This distribution aligns with the earlier description, indicating that the first principal component captures a significant portion of the data's variability, as evidenced by the dense clustering along this axis.

The second principal component, while explaining a smaller portion of the variance (as suggested by the previous eigenvalue and percentage variance explanation), still shows a spread of data points. This spread suggests that it captures additional, albeit less significant, variation in the data that the first principal component does not account for.

In summary, the scatter plot supports the description that the first principal component is dominant in explaining the variability in the data. At the same time, the second principal component does reveal additional dimensions of variability, albeit to a lesser extent. Therefore, the explanation provided matches the visual evidence from the scatter plot.

d) For the Entire Dataset

d. For the Entire Dataset

i. Display the relevant sample covariance matrix S

```
data1 <- read_excel("Assignment8_data.xlsx", col_names = c("X1", "X2", "X3", "X4"))

xx <- data1[, 1:2]
yy <- data1[, 3:4]
colnames(yy) <- c("X1", "X2")

entire <- rbind(xx, yy)
entire
```

```
## # A tibble: 306 × 2
##       X1    X2
##    <dbl> <dbl>
##  1  5.70 3.00
##  2  9.1  3.00
##  3  6.80 2.80
##  4  6.30 2.30
##  5  5.40 3.70
##  6  6.30 3.30
##  7  5.80 2.70
##  8  5.00 0.101
##  9 10.5  3.7
## 10  5.40 3.00
## # ℹ 296 more rows
```

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

```
##          X1       X2
## X1 7.398001 2.548674
## X2 2.548674 1.586312
```

The sample covariance matrix is $\begin{bmatrix} 7.398 & 2.549 \\ 2.549 & 1.586 \end{bmatrix}$

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] 8.3573441 0.6269697
##
## $vectors
##            [,1]       [,2]
## [1,] -0.9358951  0.3522789
## [2,] -0.3522789 -0.9358951
```

```
eigenvalues_entire <- eigen_entire$values

total_entire <- sum(eigenvalues_entire)

percent_variance_entire <- eigenvalues_entire / total_entire * 100

eigenvalues_entire
```

```
## [1] 8.3573441 0.6269697
```

```
percent_variance_entire
```

```
## [1] 93.021507  6.978493
```

The eigenvalue for (PC1) is 8.357 and the PC1 explains 93.02% of the total variance.
The eigenvalue for (PC2) is 0.627 and the PC2 explains 6.98% of the total variance

This means that the first principal component accounts for approximately 93.02% of the variability in the data, while the second principal component accounts for about 6.98%. The high percentage of the first eigenvalue suggests that the first principal component captures most of the variance in the dataset, indicating that it is the dominant feature in the data.

iii) Eigenvalue Criterion: According to this criterion, we should retain principal components with eigenvalues greater than 1. The first eigenvalue is significantly above this threshold at 8.36, indicating that the first principal component should be retained. The second eigenvalue, at 0.63, is below 1, which means it would typically not be retained according to this rule.
Percentage of Variance Explained: The first principal component (PC1) explains about 93.02% of the variance, and the second principal component (PC2) explains about 6.98%. Since the PC1 accounts for the vast majority of the information, retaining only the PC1 may be justified. Combining these approaches, we can conclude that the PC1 significantly captures the structure

within the data, and additional components contribute relatively little. Therefore, retaining only one principal component seems to be a reasonable approach for this dataset.

iv. Give the eigenvectors for the principal components you retain

```
eigenvectors <- eigen_entire$vectors

eigenvectors
```

```
##             [,1]       [,2]
## [1,] -0.9358951  0.3522789
## [2,] -0.3522789 -0.9358951
```

The eigenvectors for the principal components is $\begin{bmatrix} -0.9359 & 0.3523 \\ -0.3523 & -0.9359 \end{bmatrix}$

For the PC1, the eigenvector is [-0.9359, 0.3523]. This indicates that PC1 is primarily associated with the first variable with a strong negative weight and a much smaller positive weight on the second variable, suggesting that as the first variable increases, the value on PC1 decreases, and as the second variable increases, the value on PC1 slightly increases.

For the PC2, the eigenvector is [-0.3523, -0.9359]. This demonstrates that PC2 has a small negative weight on the first variable and a strong negative weight on the second variable, indicating that PC2 captures the variation in both variables in a negative direction.

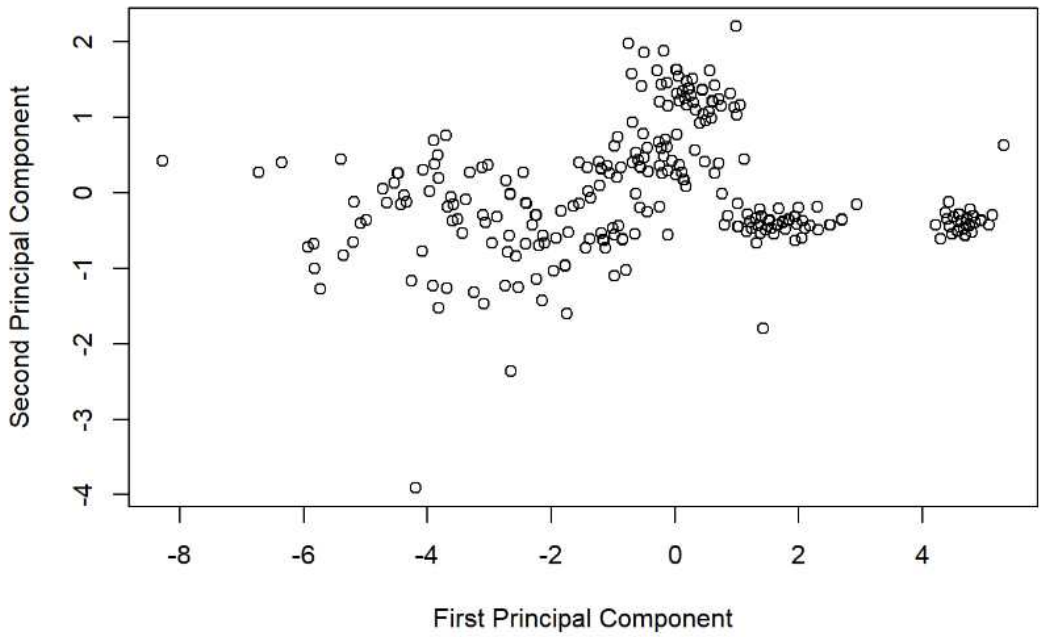
v) The PC1 has strong negative relationship on the first variable (-0.9359), indicating a dependency because it is close to 1.

The PC2 has strong negative relationship on the second variable (-0.9359), indicating a dependency because it is close to 1.

vi. 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")
```



Unlike the previous plots where most data points were clustered around the origin of the first principal component, in this plot, we observe a more dispersed distribution along both components. The data points form a curved elongated cluster that extends from the negative side of the first principal component into the positive side, with a considerable spread along the PC2 axis as well.

This pattern suggests that the PC1 still captures a significant portion of the data variability, particularly along a specific trajectory. However, the visible spread along the PC2 suggests it is also capturing a substantial amount of the variability, which could represent an underlying structure or pattern in the data that is orthogonal to the variance captured by the PC1.

In summary, while the PC1 remains a strong indicator of variability, the PC2 also plays a crucial role and should not be disregarded. The shape and distribution of the points suggest that both components are necessary to fully understand the underlying structure in the data.

e) Compare the results

Variance Explained: In all three cases, the PC1 explains a significant portion of the variance (X group: 93.49%, Y group: 97.82%, Entire dataset: 93.02%). This suggests that for all subsets of the data, PC1 is capturing a major underlying pattern or trend.
Eigenvalue Criterion: For all groups, the eigenvalue of PC1 is well above the threshold of 1, which supports the decision to retain PC1 in each analysis. The eigenvalues for PC2 are below 1 (X group: 0.331, Y group: 0.178, Entire dataset: 0.627), which typically would lead to their exclusion based on the eigenvalue criterion alone.
Contribution of PC2: Despite PC2 having a smaller eigenvalue, the scatter plots indicate that PC2 still captures meaningful variation not accounted for by PC1. Particularly in the entire dataset, the scatter plot suggests a more complex structure that may be important for a full understanding of the data.
Eigenvectors and Relationships: The eigenvectors indicate how each principal component is related to the original variables. In each group, PC1 has a strong negative relationship with the first variable, which suggests that the value of PC1 decreases as the first variable increases.