STAT 445/645 Assignment Cover Page

Student Name	Heewon Oh
SFU Student Number	301268860
SFU email address	heewono@sfu.ca
Assignment Number	7
Due Date	April 15,2022

Provide references for any data sets used in this assignment

Hand, D., Daly, F., Lunn, A., McConway, K. and Ostrowski, E. (eds), A Handbook of Small Data Sets, Chapman & Hall, London, 1994.

A. Rencher and W. Christensen, Methods of Multivariate Analysis, Table 7.2, Wiley, New York, 2012.

List software used in this assignment.

R, Rstudio

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

Notes from Class

Code from Tutorial

https://www.datanovia.com/en/blog/easy-way-to-expand-color-palettes-in-r/

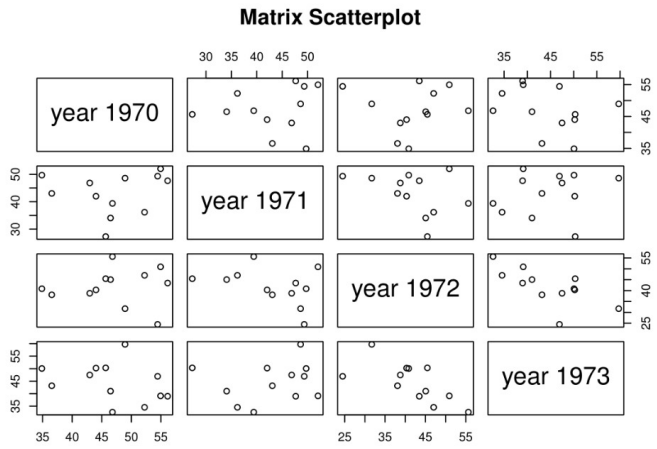
☑ I personally completed the computations and wrote the solutions submitted in this document.

Department of Statistics and Actuarial Science

SFU SIMON FRASER UNIVERSITY

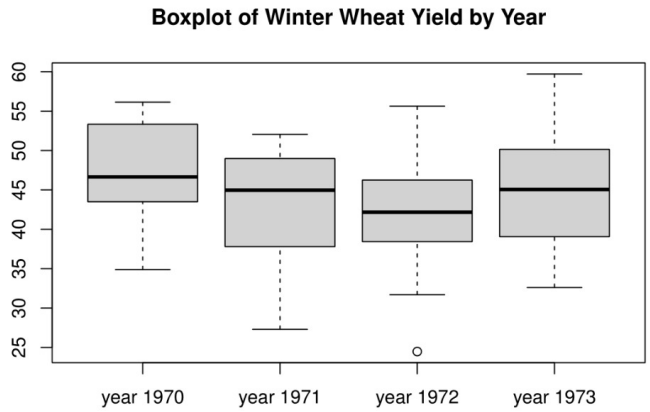
1)
A)i)

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



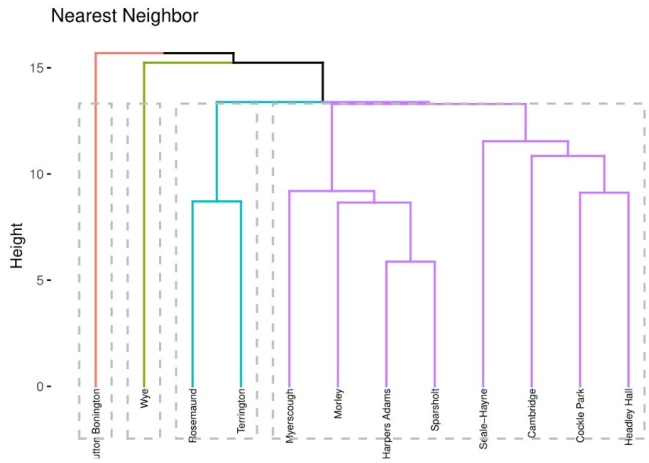
There are no apparent groups in the scatterplots.

ii)



We do not anticipate that scale may be an issue as all the variables are of similar range and distribution.

B)i)



ii)

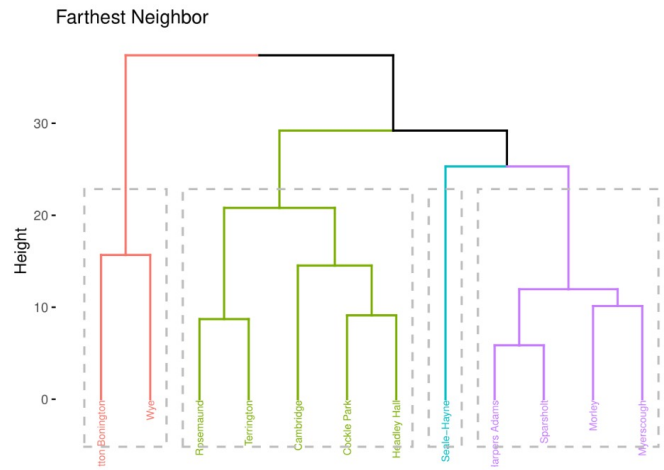
```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                1                1
##           Morley       Myerscough        Rosemaund      Seale-Hayne
##                1                1                1                1
##        Sparsholt Sutton Bonington       Terrington             Wye
##                1                2                1                3
```

Cluster 1: Cambridge, Cockle Park, Harpers Adams, Headley Hall, Morley, Myerscough, Rosemaund, Seale-Hayne, Sparsholt, Terrington

Cluster 2: Sutton Bonington

Cluster 3: Wye

C)i)



ii)

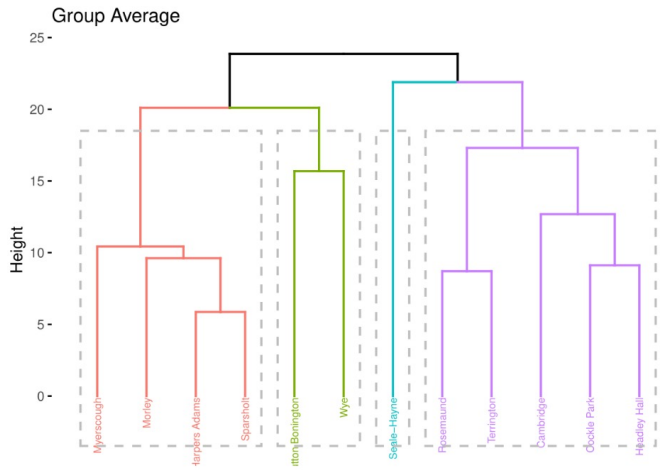
```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                2                1
##           Morley       Myerscough       Rosemaund      Seale-Hayne
##                2                2                1                2
##        Sparsholt Sutton Bonington       Terrington              Wye
##                2                3                1                3
```

Cluster 1: Cambridge, Cockle Park, Headley Hall, Rosemaund, Terrington

Cluster 2: Harpers Adams, Morley, Myerscough, Seale-Hayne, Sparsholt

Cluster 3: Sutton Bonington, Wye

D)i)



ii)

```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                2                1
##           Morley       Myerscough       Rosemaund      Seale-Hayne
##                2                2                1                3
##        Sparsholt Sutton Bonington       Terrington             Wye
##                2                2                1                2
```

Cluster 1: Cambridge, Cockle Park, Headley Hall, Rosemaund, Terrington

Cluster 2: Harpers Adams, Morley, Myerscough, Sparsholt, Sutton Bonington, Wye

Cluster 3: Seale-Hayne

E)

In the nearest neighbor method, two singleton clusters are: Sutton Bonington and Wye.

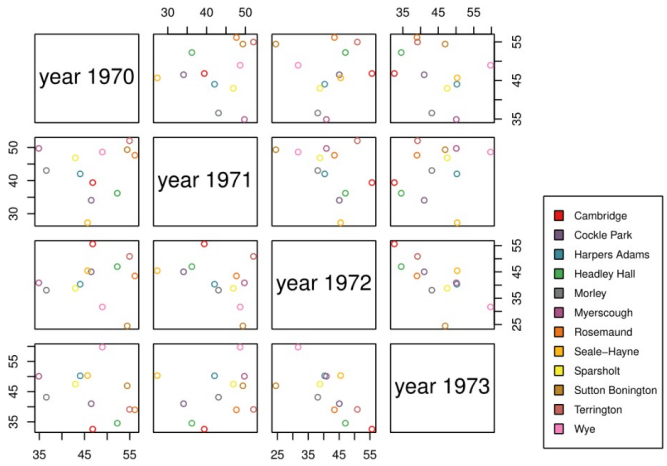
The two-item cluster is composed of: Rosemaund and Terrington.

In the furthest neighbor method, the singleton cluster is: Seale-Hayne.

The two-item cluster is composed of: Sutton Bonington and Wye.

In the group average method, the singleton cluster is: Seale-Hayne.

The two-item cluster is composed of: Sutton Bonington and Wye.



Above is a colored matrix scatterplot with legend. Below we do a visual inspection.

Visually, we note here that Sutton Bonington is somewhat of a low outlier in the year 1972 variable in the matrix scatterplot. We see that Wye is a high outlier in the variable year 1973. It is also somewhat low in the variable year 1972. In both, it is separated from the others. Seale-Hayne is a low outlier in the variable year 1972. Rosemaund has the highest value in the year 1970 variable; however, it is not a distinct outlier. Terrington has the highest value in the year 1971; however, like Rosemaund it does not appear to be a distinct outlier. Sutton Bonington and Wye have similar values overall across the variables as do Rosemaund and Terrington.

All three approaches have Sutton Bonington and Wye as either a singleton cluster or a two-item cluster, with the first method having them as singleton and the last two as a two-item cluster.

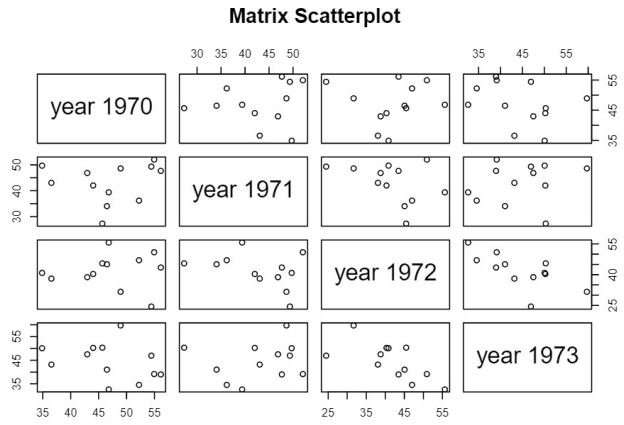
The last two methods (furthest neighbor and group average) both have Seale-Hayne as the singleton cluster and Sutton Bonington and Wye as the two-item cluster. They also both have a 4-item cluster for Harpers Adams, Sparsholt, Morley, and Myerscough. The nearest neighbor method also has these under one cluster; however, they are 4 other items in the cluster joined together at its top-level of the dendogram.

Q1

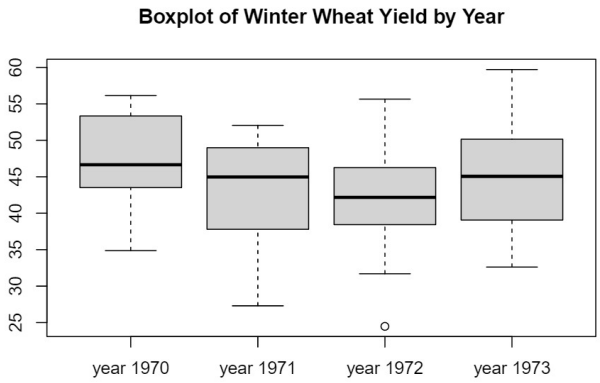
1)F) Code for 1A)

```
library(readxl)
wheat<-as.data.frame(read_excel("winter_wheat.xlsx"))
rownames(wheat)<-wheat[,1]
wheat<-wheat[,-1]
```

```
pairs(wheat,main="Matrix Scatterplot")
```



```
boxplot(wheat,main="Boxplot of Winter Wheat Yield by Year")
```



Code for 1B)

```
d=dist(wheat,method="euclidean",diag=T,upper=T)
hc_single<-hclust(d=d,method="single")
three_clust<-cutree(hc_single,k=3)
```

```
library(factoextra)
```

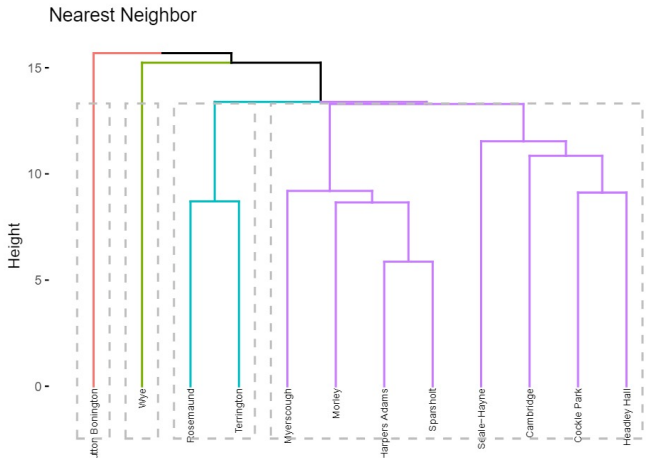
```
## Warning: package 'factoextra' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
fviz_dend(hc_single,cex=0.5,k=4,main="Nearest Neighbor",
          color_labels_by_k=FALSE,rect=TRUE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



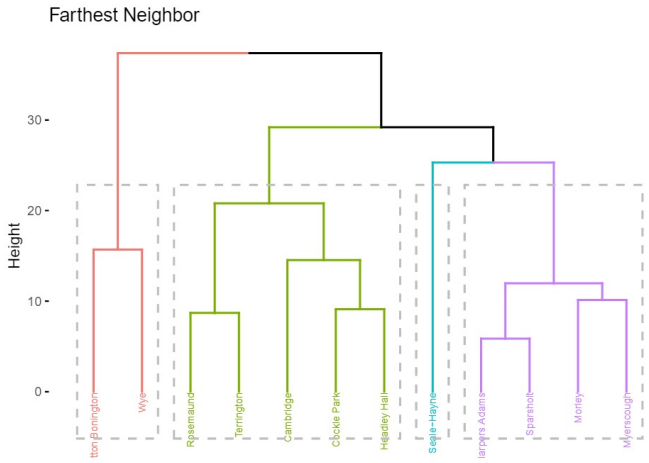

```
three_clust
```

```
##        Cambridge      Cockle Park    Harpers Adams      Headley Hall
##                1                1                1                1
##           Morley       Myerscough        Rosemaund      Seale-Hayne
##                1                1                1                1
##        Sparsholt Sutton Bonington       Terrington              Wye
##                1                2                1                3
```

Code for 1C)

```
hc_complete<-hclust(d=d,method="complete")
fviz_dend(hc_complete,cex=0.5,k=4,main="Farthest Neighbor",
          color_labels_by_k=TRUE,rect=TRUE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```



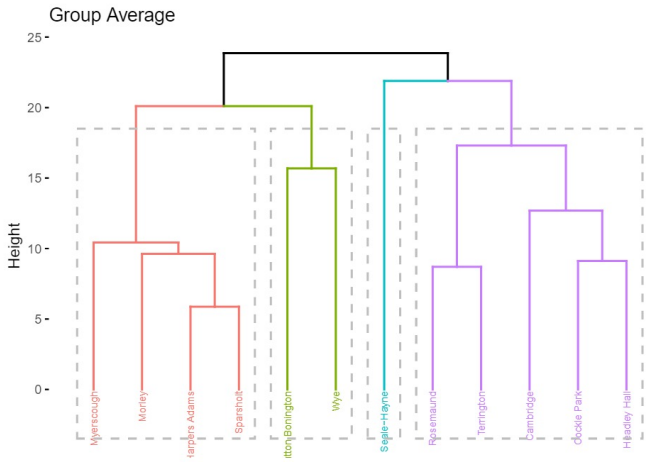

```
cutree(hc_complete,k=3)
```

```
##        Cambridge      Cockle Park    Harpers Adams      Headley Hall
##                1                1                2                1
##           Morley       Myerscough        Rosemaund      Seale-Hayne
##                2                2                1                2
##        Sparsholt Sutton Bonington       Terrington              Wye
##                2                3                1                3
```

Code for 1D)

```
hc_avg<-hclust(d=d,method="average")
fviz_dend(hc_avg,cex=0.5,k=4,main="Group Average",
          color_labels_by_k=TRUE,rect=TRUE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```




```
cutree(hc_avg,k=3)
```

```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                2                1
##           Morley       Myerscough        Rosemaund      Seale-Hayne
##                2                2                1                3
##        Sparsholt Sutton Bonington       Terrington              Wye
##                2                2                1                2
```

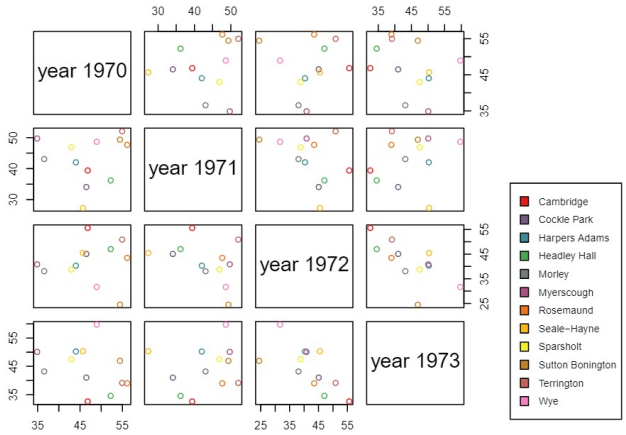
```
library(readxl)
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --

## v tibble  3.1.6     v dplyr   1.0.7
## v tidyr   1.1.4     v stringr 1.4.0
## v readr   2.1.1     v forcats 0.5.1
## v purrr   0.3.4

## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(RColorBrewer)
```

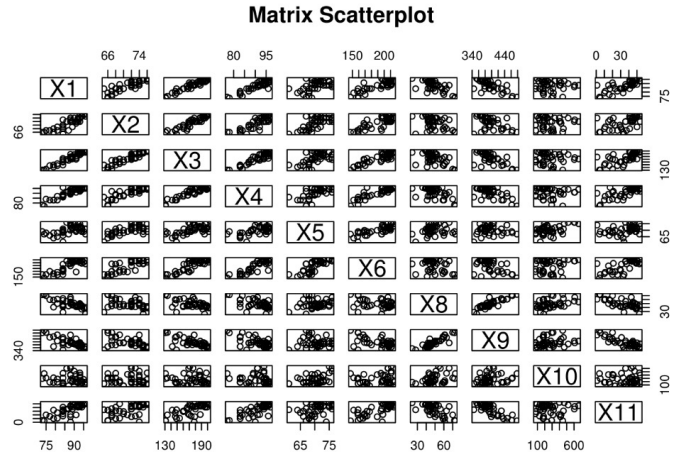
```
## Warning: package 'RColorBrewer' was built under R version 4.1.3
```

```
coul <- colorRampPalette(brewer.pal(8, "Set1"))(12)
wheat1<-as.data.frame(read_excel("winter_wheat.xlsx"))
plot(wheat1[,-1],col=coul,oma=c(3,3,3,14))
par(xpd=T)
legend("bottomright",
       fill=coul,
       legend=factor(wheat1[,1]),
       cex=0.6)
```



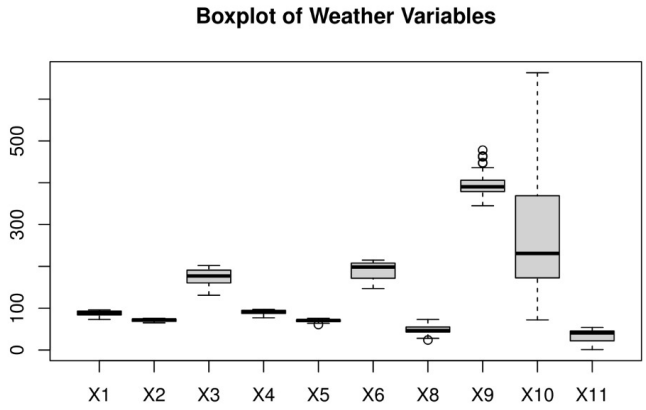
2)A)
i)

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



There are no apparent groups from the matrix scatterplot. Overall. Those involving variable X6, appear to possibly suggest groups. However, there is nothing truly indicative.

ii)



We anticipate that there may be significant scale issues as some variables, such as X10 are much larger in range and average value than the others.

iii)

First row:

```
##           X1        X2        X3        X4        X5        X6        X8
## 1 -0.5891427 -1.977605 -1.249747 -0.9485805 -2.790298 -1.568095 -1.998541
##           X9        X10         X11
## 1 -1.692119 -0.8916562 -0.03363364
```

Last row:

```
##           X1       X2       X3       X4        X5        X6         X8
## 44 1.355028 1.428662 1.385344 1.010928 0.2041681 0.8245541 -0.8503041
##             X9        X10      X11
## 44 -0.9364153 -0.8983364 1.042643
```

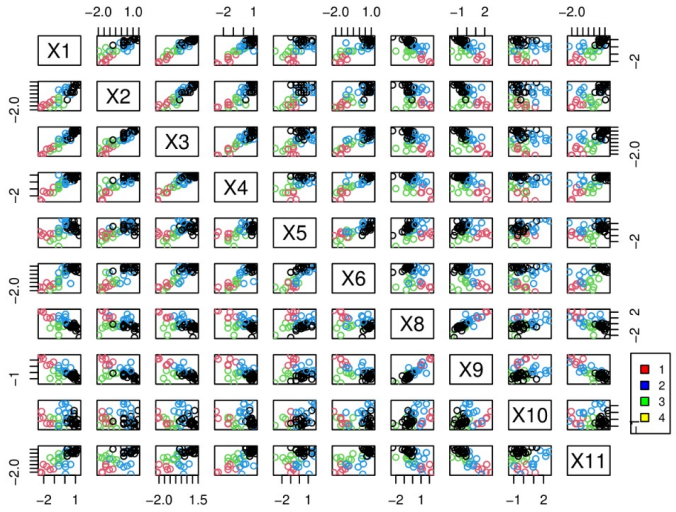
B)i)

Scaled data incorrect. **-3**

The number of observations in each cluster (in ascending order from Cluster 1 to 4) are:

```
20  6  8 10
```

ii) Below is the matrix scatterplot with cluster-membership.



iii) Overall, the clusters are not well-separated in the plots.

C)i)

List of eigenvalues:

```
[1] 6.36324008 2.08501617 0.81032361 0.31911410 0.12231773 0.11539327
[7] 0.07246213 0.06384998 0.03493530 0.01334763
```

Percent contributions of eigenvalues:

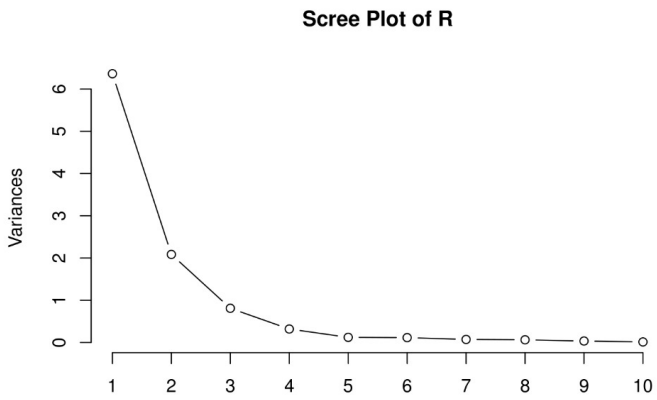
```
 [1] 63.6324008 20.8501617  8.1032361  3.1911410  1.2231773  1.1539327
 [7]  0.7246213  0.6384998  0.3493530  0.1334763
```

The eigenvalues and the percent contributions to the variance are incorrect. **-3**

Average percent contributions of eigenvalues:

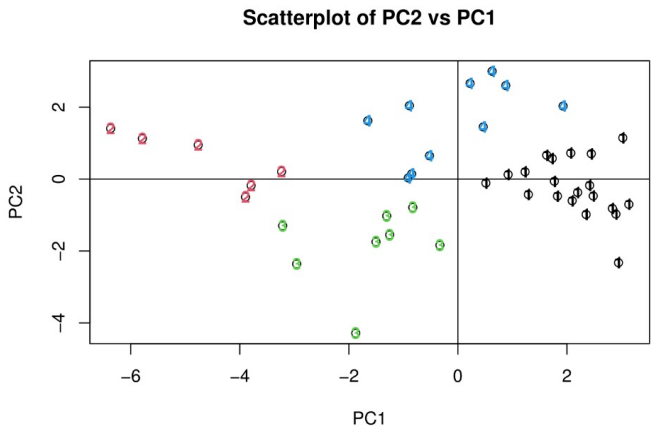
```
pvaravg=mean(100*(pca$sdev)^2/sum((pca$sdev)^2))
pvaravg
```

```
## [1] 10
```



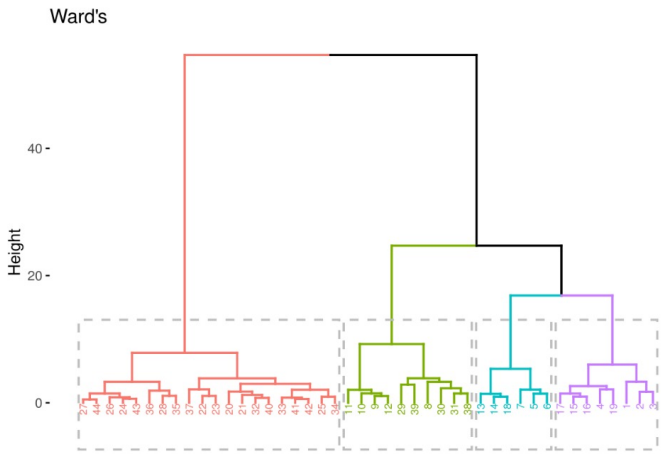
We retain 2 PCs by using the 80% rule, average percentage contribution test, and the scree plot. All three support 2 PCs. Two eigenvalues accounts for over 80% of the variance, the 3rd eigenvalue contributes less than 10% towards the variance, and the scree plot curves after point 3.

ii)



iii) Yes, k-means analysis yielded reasonably defined clusters as shown from the PC2 vs PC1 scatterplot, where the clusters are reasonably separated.

D)i)



ii)

The membership of the four clusters produced by k-means are as below (by increasing row value):

```
 [1] 3 3 3 3 2 2 2 4 4 4 4 4 2 2 3 3 3 2 3 1 1 1 1 1 1 1 1 1 4 4
[31] 4 1 1 1 1 1 1 4 4 1 1 1 1 1
```

Moreover, The memberships for the clusters match perfectly.

Below are the four cluster memberships produced by Ward's method:

```
 [1] 1 1 1 1 2 2 2 3 3 3 3 3 2 2 1 1 1 2 1 4 4 4 4 4 4 4 4 4 3 3
[31] 3 4 4 4 4 4 4 4 3 3 4 4 4 4 4
```

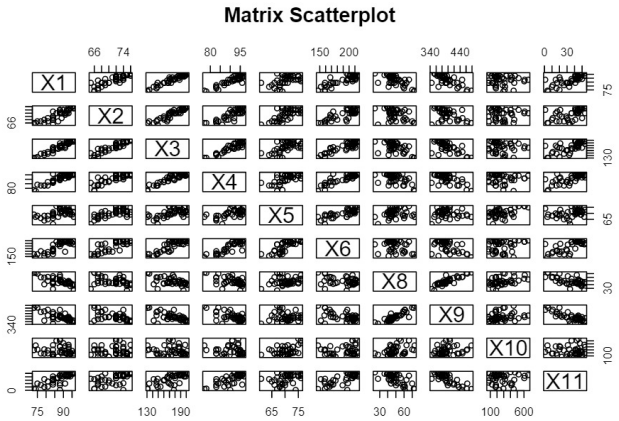
Q2

2)E)

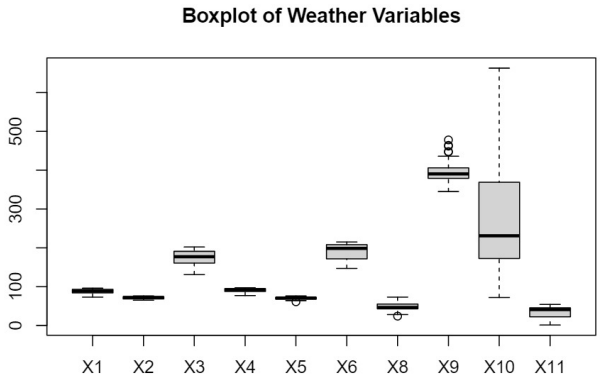
Code for 2)A)

```
library(readxl)
weather<-read_excel("temphumevap_strip.xlsx")
colnames(weather)<-c("X1","X2","X3","X4","X5","X6","X8","X9","X10","X11")
```

```
pairs(weather,main="Matrix Scatterplot")
```




```
boxplot(weather,main="Boxplot of Weather Variables")
```




```
scaledw<-as.data.frame(scale(weather))
scaledw[1,]
```

```
##           X1        X2        X3         X4        X5        X6        X8
## 1 -0.5891427 -1.977605 -1.249747 -0.9485805 -2.790298 -1.568095 -1.998541
##           X9        X10         X11
## 1 -1.692119 -0.8916562 -0.03363364
```

```
scaledw[44,]
```

```
##          X1       X2       X3       X4        X5        X6         X8
## 44 1.355028 1.428662 1.385344 1.010928 0.2041681 0.8245541 -0.8503041
##           X9        X10      X11
## 44 -0.9364153 -0.8983364 1.042643
```

Code for 2B)

```
set.seed(1234)
km_4<-kmeans(scaledw,centers=4)
km_4$cluster
```

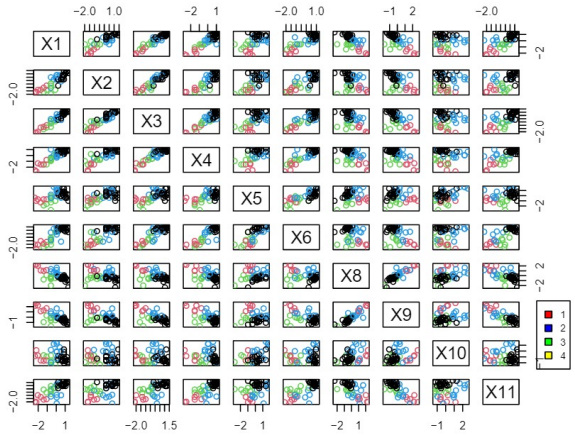
```
##  [1] 3 3 3 3 2 2 2 4 4 4 4 4 2 2 3 3 3 2 3 1 1 1 1 1 1 1 1 1 4 4 4 1 1 1 1 1 1 4
## [39] 4 1 1 1 1 1
```

```
count<-vector(length=4)
for(i in 1:4){
  for(j in 1:44){
    if(km_4$cluster[j]==i){
      count[i]<-count[i]+1
    }
  }
}
```

```
count
```

```
## [1] 20  6  8 10
```

```
scaledw$Group=km_4$cluster
plot(scaledw[,-11],col=scaledw$Group,oma=c(3,3,3,10))
par(xpd=T)
legend("bottomright",
       fill=c("red","blue","green","yellow"),
       legend=c("1","2","3","4"),
       cex=0.6)
```



Code for 2C)

```
pca=prcomp(scaledw[,-11],center=T,scale=F)
eigen<- (pca$sdev)^2
eigen
```

```
##  [1] 6.36324008 2.08501617 0.81032361 0.31911410 0.12231773 0.11539327
##  [7] 0.07246213 0.06384998 0.03493530 0.01334763
```

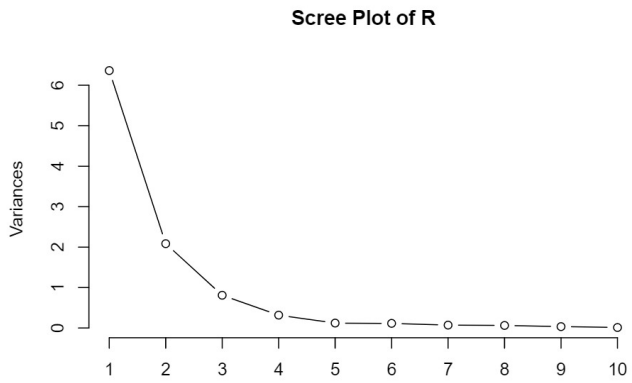
```
pvar=100*(pca$sdev)^2/sum((pca$sdev)^2)
pvar
```

```
##  [1] 63.6324008 20.8501617  8.1032361  3.1911410  1.2231773  1.1539327
##  [7]  0.7246213  0.6384998  0.3493530  0.1334763
```

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

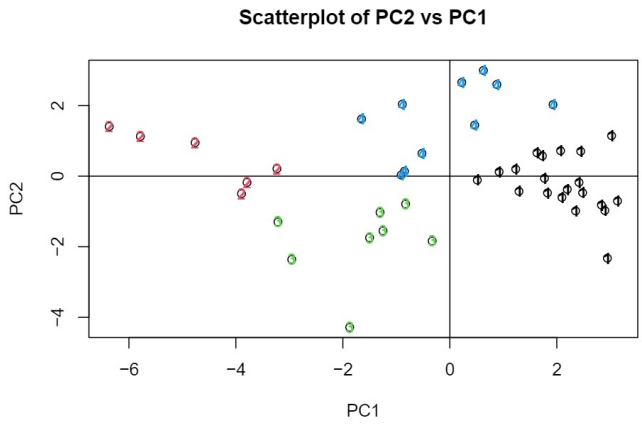
```
## [1] 10
```

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




```
PC1=pca$x[,1]
PC2=pca$x[,2]
```

```
plot(PC1,PC2,main="Scatterplot of PC2 vs PC1")
abline(v=0,h=0)
text(PC2-PC1,labels=scaledw$Group,
     data=scaledw[,-11],col=scaledw$Group,cex=0.9,font=2)
```



Code for 2D)

```
library(factoextra)
```

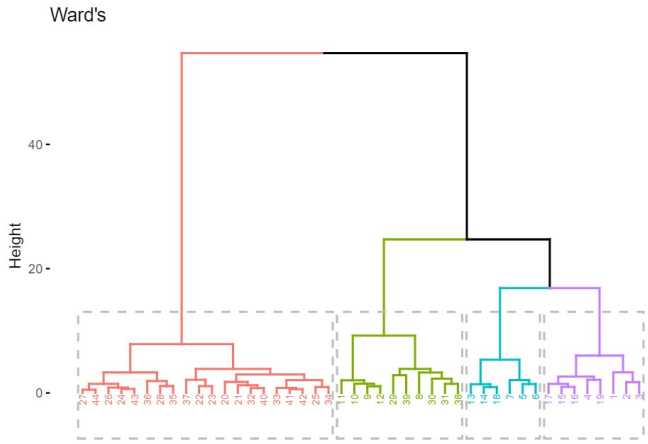
```
## Warning: package 'factoextra' was built under R version 4.1.3
```

```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
d=dist(scaledw,method="euclidean",diag=T,upper=T)
hc_ward<-hclust(d=d,method="ward.D")
fviz_dend(hc_ward,cex=0.5,k=4,main="Ward's",
          colors_labels_by_k=TRUE,rect=TRUE)
```

```
## Warning: `guides(<scale> = FALSE)` is deprecated. Please use `guides(<scale> =
## "none")` instead.
```




```
four_clust<-cutree(hc_ward,k=4)
four_clust
```

```
##  [1] 1 1 1 1 2 2 2 3 3 3 3 3 2 2 1 1 1 2 1 4 4 4 4 4 4 4 4 4 4 3 3 3 4 4 4 4 4 4 3
## [39] 3 4 4 4 4 4
```

```
count2<-vector(length=4)
for(i in 1:4){
  for(j in 1:44){
    if(four_clust[j]==i){
      count2[i]<-count2[i]+1
    }
  }
}
count2
```

```
## [1]  8  6 10 20
```