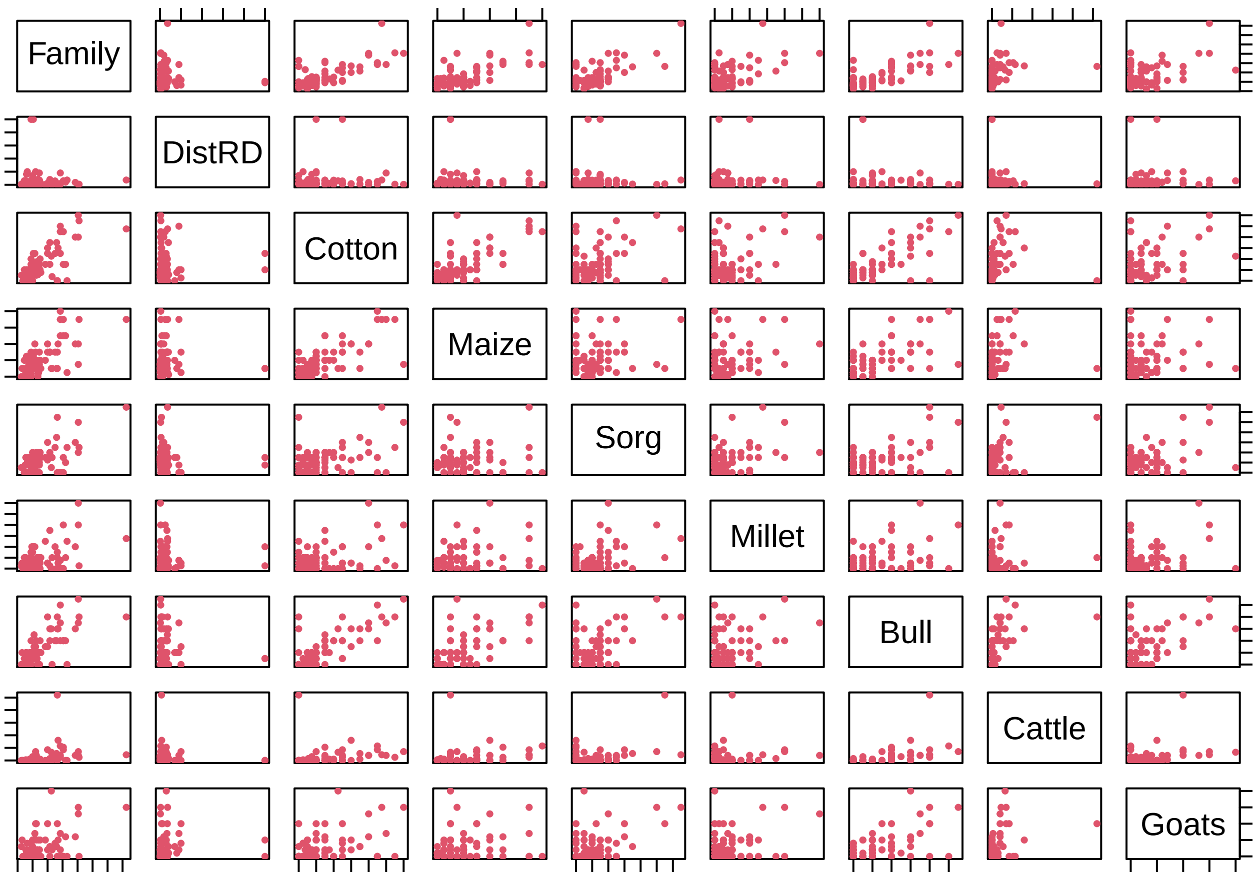
A5P3

Dawu Liu

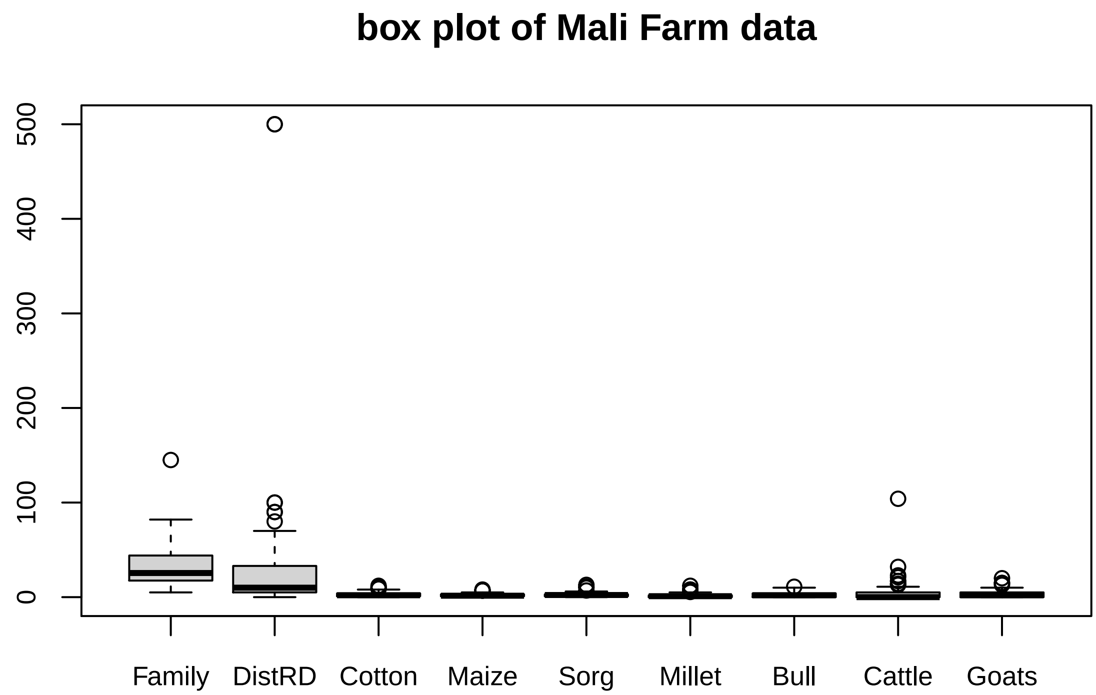
In this assignment, I will write principal component as **PC** sometimes for short.

(a)

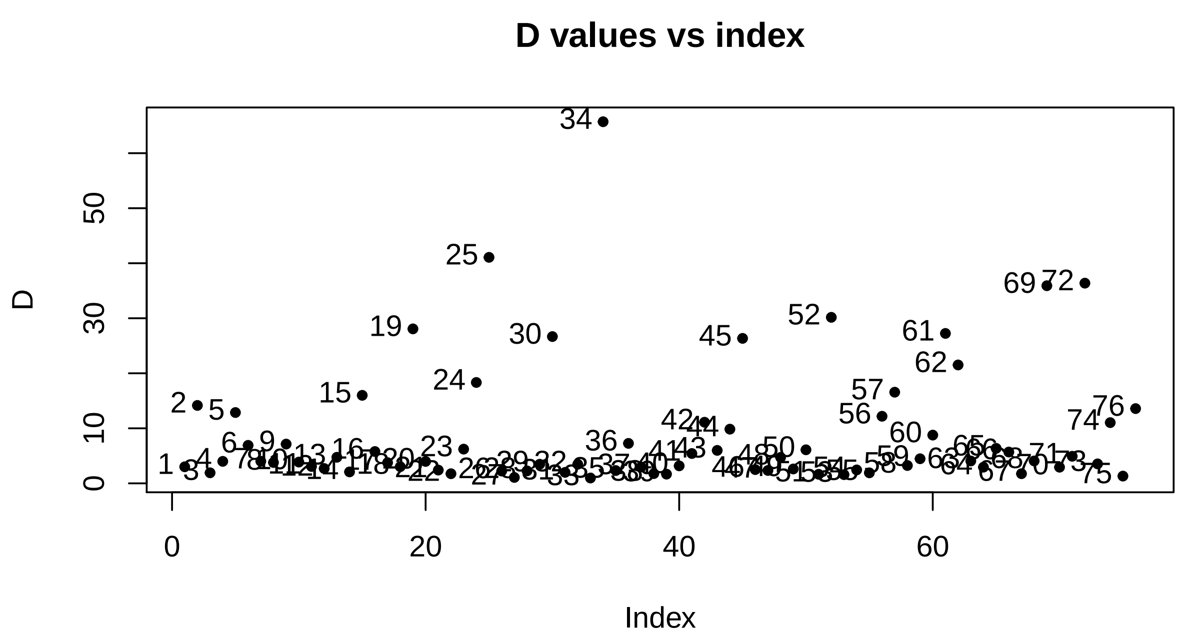
i. Matrix scatter plot of the data



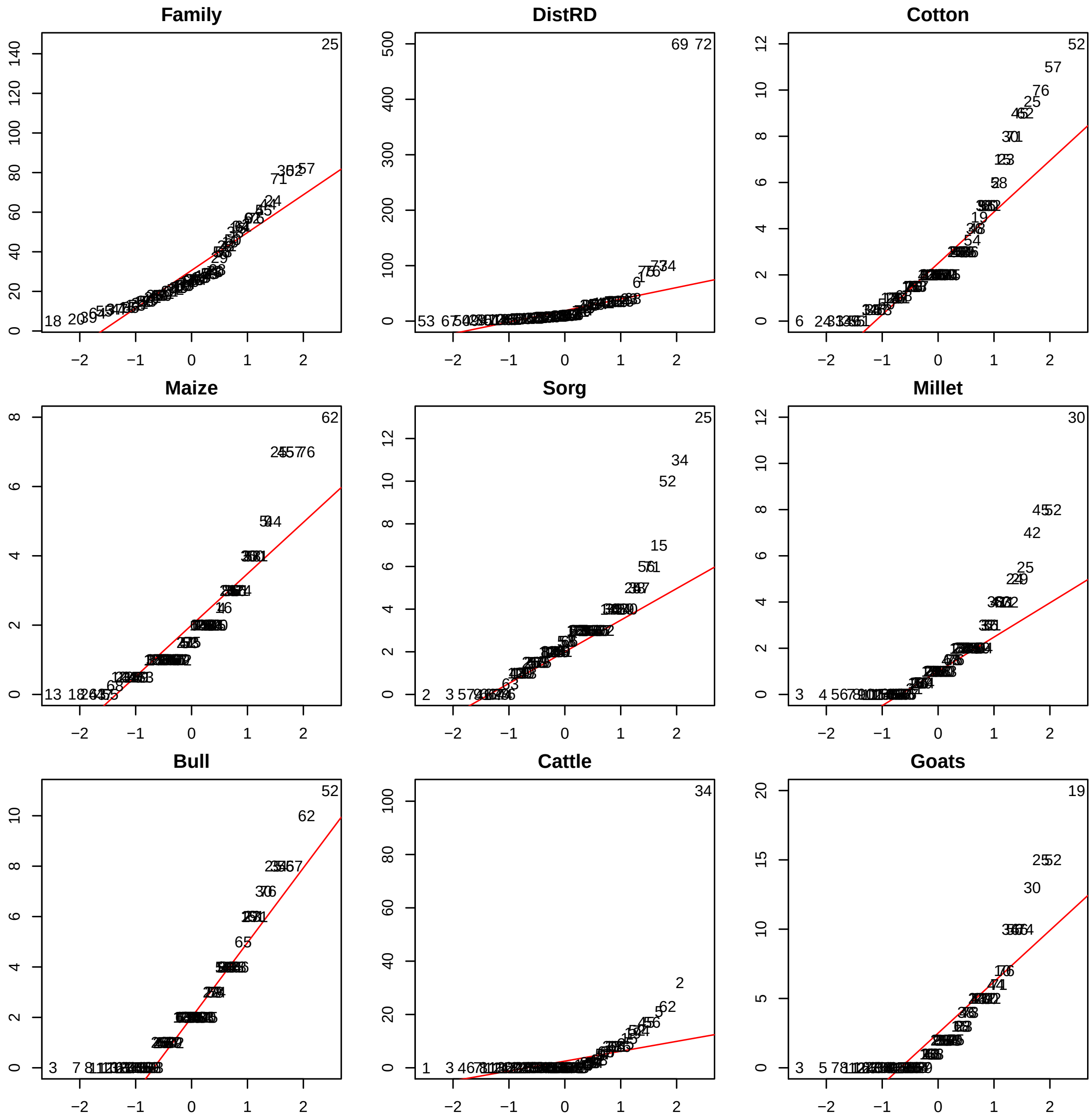
ii. Boxplot of the data



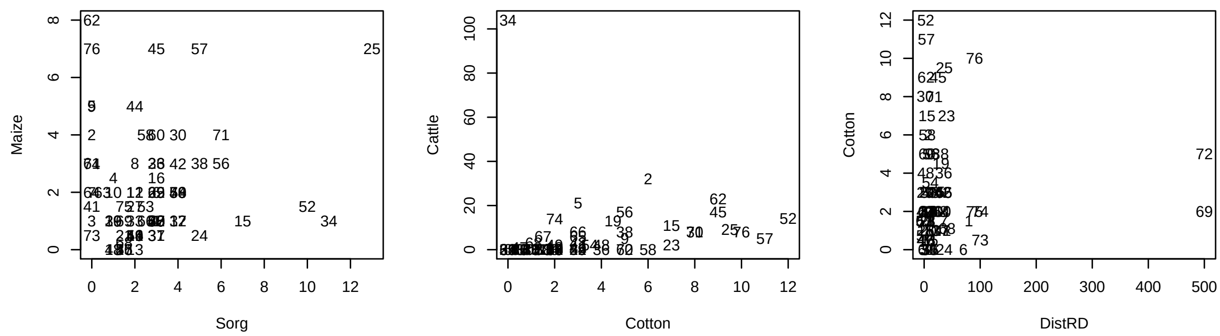
iii.



iv. I added qq-plot to help detect the outliers.



Also three scatter plots:



The outliers are 25, 34, 52, 57, 62, 69, 72.

From the three scatter plots above, Maize vs Sorg indicates 25; Cattle vs Cotton indicates 34; and Cotton vs DistRD indicates 69, 72 are the outliers.

All 7 outliers have relatively high values in the distance values plot.

In both the Boxplot and Q-Q plot, Sorg indicates 25; Cattle indicates 34; Cotton indicates 52, 57; Maize indicates 62; and DistRD indicates 69 and 72 are the outliers.

The table below shows which indicator indicates the outliers.

```
##                 25  34  52  57  62  69  72
## Scatter Plots yes yes             yes yes
## Distance Plot yes yes yes yes yes yes yes
## Box Plot      yes yes yes yes yes yes yes
## Q-Q Plots     yes yes yes yes yes yes yes
```

v. Create a data matrix $\tilde{X}$ by removing the outliers.

New data set $\tilde{X}$ is created by removing row 25,34,52,57,62,69,72 and will be used in the second PCA of part (b).

analyses starts on the next page…

(b)

PCA on the original data set X

i. Sample covariance matrix **S**

```
##          Family   DistRD Cotton  Maize    Sorg Millet    Bull  Cattle  Goats
## Family  550.876 -158.768 48.117 29.539  31.837 26.393  45.458 103.754 46.810
## DistRD -158.768 6533.751  6.436 -8.105 -13.692  3.941 -19.025 -67.355 10.363
## Cotton   48.117    6.436  8.012  3.832   2.585  2.446   5.763   6.504  4.654
## Maize    29.539   -8.105  3.832  3.434   0.481  0.894   3.074   4.809  1.042
## Sorg     31.837  -13.692  2.585  0.481   5.700  2.029   2.816  12.699  4.171
## Millet   26.393    3.941  2.446  0.894   2.029  4.942   2.091   2.366  2.801
## Bull     45.458  -19.025  5.763  3.074   2.816  2.091   7.089  18.206  6.150
## Cattle  103.754  -67.355  6.504  4.809  12.699  2.366  18.206 173.081 19.364
## Goats    46.810   10.363  4.654  1.042   4.171  2.801   6.150  19.364 17.013
```

ii. The eigenvalues are:

```
## [1] 6538.8594  590.1075  147.5506   12.7110    5.8905    3.9910    2.9594
## [8]    1.1372    0.6913
```

The first eigenvalue accounts for the majority of the total variance. PC4 to PC9 have very little on the proportion of the total variance. (table shown in iii)

iii.

Criteria 1, eigenvalues and their cumulative proportions table

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  6538.8594          89.5256                     89.5256
## Dim.2   590.1075           8.0794                     97.6050
## Dim.3   147.5506           2.0202                     99.6251
## Dim.4    12.7110           0.1740                     99.7992
## Dim.5     5.8905           0.0806                     99.8798
## Dim.6     3.9910           0.0546                     99.9344
## Dim.7     2.9594           0.0405                     99.9750
## Dim.8     1.1372           0.0156                     99.9905
## Dim.9     0.6913           0.0095                    100.0000
```

This method suggests to keep **1** principal component which gives 89.5% of the total variance.

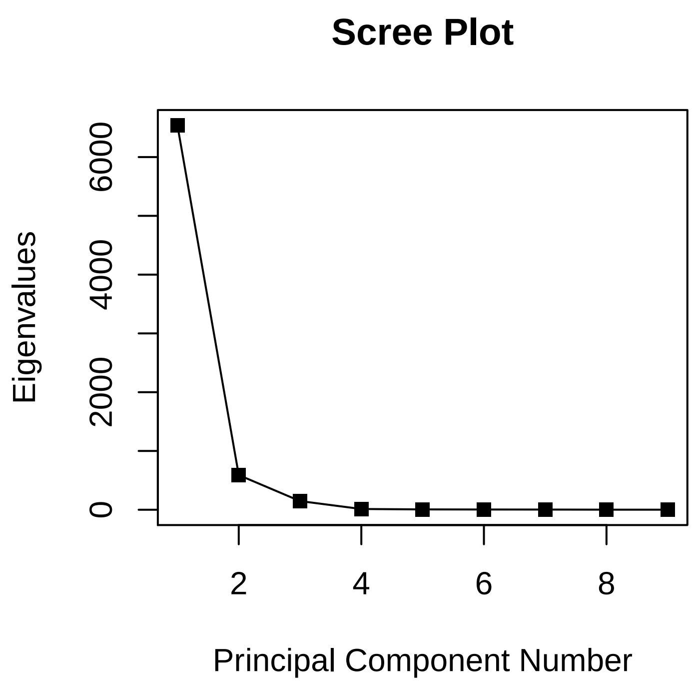
Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

The mean of the eigenvalues is:

```
## [1] 811.5442
```

This method suggests to keep **1** principal component.

Criteria3, scree plot



The "bend" occurs at PC2, indicating from PC2 and on, the the eigenvalues are relatively small. This method suggests to keep **1** principal component.

Overall, the criteria suggest **1** principal component should be retained.

But in order to make a scatter plot, we will use 2 principal components.

iv. The eigenvectors for the principal components:

$(\tilde{\hat{e}}_1)^T =$

```
##  Family  DistRD  Cotton   Maize    Sorg  Millet    Bull  Cattle   Goats
##  0.0267 -0.9996 -0.0008  0.0014  0.0022 -0.0005  0.0031  0.0110 -0.0014
```

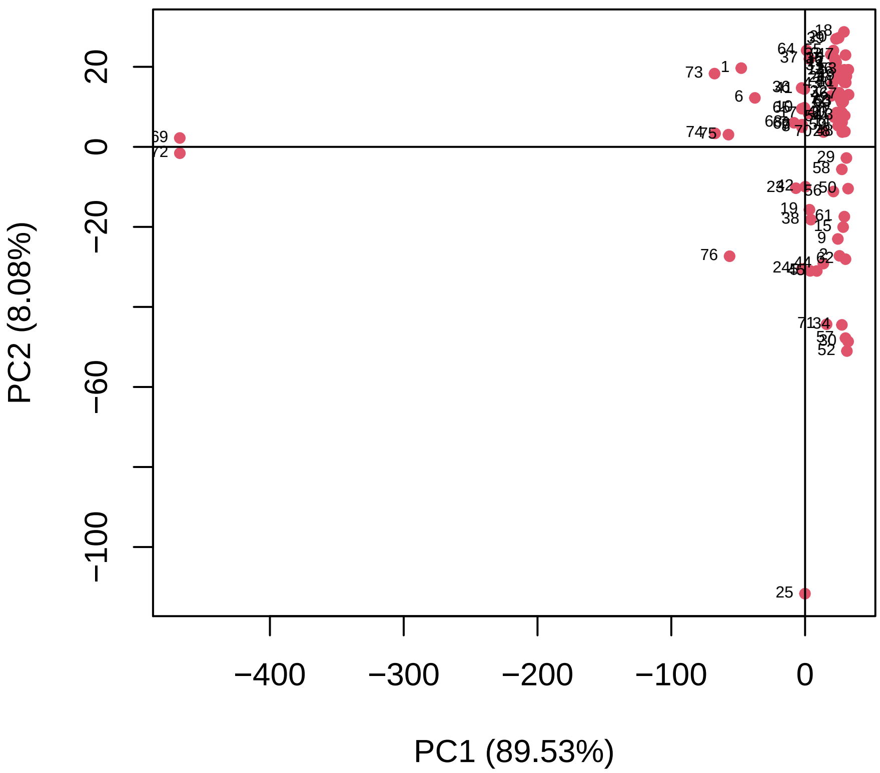
$(\tilde{\hat{e}}_2)^T =$

```
##  Family  DistRD  Cotton   Maize    Sorg  Millet    Bull  Cattle   Goats
## -0.9536 -0.0284 -0.0842 -0.0509 -0.0582 -0.0455 -0.0835 -0.2444 -0.0890
```

v.

PC1 almost entirely depends on DistRD, it's a component representing DistRD. PC2 almost entirely depends on Family, with some dependency on cattle. For both PCs, all the crops and farm animals have very little loads on them.

vi. Scatter plot for PC2 vs PC1



Observation number 69 and 72 on the far left have extreme small PC1 values compare to the rest of the data, means they are outliers with larges values in DistRD(negative coefficients). Based on this plot, it appears most of the observations are grouped together in terms of PC1 values, but this might be a scaling issue caused by the outliers.

PCA on the new data set $\tilde{X}$ where the outliers are removed

i. Sample covariance matrix **S**

```
##          Family  DistRD Cotton  Maize   Sorg Millet   Bull Cattle  Goats
## Family 318.587  16.958 27.505 18.451  8.021 19.311 25.023 55.753 22.746
## DistRD  16.958 595.620  3.786  7.585 -8.188 -3.644  7.347 17.977 18.546
## Cotton  27.505   3.786  5.179  2.629  1.038  1.656  3.543  7.581  3.065
## Maize   18.451   7.585  2.629  2.474 -0.038  0.995  2.005  4.570  0.892
## Sorg     8.021  -8.188  1.038 -0.038  2.554  0.827  0.657  0.137  0.405
## Millet  19.311  -3.644  1.656  0.995  0.827  4.467  1.446  1.321  0.969
## Bull    25.023   7.347  3.543  2.005  0.657  1.446  4.467  7.761  4.186
## Cattle  55.753  17.977  7.581  4.570  0.137  1.321  7.761 35.428  8.794
## Goats   22.746  18.546  3.065  0.892  0.405  0.969  4.186  8.794 13.209
```

ii. The eigenvalues are:

```
## [1] 598.6558 336.1485  26.5261  10.1536   3.6717   2.9183   2.2307   1.1355
## [9]   0.5448
```

The first and second eigenvalues account for the majority of the total variance. PC3 to PC9 have very little on the proportion of the total variance. (table shown in iii)

iii.

Criteria 1, eigenvalues and their cumulative proportions table

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   598.6558          60.9638                     60.9638
## Dim.2   336.1485          34.2315                     95.1954
## Dim.3    26.5261           2.7013                     97.8967
## Dim.4    10.1536           1.0340                     98.9306
## Dim.5     3.6717           0.3739                     99.3045
## Dim.6     2.9183           0.2972                     99.6017
## Dim.7     2.2307           0.2272                     99.8289
## Dim.8     1.1355           0.1156                     99.9445
## Dim.9     0.5448           0.0555                    100.0000
```

This method suggests to keep **2** principal component which gives 95.2% of the total variance.

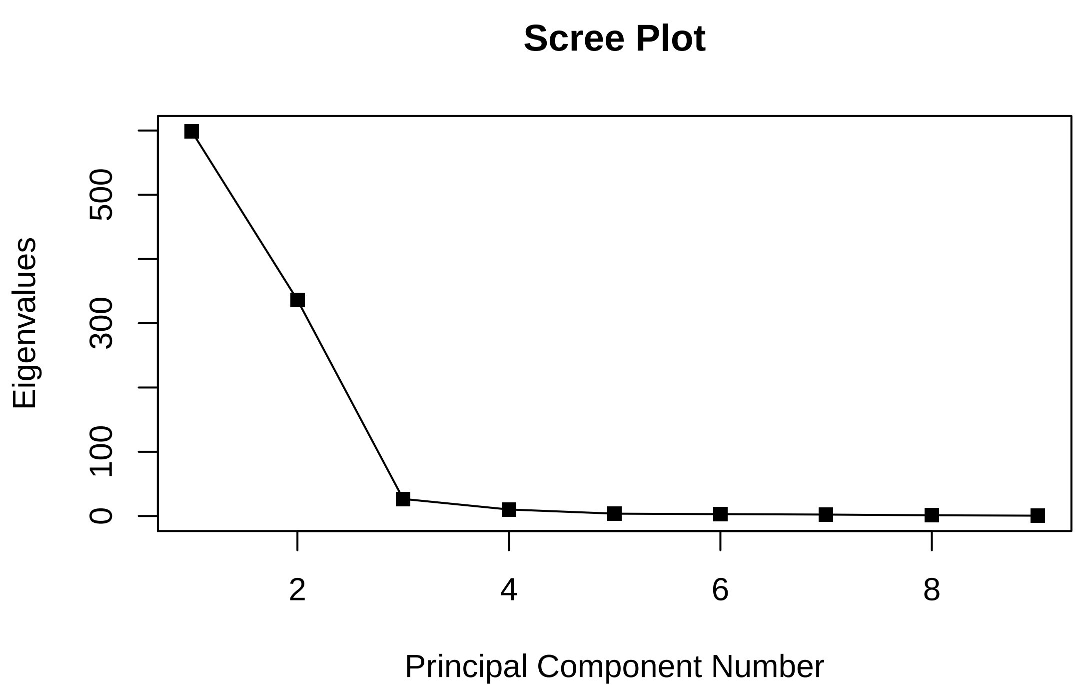
Criteria 2, check which eigenvalue(s) is greater than the mean of eigenvalues

The mean of the eigenvalues is:

```
## [1] 109.1094
```

This method suggests to keep **2** principal component.

Criteria3, scree plot



The "bend" occurs at PC3, indicating from PC3 and on, the the eigenvalues are relatively small. This method suggests to keep **2** principal component.

Overall, all three criteria suggest **2** principal components should be retained.

iv. The eigenvectors for the principal components:

$(\tilde{\hat{e_1}})^T =$

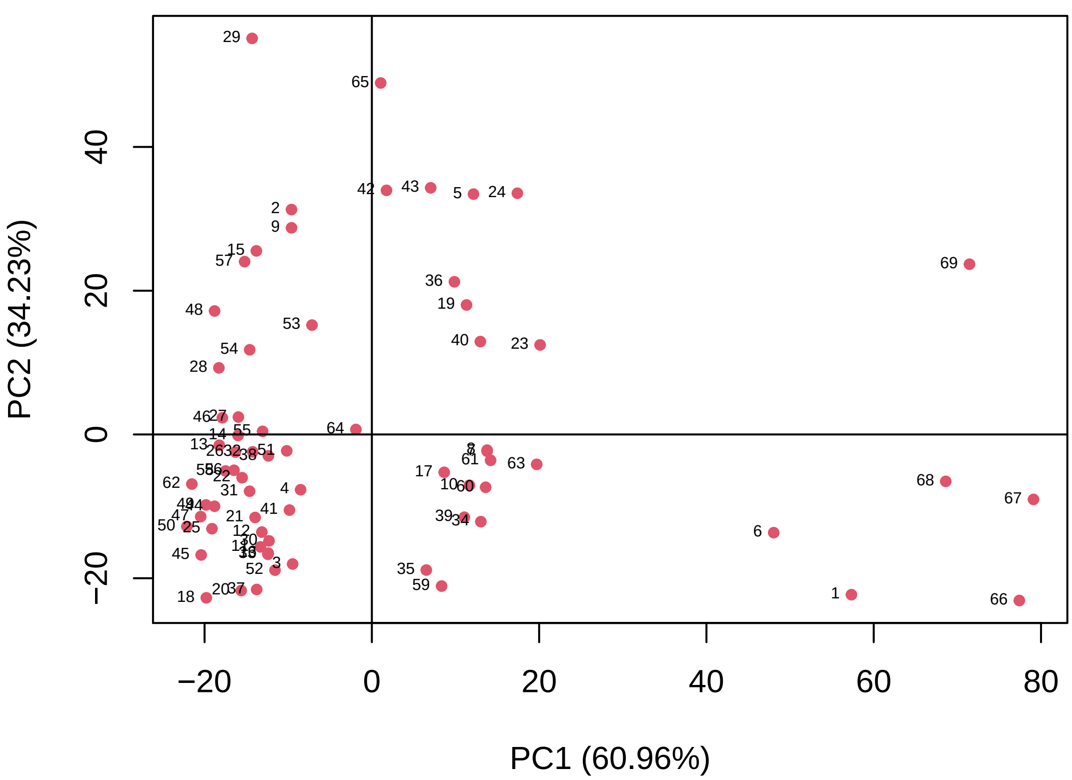
```
## Family DistRD  Cotton  Maize   Sorg    Millet  Bull   Cattle Goats
## 0.0740 0.9954  0.0106  0.0154  -0.0126 -0.0035 0.0163 0.0401 0.0352
```

$(\tilde{\hat{e_2}})^T =$

```
## Family DistRD  Cotton  Maize   Sorg    Millet  Bull   Cattle Goats
## 0.9666 -0.0842 0.0858  0.0555  0.0260  0.0591  0.0778 0.1815 0.0704
```

v. Similar to before, PC1 almost entirely depends on DistRD, with rest of the variables having loads about equal to zero. PC2 almost entirely depends on family, and some dependency on cattle. For both PCs, all the crops and farm animals have very little loads on them.

vi. Scatter plot for PC2 vs PC1



In the content, DistRD appears to be the distance to the road. There is a cluster on the left button area with relatively small PC1 and PC2 values, showing a significant portion of farms with fewer family members tend to live closer to the road. Farm 1, 6, 66, 67, and 68 on the far right shows farms are far away from the road tend to have relatively low family members. Also, there is a gap along the vertical axis.

(c) Comparaison on the results for the two analyses

Removing the outliers has relatively little effect on the eigenvectors(coefficients) for the first and second principal components, i.e. what they represent. But it significantly increases the weighing(eigenvalue) of PC2 on the total variance, which was overshadowed by PC1 before removing the outliers. Therefore, the second test suggests to keep two principal components instead of one in the first test. It also helps with scaling issue on the scatter plot, allows us to obtain information from it clearly.

I like the result of the analysis after removing the outliers better, because it gives us more clear images on the analysis, while removing very little information from it. Also, it shows some information we might ignore before.

P.S. Due to the scaling issue of Family and DistRD having dominating values, we were not able to obtain much information on the crops and farm animals from both analyses. I looked at the eigenvectors for PC3 to PC9, 5 out 6 are dominated by only one coefficient, meaning we can't obtain relationships between the crops and farm animals from them either. Running PCA using sample correlation matrix R might be a better choice, but we might end up choosing more PCs.

Code used to solve the questions(graphs are hidden):

```
rm(list = ls())
library(readxl)
library(factoextra)
X <- read_excel("C:/Users/John/Desktop/STAT 445/Data/malifarmdata.xlsx")

# a
pairs(X,pch=20, col = 2, oma = c(0,0,0,0),cex=0.8)
```

```
par(cex.axis=0.8)
boxplot(X, main = "box plot of Mali Farm data")
```

```
S <- cov(X)
x_bar <- colMeans(X)
D <- c()
for (i in 1:nrow(X)) {
  x_i <- t(X[i,])
  D[i]=t(x_i-x_bar)%*%solve(S)%*%(x_i-x_bar)
}
plot(D, main = "D values vs index", pch=20);text(D, pos = 2, offset = 0.3)
```

```
par(mfrow=c(3, 3),mar= c(2,2,2,2), oma=c(0,0,0,0))
q <- vector("list", length = 9)
for (i in 1:9) {
  q[[i]]=qqnorm(X[[i]],type="n", main = names(X)[i]);qqline(X[[i]],col="red")
  text(q[[i]]$x, q[[i]]$y)
}
```

```r
dtable <- data.frame(matrix(ncol = 7, nrow = 4))
colnames(dtable) <- c(25,34,52,57,62,69,72)
rownames(dtable) <- c("Scatter Plots", "Q-Q Plots", "Distance Plot", "Box Plot")
dtable[1,] <- c("yes","yes"," "," "," ","yes","yes")
dtable[2,] <- c("yes","yes","yes","yes","yes","yes","yes")
dtable[3,] <- c("yes","yes","yes","yes","yes","yes","yes")
dtable[4,] <- c("yes","yes","yes","yes","yes","yes","yes")
dtable
```

```
##                25  34  52  57  62  69  72
## Scatter Plots yes yes             yes yes
## Q-Q Plots     yes yes yes yes yes yes yes
## Distance Plot yes yes yes yes yes yes yes
## Box Plot      yes yes yes yes yes yes yes
```

```r
Y <- X[-c(25,34,52,57,62,69,72),]

# b
# analysis 1
X_pr = prcomp(X, center = TRUE, scale. = FALSE)
X_eigen_table = get_eigenvalue(X_pr)
S <- cov(X)
round(S,3)
```

```
##          Family   DistRD Cotton  Maize    Sorg Millet    Bull  Cattle  Goats
## Family  550.876 -158.768 48.117 29.539  31.837 26.393  45.458 103.754 46.810
## DistRD -158.768 6533.751  6.436 -8.105 -13.692  3.941 -19.025 -67.355 10.363
## Cotton   48.117    6.436  8.012  3.832   2.585  2.446   5.763   6.504  4.654
## Maize    29.539   -8.105  3.832  3.434   0.481  0.894   3.074   4.809  1.042
## Sorg     31.837  -13.692  2.585  0.481   5.700  2.029   2.816  12.699  4.171
## Millet   26.393    3.941  2.446  0.894   2.029  4.942   2.091   2.366  2.801
## Bull     45.458  -19.025  5.763  3.074   2.816  2.091   7.089  18.206  6.150
## Cattle  103.754  -67.355  6.504  4.809  12.699  2.366  18.206 173.081 19.364
## Goats    46.810   10.363  4.654  1.042   4.171  2.801   6.150  19.364 17.013
```

```r
round(X_eigen_table[,1], 4)
```

```
## [1] 6538.8594  590.1075  147.5506   12.7110    5.8905    3.9910    2.9594
## [8]    1.1372    0.6913
```

```r
round(X_eigen_table,4)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1  6538.8594          89.5256                     89.5256
## Dim.2   590.1075           8.0794                     97.6050
## Dim.3   147.5506           2.0202                     99.6251
## Dim.4    12.7110           0.1740                     99.7992
## Dim.5     5.8905           0.0806                     99.8798
## Dim.6     3.9910           0.0546                     99.9344
## Dim.7     2.9594           0.0405                     99.9750
## Dim.8     1.1372           0.0156                     99.9905
## Dim.9     0.6913           0.0095                    100.0000
```

```r
mean(X_eigen_table[,1])
```

```
## [1] 811.5442
```

```r
plot(X_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
e_X_1 <- X_pr$rotation[,1]
e_X_2 <- X_pr$rotation[,2]
round(e_X_1, 4)
```

```
##  Family  DistRD  Cotton   Maize    Sorg  Millet    Bull  Cattle   Goats
##  0.0267 -0.9996 -0.0008  0.0014  0.0022 -0.0005  0.0031  0.0110 -0.0014
```

```r
round(e_X_2, 4)
```

```
##  Family  DistRD  Cotton   Maize    Sorg  Millet    Bull  Cattle   Goats
## -0.9536 -0.0284 -0.0842 -0.0509 -0.0582 -0.0455 -0.0835 -0.2444 -0.0890
```

```r
plot(X_pr$x[,1], X_pr$x[,2], xlab = "PC1 (89.53%)", ylab = "PC2 (8.08%)", pch = 20, col = 2)
text(X_pr$x[,1], X_pr$x[,2], pos = 2, offset = 0.3, cex = 0.5)
abline(v = 0, h = 0)

# analysis 2
Y_pr = prcomp(Y, center = TRUE, scale. = FALSE)
Y_eigen_table = get_eigenvalue(Y_pr)
Sy <- cov(Y)
round(Sy,3)
```

```
##         Family  DistRD Cotton  Maize   Sorg Millet   Bull Cattle  Goats
## Family 318.587  16.958 27.505 18.451  8.021 19.311 25.023 55.753 22.746
## DistRD  16.958 595.620  3.786  7.585 -8.188 -3.644  7.347 17.977 18.546
## Cotton  27.505   3.786  5.179  2.629  1.038  1.656  3.543  7.581  3.065
## Maize   18.451   7.585  2.629  2.474 -0.038  0.995  2.005  4.570  0.892
## Sorg     8.021  -8.188  1.038 -0.038  2.554  0.827  0.657  0.137  0.405
## Millet  19.311  -3.644  1.656  0.995  0.827  4.467  1.446  1.321  0.969
## Bull    25.023   7.347  3.543  2.005  0.657  1.446  4.467  7.761  4.186
## Cattle  55.753  17.977  7.581  4.570  0.137  1.321  7.761 35.428  8.794
## Goats   22.746  18.546  3.065  0.892  0.405  0.969  4.186  8.794 13.209
```

```r
round(Y_eigen_table[,1], 4)
```

```
## [1] 598.6558 336.1485  26.5261  10.1536   3.6717   2.9183   2.2307   1.1355
## [9]   0.5448
```

```r
round(Y_eigen_table,4)
```

```
##       eigenvalue variance.percent cumulative.variance.percent
## Dim.1   598.6558          60.9638                     60.9638
## Dim.2   336.1485          34.2315                     95.1954
```

```
## Dim.3    26.5261           2.7013                          97.8967
## Dim.4    10.1536           1.0340                          98.9306
## Dim.5     3.6717           0.3739                          99.3045
## Dim.6     2.9183           0.2972                          99.6017
## Dim.7     2.2307           0.2272                          99.8289
## Dim.8     1.1355           0.1156                          99.9445
## Dim.9     0.5448           0.0555                         100.0000
```

```r
mean(Y_eigen_table[,1])
```

```
## [1] 109.1094
```

```r
plot(Y_eigen_table[,1], type = "o", pch = 15, main = "Scree Plot",
     xlab = "Principal Component Number", ylab = "Eigenvalues")
e_Y_1 <- Y_pr$rotation[,1]
e_Y_2 <- Y_pr$rotation[,2]
round(e_Y_1, 4)
```

```
##  Family  DistRD  Cotton   Maize    Sorg  Millet    Bull  Cattle   Goats
##  0.0740  0.9954  0.0106  0.0154 -0.0126 -0.0035  0.0163  0.0401  0.0352
```

```r
round(e_Y_2, 4)
```

```
##  Family  DistRD  Cotton   Maize    Sorg  Millet    Bull  Cattle   Goats
##  0.9666 -0.0842  0.0858  0.0555  0.0260  0.0591  0.0778  0.1815  0.0704
```

```r
plot(Y_pr$x[,1], Y_pr$x[,2], xlab = "PC1 (60.96%)", ylab = "PC2 (34.23%)", pch = 20, col = 2)
text(Y_pr$x[,1], Y_pr$x[,2], pos = 2, offset = 0.3, cex = 0.5)
abline(v = 0, h = 0)
```