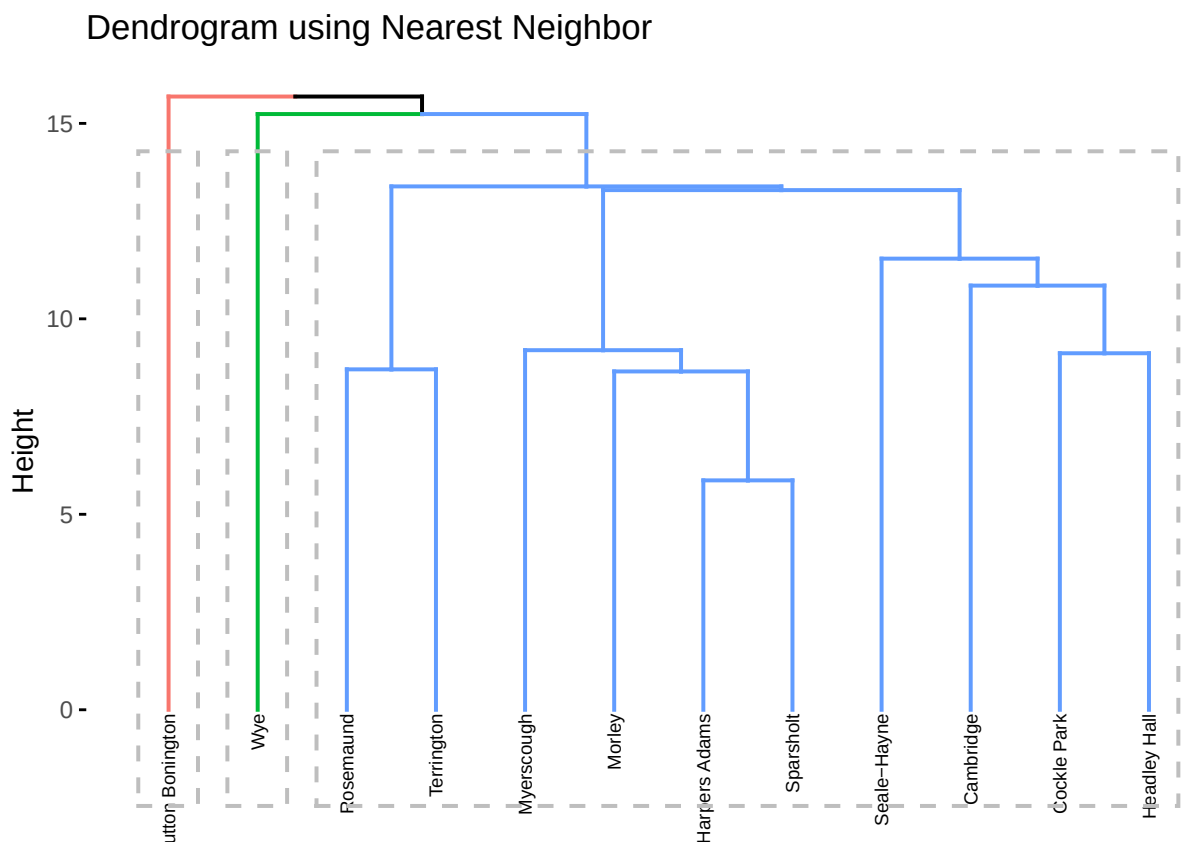
Assignment07

Dawu Liu

Question 1

(a) Nearest Neighbor

i.



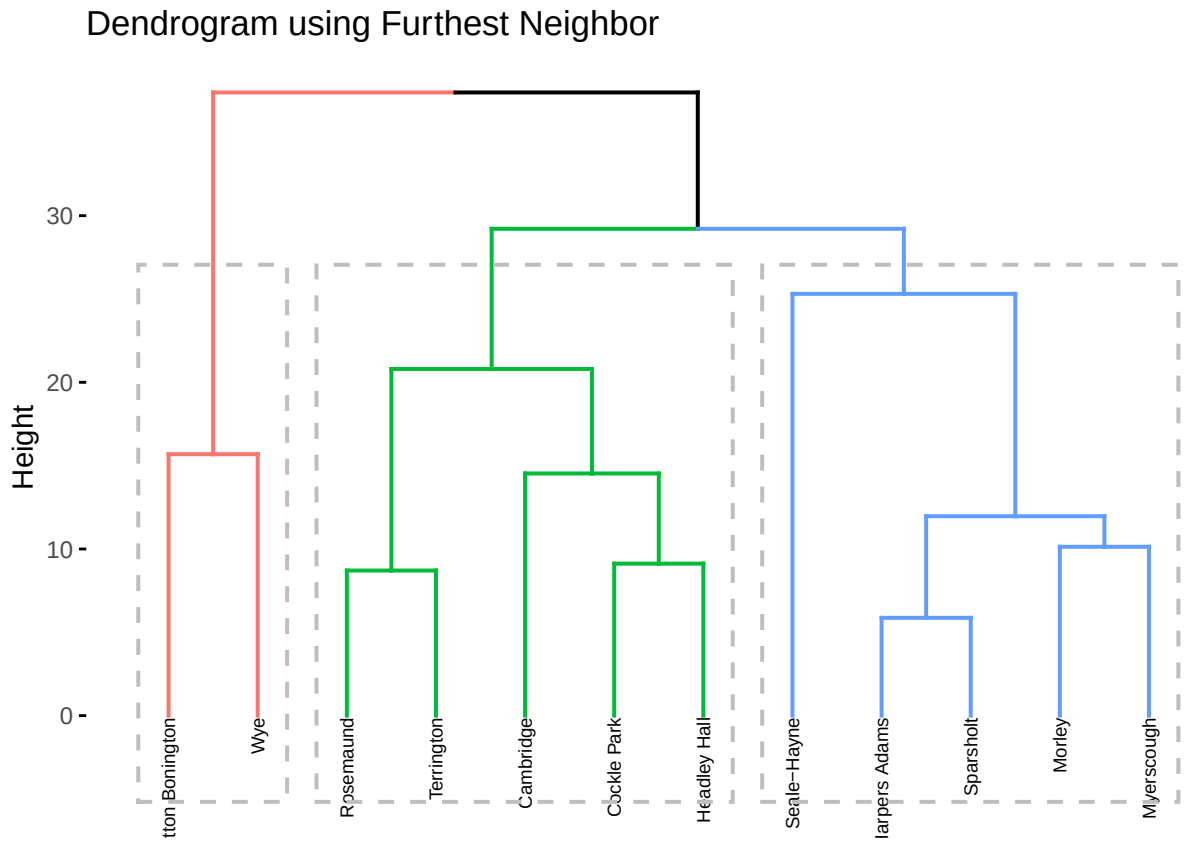

```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                1                1
##           Morley       Myerscough       Rosemaund      Seale-Hayne
##                1                1                1                1
##        Sparsholt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

ii. Cluster list

Clusters	Observations
Cluster 1	Cambridge, Cockle Park, Harpers Adams, Headley Hall, Morley, Myerscough, Rosemaund, Seale-Hayne, Sparsholt, Terrington
Cluster 2	Sutton Bonington
Cluster 3	Wye

(b) Furthest Neighbor

i.



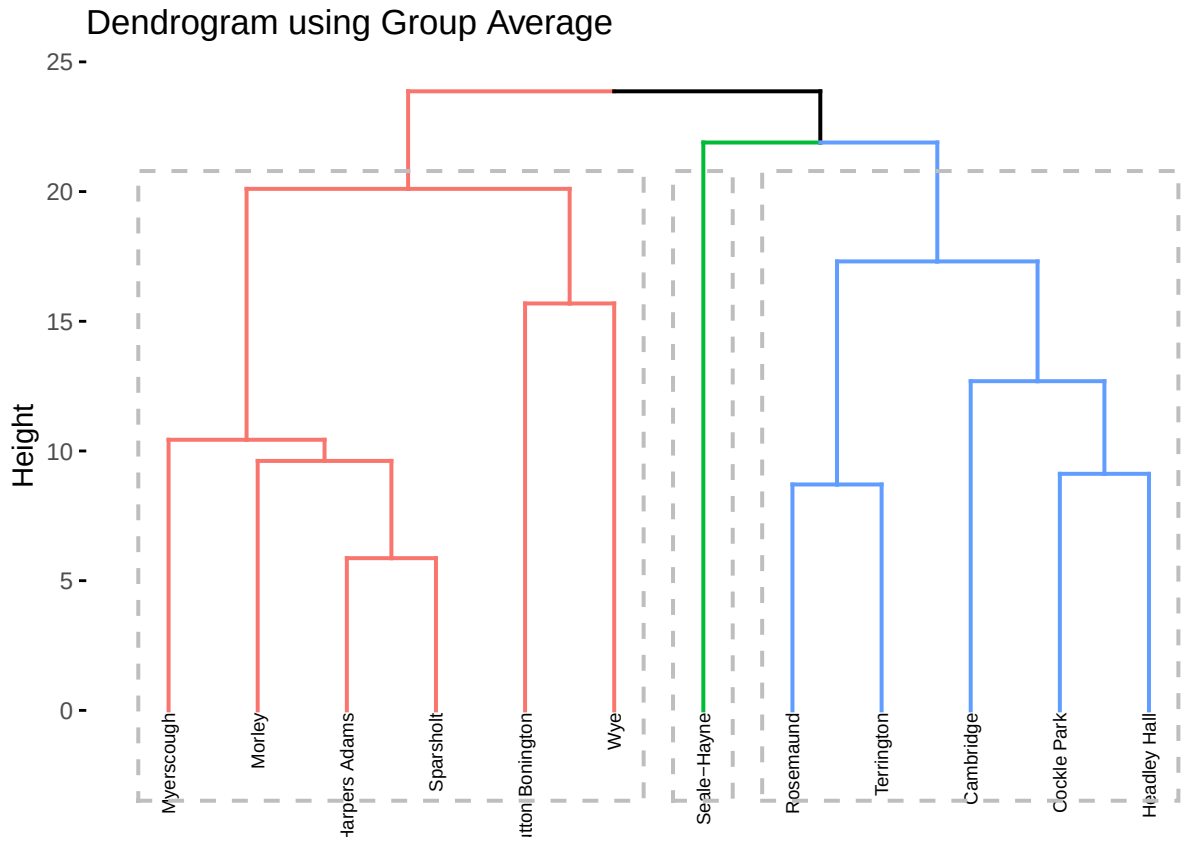

```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                2                1
##           Morley      Myerscough       Rosemaund      Seale-Hayne
##                2                2                1                2
##        Sparsholt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

ii. Cluster list

Clusters	Observations
Cluster 1	Cambridge, Cockle Park, Headley Hall, Rosemaund, Terrington
Cluster 2	Harpers Adams, Morley, Myerscough, Seale-Hayne, Sparsholt
Cluster 3	Sutton Bonington, Wye

(c) Group Average

i.



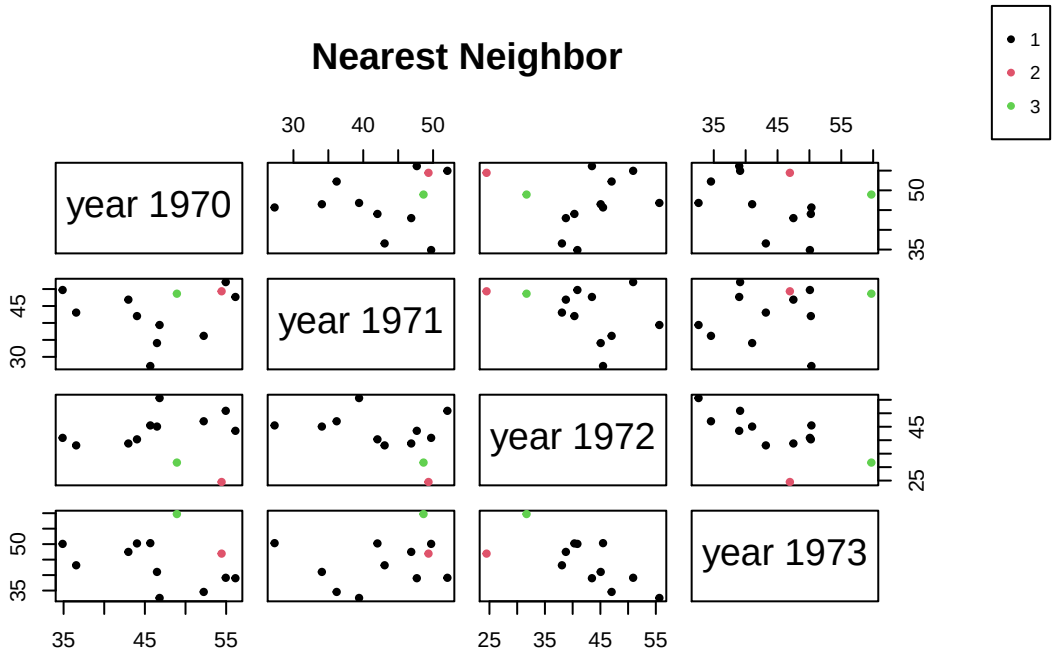

```
##        Cambridge      Cockle Park    Harpers Adams     Headley Hall
##                1                1                2                1
##           Morley       Myerscough       Rosemaund      Seale-Hayne
##                2                2                1                3
##        Sparsholt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

ii. Cluster list

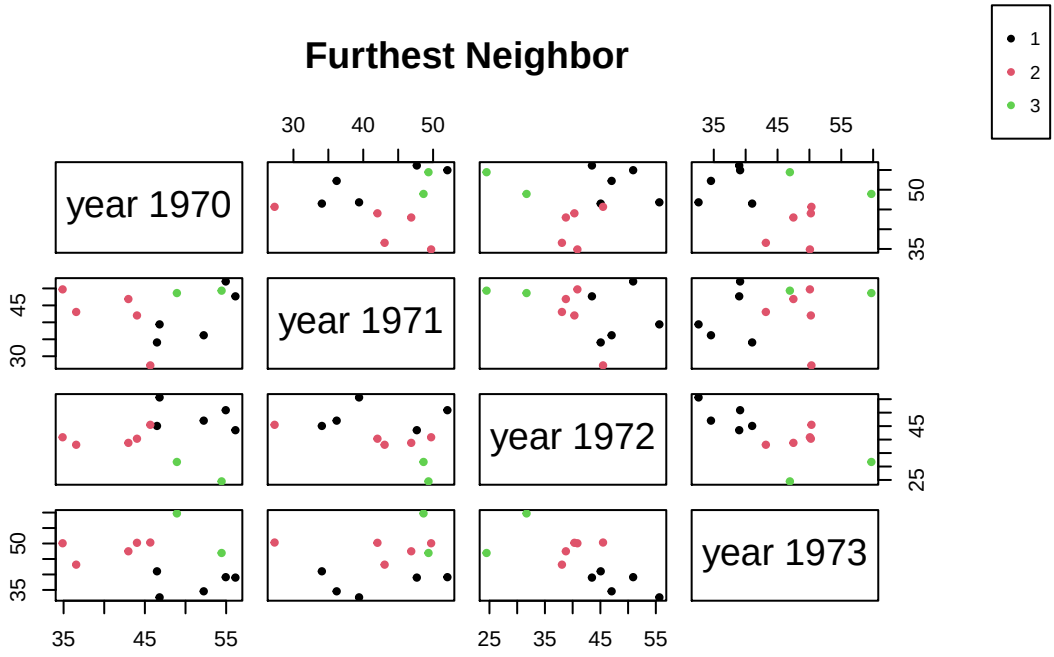
Clusters	Observations
Cluster 1	Cambridge, Cockle Park, Headley Hall, Rosemaund, Terrington
Cluster 2	Harpers Adams, Morley, Myerscough, Sparsholt, Sutton Bonington, Wye
Cluster 3	Seale-Hayne

(d) Matrix scatter plots of the data

In all three matrix scatter plots, the cluster assignments are labeled by colors

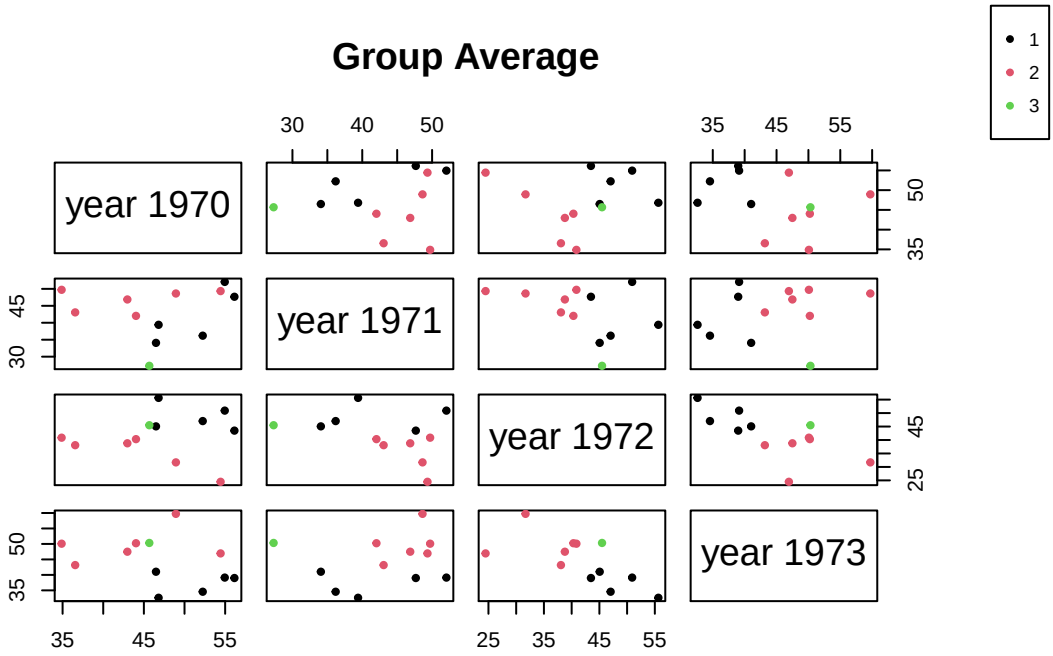


In Nearest Neighbor method, cluster 2, and 3 are singleton clusters.
On year 1970 vs year 1971, cluster 2 and 3 are overlapped with cluster 1.
On year 1971 vs year 1973, cluster 2 is overlapped with cluster 1.



In Furthest Neighbor method, there are some obvious overlapping.

On year 1970 vs year 1971, cluster 2 and 3 are overlapped with cluster 1.
On year 1971 vs year 1972, cluster 2 is overlapped with cluster 1.
On year 1971 vs year 1973, cluster 2 is overlapped with cluster 3.



In Group Average method, cluster 3 is singleton cluster.
On year 1970 vs year 1971, cluster 2 is overlapped with cluster 1.
On year 1970 vs year 1973, cluster 2 is overlapped with cluster 3.

Some comparisons of the three methods:

- In Nearest Neighbor, Sutton Bonington and Wye are both singleton cluster, and the rest of the observations are in one cluster. But in Furthest Neighbor, Sutton Bonington and Wye are grouped as one cluster, and the rest of the observations are splitted into 2 clusters.
- Furthest Neighbor and Group Average have identical cluster 1. But Seale-Hayne is in cluster 2, Sutton Bonington and Wye are in cluster 3 for Furthest Neighbor. Seale-Hayne is in cluster 3, Sutton Bonington and Wye are in cluster 2 for Group Average. (comparison shown below)

Clusters#	Furthest Neighbor	Group Average
Cluster 2	Harpers Adams, Morley, Myerscough, Seale-Hayne, Sparsholt	Harpers Adams, Morley, Myerscough, Sparsholt, Sutton Bonington, Wye
Cluster 3	Sutton Bonington, Wye	Seale-Hayne

Question 2

(a) Standardize the data

```
# the first row
round(data[1,], 5)
```

```
##     ...1     ...2     ...3     ...4     ...5     ...6     ...7     ...8
## -0.58042 -1.87444 -1.30767 -1.11554 -2.98495 -1.85799 -0.83419  0.04875
##     ...9    ...10
## -0.00315 -0.29909
```

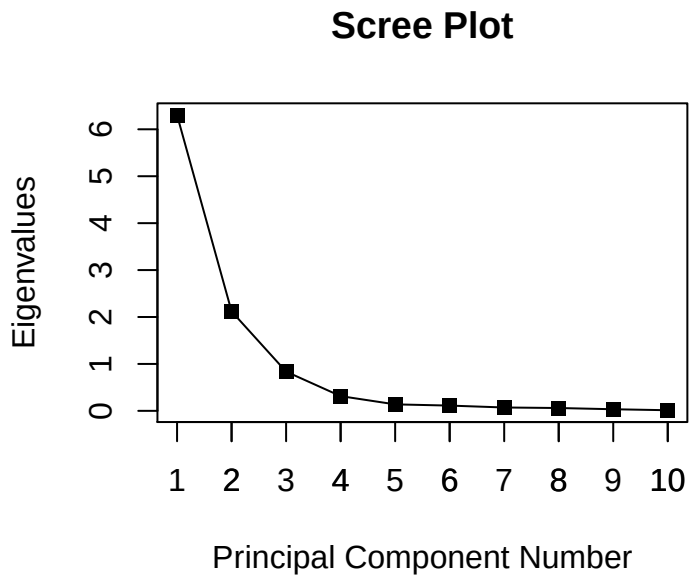
```
# the last row
round(data[nrow(data),], 5)
```

```
##     ...1     ...2     ...3     ...4     ...5     ...6     ...7     ...8
##  1.37851  1.42751  1.40295  1.03307  0.25174  0.84215 -0.83419 -0.94832
##     ...9    ...10
## -0.90865  1.06040
```

(b) Number of observations in each cluster:

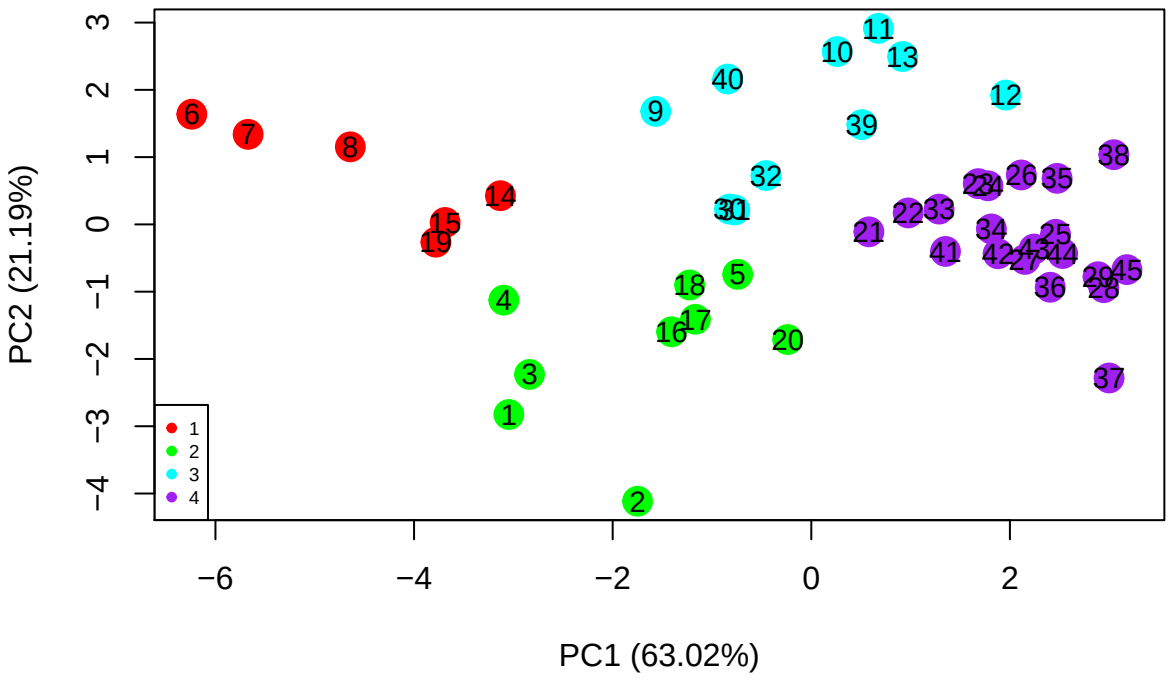
Cluster 1: 6
Cluster 2: 9
Cluster 3: 10
Cluster 4: 20

(c) Conduct a principal components analysis and display the scree Plot



Two principal components are sufficient, as the plot is suggesting keep 2 PCs. From PC3 and on, PCs have relatively small variances.

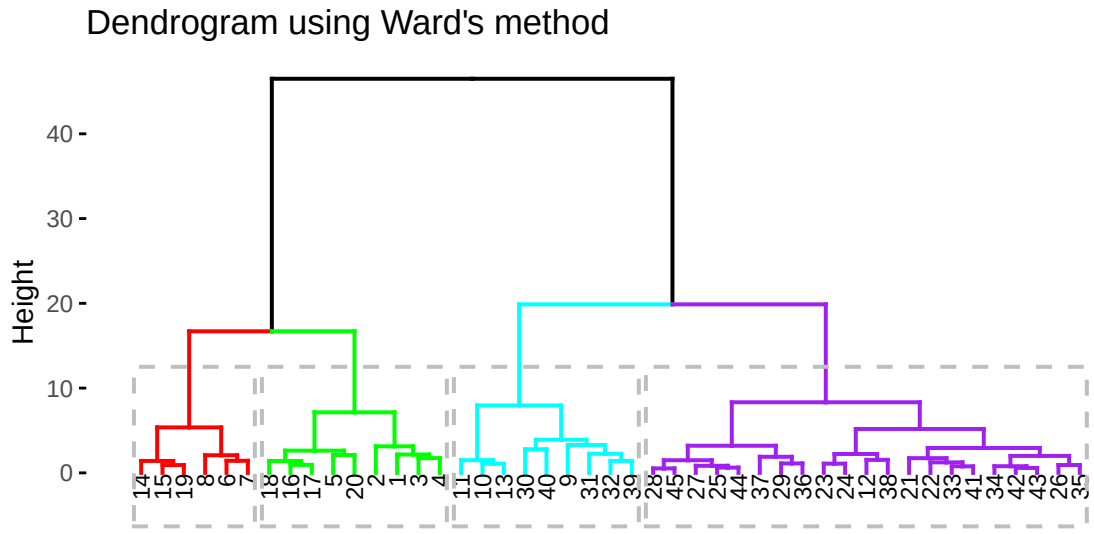
(d) PC1 vs PC2 scatter plot with clusters colored



The cluster assignments are labeled by four colors, the numbers on data points are the observation numbers.

The k-means analysis yields well-defined clusters. There are clear divides between each cluster, with no overlapping.

(e)



For easier comparison, I've matched the colors and order for each of the 4 clusters in both methods.

By comparing the results of (d) and (e), the first (red) and second (green) clusters are identical between

the two methods. The only difference is that, in k-means one observation (observation 12) is in cluster 3 (cyan), but in Ward's method observation 12 is in cluster 4 (purple). And the rest of the clusters are the same. Both methods yield very similar results.

Code used to solve the question:

```
rm(list = ls())
library(readxl)
library(MESS)
library(factoextra)

# Question 1
wheat <- as.data.frame(read_excel("C:/Users/John/Desktop/STAT 445/Data/winter_wheat.xlsx"))
rownames(wheat) <- wheat[,1]
wheat <- wheat[,-1]
d=dist(wheat,method="euclidean")

# a
a <- hclust(d=d, method="single")
fviz_dend(a, cex=0.4, k=3, main="Dendrogram using Nearest Neighbor",
          color_labels_by_k=TRUE, rect=TRUE)
```

```
clustera <- cutree(a, k=3)
clustera
```

```
##        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
```

```
# b
b <- hclust(d=d, method="complete")
fviz_dend(b, cex=0.4, k=3, main="Dendrogram using Furthest Neighbor",
          color_labels_by_k=TRUE, rect=TRUE)
```

```
clusterb <- cutree(b, k=3)
clusterb
```

```
##        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
```

```
# c
c <- hclust(d=d, method="average")
fviz_dend(c, cex=0.4, k=3, main="Dendrogram using Group Average",
          color_labels_by_k=TRUE, rect=TRUE)
```

```
clusterc <- cutree(c, k=3)
clusterc
```

```
##        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
```

```
# d
pairs(wheat, col = clustera, pch = 20, oma = c(3,3,9,9),
      main ="Nearest Neighbor")
par(xpd = TRUE)
legend("topright", col = unique(clustera), pch=c(20,20,20),
       legend = levels(factor(clustera)),
       cex = 0.6)
```

```
pairs(wheat, col = clusterb, pch = 20, oma = c(3,3,9,9),
      main ="Furthest Neighbor")
par(xpd = TRUE)
legend("topright", col = unique(clusterb), pch=c(20,20,20),
       legend = levels(factor(clusterb)),
       cex = 0.6)
```

```
pairs(wheat, col = clusterc, pch = 20, oma = c(3,3,9,9),
      main ="Group Average")
par(xpd = TRUE)
legend("topright", col = unique(clusterc), pch=c(20,20,20),
       legend = levels(factor(clusterc)),
       cex = 0.6)
```

Code used to solve the question:

```
# Question 2
data <- read_excel("C:/Users/John/Desktop/STAT 445/Data/temphumevap_strip.xlsx", col_names = FALSE)
data <- scale(data, center = TRUE, scale = TRUE)

# a
round(data[1,], 5)
```

```
##     ...1     ...2     ...3     ...4     ...5     ...6     ...7     ...8
## -0.58042 -1.87444 -1.30767 -1.11554 -2.98495 -1.85799 -0.83419  0.04875
##     ...9    ...10
## -0.00315 -0.29909
```

```
round(data[nrow(data),], 5)
```

```
##     ...1     ...2     ...3     ...4     ...5     ...6     ...7     ...8
##  1.37851  1.42751  1.40295  1.03307  0.25174  0.84215 -0.83419 -0.94832
##     ...9    ...10
## -0.90865  1.06040
```

```
# b
set.seed(445)
km <- kmeans(data, centers=4)
length(which(km$cluster==1))
```

```
## [1] 6
```

```
length(which(km$cluster==2))
```

```
## [1] 20
```

```
length(which(km$cluster==3))
```

```
## [1] 9
```

```
length(which(km$cluster==4))
```

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

```
km$cluster
```

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

```
# c
pc <- prcomp(data,center=T,scale=T)
screeplot(pc,type="l", main = "Scree Plot")
```

```r
# d
table <- as.data.frame(pc$x)
table$cluster <- factor(km$cluster)
get_eigenvalue(pc)$variance.percent
```

```
##  [1] 63.0206215 21.1915813  8.3608564  3.1395437  1.3800352  1.1178709
##  [7]  0.7179254  0.6007182  0.3408665  0.1299810
```

```r
plot(x=table$PC1,y=table$PC2,cex=2,pch=19,
     col=c("red","purple","green","cyan")[table$cluster],
     xlab="PC1 (63.02%)",ylab="PC2 (21.19%)")
text(PC2~PC1, labels=rownames(table),data=table, cex=0.9, font=0.5)
legend("bottomleft", col = c("red","green","cyan","purple"),
       pch=c(19,19,19),
       legend = c(1,2,3,4), cex = 0.6)
```

```r
# e
di=dist(data,method="euclidean")
ward <- hclust(d=di, method="ward.D")
fviz_dend(ward, cex=0.5, k=4, main="Dendrogram using Ward's method",
          k_colors = c("red","green","cyan","purple"),
          color_labels_by_k=TRUE, rect=TRUE)
```