STAT 445/645 Assignment Cover Page

Student Name	Heewon Oh
SFU Student Number	301268860
SFU email address	heewono@sfu.ca
Assignment Number	5
Due Date	March 24,2022

Provide references for any data sets used in this assignment

A. Rencher and W. Christensen, Methods of Multivariate Analysis, Wiley, New York, 2012.

Lubischew, A., On the use of discriminant functions in taxonomy, Biometrics 18 (1962), 455-477.

R. Johnson and D. Wichern, Applied Multivariate Statistical Analysis, Pearson, New Jersey, 2019.

List software used in this assignment.

R, Rstudio

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

Code from Tutorials,
Notes from Class

https://www.rdocumentation.org/packages/readxl/versions/1.3.1/topics/read_excel

https://www.gastonsanchez.com/visually-enforced/how-to/2014/01/15/Center-data-in-R/

https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp

https://jules32.github.io/r-for-excel-users/readxl.html

Knowledge from previous classes on R regarding the usage of dplyr package.

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

Department of Statistics and Actuarial Science

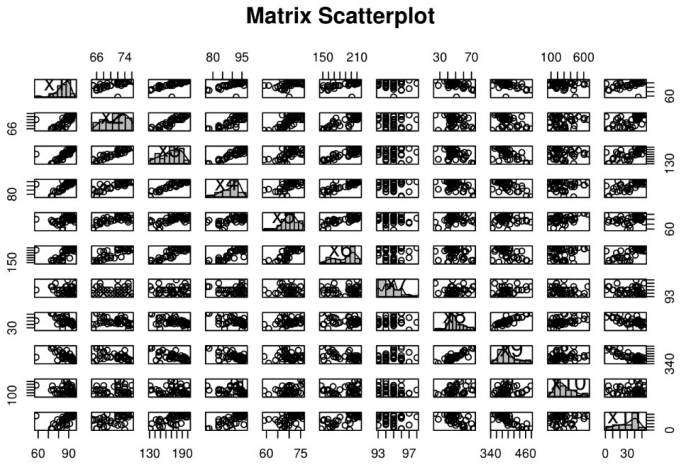
SFU SIMON FRASER UNIVERSITY

1)A)

We had to skip the first row of the file containing a comment and use the 2nd row as the column names.

The different variables in the data measure different properties, such as total wind, humidity, and temperature, which means that the different variables are using different units. There is no indication of groups. The PCA expects us to center the data around the sample mean, so while we did not subtract the sample mean from the data as part of the initial assessment, we do so later in the PCA.

Below is the matrix scatterplot. It can be read as the independent variable being that belonging to the variable associated with the column and the dependent variable belonging to the variable associated with the row.



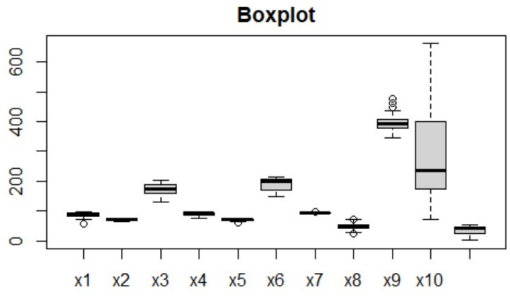
There are indications of outliers as there are numerous histograms that are heavily skewed due to extremes, such as X1, which has a very large negative outlier, which we can also see on the scatterplots. We can similarly see this pronounced on X4 and X5, as well as X7 although it is due to several positive outliers.

There are some indications of dependencies, such as many of the scatterplots appear to have a relationship, rather than being randomly scattered. For instance, the scatterplots for X1 and X2, X1 and X3, and X1 and X4 appear to have somewhat of a positive linear relationship. Similarly,

X3 and X4, X5 and X6, X4 and X6, and X8 and X9 appear to have a somewhat positive linear relationship. Conversely, X9 and X11 appear to have a somewhat negative linear relationship.

For majority of the scatterplots, the data does not cluster heavily. However, we note that there are many notable exceptions such as some of the ones involving X1, X6, X8 and X9, such as X1 and X2, X1 and X8, X1 and X9, X1 and X10, etc. Many involving X6, such as X6 and X10 appear to have two clusters. Furthermore, the data does not appear to be consistent with a normal distribution.

Below is a boxplot:



Based on the boxplot and matrix scatterplot, there is the potential for scale issues because of the difference in range and values from variables such as X10 and X2 being very large.

B)

The sample Covariance matrix S is:

```
##             x1         x2          x3          x4         x5          x6
## x1   55.681159  16.489855  117.553623  26.9768116  10.383575   97.729469
## x2   16.489855  10.863768   61.672464  14.0251208   8.227053   56.478261
## x3  117.553623  61.672464  402.699517  92.9497585  43.457005  365.265700
## x4   26.976812  14.025121   92.949758  25.6637681  10.595169   92.810628
## x5   10.383575   8.227053   43.457005  10.5951691  13.440097   59.925121
## x6   97.729469  56.478261  365.265700  92.8106280  59.925121  438.253623
## x7   -1.558454  -0.657971   -3.861836  -0.6309179  -0.536715   -1.023671
## x8  -42.356522 -10.797101 -106.894686 -26.6695652   8.461836  -59.944928
## x9 -128.272464 -44.404831 -387.723671 -94.3285024  -4.362319 -313.469565
## x10 -209.095652  14.931401 -294.551691 -67.9314010 224.577295  386.818841
## x11   61.367150  25.987440  201.869565  53.6125604  17.580193  217.061353
##              x7          x8          x9         x10         x11
## x1   -1.5584541  -42.356522 -128.272464  -209.09565   61.367150
## x2   -0.6579710  -10.797101  -44.404831    14.93140   25.987440
## x3   -3.8618357 -106.894686 -387.723671  -294.55169  201.869565
## x4   -0.6309179  -26.669565  -94.328502   -67.93140   53.612560
## x5   -0.5367150    8.461836   -4.362319   224.57729   17.580193
## x6   -1.0236715  -59.944928 -313.469565   386.81884  217.061353
## x7    1.4516908    1.902415    9.970048   -26.38309   -3.240097
## x8    1.9024155  106.197101  271.649275   597.66860  -96.854106
```

The Sample Correlation matrix R is:

```
##             x1          x2          x3          x4          x5          x6
## x1   1.0000000  0.67045936  0.78503918  0.71363522  0.37956973  0.62561715
## x2   0.6704594  1.00000000  0.93241768  0.83995526  0.68085223  0.81851815
## x3   0.7850392  0.93241768  1.00000000  0.91431848  0.59070146  0.86947196
## x4   0.7136352  0.83995526  0.91431848  1.00000000  0.57048796  0.87513437
## x5   0.3795697  0.68085223  0.59070146  0.57048796  1.00000000  0.78080928
## x6   0.6256171  0.81851815  0.86947196  0.87513437  0.78080928  1.00000000
## x7  -0.1733416 -0.16568357 -0.15972253 -0.10336549 -0.12150816 -0.04058455
## x8  -0.5508202 -0.31787810 -0.51690321 -0.51085705  0.22397888 -0.27786468
## x9  -0.5776349 -0.45270372 -0.64924087 -0.62568655 -0.03998437 -0.50316070
## x10 -0.1879528  0.03038564 -0.09845301 -0.08994321  0.41088727  0.12393755
## x11  0.5621001  0.53889617  0.68756243  0.72333284  0.32775875  0.70868231
##              x7          x8          x9         x10         x11
## x1  -0.17334161 -0.5508202 -0.57763485 -0.18795284  0.56210010
## x2  -0.16568357 -0.3178781 -0.45270372  0.03038564  0.53889617
## x3  -0.15972253 -0.5169032 -0.64924087 -0.09845301  0.68756243
## x4  -0.10336549 -0.5108571 -0.62568655 -0.08994321  0.72333284
## x5  -0.12150816  0.2239789 -0.03998437  0.41088727  0.32775875
## x6  -0.04058455 -0.2778647 -0.50316070  0.12393755  0.70868231
## x7   1.00000000  0.1532188  0.27805729 -0.14687459 -0.18380306
## x8   0.15321877  1.0000000  0.88578089  0.38901123 -0.64238228
## x9   0.27805729  0.8857809  1.00000000  0.21884501 -0.81623034
## x10 -0.14687459  0.3890112  0.21884501  1.00000000  0.04336362
## x11 -0.18380306 -0.6423823 -0.81623034  0.04336362  1.00000000
```

C)

i)

List of Eigenvalues:

```
##  [1] 2.230350e+04 1.590679e+03 3.580457e+02 6.336652e+01 2.932695e+01
##  [6] 1.711486e+01 1.274780e+01 2.832997e+00 1.906924e+00 8.769315e-01
## [11] 7.028301e-01
```

List of percent contributions to variance:

```
##  [1] 91.478642661  6.524230092  1.468538184  0.259900205  0.120285601
##  [6]  0.070197257  0.052285579  0.011619646  0.007821322  0.003596768
## [11]  0.002882684
```

ii)

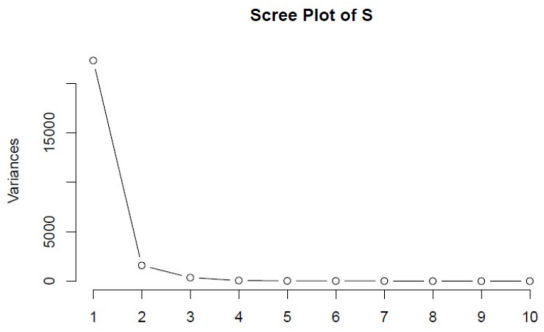
The number of principal components we retain is 1. This is because of 3 methods:

The first method is the sufficient percentage of variance method. We set the threshold arbitrarily at 80%. The first principal component has greater percentage contribution to variance than this.

The second method is maintaining more than the average of eigenvalue method.

The average is: 9.090909%. Only the first principal component passes the threshold.

The last method we use is the Scree Plot. Below, we can see that the bend changes on point 2, which means that the first principal component is the only important one here as well.



iii)

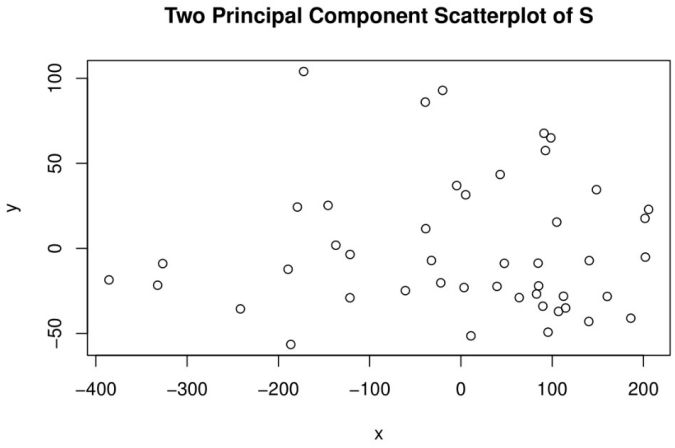
The eigenvectors for the principal components we retain (the first principal component) is:

```
##           x1           x2           x3           x4           x5
##  0.0096876517 -0.0005662549  0.0141012748  0.0032575176 -0.0100731934
##           x6           x7           x8           x9          x10
## -0.0167008155  0.0011564711 -0.0274697393 -0.0456252296 -0.9982317565
##          x11
## -0.0034292494
```

No formula for the principal component. -1

iv) We can see that the coefficient for x10 is much larger than the rest, which means that the first principal component has a strong de... wind, miles per day). It is not dependent on the remaining variable...

v) We use two principal components for the scatterplot display only. In our earlier analysis, we noted that only the first principal component was necessary. We see that there is no obvious dependency between PC1 (x-axis) and PC2 (y-axis). However, there is a greater concentration to the lower-right. There are no notable outliers.



D)

i)

The eigenvalues are:

```
##  [1] 6.02024515 2.11933612 1.13029100 0.76001708 0.35535540 0.25934244
##  [7] 0.12207563 0.11048840 0.05980829 0.04218515 0.02085533
```

Their percentage contributions to variance are:

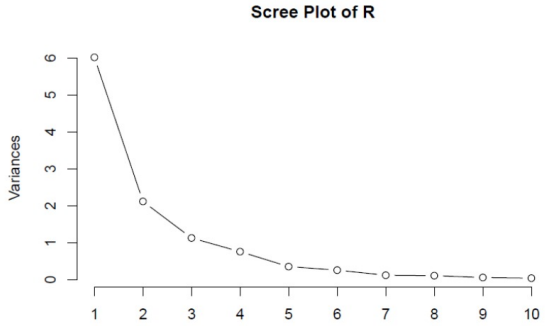
```
##  [1] 54.7295013 19.2666920 10.2753727  6.9092462  3.2305037  2.3576586
##  [7]  1.1097785  1.0044400  0.5437117  0.3835014  0.1895939
```

ii) The number of principal components we retain are 3.

We can see that the arbitrary percentage threshold method of 80% is not satisfied by only 2 components (~74%), but is by 3 (~84%).

We can also see from the average eigenvalue contribution method, the average is: 9.090909%, which means that the first 3 eigenvalues are above this and not the remaining ones.

We also see from the Scree plot, the result is rather inconclusive between supporting 2 and 3 principal components as it can be argued the bend starts either between points 3 and 4. However, this is not surprising given that the 3rd principal value is close to the average we saw from the earlier method. Thus, we have good support for using 3 principal values.



iii)

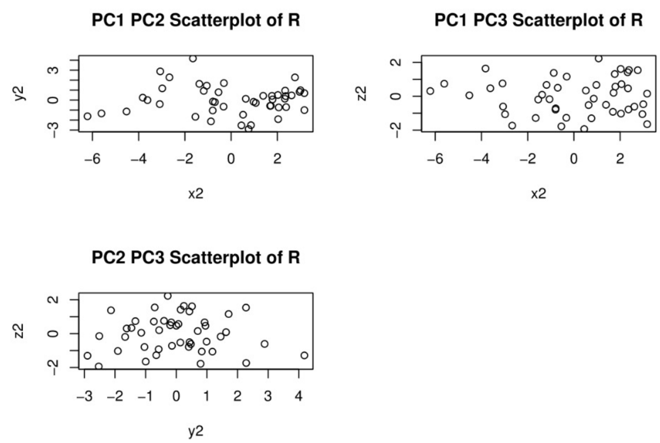
The eigenvectors for the 3 principal components we retain are:

	PC1	PC2	PC3
x1	0.33042817	0.07872408	0.08800766
x2	0.35415881	-0.19280098	0.10705532
x3	0.39232582	-0.05181668	0.11048102
x4	0.38204564	-0.04738017	0.13335408
x5	0.23230571	-0.53031822	0.01542079
x6	0.36212305	-0.23605654	0.11982646
x7	-0.08843948	-0.02126463	0.79460449
x8	-0.25005597	-0.50229576	0.08261299
x9	-0.31110797	-0.35947297	0.21358474
x10	-0.02426425	-0.46848762	-0.46693016
x11	0.33568563	0.11526346	-0.18532362

No formula for the principal components. -2

iv) The first principal component is dependent on all the variables fairly similarly except x7 and x10 (maximum daily humidity and total wind, miles per day); albeit, x5 and x8 (minimum daily soil temperature and minimum daily relative humidity) less so than the others not mentioned. The second principal component is dependent on x5, x8, and x10 most heavily (minimum daily soil temperature, minimum daily relative humidity, and total wind (miles per day)). It is more not dependent on x1, x3, x4, and x7 (maximum daily air temperature, integrated area under daily air temperature curve, maximum daily soil temperature, and maximum daily relative humidity). The third principal component is dependent on x7 (particularly strongly) and x10 heavily (maximum daily relative humidity and total wind (miles per day)). It is not dependent on x1, x5, and x8 (maximum daily air temperature, minimum daily soil temperature, and minimum daily relative humidity).

v) We can see that there are potential outliers, most apparent in the right-most point in the PC2 and PC3 Scatterplot of R and the two leftmost points in the PC1 PC2 and PC1 PC3 scatterplots. The PC1 PC2 scatterplot is relatively clustered to the right, the PC2 and PC3 scatterplot to the middle and the PC1 PC3 scatterplot is relatively dispersed. None of the 3 have any obvious dependencies.



e) The PCA for R is better because as we can see from the matrix scatterplots; there are no apparent outliers and thus there is a more normal distribution as we can see from the histograms in comparison.

f) There are 3 principal components for PCA that we find are significant for R (the better PCA). All 11 variables had a reasonably large effect on at least one of the 3 principal components. However, as mentioned in 1)D)V), each PCA has stronger and weaker relative dependencies amongst different variables. Thus, this suggests that at 11 variables are important. The principal components do not appear to show any proof of linear dependencies between the variables here. However, as the proportional contributions to variance are higher in PC1 and PC2 than PC3 significantly, we can infer that x7 (maximum daily relative humidity) has a very small effect on the variance when compared to variables like x10 (total wind, miles per day).

Q1

G) Code for 1a)

```
library(readxl)
temperature=read_excel("temperaturedata.xlsx",col_names=TRUE, skip=1)
```

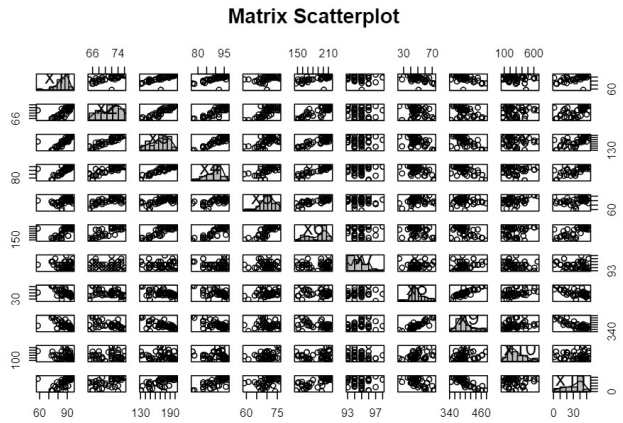
```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.3
```

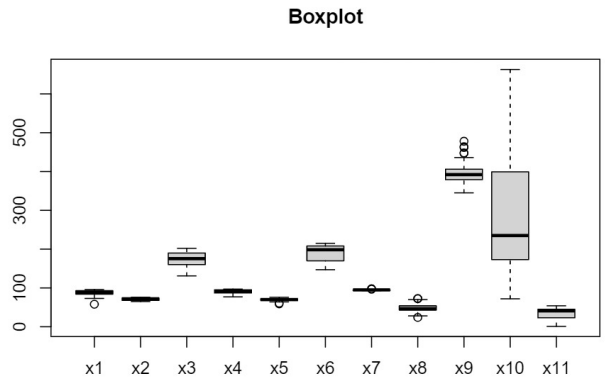
```
library(MESS)
```

```
## Warning: package 'MESS' was built under R version 4.1.3
```

```
pairs(temperature[,c(1:11)], diag.panel=panel.hist,main="Matrix Scatterplot")
```



```
boxplot(temperature,main="Boxplot")
```



Code for 1b)

```
S <- cov(temperature)
S
```

```
##             x1         x2          x3          x4         x5          x6
## x1    55.681159  16.489855  117.553623  26.9768116  10.383575   97.729469
## x2    16.489855  10.863768   61.672464  14.0251208   8.227053   56.478261
## x3   117.553623  61.672464  402.699517  92.9497585  43.457005  365.265700
## x4    26.976812  14.025121   92.949758  25.6637681  10.595169   92.810628
## x5    10.383575   8.227053   43.457005  10.5951691  13.440097   59.925121
## x6    97.729469  56.478261  365.265700  92.8106280  59.925121  438.253623
## x7    -1.558454  -0.657971   -3.861836  -0.6309179  -0.536715   -1.023671
## x8   -42.356522 -10.797101 -106.894686 -26.6695652   8.461836  -59.944928
## x9  -128.272464 -44.404831 -387.723671 -94.3285024  -4.362319 -313.469565
## x10 -209.095652  14.931401 -294.551691 -67.9314010 224.577295  386.818841
## x11   61.367150  25.987440  201.869565  53.6125604  17.580193  217.061353
##              x7          x8          x9         x10         x11
## x1   -1.5584541  -42.356522 -128.272464  -209.09565   61.367150
## x2   -0.6579710  -10.797101  -44.404831    14.93140   25.987440
## x3   -3.8618357 -106.894686 -387.723671  -294.55169  201.869565
## x4   -0.6309179  -26.669565  -94.328502   -67.93140   53.612560
## x5   -0.5367150    8.461836   -4.362319   224.57729   17.580193
```

```
## x6   -1.0236715  -59.944928 -313.469565   386.81884  217.061353
## x7    1.4516908    1.902415    9.970048   -26.38309   -3.240097
## x8    1.9024155  106.197101  271.649275   597.66860  -96.854106
## x9    9.9700483  271.649275  885.628986   970.96715 -355.391304
## x10 -26.3830918  597.668599  970.967150 22227.15797   94.587923
## x11  -3.2400966  -96.854106 -355.391304    94.58792  214.060386
```

```
R<-cor(temperature)
R
```

```
##             x1          x2          x3          x4          x5          x6
## x1   1.0000000  0.67045936  0.78503918  0.71363522  0.37956973  0.62561715
## x2   0.6704594  1.00000000  0.93241768  0.83995526  0.68085223  0.81851815
## x3   0.7850392  0.93241768  1.00000000  0.91431848  0.59070146  0.86947196
## x4   0.7136352  0.83995526  0.91431848  1.00000000  0.57048796  0.87513437
## x5   0.3795697  0.68085223  0.59070146  0.57048796  1.00000000  0.78080928
## x6   0.6256171  0.81851815  0.86947196  0.87513437  0.78080928  1.00000000
## x7  -0.1733416 -0.16568357 -0.15972253 -0.10336549 -0.12150816 -0.04058455
## x8  -0.5508202 -0.31787810 -0.51690321 -0.51085705  0.22397888 -0.27786468
## x9  -0.5776349 -0.45270372 -0.64924087 -0.62568655 -0.03998437 -0.50316070
## x10 -0.1879528  0.03038564 -0.09845301 -0.08994321  0.41088727  0.12393755
## x11  0.5621001  0.53889617  0.68756243  0.72333284  0.32775875  0.70868231
##              x7         x8          x9         x10         x11
## x1  -0.17334161 -0.5508202 -0.57763485 -0.18795284  0.56210010
## x2  -0.16568357 -0.3178781 -0.45270372  0.03038564  0.53889617
## x3  -0.15972253 -0.5169032 -0.64924087 -0.09845301  0.68756243
## x4  -0.10336549 -0.5108571 -0.62568655 -0.08994321  0.72333284
## x5  -0.12150816  0.2239789 -0.03998437  0.41088727  0.32775875
## x6  -0.04058455 -0.2778647 -0.50316070  0.12393755  0.70868231
## x7   1.00000000  0.1532188  0.27805729 -0.14687459 -0.18380306
## x8   0.15321877  1.0000000  0.88578089  0.38901123 -0.64238228
## x9   0.27805729  0.8857809  1.00000000  0.21884501 -0.81623034
## x10 -0.14687459  0.3890112  0.21884501  1.00000000  0.04336362
## x11 -0.18380306 -0.6423823 -0.81623034  0.04336362  1.00000000
```

Code for 1C)

```
pca=prcomp(temperature[1:11],scale=FALSE,center=TRUE)
eigen<- (pca$sdev)^2
eigen
```

```
##  [1] 2.230350e+04 1.590679e+03 3.580457e+02 6.336652e+01 2.932695e+01
##  [6] 1.711486e+01 1.274780e+01 2.832997e+00 1.906924e+00 8.769315e-01
## [11] 7.028301e-01
```

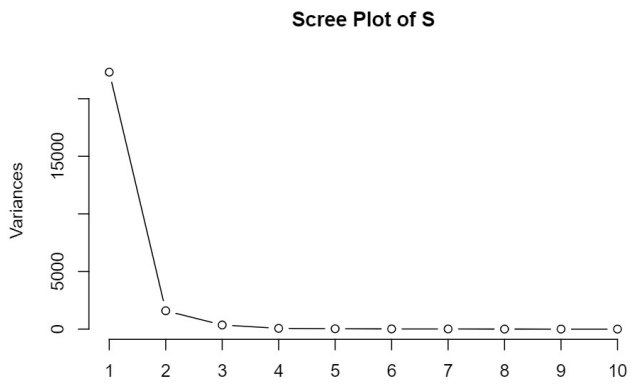
```
pvar=100*(pca$sdev)^2/sum((pca$sdev)^2)
pvar
```

```
##  [1] 91.478642661  6.524230092  1.468538184  0.259900205  0.120285601
##  [6]  0.070197257  0.052285579  0.011619646  0.007821322  0.003596768
## [11]  0.002882684
```

```
pvaravg=mean(100*(pca$sdev)^2/sum((pca$sdev)^2))
pvaravg
```

```
## [1] 9.090909
```

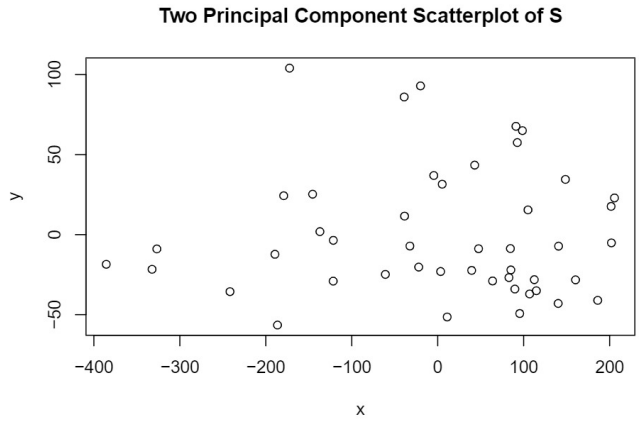
```
screeplot(pca,type="lines",main="Scree Plot of S")
```




```
principalcomp<- pca$rotation[,1]
principalcomp
```

```
##            x1            x2            x3            x4            x5
##  0.0096876517 -0.0005662549  0.0141012748  0.0032575176 -0.0100731934
##            x6            x7            x8            x9           x10
## -0.0167008155  0.0011564711 -0.0274697393 -0.0456252296 -0.9982317565
##           x11
## -0.0034292494
```

```
pc1<-pca$x[,1]#pc1
pc2<-pca$x[,2]#pc2
x=pc1
y=pc2
plot(x,y,main="Two Principal Component Scatterplot of S")
```



Code for 1D)

```
pca2=prcomp(temperature[1:11],scale=TRUE,center=TRUE)
eigen2<- (pca2$sdev)^2
eigen2
```

```
##  [1] 6.02024515 2.11933612 1.13029100 0.76001708 0.35535540 0.25934244
##  [7] 0.12207563 0.11048840 0.05980829 0.04218515 0.02085533
```

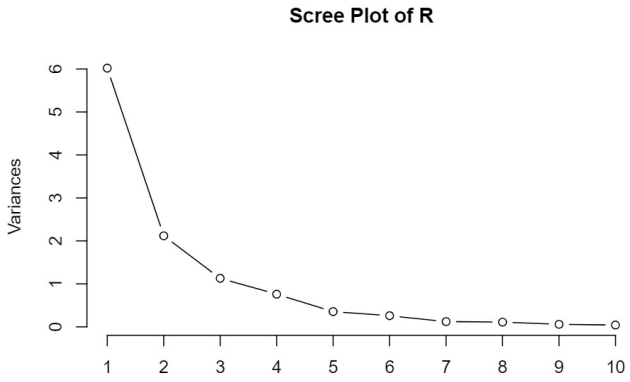
```
pvar2=100*(pca2$sdev)^2/sum((pca2$sdev)^2)
pvar2
```

```
##  [1] 54.7295013 19.2666920 10.2753727  6.9092462  3.2305037  2.3576586
##  [7]  1.1097785  1.0044400  0.5437117  0.3835014  0.1895939
```

```
pvaravg2=mean(100*(pca2$sdev)^2/sum((pca2$sdev)^2))
pvaravg2
```

```
## [1] 9.090909
```

```
screeplot(pca2,type="lines", main= "Scree Plot of R")
```

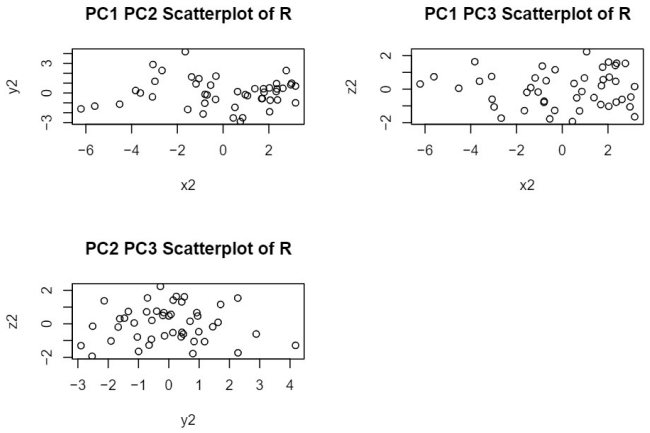



```
principalcomp2<- pca2$rotation[,1:3]
principalcomp2
```

```
##             PC1         PC2         PC3
## x1   0.33042817  0.07872408  0.08800766
## x2   0.35415881 -0.19280098  0.10705532
## x3   0.39232582 -0.05181668  0.11048102
## x4   0.38204564 -0.04738017  0.13335408
## x5   0.23230571 -0.53031822  0.01542079
## x6   0.36212305 -0.23605654  0.11982646
## x7  -0.08843948 -0.02126463  0.79460449
## x8  -0.25005597 -0.50229576  0.08261299
## x9  -0.31110797 -0.35947297  0.21358474
## x10 -0.02426425 -0.46848762 -0.46693016
## x11  0.33568563  0.11526346 -0.18532362
```

```
pc3<-pca2$x[,1]#pc1
pc4<-pca2$x[,2]#pc2
pc5<-pca2$x[,3]#pc3
x2=pc3
y2=pc4
z2=pc5
par(mfrow=c(2,2))
plot(x2,y2,main="PC1 PC2 Scatterplot of R")
```

```
plot(x2,z2,main="PC1 PC3 Scatterplot of R")
plot(y2,z2,main="PC2 PC3 Scatterplot of R")
```

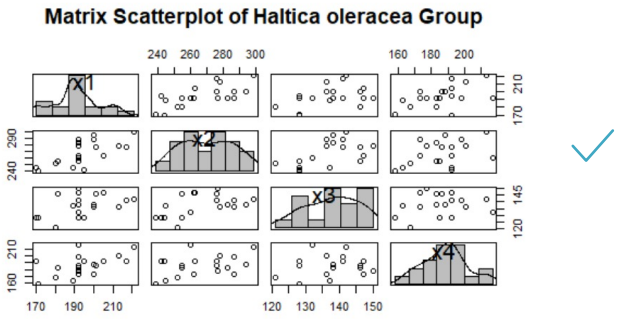


2)A)

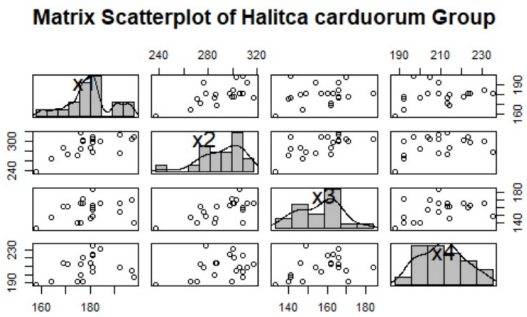
We had to skip the first two rows of the file containing a comment and the group names use the 3rd row as the column names. We also assign a new column for groups 1 and 2 (1 being oleracea and 2 being carduorum) and create two different data frames and a third one with them row-bound. We do not maintain the experiment number and thus have 19 rows in group 1 and 20 in group 2.

The different variables in the data measure different lengths of the beetles, which we assume are in the same units. Like mentioned above, the data file indicates two groups: oleracea and carduorum. The PCA expects us to center the data around the sample mean, so while we did not subtract the sample mean from the data as part of the initial assessment, we do so later in the PCA.

Below are the matrix scatterplots. They can be read as the independent variable being that belonging to the variable associated with the column and the dependent variable belonging to the variable associated with the row.

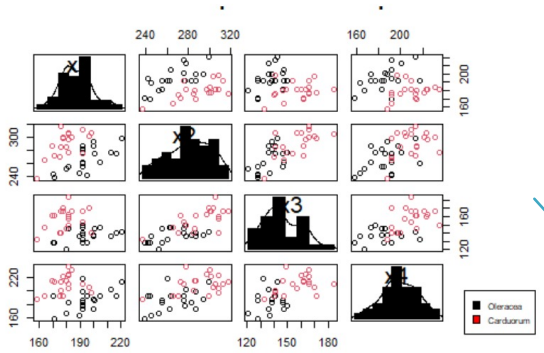


The histograms in the matrix scatterplot for oleracea shows that the data is largely, not normally distributed; however, there are no obvious indicators of outliers. There are no obvious clusters in any of the scatterplots. There are very weak indications of potentially weak positive linear dependencies in several scatterplots such as the one for x1 and x2, x2 and x3, and x2 and x4.



The histograms in the matrix scatterplot for carduorum shows that the data is largely, not normally distributed and is relatively more skewed than the histograms in oleracea. The scatterplots suggests that there is a point on x2 that may be a low outlier. There are no obvious clusters in any of the scatterplots. There are very weak indications of potentially weak positive linear dependencies in several scatterplots such as the one for x1 and x2, x2 and x3, and x3 and x4. Some scatterplots appear to have stronger indications than in oleracea and some weaker.

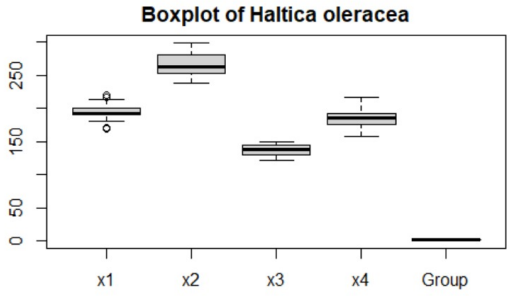
Matrix Plot for Combined Group

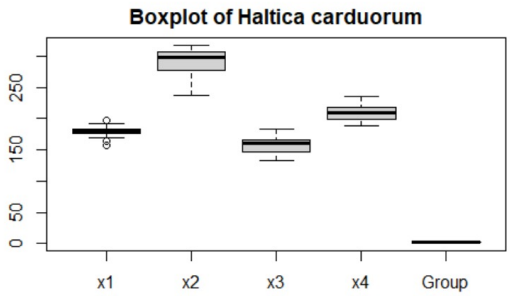


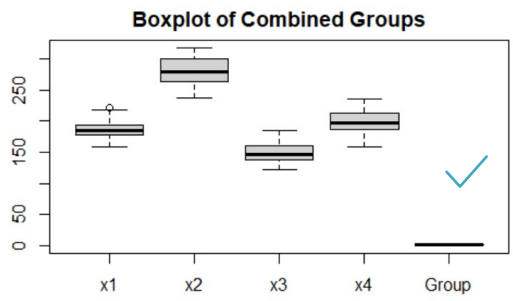
Here we see that from the matrix scatterplots, there is relative separation between the two groups. This is apparent in all the scatterplots. Thus, we can infer that the groups behave differently from each other. There are no obvious large clusters. Interestingly, while the histograms are certainly not normally distributed, they are much closer to a normal distribution

a. The plots do not provide strong evidence of an outlier. -1

the carduorum matrix scatterplot, but stronger than in the oleracea one. The most apparent weaker in some plots such as that for x1 and x2 (compared to when they are separated).
However, it is also stronger in some plots such as those for x3 and x4, x2 and x4, and x2 and x3.


Boxplot of Haltica oleracea
0
50
150
250
x1
x2
x3
x4
Group


Boxplot of Haltica carduorum
0
50
150
250
x1
x2
x3
x4
Group



All 3 boxplots show that there is a different in range and relative size between the 4 variables. However, this does not different greatly from the 3 boxplots (the two groups a... relative differences between the variables are also not extreme. Thus, we do n... scale issues.

a. There is a potential for scale issues. -1

B)

i) Sample covariance matrix S:

```
##           x1        x2       x3        x4
## x1 187.59649 176.86257 48.37135 113.58187
## x2 176.86257 345.38596 75.97953 118.78070
## x3  48.37135  75.97953 66.35673  16.24269
## x4 113.58187 118.78070 16.24269 239.94152
```

ii)

Eigenvalues:

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

Percent contributions to variance:

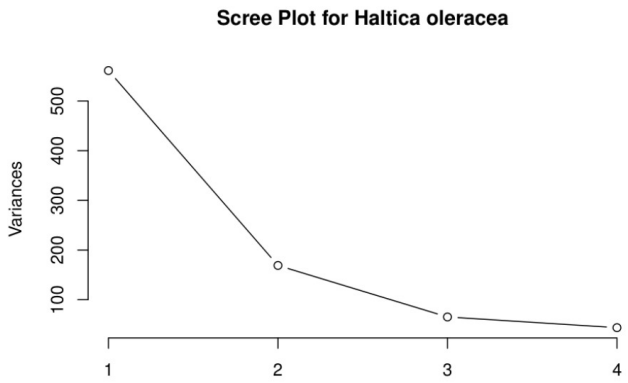
```
## [1] 66.879382 20.134603  7.777743  5.208273
```

iii) Number of principal components is 2. This is justified by 3 methods.

The first method is the cumulative proportion threshold, which we arbitrarily set to be 80%. We can see that this is reached by using 2 principal components.

The 2nd method is the average proportional eigenvalue method. The average proportional eigenvalue is 25%. The 2nd principal component under this method is a bit shy of this. However, we note that critically, there are only 4 eigenvalues. Thus, under such a situation with so few eigenvalues, this method requires some flexibility. Thus, this method results in an inconclusive result.

The 3rd method is the scree plot.



The scree plot suggests 2 principal components as the bend changes past point 3. Thus, the evidence points towards 2 PCs.

iv)

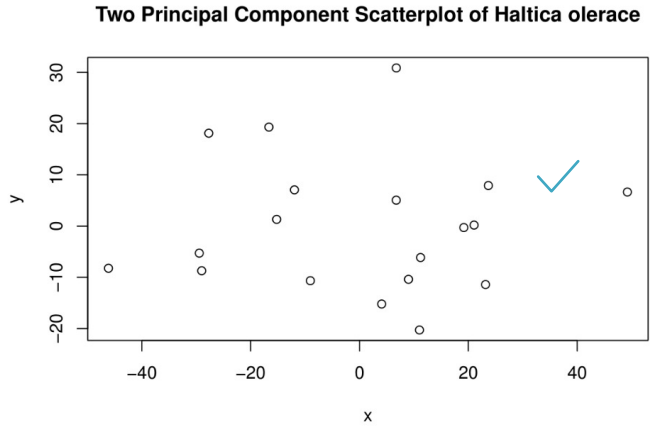
The eigenvectors for PC1 and PC2 we retain are:

```
##           PC1          PC2
## x1 0.4997445  0.009204574
## x2 0.7187015 -0.484408702
## x3 0.1739702 -0.220296505
## x4 0.4510631  0.846600812
```

v) For PC1, the x2 coefficient is very large, and the x3 very small. That for x1 and x4 are similar. Thus, we can see that PC1 is heavily dependent on x2 (length of elytra) and not on x3 (length of

second antennal joint). We can see that for PC2, the coefficient for x4 is very large and that for x1 very small (x2 and x3 are moderate with x2 being about double that of x3). Thus, PC2 is heavily dependent on x4 (length of third antennal joint) and not dependent of x1 (distance of transverse groove).

vi)



There is very little to note from this scatterplot as it is relatively well dispersed. However, we note that there are no apparent dependencies, nor outliers.

C)

i) The sample covariance matrix of the Carduorum group:

```
##           x1        x2        x3        x4
## x1 101.83947 128.06316  36.98947  32.59211
## x2 128.06316 389.01053 165.35789  94.36842
## x3  36.98947 165.35789 167.53684  66.52632
## x4  32.59211  94.36842  66.52632 177.88158
```

ii) The eigenvalues:

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

The percent contributions to the variance:

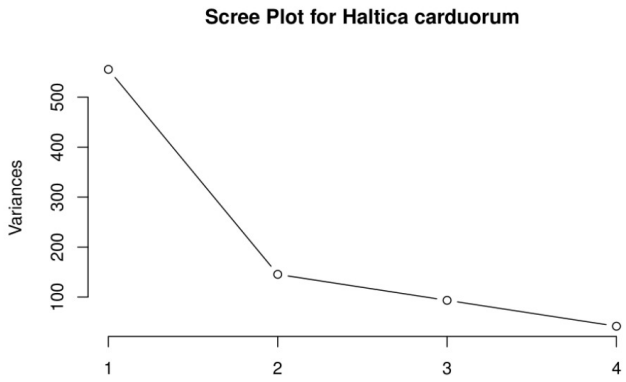
```
## [1] 66.44914 17.39230 11.17628  4.98228
```

iii) We determine 2 principal components. We justify it using three methods:

The first method is the cumulative proportion threshold, which we arbitrarily set to be 80%. We can see that this is reached by using 2 principal components.

The 2nd method is the average proportional eigenvalue method. The average proportional eigenvalue is 25%. The 2nd principal component under this method is a shy of this. However, we note that critically, there are only 4 eigenvalues. Thus, under such a situation with so few eigenvalues, this method requires some flexibility. Thus, this method results in an inconclusive result.

The 3rd method is the scree plot.



The scree plot also gives a relatively inconclusive result that suggests anywhere from 1 to 3 PCs as there is a bend going to points 2 to 3, and a similar one from 3 to 4, that is notably sloped somewhat downwards. However, like the average method, having only 4 values in the eigenvector makes this method weak. Thus, we default back to the first test and decide on 2 PCs.

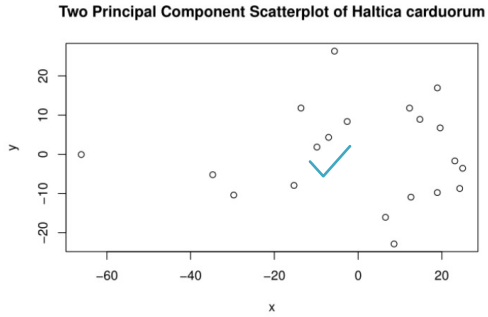
iv) The eigenvectors of the 2 PCs we maintain

```
##           PC1        PC2
## x1 0.2836552 -0.2007357
## x2 0.8068689 -0.3389760
## x3 0.4222422  0.1359900
## x4 0.3003563  0.9090144
```

v)

For PC1, the x2 coefficient is very large, and the others somewhat similar. Thus, we can see that PC1 is heavily dependent on x2 (length of elytra) and still dependent on the other 3 variables. We can see that for PC2, the coefficient for x4 (length of third antennal joint) is very large and that for x3 (length of 2nd antennal joint) and to a lesser degree, x1 (distance of transverse groove) fairly small. Thus, PC2 is heavily dependent on x4 (length of third antennal joint) and not dependent of x3 and very minimally dependent on x1.

vi)



Like the results in Oleracea, there very little to note from this scatterplot as there is no apparent relationship However, here, we see potential evidence for an outlier on the left-most point.

D)

i) Sample Matrix of Combined Group

```
##           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
```

ii) List of eigenvalues

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

Percent contributions to variance:

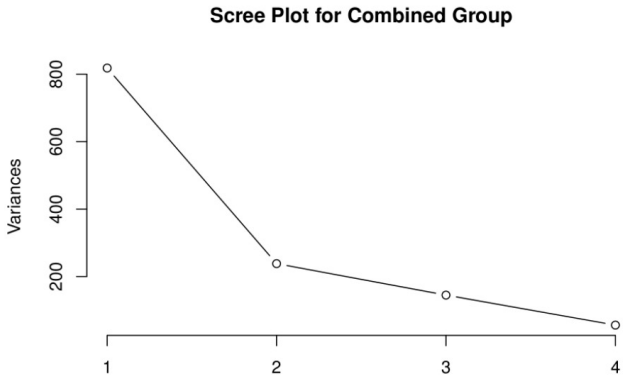
```
## [1] 65.072875 18.945102 11.527960  4.454063
```

iii) The number of Principal components we maintain are 2. We justify it using three methods:

The first method is the cumulative proportion threshold, which we arbitrarily set to be 80%. We can see that this is reached by using 2 principal components.

The 2nd method is the average proportional eigenvalue method. The average proportional eigenvalue is 25%. The 2nd principal component under this method is a shy of this. However, we note that critically, there are only 4 eigenvalues. Thus, under such a situation with so few eigenvalues, this method requires some flexibility. Thus, this method results in an inconclusive result.

The 3rd method is the scree plot.



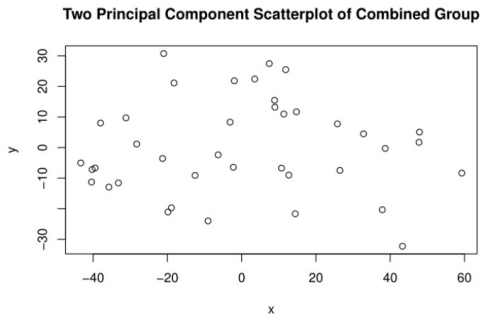
The scree plot also gives a relatively inconclusive result that suggests anywhere from 1 to 3 PCs as there is a bend going to points 2 to 3, and a similar one from 3 to 4, that is notably sloped somewhat downwards. However, like the average method, having only 4 values in the eigenvector makes this method weak. Thus, we default back to the first test and decide on 2 PCs

iv) The eigenvectors of the 2 PCs we maintain

```
##            PC1        PC2
## x1 -0.0276432  0.8303372
## x2 -0.7365338  0.3547644
## x3 -0.4294145 -0.1990933
## x4 -0.5218784 -0.3808467
```

v) For PC1, the x2 coefficient is very large, and x1 very small (the other two are somewhat similar and significant). Thus, we can see that PC1 is heavily dependent on x2 (length of elytra) and still dependent on the other 2 variables. However, it is not dependent on x1(distance of transverse groove). We can see that for PC2, the coefficient for x1(distance of transverse groove) is very large and that for x3 (length of second antennal joint) relatively small. Thus, PC2 is heavily dependent on x1 (length of third antennal joint) and not dependent of x3 and not dependent on x3.

vi)



There is very little to note from this scatterplot as it is relatively well dispersed. However, we note that there are no apparent dependencies, nor outliers.

E) Overall, the 3 PCAs yielded varying results. However, all 3 suggested that the first component is highly dependent on x2 (length of elytra). In contrast, all 3 suggested different non-dependent variables for P1. For P2, both groups were heavily dependent on x4; however, the combined group was dependent on x1. Similarly, in P2, both groups were not dependent on x1, but in the combined group, x3 was the non-dependent variable. We can infer that from the matrix scatterplot from the combined group done in the initial investigation (showing separation of variables by group) and the change in the non-dependent variables in both principal components when combined compared to when separate as the principal components explain variation and as we saw in the initial investigation matrix scatterplot, the combined group often suggested stronger and weaker linear dependencies between different variables. While the two groups have similar variance in traits within themselves, the shift in dependencies in the principal components suggest that speciation may have occurred from disruptive equilibrium pressure. They share the same genus and are different species, suggesting a close common ancestor (or that one came from the other). The first component's large dependent factor remains the same throughout, suggesting that the greatest variation can be explained from the distance of the transverse groove.

e. The second principle component y2 for all of the data has the same dependency as the second group. -2

-2

Q2

2)F) (unlisted but named as code required) Code for 2A)

```
library(readxl)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
oleracea<- read_excel("fleabeetledata.xlsx",range="B3:E22",col_names=TRUE,skip=2)%>%
  mutate(Group=1)

carduorum <- read_excel("fleabeetledata.xlsx",range=c("F3:I23"),col_names=TRUE,skip=2)%>%
  mutate(Group=2)
```

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

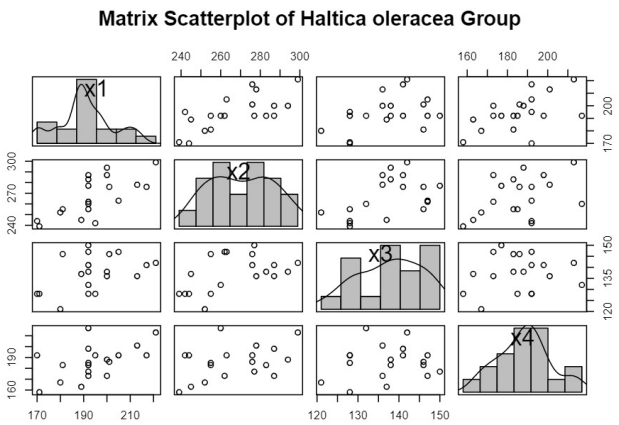
```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.3
```

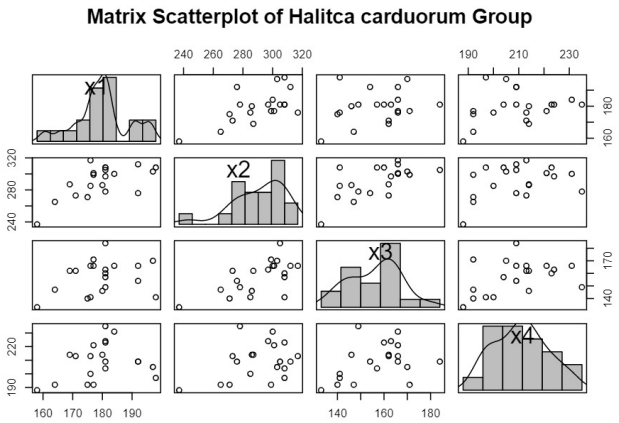
```
library(MESS)
```

```
## Warning: package 'MESS' was built under R version 4.1.3
```

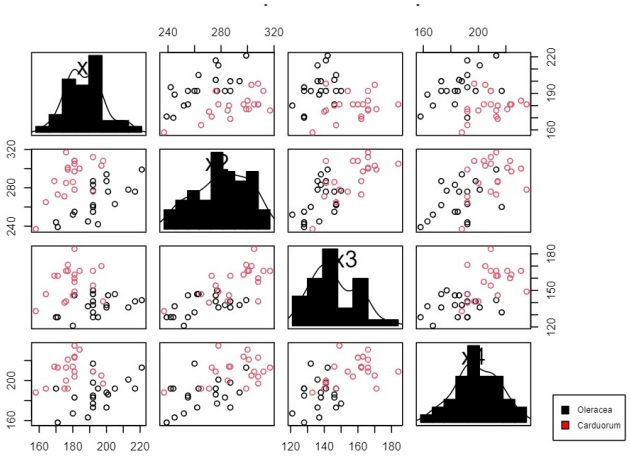
```
pairs(oleracea[,c(1:4)], diag.panel=panel.hist,main= "Matrix Scatterplot of Haltica oleracea Group")
```



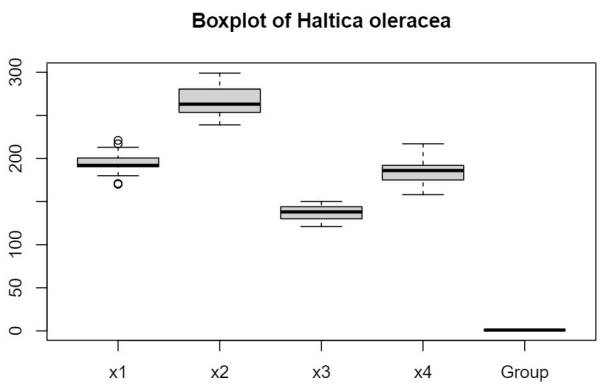

```
library(psych)
library(MESS)
pairs(carduorum[,c(1:4)], diag.panel=panel.hist,main="Matrix Scatterplot of Halitca carduorum Group")
```



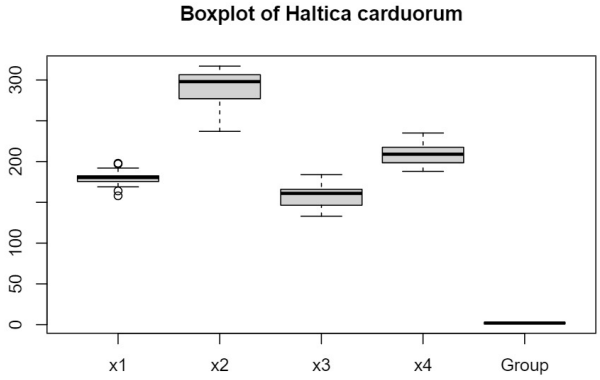

```
library(psych)
library(MESS)

data$Group<-as.factor(data$Group)
plot(data[,c(1:4)],col=data$Group,oma=c(3,3,3,9),pch=21, diag.panel=panel.hist,main= "Combined Group Ma
par(xpd=T)
legend("bottomright", fill=c("black","red"),legend=c("Oleracea","Carduorum"), cex=0.5)
```



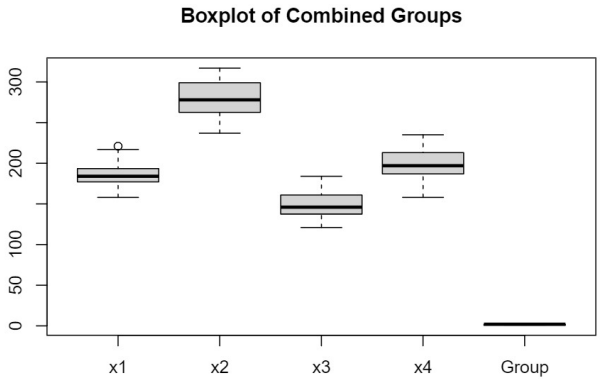

```
boxplot(oleracea,main="Boxplot of Haltica oleracea")
```




```
boxplot(carduorum,main="Boxplot of Haltica carduorum")
```




```
boxplot(data,main="Boxplot of Combined Groups")
```



Code for 2b)

```
S <- cov(oleracea[,1:4])
S
```

```
##           x1        x2       x3        x4
## x1 187.59649 176.86257 48.37135 113.58187
## x2 176.86257 345.38596 75.97953 118.78070
## x3  48.37135  75.97953 66.35673  16.24269
## x4 113.58187 118.78070 16.24269 239.94152
```

```
pca=prcomp(oleracea[1:4],scale=FALSE,center=TRUE)
eigen<- (pca$sdev)^2
eigen
```

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

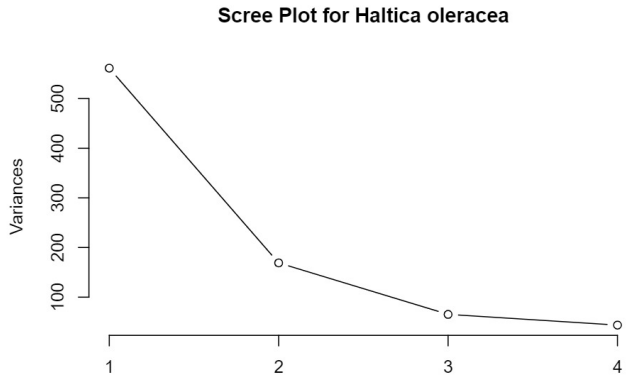
```
pvar=100*(pca$sdev)^2/sum((pca$sdev)^2)
pvar
```

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

```
pvaravg=mean(100*(pca$sdev)^2/sum((pca$sdev)^2))
pvaravg
```

```
## [1] 25
```

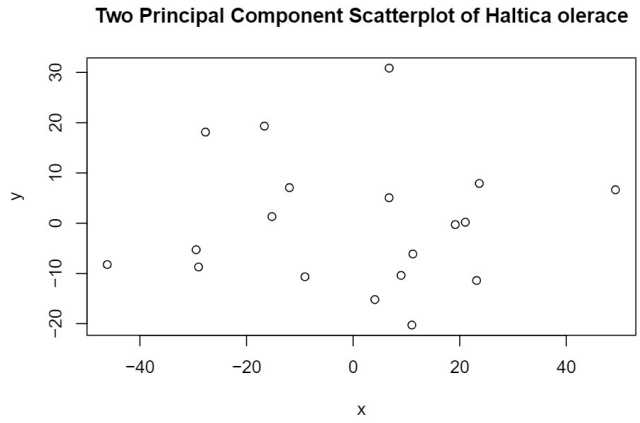
```
screeplot(pca,type="lines",main="Scree Plot for Haltica oleracea")
```




```
principalcomp<- pca$rotation[,1:2]
principalcomp
```

```
##          PC1          PC2
## x1 0.4997445  0.009204574
## x2 0.7187015 -0.484408702
## x3 0.1739702 -0.220296505
## x4 0.4510631  0.846600812
```

```
pc1<-pca$x[,1]#pc1
pc2<-pca$x[,2]#pc2
x=pc1
y=pc2
plot(x,y,main="Two Principal Component Scatterplot of Haltica olerace")
```



Code for 2C)

```
S2 <- cov(carduorum[,1:4])
S2
```

```
##           x1        x2        x3        x4
## x1 101.83947 128.06316  36.98947  32.59211
## x2 128.06316 389.01053 165.35789  94.36842
## x3  36.98947 165.35789 167.53684  66.52632
## x4  32.59211  94.36842  66.52632 177.88158
```

```
pca2=prcomp(carduorum[1:4],scale=FALSE,center=TRUE)
eigen2<- (pca2$sdev)^2
eigen2
```

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

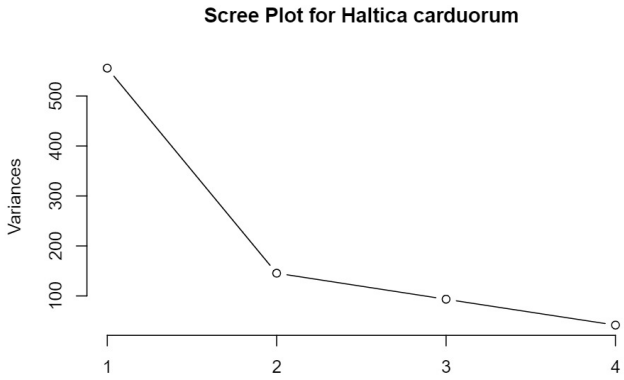
```
pvar2=100*(pca2$sdev)^2/sum((pca2$sdev)^2)
pvar2
```

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

```
pvaravg2=mean(100*(pca2$sdev)^2/sum((pca2$sdev)^2))
pvaravg2
```

```
## [1] 25
```

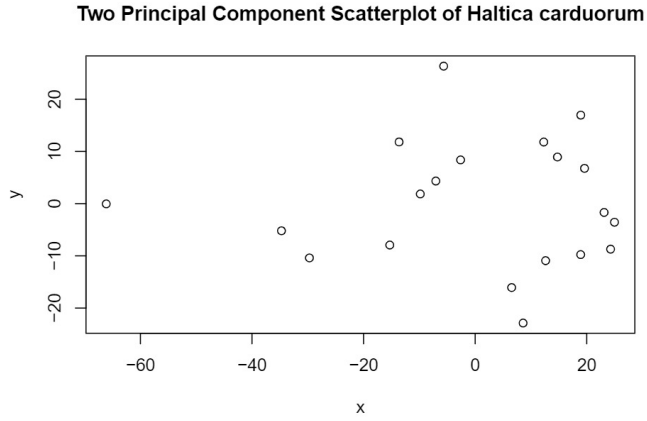
```
screeplot(pca2,type="lines",main="Scree Plot for Haltica carduorum")
```




```
principalcomp2<- pca2$rotation[,1:2]
principalcomp2
```

```
##          PC1        PC2
## x1 0.2836552 -0.2007357
## x2 0.8068689 -0.3389760
## x3 0.4222422  0.1359900
## x4 0.3003563  0.9090144
```

```
pc3<-pca2$x[,1]#pc1
pc4<-pca2$x[,2]#pc2
x=pc3
y=pc4
plot(x,y,main="Two Principal Component Scatterplot of Haltica carduorum")
```



Code for 2D)

```
S3 <- cov(data[,1:4])
S3
```

```
##           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
```

```
pca3=prcomp(data[1:4],scale=FALSE,center=TRUE)
eigen3<- (pca3$sdev)^2
eigen3
```

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

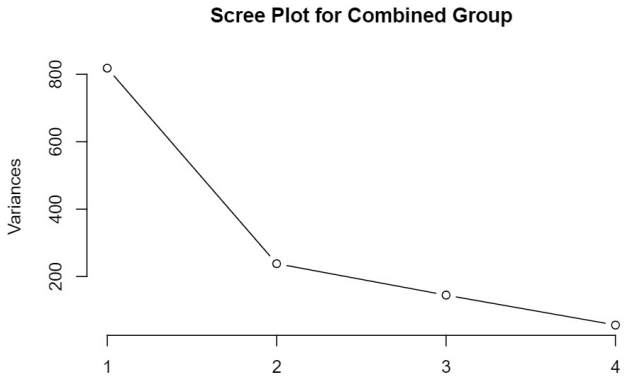
```
pvar3=100*(pca3$sdev)^2/sum((pca3$sdev)^2)
pvar3
```

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

```
pvaravg3=mean(100*(pca3$sdev)^2/sum((pca3$sdev)^2))
pvaravg3
```

```
## [1] 25
```

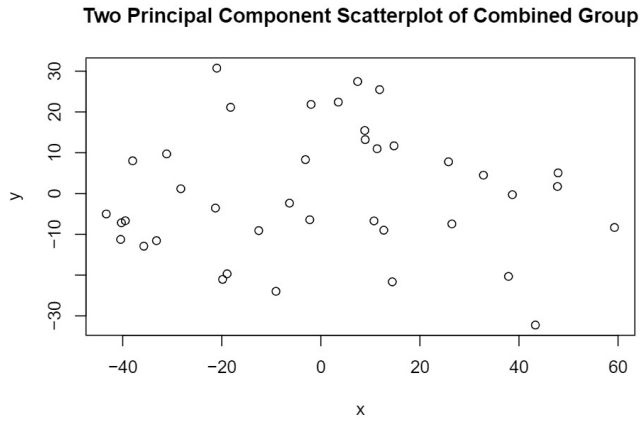
```
screeplot(pca3,type="lines",main="Scree Plot for Combined Group")
```




```
principalcomp3<- pca3$rotation[,1:2]
principalcomp3
```

```
##           PC1        PC2
## x1 -0.0276432  0.8303372
## x2 -0.7365338  0.3547644
## x3 -0.4294145 -0.1990933
## x4 -0.5218784 -0.3808467
```

```
pc5<-pca3$x[,1]#pc1
pc6<-pca3$x[,2]#pc2
x=pc5
y=pc6
plot(x,y,main="Two Principal Component Scatterplot of Combined Group")
```


Two Principal Component Scatterplot of Combined Group
y
x
30
20
10
0
−10
−30
−40
−20
0
20
40
60

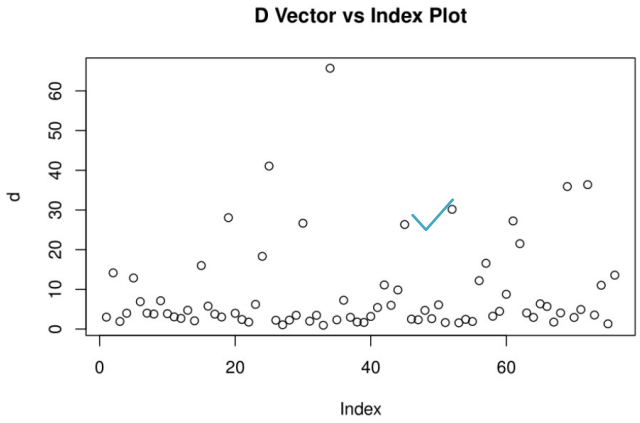
13

3)
A) Done – Reference Code

B)

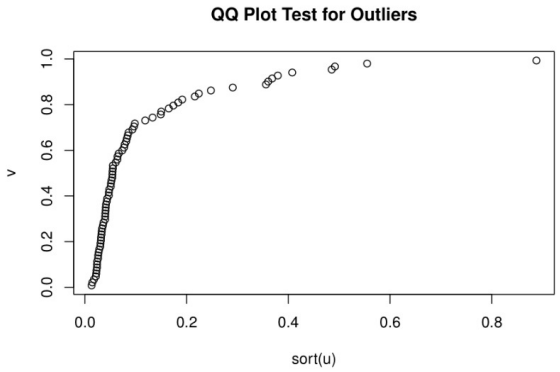
i) We calculate Vector D and Graph it vs Index below

```
##  [1]  3.0079 14.1601  1.9184  3.9998 12.8689  6.9069  4.0184  3.7855  7.1387
## [10]  3.8688  3.0687  2.6988  4.7409  2.0824 15.9957  5.7844  3.7633  3.0341
## [19] 28.0707  3.9822  2.4127  1.7518  6.2184 18.3308 41.0702  2.2351  1.0779
## [28]  2.2623  3.4780 26.6744  1.9923  3.4740  0.9682 65.7133  2.3252  7.2624
## [37]  2.9656  1.7749  1.6750  3.1795  5.4254 11.1079  5.9916  9.8501 26.3452
## [46]  2.4995  2.3426  4.7184  2.6145  6.0975  1.6448 30.1726  1.5845  2.4426
## [55]  1.9109 12.1924 16.5667  3.2427  4.4718  8.7679 27.2435 21.5167  4.0658
## [64]  2.9407  6.3572  5.6947  1.7569  4.0698 35.9143  2.9333  4.9218 36.3801
## [73]  3.5395 11.0358  1.3209 13.5826
```



As we can see, the data in the rows 25,34,52,57,62,69, and 72 have the highest D values and are visually separated from the rest of the data.

We use QQ Plot Test for Outliers Below:



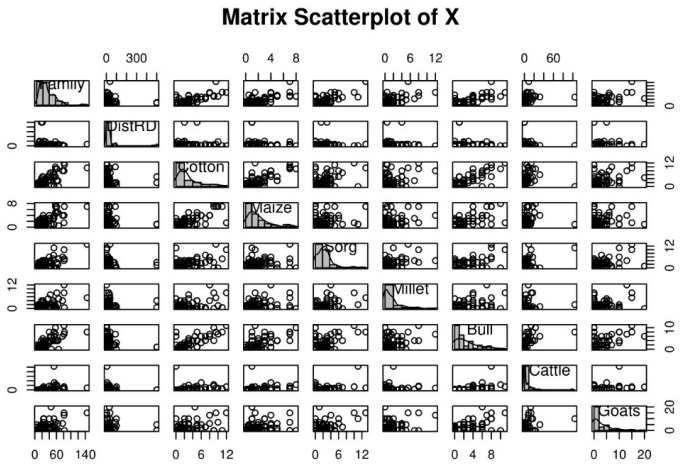
The QQ Plot Test for Outliers suggests anywhere from 5 to 9 outliers (depending on interpretation). 7 Outliers fit well here. Thus, the two indicators support the claim.

ii)

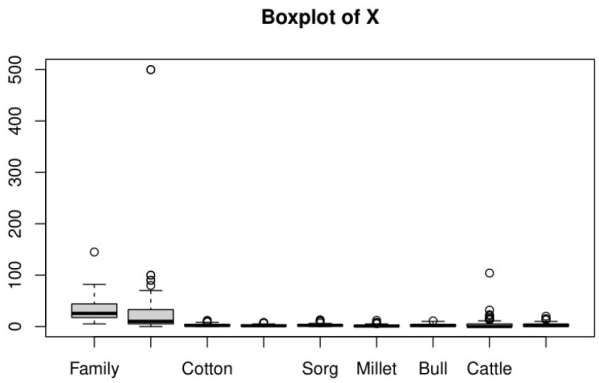
There is no treatment done to the data. The different variables in the data measure different properties, such as number of people and hectares planted, which means that the different variables are using different units. There are no apparent groups.

The PCA expects us to center the data around the sample mean, so while we did not subtract the sample mean from the data as part of the initial assessment, we do so later in the PCA.

Below are the matrix scatterplots. They can be read as the independent variable being that belonging to the variable associated with the column and the dependent variable belonging to the variable associated with the row.



The histograms in the matrix scatterplot for X shows that the data is largely heavily skewed and certainly, not normally distributed, greatly suggesting the existence of outliers for all variables. Most of the scatterplots have points greatly concentrated in a single cluster (such as the scatterplot between family and DistRD). There are very weak indications of potentially weak positive linear dependencies in a few scatterplots such as between Cotton and Bull and Family and Bull. The scatterplots are certainly not randomly dispersed.



The boxplot shows evidence that there are likely several extreme outliers. Between the outliers and the large variance in range between the different variables (as they are measuring different units), it is apparent that we may have scale issues.

C) Done in Code

D)

i) Sample Covariance Matrix of X

```
##            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
```

ii) List of eigenvalues

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

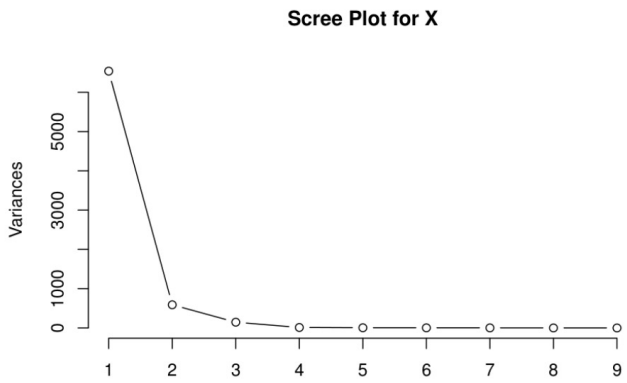
Percent contributions to Variance:

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

iii) The number of principal components to retain is 1. We can see that this is reached by using the 3 methods.

The first method is the cumulative proportion threshold, which we arbitrarily set to 80%. We can see that the first value surpasses this threshold alone.

The 2nd method is the average proportional eigenvalue method. The average proportional eigenvalue is 11.11111%. The 2nd principal component under this method is a below this level and there are many eigenvalues. The 3rd method is the scree plot.



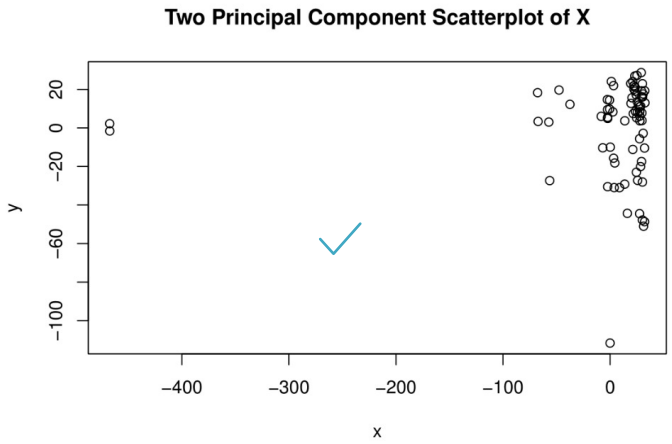
The scree plot clearly suggests 1 principal component as the bend shifts heavily on point 2. Thus, this is clear evidence for 1 PC.

iv) The eigenvectors for the principal component

```
##         Family        DistRD        Cotton         Maize          Sorg
##  0.0267175675 -0.9995725665 -0.0007739567  0.0013694250  0.0022466401
##         Millet          Bull        Cattle         Goats
## -0.0004899182  0.0031275792  0.0110210195 -0.0013599578
```

v) The coefficient for DistRD is very large and the others small. Thus, PC1 is dependent on DistRD and not dependent on the other variables.

vi) Only the first principal component is significant, but to do the scatterplot, we use the two PCs.



Most of the data is clustered/concentrated together on the right (somewhat topright); however, there are three notable potential outliers on the scatterplots – two PC1 and one PC2 values.

E)

i) Sample Covariance Matrix of X with outliers removed

```
##            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
```

ii)

Eigenvalues of X with outliers removed

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

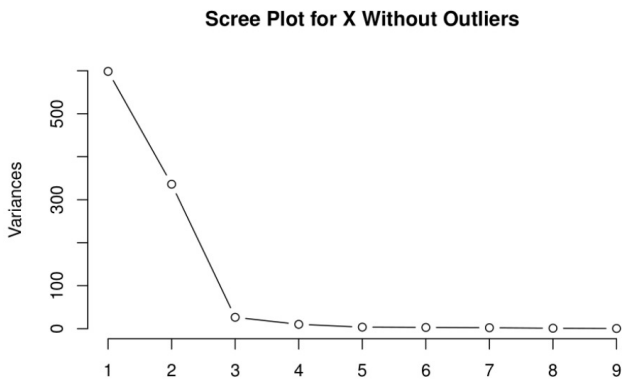
Percent Contributions to Variance

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

iii) There are 2 principal components. We can see that this is reached by using the 3 methods.

The first method is the cumulative proportion threshold, which we arbitrarily set to 80%. We can see that only after 2 principal components does the value surpasses this threshold.

The 2nd method is the average proportional eigenvalue method. The average proportional eigenvalue is 11.11111%. The 2nd principal component is above this level and the remainders are significantly lower. The 3rd method is the scree plot.



The scree plot clearly suggests 2 principal components as the bend shifts heavily on point 3.

Thus, we see clear evidence for 2 PCs.

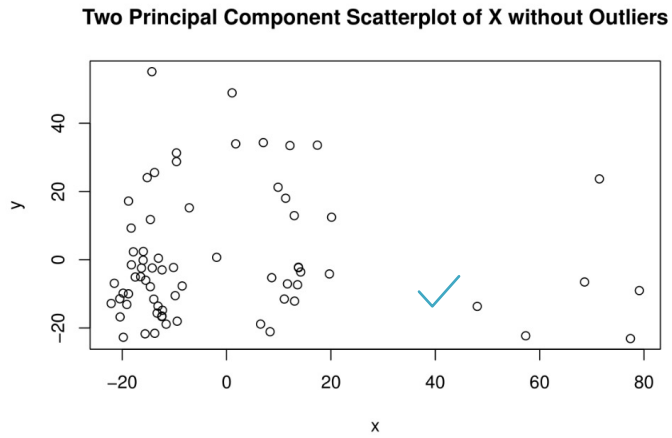
iv) The eigenvectors for the retained PCs.

```
##                PC1         PC2
## Family  0.074032153  0.96661143
## DistRD  0.995431316 -0.08420571
## Cotton  0.010609767  0.08582783
## Maize   0.015413007  0.05552771
## Sorg   -0.012613666  0.02602851
## Millet -0.003474341  0.05913095
## Bull    0.016290722  0.07775609
## Cattle  0.040130538  0.18151903
## Goats   0.035193537  0.07037673
```

v)

For PC1, the coefficient for DistRD is very large and the rest small. For PC2, the coefficient for Family is very large and the rest are small. Thus, PC1 is dependent only on DistRD and PC2 on Family.

vi)



Most of the data is somewhat clustered/concentrated together on the left (somewhat bottom-left); however, there are several notable potential outliers on the scatterplots – both for PC1 and PC2. However, the outliers are less apparent than they were in the original PC1 PC2 scatterplot (without outliers removed).

F)

The more centralized histograms on the PC matrix scatterplot without the outliers suggests lower covariance, supporting that dropping the outliers is beneficial. We also note that we went from 1 to 2 PCs by dropping the outliers, and consequently a larger aggregate percentage explanation for the variance. We went from just under 90% to over 95%. Thus, because of these improvements, we prefer the result without the outliers.

f. The dependency of the first two principal components on the original random variables did not change significantly when the outliers were removed. **-1**

Q3

3)G)

Code for 3a)

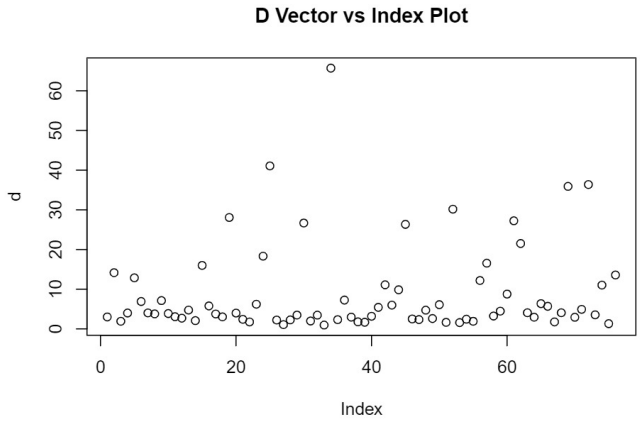
```
library(readxl)
X<-read_excel("malifarmdata.xlsx",col_names=TRUE)
```

3b)

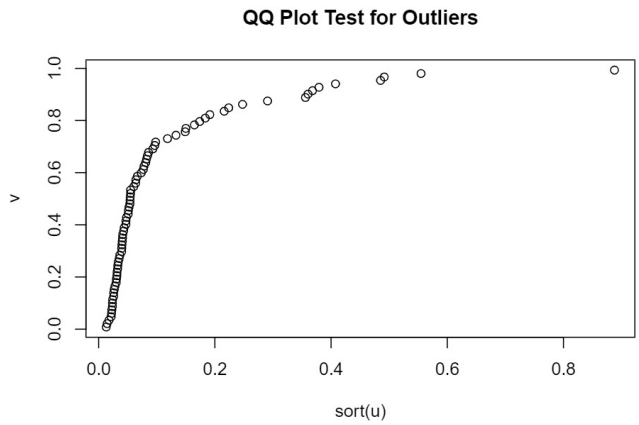
```
x_bar <- as.numeric(colMeans(X))
S<-cov(X)
d<-mahalanobis(X,center=x_bar,cov=S)
round(d,4)
```

```
##  [1]  3.0079 14.1601  1.9184  3.9998 12.8689  6.9069  4.0184  3.7855  7.1387
## [10]  3.8688  3.0687  2.6988  4.7409  2.0824 15.9957  5.7844  3.7633  3.0341
## [19] 28.0707  3.9822  2.4127  1.7518  6.2184 18.3308 41.0702  2.2351  1.0779
## [28]  2.2623  3.4780 26.6744  1.9923  3.4740  0.9682 65.7133  2.3252  7.2624
## [37]  2.9656  1.7749  1.6750  3.1795  5.4254 11.1079  5.9916  9.8501 26.3452
## [46]  2.4995  2.3426  4.7184  2.6145  6.0975  1.6448 30.1726  1.5845  2.4426
## [55]  1.9109 12.1924 16.5667  3.2427  4.4718  8.7679 27.2435 21.5167  4.0658
## [64]  2.9407  6.3572  5.6947  1.7569  4.0698 35.9143  2.9333  4.9218 36.3801
## [73]  3.5395 11.0358  1.3209 13.5826
```

```
plot(d,main="D Vector vs Index Plot")
```




```
#QQ plot
n=nrow(X)
u=n/(n-1)^2*d
p=ncol(X)
alpha=(p-2)/(2*p)
beta=(n-p-3)/(2*(n-p-1))
v=c()
for(i in 1:n) {
  v[i] = (i-alpha)/(n-alpha-beta+1)
}

plot(x=sort(u),y=v, main="QQ Plot Test for Outliers")
```



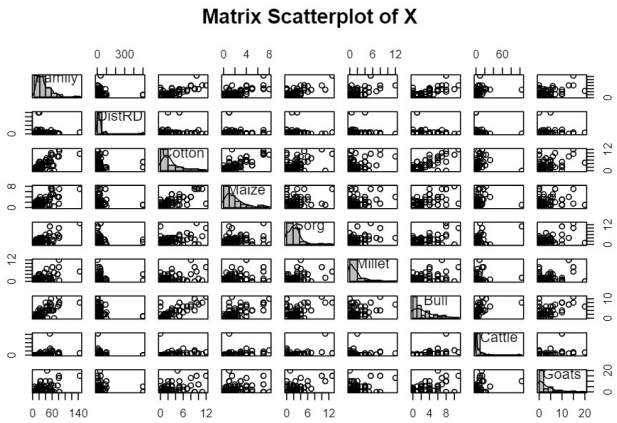

```
library(psych)
```

```
## Warning: package 'psych' was built under R version 4.1.3
```

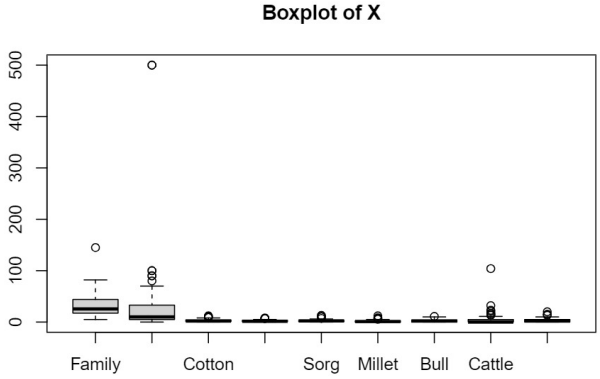
```
library(MESS)
```

```
## Warning: package 'MESS' was built under R version 4.1.3
```

```
pairs(X[,c(1:9)], diag.panel=panel.hist,main="Matrix Scatterplot of X")
```




```
boxplot(X,main="Boxplot of X")
```



Code for 3C)

```
Xed<- X[-25,]
Xed<- Xed[-33,]
Xed<-Xed[-50,]
Xed<-Xed[-54,]
Xed<-Xed[-58,]
Xed<-Xed[-64,]
Xed<-Xed[-66,]
head(Xed)
```

```
## # A tibble: 6 x 9
##   Family DistRD Cotton Maize  Sorg Millet  Bull Cattle Goats
##    <dbl>  <dbl>  <dbl> <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>
## 1     12     80    1.5   1       3   0.25     2      0     1
## 2     54      8    6     4       0   1        6     32     5
## 3     11     13    0.5   1       0   0        0      0     0
## 4     21     13    2     2.5     1   0        1      0     5
## 5     61     30    3     5       0   0        4     21     0
## 6     20     70    0     2       3   0        2      0     3
```

Code for 3D)

```
S <- cov(X)
S
```

```
##            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
```

```
pca=prcomp(X[1:9],scale=FALSE,center=TRUE)
eigen<- (pca$sdev)^2
eigen
```

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

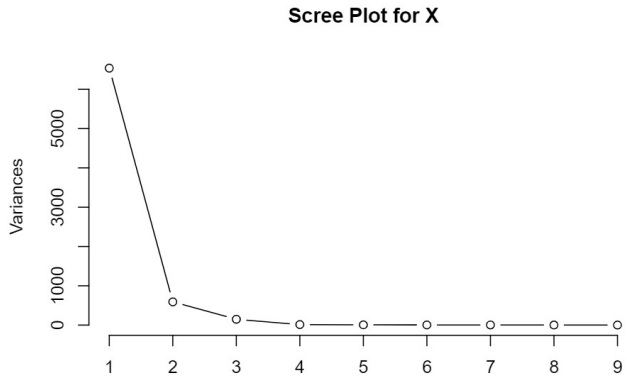
```
pvar=100*(pca$sdev)^2/sum((pca$sdev)^2)
pvar
```

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

```
pvaravg=mean(100*(pca$sdev)^2/sum((pca$sdev)^2))
pvaravg
```

```
## [1] 11.11111
```

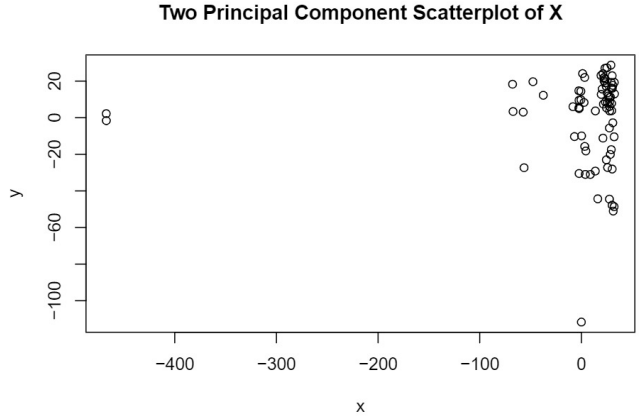
```
screeplot(pca,type="lines",main="Scree Plot for X")
```




```
principalcomp<- pca$rotation[,1]
principalcomp
```

```
##        Family        DistRD        Cotton         Maize          Sorg
##  0.0267175675 -0.9995725665 -0.0007739567  0.0013694250  0.0022466401
##        Millet          Bull        Cattle         Goats
## -0.0004899182  0.0031275792  0.0110210195 -0.0013599578
```

```
pc1<-pca$x[,1]#pc1
pc2<-pca$x[,2]#pc2
x=pc1
y=pc2
plot(x,y,main="Two Principal Component Scatterplot of X")
```



Code for 3)E)

```
Sed <- cov(Xed)
Sed
```

```
##            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
```

```
pca2=prcomp(Xed[1:9],scale=FALSE,center=TRUE)
eigen2<- (pca2$sdev)^2
eigen2
```

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

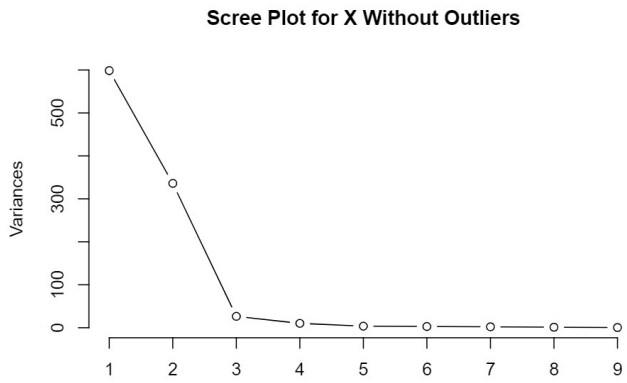
```
pvar2=100*(pca2$sdev)^2/sum((pca2$sdev)^2)
pvar2
```

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

```
pvaravg2=mean(100*(pca2$sdev)^2/sum((pca2$sdev)^2))
pvaravg2
```

```
## [1] 11.11111
```

```
screeplot(pca2,type="lines",main="Scree Plot for X Without Outliers")
```




```
principalcomp2<- pca2$rotation[,1:2]
principalcomp2
```

```
##                PC1         PC2
## Family  0.074032153  0.96661143
## DistRD  0.995431316 -0.08420571
## Cotton  0.010609767  0.08582783
## Maize   0.015413007  0.05552771
## Sorg   -0.012613666  0.02602851
## Millet -0.003474341  0.05913095
## Bull    0.016290722  0.07775609
## Cattle  0.040130538  0.18151903
## Goats   0.035193537  0.07037673
```

```
pc3<-pca2$x[,1]#pc1
pc4<-pca2$x[,2]#pc2
x=pc3
y=pc4
plot(x,y,main="Two Principal Component Scatterplot of X without Outliers")
```