Assignment-05

Sumanth Donthula

2023-04-29

Recitation Exercises Chapter 12

Exercise1.a)

Exercise1.b)

We calculate the cluster centroid for each of the K clusters in step 2 of algorithm 12.2, and we then allocate each observation to the cluster whose centroid is closest. The value of RHS will decline with each iteration. This is due to the fact that it represents the sum of the squared deviations of each observation from the mean. As a result, we can see that each iteration of the Kmeans clustering algorithm results in a lower objective.

Exercise 2.a)

```
my_matrix = matrix(c(0, 0.3, 0.4, 0.7,
                     0.3, 0, 0.5, 0.8,
                     0.4, 0.5, 0, 0.45,
                     0.7, 0.8, 0.45, 0), nrow = 4)

my_dist = as.dist(my_matrix)
my_tree = hclust(my_dist, method = "complete")
plot(my_tree)
```

Q2.a)

Step-1

We already have

$$\begin{bmatrix} & 0.3 & 0.44 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.44 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

Step-2 :-

i = 4

We may see 0.3 is minimum dissimilarity so we fuse 1 & 2 to form (1,2) cluster at height 0.3. Now dissimilarity matrix will become.

$$\begin{bmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{bmatrix}$$

i = 3

We see min dissimilarity is 0.45, so we can fuse observations 3 & 4 to form cluster (3,4) at height 0.45. We now have dissimilarity matrix

$$\begin{bmatrix} & 0.8 \\ 0.8 & \end{bmatrix}$$

Figure 1: Solution

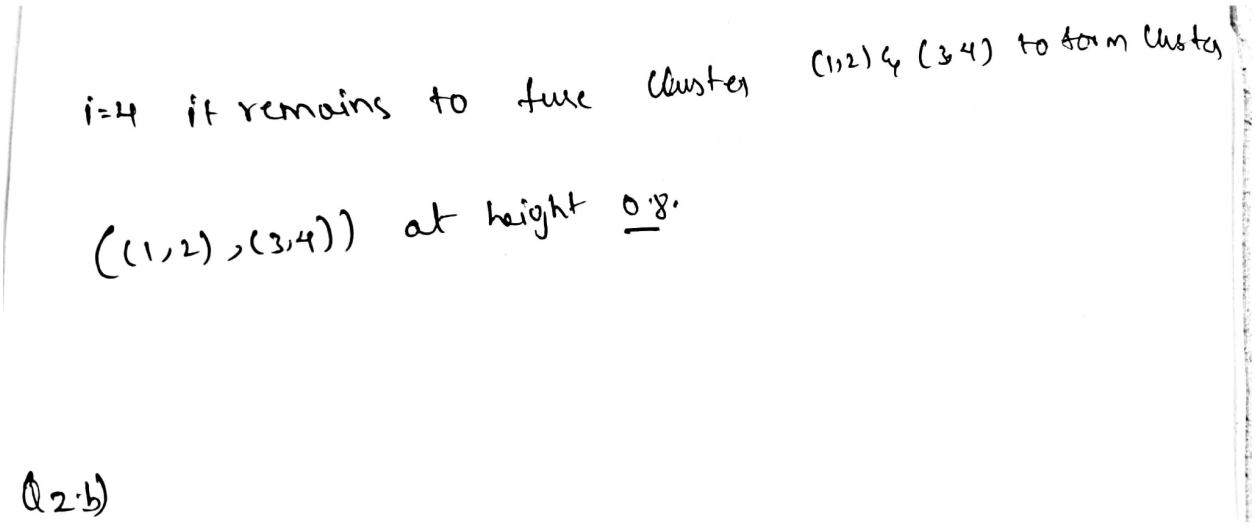
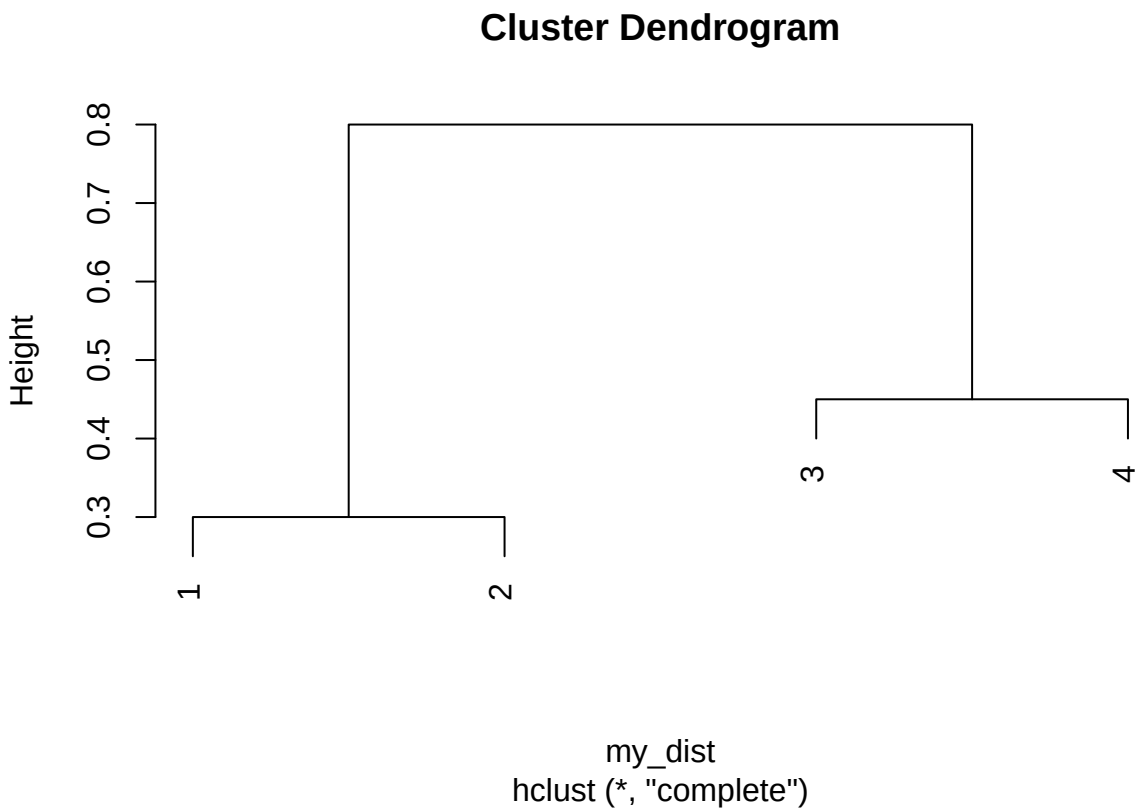


Figure 2: Solution



Exercise 2.b)

```
# Create a distance matrix
dist_matrix = as.dist(matrix(c(0, 0.3, 0.4, 0.7,
                               0.3, 0, 0.5, 0.8,
                               0.4, 0.5, 0, 0.45,
```

Q2.b)

Step-1:-

We already have

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.245 \\ 0.7 & 0.8 & 0.245 & \end{bmatrix}$$

Step-2:-

i=1: We maysee that 0.3 is minimum dissimilarity so, we fuse Observations 1 & 2 to form cluster (1,2) at height 0.3, We now have dissimilarity matrix,

$$\begin{bmatrix} & 0.4 & 0.7 \\ 0.4 & & 0.245 \\ 0.7 & 0.245 & \end{bmatrix}$$

Step3:-

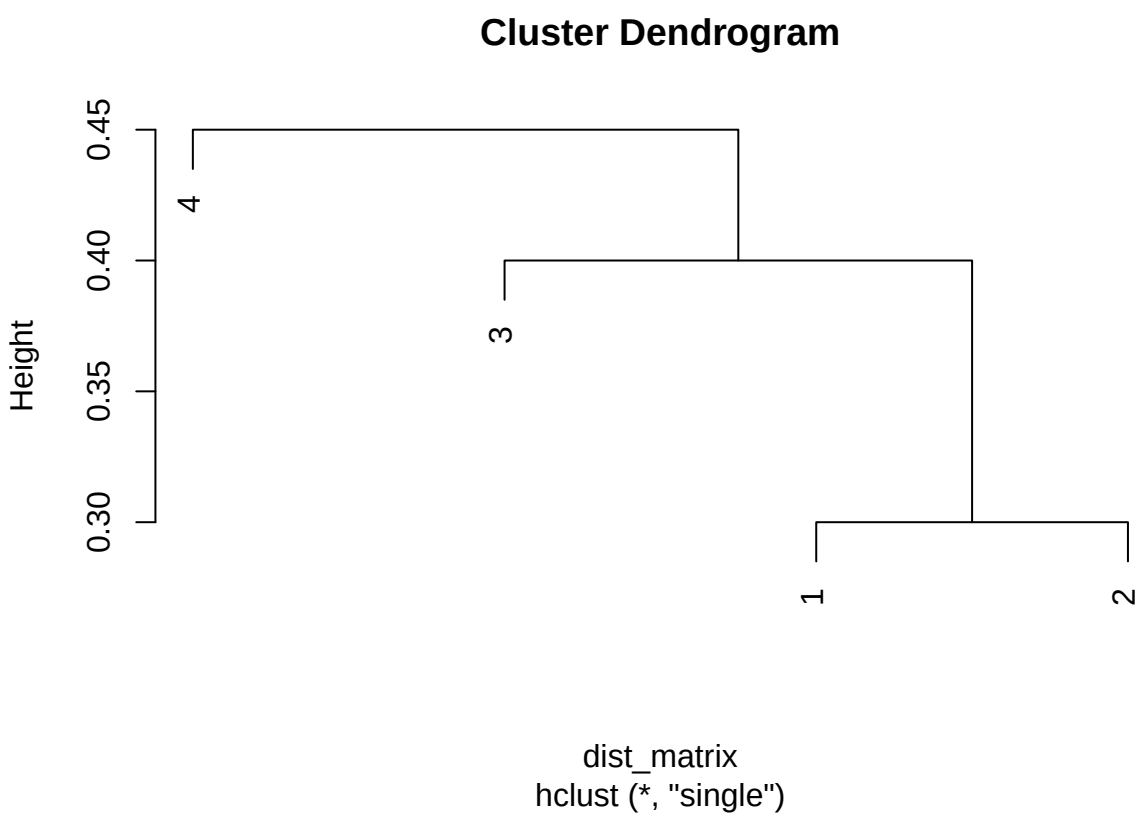
i= 3 We see dissimilarity is 0.4, we cluster (1,2) & Observatio 3 to form ((1,2),3) at height 0.4. We now have dissimilarity matrix.

$$\begin{bmatrix} & 0.245 \\ 0.245 & \end{bmatrix}$$

i=4 it remains to fuse cluster ((1,2),3) & Observation-4 to form cluster (((1,2),3),4) at height 0.245.

Figure 3: Solution

```
                                0.7, 0.8, 0.45, 0), nrow = 4))

# Generate a dendrogram using single linkage method
dendrogram = hclust(dist_matrix, method = "single")

# Plot the dendrogram
plot(dendrogram)
```



Exercise 2.c)

We will have clusters (1,2) and (3,4).

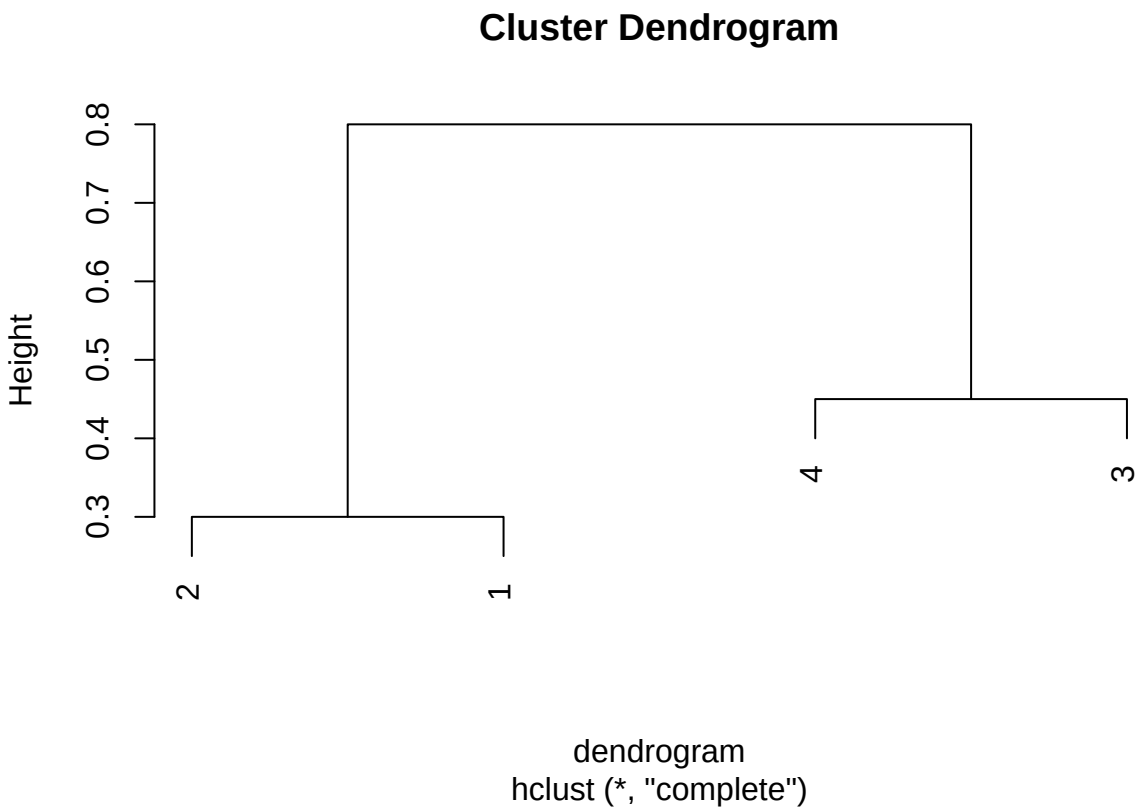
Exercise 2.d)

We have clusters ((1,2),3) and (4)

Exercise 2.e)

The position of the two clusters being fused can be switched at each fusion point in the dendrogram without altering the dendrogram's interpretation, as it is explained in the chapter. Draw a dendrogram that is comparable to the one in (a), with at least two of the leaves in a different position, but the dendrogram's meaning being the same.

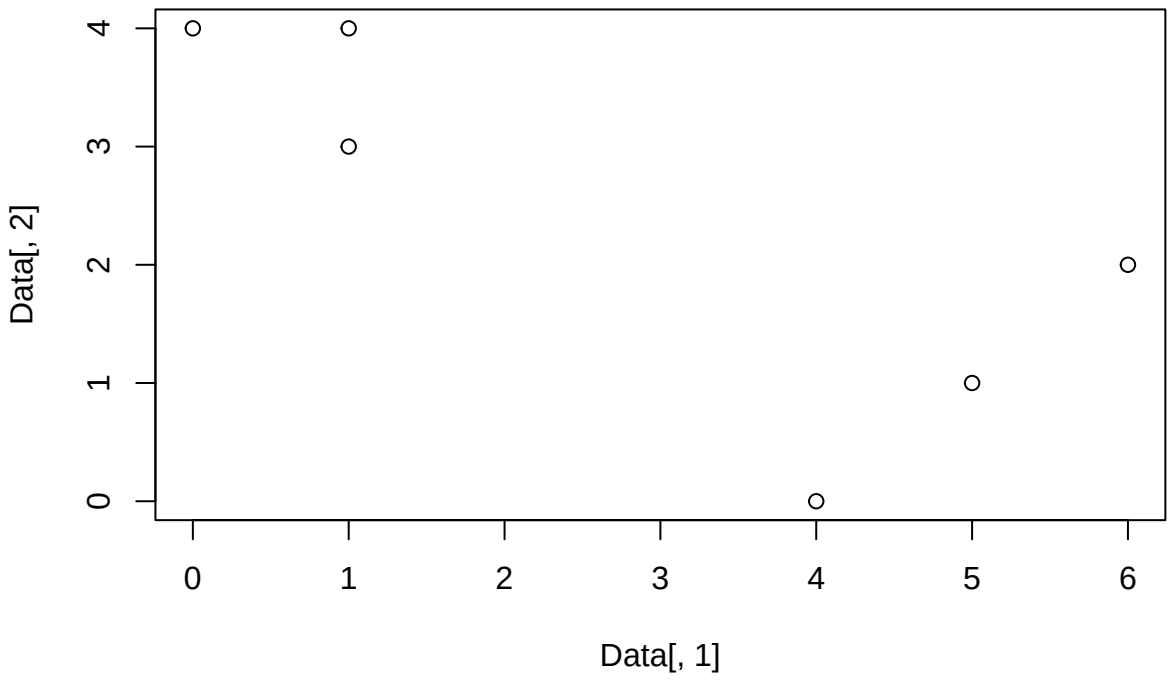
```
dendrogram = as.dist(matrix(c(0,0.3,0.4,0.7,
                              0.3,0,0.5,0.8,
                              0.4,0.5,0,0.45,
                              0.7,0.8,0.45,0), nrow = 4))
hclust_res = hclust(dendrogram, method = "complete")
plot(hclust_res, labels = c(2, 1, 4, 3))
```



Exercise 3.a)

Plotting the observations

```
Data = cbind(c(1,1,0,5,6,4),c(4,3,4,1,2,0))
plot(Data[,1],Data[,2])
```



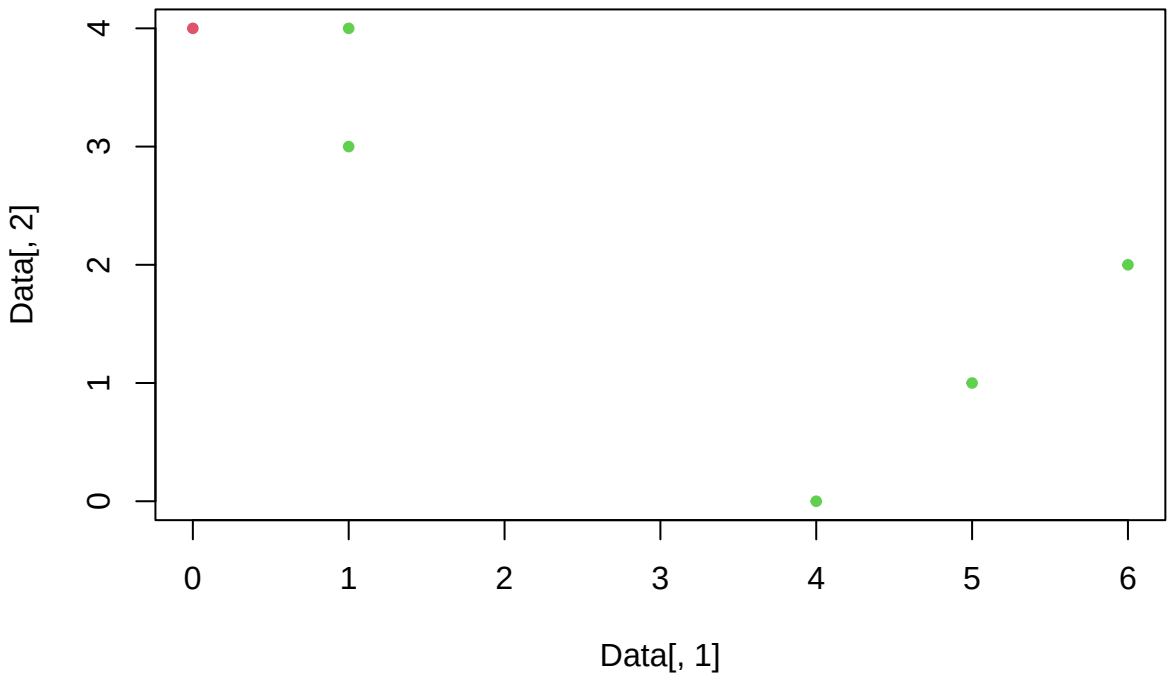
Exercise 3.b)

Randomly assigning a cluster label to each observation

```
clusterLab = sample(2,nrow(Data),replace = T)
clusterLab
```

```
## [1] 2 2 1 2 2 2
```

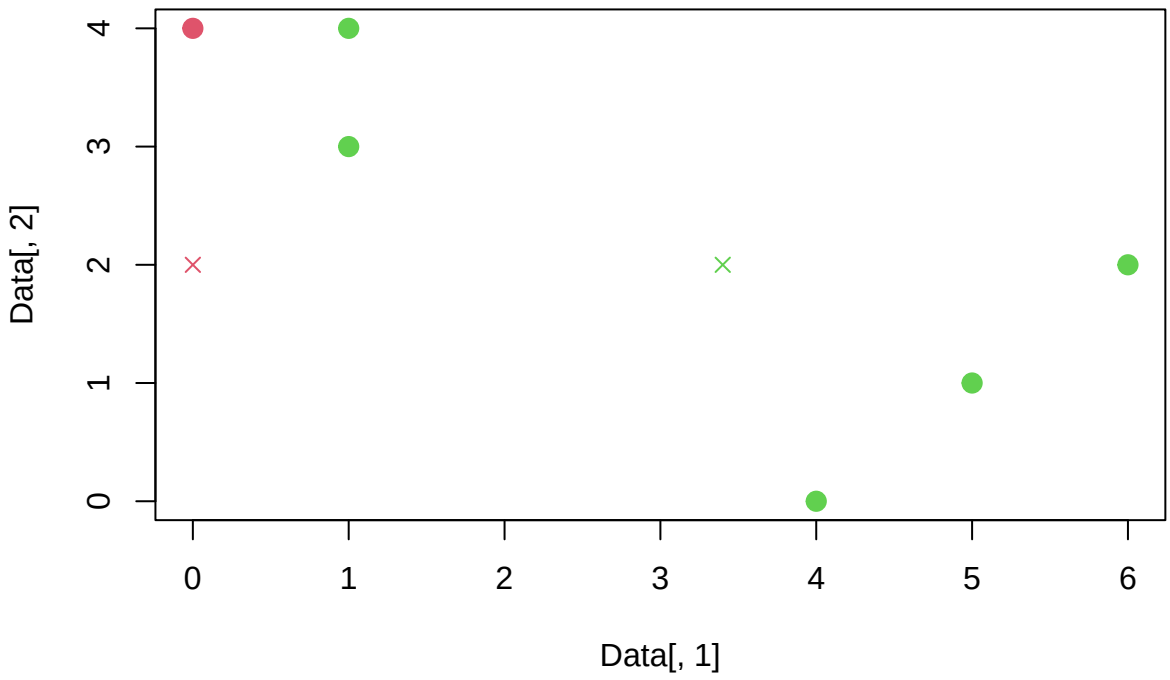
```
plot(Data[,1],Data[,2],col=(clusterLab+1),pch=20)
```



Exercise 3.c)

Computing the centroid for each cluster.

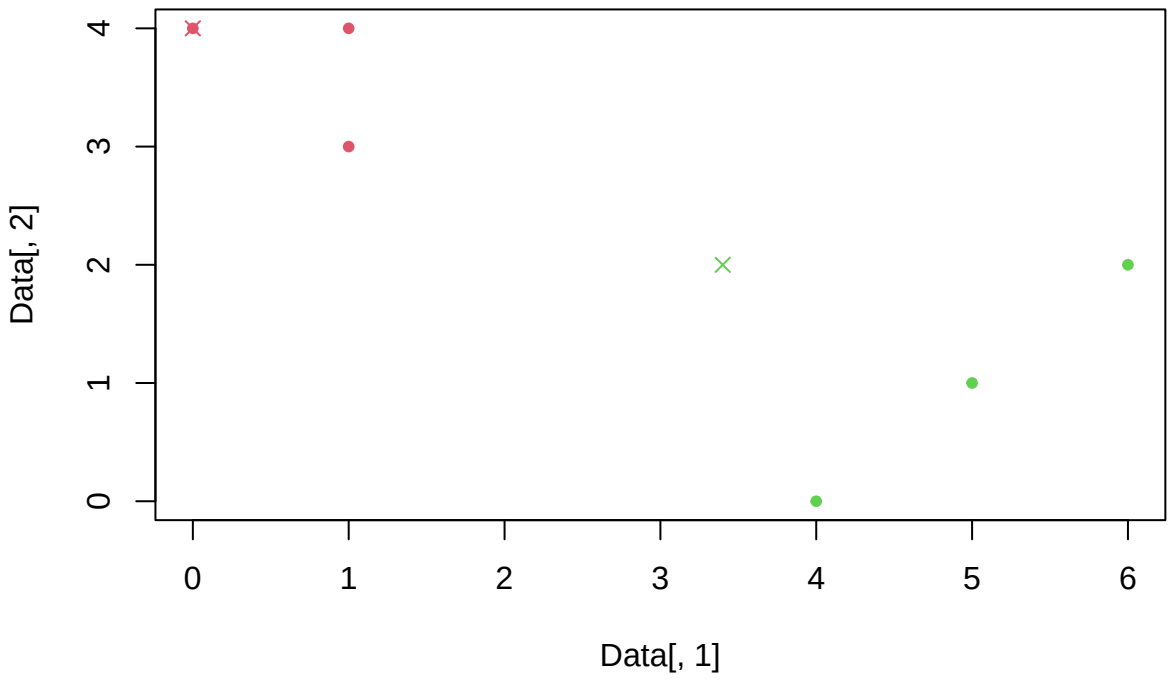
```
cent1 = c(mean(Data[clusterLab == 1, 1]), mean(Data[clusterLab == 1, 2]))
cent2 = c(mean(Data[clusterLab == 2, 1]), mean(Data[clusterLab == 2, 2]))
plot(Data[,1], Data[,2], col=(clusterLab + 1), pch = 20, cex = 2)
points(cent1[1], cent2[2], col = 2, pch = 4)
points(cent2[1], cent2[2], col = 3, pch = 4)
```



Exercise 3.d)

Assigning each observation to the centroid to which it is closest, in terms of Euclidean distance

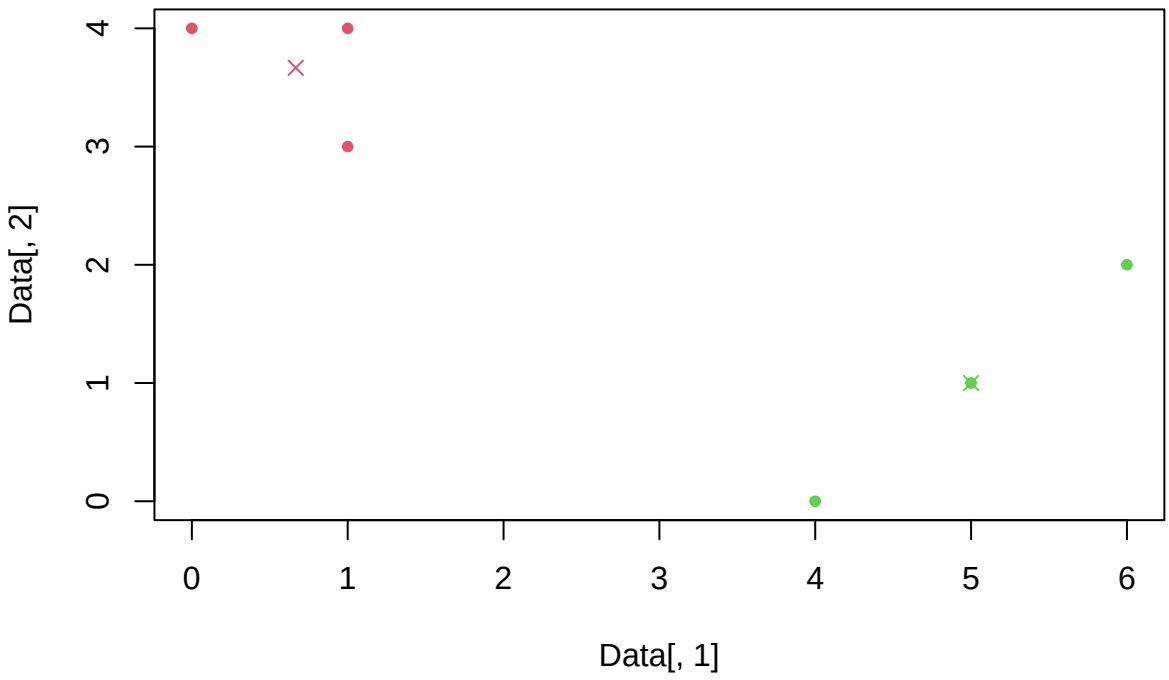
```
clusterLab = c(1,1,1,2,2,2)
plot(Data[, 1], Data[, 2], col = (clusterLab + 1), pch = 20)
points(cent1[1], cent1[2], col = 2, pch = 4)
points(cent2[1], cent2[2], col = 3, pch = 4)
```



Exercise 3.e)

On assigning each observation to the centroid to which it is closest, we see that nothing changes.

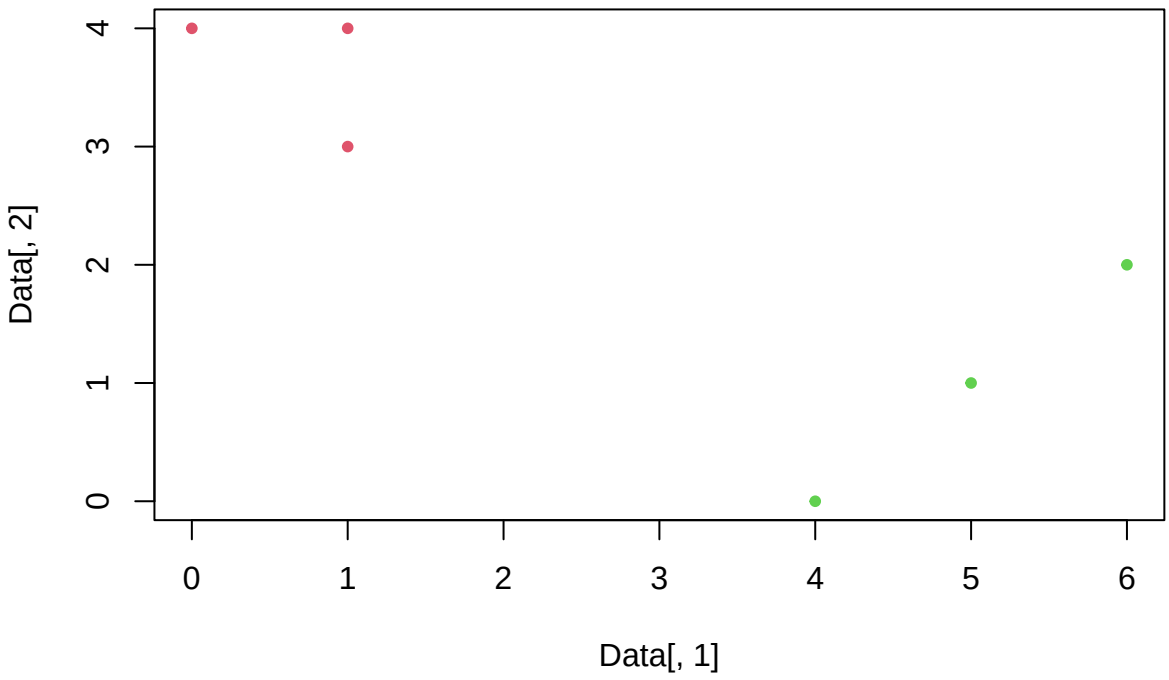
```
cent1 = c(mean(Data[clusterLab == 1, 1]), mean(Data[clusterLab == 1, 2]))
cent2 = c(mean(Data[clusterLab == 2, 1]), mean(Data[clusterLab == 2, 2]))
plot(Data[,1], Data[,2], col=(clusterLab + 1), pch = 20)
points(cent1[1], cent1[2], col = 2, pch = 4)
points(cent2[1], cent2[2], col = 3, pch = 4)
```



Exercise 3.f)

We color the observations in a) according to the clusters obtained:

```
plot(Data[, 1], Data[, 2], col=(clusterLab + 1), pch = 20)
```



Exercise 4.a)

Single linkage utilizes the least inter-observation distance, whereas Complete Linkage utilizes the largest. The fusion will appear higher on the tree than the single linkage in the case of complete linkage. Both linkages will only occur at the same height if all the distances are equal.

Exercise 4.b)

Since there are only single elements, the minimal and maximal distances for a single linkage and a complete linkage are both the same. The fusion will therefore take place at the same height.

Practicum Problems

Problem 1

Data is loaded and required operations are implemented.

Answers to Questions in Problems:

The biplot reveals that Malic is the feature that stands in opposition to Hue. Since the two features' directions are in opposition to one another, it follows that their response profiles and intended meanings will differ in the context created by the data. We estimate the PCA component loadings to support this:

We can see from the below-mentioned loadings that Hue has a loading of 0.297 and Malic has a loading of -0.309, demonstrating their statistical oppositeness.

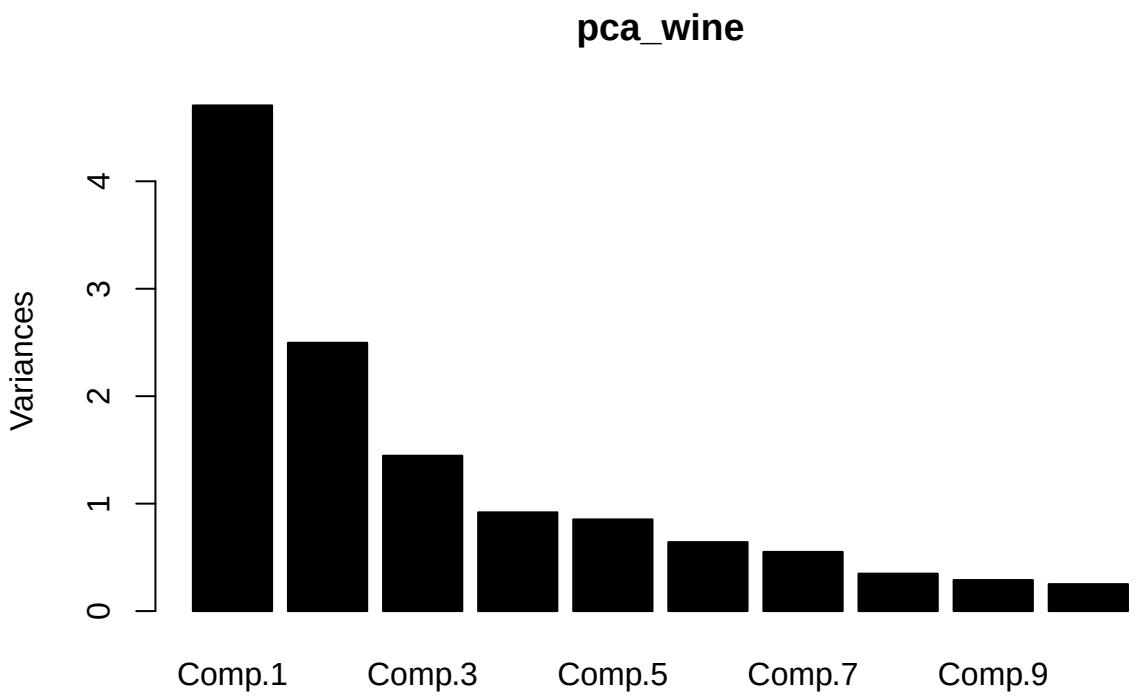
The screeplot shows that component 4 is where the slope changes. Additionally, the variance explained by PC1 is 0.3619885 and by PC2 is 0.1920749, according to the summary.

```
# Load data from URL
url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine/wine.data"
```

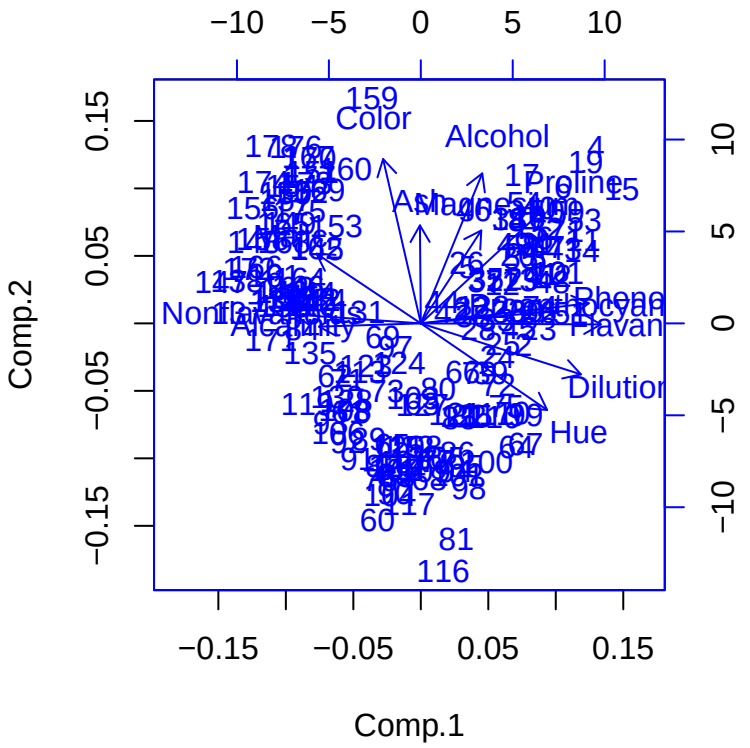
```r
raw_data = read.csv(url, header = FALSE)

# Assign column names to data
colnames(raw_data) = c('Type', 'Alcohol', 'Malic', 'Ash', 'Alcalinity',
                       'Magnesium', 'Phenols', 'Flavanoids', 'Nonflavanoids',
                       'Proanthocyanins', 'Color', 'Hue', 'Dilution', 'Proline')

# Create a copy of the data
wine_data = raw_data

# Perform principal component analysis
pca_wine = princomp(wine_data[,-1], cor = TRUE, scores = TRUE, covmat = NULL)

# Display summary of results
summary(pca_wine)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
## Standard deviation     2.1692972 1.5801816 1.2025273 0.9586313 0.92370351
## Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294
## Cumulative Proportion  0.3619885 0.5540634 0.6652997 0.7359900 0.80162293
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.80103498 0.74231281 0.59033665 0.53747553 0.50090167
## Proportion of Variance 0.04935823 0.04238679 0.02680749 0.02222153 0.01930019
## Cumulative Proportion  0.85098116 0.89336795 0.92017544 0.94239698 0.96169717
##                           Comp.11    Comp.12     Comp.13
## Standard deviation     0.47517222 0.41081655 0.321524394
## Proportion of Variance 0.01736836 0.01298233 0.007952149
## Cumulative Proportion  0.97906553 0.99204785 1.000000000
```

```r
# Plot the PCA results
plot(pca_wine, col = 'black')
```




```
# Display biplot of PCA loadings and scores
biplot(pca_wine, col = 'blue')
```

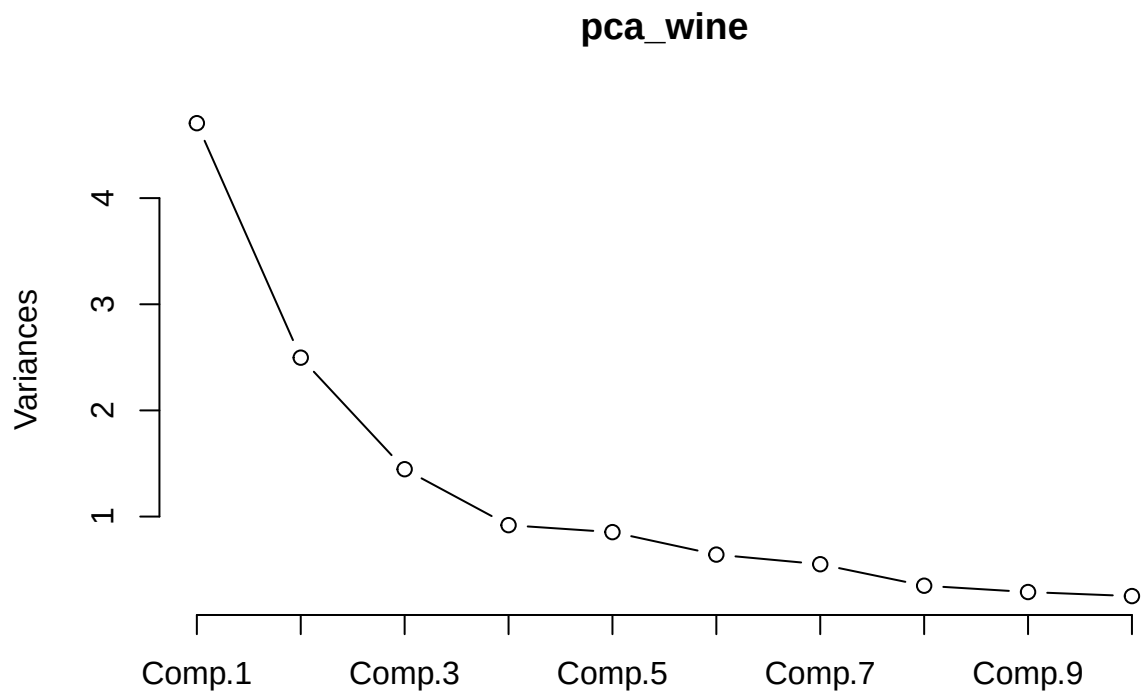



```
# Display the loadings for each principal component
pca_wine$loadings
```

```
##
## Loadings:
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## Alcohol         0.144  0.484  0.207         0.266  0.214         0.396  0.509
## Malic          -0.245  0.225        -0.537         0.537 -0.421
## Ash                    0.316 -0.626  0.214  0.143  0.154  0.149 -0.170 -0.308
## Alcalinity     -0.239        -0.612                -0.101  0.287  0.428  0.200
## Magnesium       0.142  0.300 -0.131  0.352 -0.727        -0.323 -0.156  0.271
## Phenols         0.395        -0.146 -0.198  0.149                -0.406  0.286
## Flavanoids      0.423        -0.151 -0.152  0.109                -0.187
## Nonflavanoids  -0.299        -0.170  0.203  0.501 -0.259 -0.595 -0.233  0.196
## Proanthocyanins 0.313        -0.149 -0.399 -0.137 -0.534 -0.372  0.368 -0.209
## Color                  0.530  0.137                -0.419  0.228
## Hue             0.297 -0.279         0.428  0.174  0.106 -0.232  0.437
## Dilution        0.376 -0.164 -0.166 -0.184  0.101  0.266                0.137
## Proline         0.287  0.365  0.127  0.232  0.158  0.120         0.120 -0.576
##                Comp.10 Comp.11 Comp.12 Comp.13
## Alcohol         0.212   0.226   0.266
## Malic          -0.309          -0.122
## Ash                     0.499          -0.141
## Alcalinity             -0.479
## Magnesium
## Phenols        -0.320  -0.304   0.304  -0.464
```

```
## Flavanoids      -0.163                   0.832
## Nonflavanoids    0.216  -0.117           0.114
## Proanthocyanins  0.134   0.237          -0.117
## Color           -0.291          -0.604
## Hue             -0.522          -0.259
## Dilution         0.524          -0.601  -0.157
## Proline          0.162  -0.539
##
##                Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7 Comp.8 Comp.9
## SS loadings     1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000  1.000
## Proportion Var  0.077  0.077  0.077  0.077  0.077  0.077  0.077  0.077  0.077
## Cumulative Var  0.077  0.154  0.231  0.308  0.385  0.462  0.538  0.615  0.692
##                Comp.10 Comp.11 Comp.12 Comp.13
## SS loadings      1.000   1.000   1.000   1.000
## Proportion Var   0.077   0.077   0.077   0.077
## Cumulative Var   0.769   0.846   0.923   1.000
```

```r
# Display scree plot of eigenvalues
screeplot(pca_wine, type = "lines", col = 'black')
```




```r
# Display summary of PCA results
summary(pca_wine)
```

```
## Importance of components:
##                           Comp.1    Comp.2    Comp.3    Comp.4     Comp.5
```

```
## Standard deviation     2.1692972 1.5801816 1.2025273 0.9586313 0.92370351
## Proportion of Variance 0.3619885 0.1920749 0.1112363 0.0706903 0.06563294
## Cumulative Proportion  0.3619885 0.5540634 0.6652997 0.7359900 0.80162293
##                            Comp.6     Comp.7     Comp.8     Comp.9    Comp.10
## Standard deviation     0.80103498 0.74231281 0.59033665 0.53747553 0.50090167
## Proportion of Variance 0.04935823 0.04238679 0.02680749 0.02222153 0.01930019
## Cumulative Proportion  0.85098116 0.89336795 0.92017544 0.94239698 0.96169717
##                           Comp.11    Comp.12     Comp.13
## Standard deviation     0.47517222 0.41081655 0.321524394
## Proportion of Variance 0.01736836 0.01298233 0.007952149
## Cumulative Proportion  0.97906553 0.99204785 1.000000000
```

Problem 2

Data is loaded and required operations are implemented.

Answers to Questions in Problems:

We can see that the variables' means and variances are highly varied. So that the k-means method won't be dependent on arbitrary unit value, we shall scale the data.

```
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
dataStatistics=data.frame(Min=apply(USArrests,2,min), Med=apply(USArrests,2,median),
Mean=apply(USArrests,2,mean), SD=apply(USArrests,2,sd), Max=apply(USArrests,2,max))
dataStatistics=round(dataStatistics,1)
head(dataStatistics)
```

```
##           Min   Med  Mean   SD   Max
## Murder    0.8   7.2   7.8  4.4  17.4
## Assault  45.0 159.0 170.8 83.3 337.0
## UrbanPop 32.0  66.0  65.5 14.5  91.0
## Rape      7.3  20.1  21.2  9.4  46.0
```

```
scaledData=as.data.frame(scale(USArrests))
head(scaledData)
```

```
##                Murder   Assault   UrbanPop         Rape
## Alabama    1.24256408 0.7828393 -0.5209066 -0.003416473
## Alaska     0.50786248 1.1068225 -1.2117642  2.484202941
## Arizona    0.07163341 1.4788032  0.9989801  1.042878388
## Arkansas   0.23234938 0.2308680 -1.0735927 -0.184916602
## California 0.27826823 1.2628144  1.7589234  2.067820292
## Colorado   0.02571456 0.3988593  0.8608085  1.864967207
```

```r
kmeansResult2=kmeans(scaledData,2,nstart = 25)
kmeansResult2
```

```
## K-means clustering with 2 clusters of sizes 20, 30
##
## Cluster means:
##      Murder    Assault   UrbanPop       Rape
## 1  1.004934  1.0138274  0.1975853  0.8469650
## 2 -0.669956 -0.6758849 -0.1317235 -0.5646433
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              2              2              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              2              2              1              2              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              2              1              2              1              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              1              2              2
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              2
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              2              2              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              2              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              2              2              2              2
##
## Within cluster sum of squares by cluster:
## [1] 46.74796 56.11445
##  (between_SS / total_SS =  47.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
kmeansResult3=kmeans(scaledData,3,nstart = 25)
kmeansResult3
```

```
## K-means clustering with 3 clusters of sizes 20, 13, 17
##
## Cluster means:
##       Murder    Assault   UrbanPop       Rape
## 1  1.0049340  1.0138274  0.1975853  0.8469650
## 2 -0.9615407 -1.1066010 -0.9301069 -0.9667633
## 3 -0.4469795 -0.3465138  0.4788049 -0.2571398
##
## Clustering vector:
```

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              3              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              1              3              3              1              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              2              1              3              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              2              1              2              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              1              2              1              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              1              2              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              1              2              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              1              3              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              2              2              3
##
## Within cluster sum of squares by cluster:
## [1] 46.74796 11.95246 19.62285
##  (between_SS / total_SS =  60.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeansResult4=kmeans(scaledData,4,nstart = 25)
kmeansResult4
```

```
## K-means clustering with 4 clusters of sizes 16, 8, 13, 13
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1 -0.4894375 -0.3826001  0.5758298 -0.26165379
## 2  1.4118898  0.8743346 -0.8145211  0.01927104
## 3 -0.9615407 -1.1066010 -0.9301069 -0.96676331
## 4  0.6950701  1.0394414  0.7226370  1.27693964
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              4              4              2              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              1              1              4              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              3              4              1              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              3              2              3              4
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              4              3              2              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
```

```
##              3              3              4              3              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              4              4              2              3              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              4              1              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              3              3              1
##
## Within cluster sum of squares by cluster:
## [1] 16.212213  8.316061 11.952463 19.922437
##  (between_SS / total_SS =  71.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeansResult5=kmeans(scaledData,5,nstart = 25)
kmeansResult5
```

```
## K-means clustering with 5 clusters of sizes 10, 11, 12, 7, 10
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1 -0.6286291 -0.4086988  0.9506200 -0.38883734
## 2 -0.1642225 -0.3658283 -0.2822467 -0.11697538
## 3  0.7298036  1.1188219  0.7571799  1.32135653
## 4  1.5803956  0.9662584 -0.7775109  0.04844071
## 5 -1.1727674 -1.2078573 -1.0045069 -1.10202608
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              4              3              3              2              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              1              1              3              4
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              5              3              2              5
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              4              5              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              3              5              4              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              2              3              5              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              4              5              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              1              1              4
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              5              4              3              1              5
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              1              5              5              2
##
```

```
## Within cluster sum of squares by cluster:
## [1]  9.326266  7.788275 18.257332  6.128432  7.443899
##  (between_SS / total_SS =  75.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeansResult6=kmeans(scaledData,6,nstart = 25)
kmeansResult6
```

```
## K-means clustering with 6 clusters of sizes 8, 4, 10, 10, 11, 7
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1  0.8666035  1.2103171  0.8262657  0.84936722
## 2  0.4562038  0.9358314  0.6190084  2.26533514
## 3 -1.1727674 -1.2078573 -1.0045069 -1.10202608
## 4 -0.6286291 -0.4086988  0.9506200 -0.38883734
## 5 -0.1642225 -0.3658283 -0.2822467 -0.11697538
## 6  1.5803956  0.9662584 -0.7775109  0.04844071
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              6              2              1              5              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              4              4              1              6
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              4              3              1              5              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              5              5              6              3              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              4              1              3              6              5
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              5              5              2              3              4
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              6              3              4
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              5              5              4              4              6
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              6              1              4              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              5              4              3              3              5
##
## Within cluster sum of squares by cluster:
## [1] 5.888384 6.257771 7.443899 9.326266 7.788275 6.128432
##  (between_SS / total_SS =  78.1 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
kmeansResult7=kmeans(scaledData,7,nstart = 25)
kmeansResult7
```

```
## K-means clustering with 7 clusters of sizes 7, 1, 7, 11, 8, 3, 13
##
## Cluster means:
##       Murder    Assault    UrbanPop        Rape
## 1  1.5803956  0.9662584 -0.77751086  0.04844071
## 2  0.5078625  1.1068225 -1.21176419  2.48420294
## 3 -0.6958674 -0.5679476  1.12728218 -0.55096728
## 4 -1.1034717 -1.1654231 -0.99194587 -1.04874074
## 5  0.8666035  1.2103171  0.82626566  0.84936722
## 6  0.4389842  0.8788344  1.22926592  2.19237920
## 7 -0.2162425 -0.2611064 -0.04793489 -0.06172647
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              2              5              7              6
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              6              3              7              5              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              4              5              7              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              7              7              1              4              5
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              5              4              1              7
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              4              7              6              4              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              5              5              1              4              7
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              7              7              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              4              1              5              3              4
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              7              7              4              4              7
##
## Within cluster sum of squares by cluster:
## [1]  6.128432  0.000000  5.244931  8.499862  5.888384  1.682387 10.860162
##  (between_SS / total_SS =  80.5 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```r
kmeansResult8=kmeans(scaledData,8,nstart = 25)
kmeansResult8
```

```
## K-means clustering with 8 clusters of sizes 7, 5, 7, 8, 10, 3, 1, 9
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
```

```
## 1  1.5803956  0.9662584 -0.7775109  0.04844071
## 2 -1.1176648 -1.2258563 -1.6124616 -1.23334676
## 3 -1.0500985 -1.0736357 -0.4419515 -0.83923219
## 4  0.8666035  1.2103171  0.8262657  0.84936722
## 5 -0.1028582 -0.1651114 -0.1547521 -0.08455771
## 6  0.4389842  0.8788344  1.2292659  2.19237920
## 7  0.5078625  1.1068225 -1.2117642  2.48420294
## 8 -0.6503130 -0.5437584  1.0066563 -0.36760301
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              7              4              5              6
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              6              8              5              4              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              8              3              4              5              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              5              5              1              2              4
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              8              4              3              1              5
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              6              3              8
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              4              4              1              2              8
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              5              5              8              8              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              1              4              8              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              5              8              2              3              5
##
## Within cluster sum of squares by cluster:
## [1] 6.128432 2.196512 2.746293 5.888384 7.897361 1.682387 0.000000 7.319063
##  (between_SS / total_SS =  82.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeansResult9=kmeans(scaledData,9,nstart = 25)
kmeansResult9
```

```
## K-means clustering with 9 clusters of sizes 8, 4, 7, 5, 3, 12, 7, 3, 1
##
## Cluster means:
##       Murder    Assault    UrbanPop        Rape
## 1  0.8666035  1.2103171  0.82626566  0.84936722
## 2  1.6099149  0.6028487 -0.33092078  0.29819404
## 3 -0.6958674 -0.5679476  1.12728218 -0.55096728
## 4 -1.1176648 -1.2258563 -1.61246159 -1.23334676
## 5  0.4389842  0.8788344  1.22926592  2.19237920
## 6 -0.1675273 -0.2141089 -0.03154916 -0.02476943
## 7 -1.0500985 -1.0736357 -0.44195146 -0.83923219
```

```
## 8  1.5410366  1.4508047 -1.37296430 -0.28456373
## 9  0.5078625  1.1068225 -1.21176419  2.48420294
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              9              1              6              5
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              5              3              6              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              7              1              6              7
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              6              6              2              4              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              1              7              8              6
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              7              7              5              7              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              8              4              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              6              6              3              3              8
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              4              2              1              3              4
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              6              6              4              7              6
##
## Within cluster sum of squares by cluster:
## [1] 5.888384 1.405705 5.244931 2.196512 1.682387 9.890427 2.746293 1.038324
## [9] 0.000000
##  (between_SS / total_SS =  84.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeansResult10=kmeans(scaledData,10,nstart = 25)
kmeansResult10
```

```
## K-means clustering with 10 clusters of sizes 5, 3, 8, 3, 6, 4, 1, 4, 7, 9
##
## Cluster means:
##          Murder    Assault   UrbanPop       Rape
## 1  -1.117664812 -1.2258563 -1.6124616 -1.2333468
## 2   1.541036610  1.4508047 -1.3729643 -0.2845637
## 3   0.866603499  1.2103171  0.8262657  0.8493672
## 4   0.438984207  0.8788344  1.2292659  2.1923792
## 5  -1.156695834 -1.1290614 -0.3712208 -0.8931230
## 6   0.008494987 -0.3421022 -0.8145211 -0.4571668
## 7   0.507862482  1.1068225 -1.2117642  2.4842029
## 8   1.609914885  0.6028487 -0.3309208  0.2981940
## 9  -0.695867374 -0.5679476  1.1272822 -0.5509673
## 10 -0.272757970 -0.2157755  0.2236843  0.1128385
##
## Clustering vector:
```

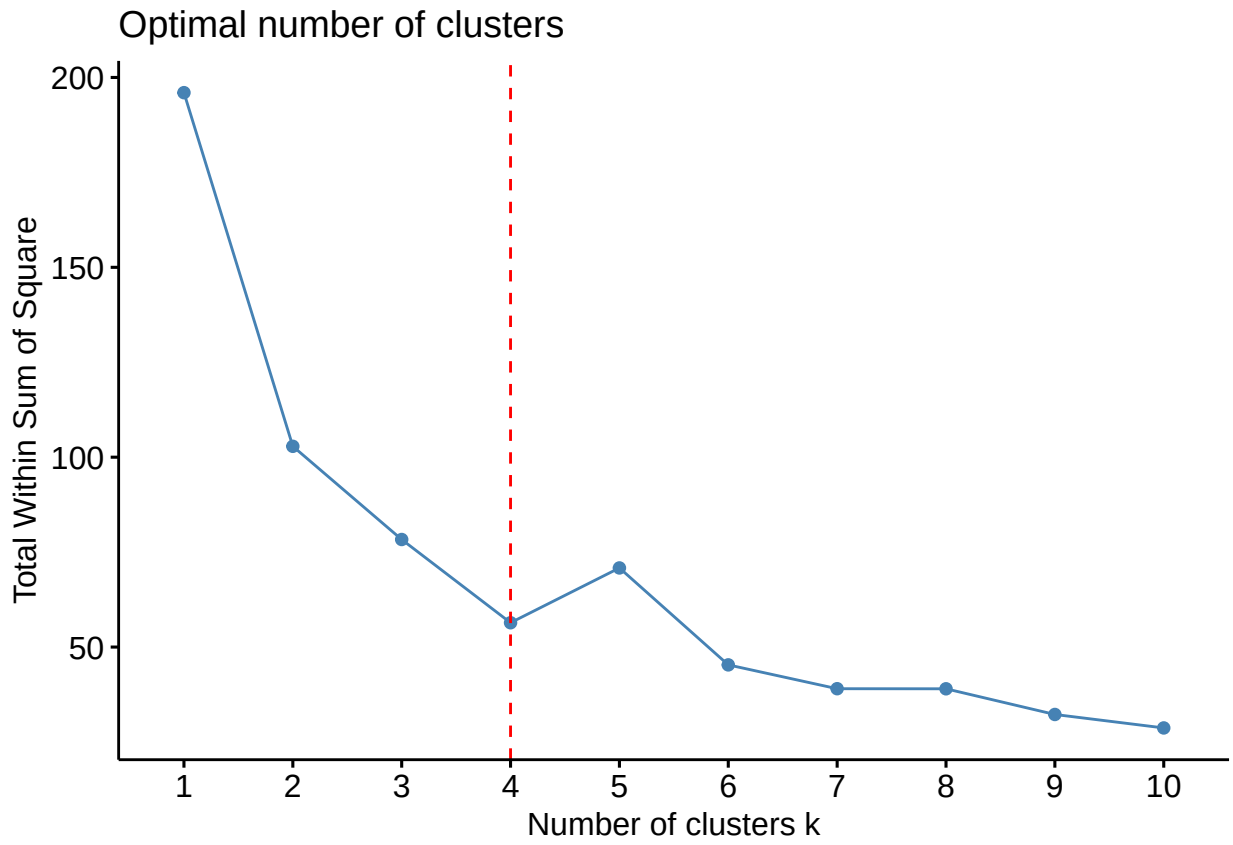
```
##        Alabama         Alaska        Arizona       Arkansas     California
##              8              7              3              6              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              9             10              3              8
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              9              5              3             10              5
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##             10              6              8              1              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              9              3              5              2             10
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              6              5              4              5              9
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              2              1             10
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##             10             10              9              9              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              1              8              3              9              1
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##             10             10              1              5              6
##
## Within cluster sum of squares by cluster:
##  [1] 2.196512 1.038324 5.888384 1.682387 1.807927 1.537684 0.000000 1.405705
##  [9] 5.244931 5.381629
##  (between_SS / total_SS =  86.6 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
library("factoextra")
```

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

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

```
fviz_nbclust(scaledData, kmeans, method = "wss") + geom_vline(xintercept = 4, linetype = 2, col='red')
```

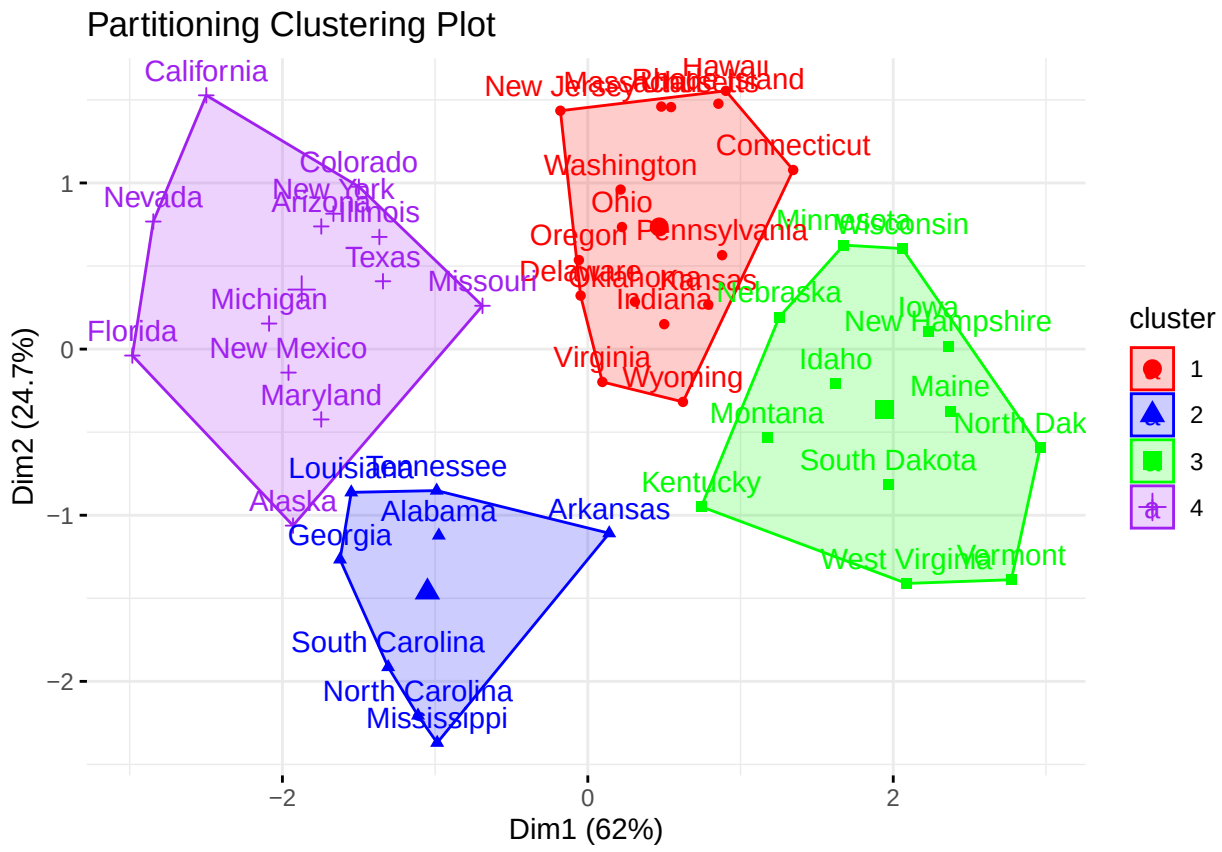



```
kmeansResult4=kmeans(scaledData,4,nstart = 25)
kmeansResult4
```

```
## K-means clustering with 4 clusters of sizes 16, 8, 13, 13
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1 -0.4894375 -0.3826001  0.5758298 -0.26165379
## 2  1.4118898  0.8743346 -0.8145211  0.01927104
## 3 -0.9615407 -1.1066010 -0.9301069 -0.96676331
## 4  0.6950701  1.0394414  0.7226370  1.27693964
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              4              4              2              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              1              1              4              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              3              4              1              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              3              2              3              4
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              1              4              3              2              4
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              4              3              1
##     New Mexico       New York North Carolina   North Dakota           Ohio
```

```
##              4              4              2              3              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              1              1              2
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              2              4              1              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              3              3              1
##
## Within cluster sum of squares by cluster:
## [1] 16.212213  8.316061 11.952463 19.922437
##  (between_SS / total_SS =  71.2 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
fviz_cluster(kmeansResult4, data = scaledData,palette = c("red", "blue", "green", "purple"), ggtheme = t
```



Problem-3

Data is loaded and required functions are implemented.

Answers to Questions in Problems:

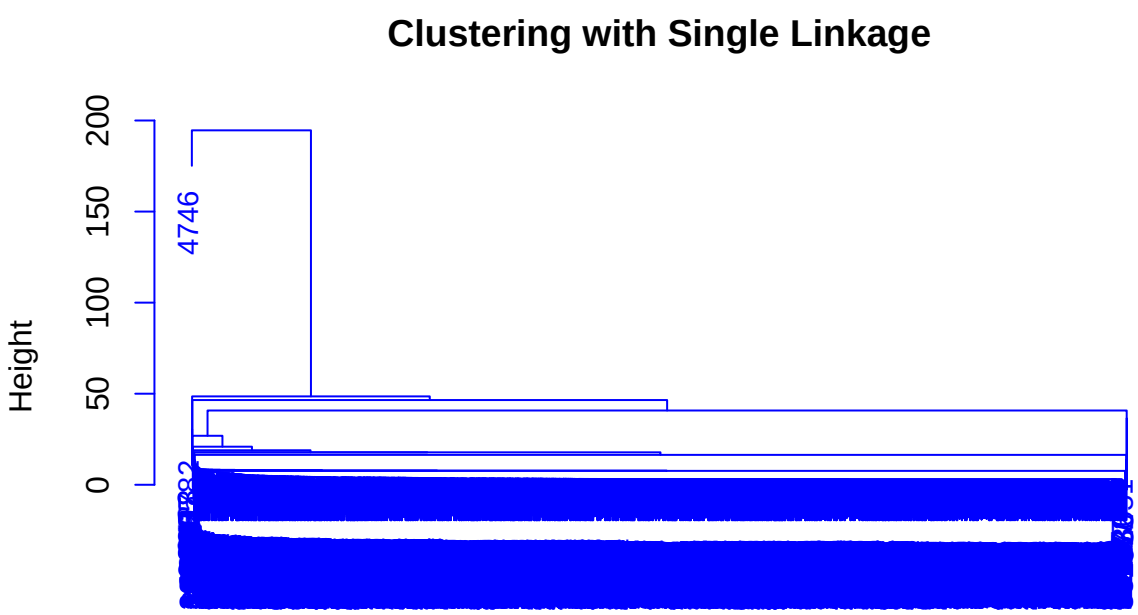
We will scale the observations as there are different types of features like pH, density and alcohol.

We can see that residual sugar, total sulfur dioxide, and free sulfur dioxide are the features that differ the most. The complete linkage is more stable because the differences are comparably less.

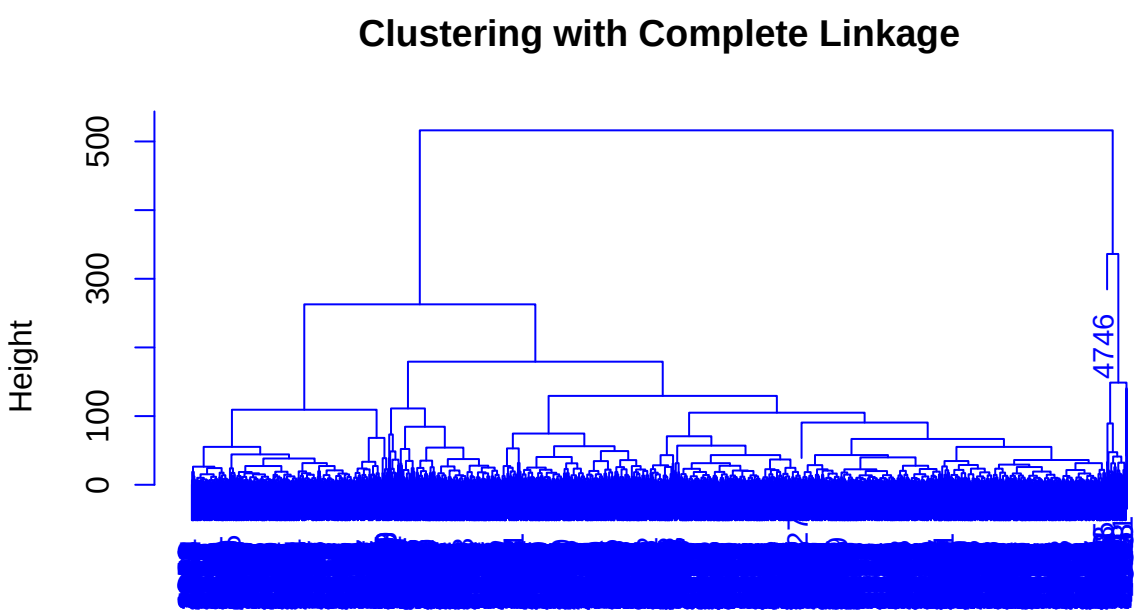
```r
# Set URL for white wine dataset
white_wine_url = "https://archive.ics.uci.edu/ml/machine-learning-databases/wine-quality/winequality-wh

# Read in the white wine dataset
white_raw_data = read.csv(white_wine_url, header=TRUE, sep=";")
summary(white_raw_data)
```

```
##  fixed.acidity    volatile.acidity  citric.acid     residual.sugar
##  Min.   : 3.800   Min.   :0.0800   Min.   :0.0000   Min.   : 0.600
##  1st Qu.: 6.300   1st Qu.:0.2100   1st Qu.:0.2700   1st Qu.: 1.700
##  Median : 6.800   Median :0.2600   Median :0.3200   Median : 5.200
##  Mean   : 6.855   Mean   :0.2782   Mean   :0.3342   Mean   : 6.391
##  3rd Qu.: 7.300   3rd Qu.:0.3200   3rd Qu.:0.3900   3rd Qu.: 9.900
##  Max.   :14.200   Max.   :1.1000   Max.   :1.6600   Max.   :65.800
##    chlorides       free.sulfur.dioxide total.sulfur.dioxide    density
##  Min.   :0.00900   Min.   :  2.00      Min.   :  9.0        Min.   :0.9871
##  1st Qu.:0.03600   1st Qu.: 23.00      1st Qu.:108.0        1st Qu.:0.9917
##  Median :0.04300   Median : 34.00      Median :134.0        Median :0.9937
##  Mean   :0.04577   Mean   : 35.31      Mean   :138.4        Mean   :0.9940
##  3rd Qu.:0.05000   3rd Qu.: 46.00      3rd Qu.:167.0        3rd Qu.:0.9961
##  Max.   :0.34600   Max.   :289.00      Max.   :440.0        Max.   :1.0390
##        pH          sulphates         alcohol         quality
##  Min.   :2.720   Min.   :0.2200   Min.   : 8.00   Min.   :3.000
##  1st Qu.:3.090   1st Qu.:0.4100   1st Qu.: 9.50   1st Qu.:5.000
##  Median :3.180   Median :0.4700   Median :10.40   Median :6.000
##  Mean   :3.188   Mean   :0.4898   Mean   :10.51   Mean   :5.878
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000
```

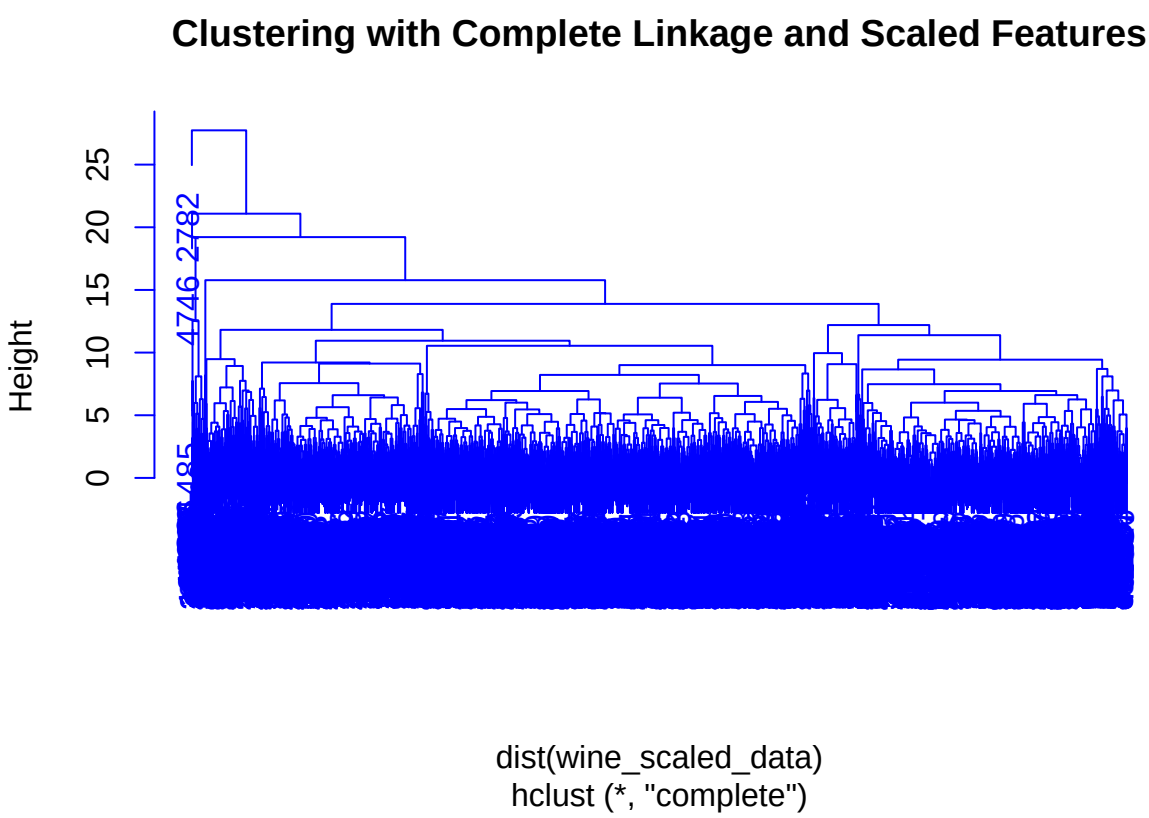
```r
# Subset the white wine dataset to remove the quality column
white_wine_filtered_data = subset(white_raw_data, select=-(quality))

# Perform hierarchical clustering with single linkage
hclust_single_linkage = hclust(dist(white_wine_filtered_data), method="single")
plot(hclust_single_linkage, main="Clustering with Single Linkage", xlab="", sub="", cex=0.9, col='blue'
```



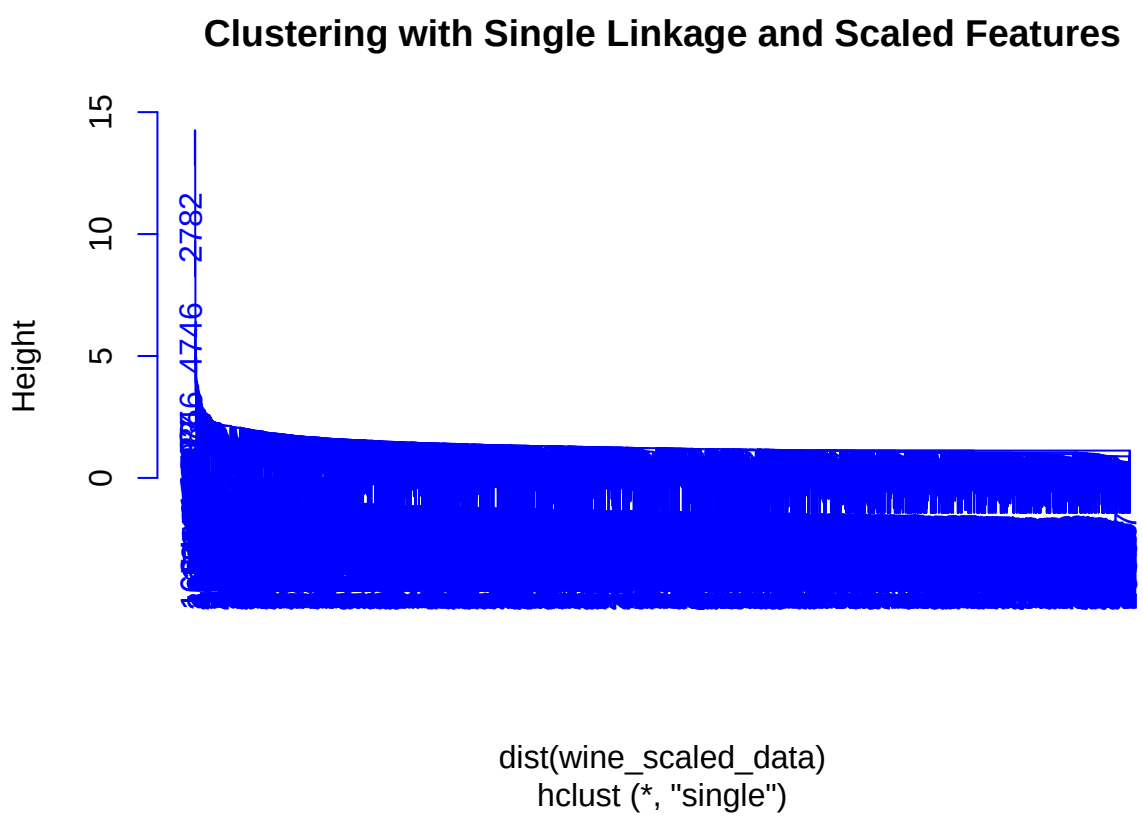

```
# Perform hierarchical clustering with complete linkage
hclust_complete_linkage = hclust(dist(white_wine_filtered_data), method="complete")
plot(hclust_complete_linkage, main="Clustering with Complete Linkage", xlab="", sub="", cex=0.9, col='bl
```




```r
# Scale the white wine dataset
wine_scaled_data = scale(white_wine_filtered_data)

# Plot hierarchical clustering with complete linkage and scaled features
plot(hclust(dist(wine_scaled_data), method="complete"), main="Clustering with Complete Linkage and Scal
```




```
# Plot hierarchical clustering with single linkage and scaled features
plot(hclust(dist(wine_scaled_data), method="single"), main="Clustering with Single Linkage and Scaled F
```




```r
# Perform clustering with complete linkage and 2 clusters
complete_linkage_cutree = cutree(hclust_complete_linkage, k=2)
complete_linkage_data = cbind(white_wine_filtered_data, cluster=complete_linkage_cutree)

# Perform clustering with single linkage and 2 clusters
single_linkage_cutree = cutree(hclust_single_linkage, k=2)
single_linkage_data = cbind(white_wine_filtered_data, cluster=single_linkage_cutree)

# Show the first few rows of the complete linkage data
head(complete_linkage_data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   cluster
```

```
## 1       1
## 2       1
## 3       1
## 4       1
## 5       1
## 6       1
```

```
# Show the first few rows of the single linkage data
head(single_linkage_data)
```

```
##   fixed.acidity volatile.acidity citric.acid residual.sugar chlorides
## 1           7.0             0.27        0.36           20.7     0.045
## 2           6.3             0.30        0.34            1.6     0.049
## 3           8.1             0.28        0.40            6.9     0.050
## 4           7.2             0.23        0.32            8.5     0.058
## 5           7.2             0.23        0.32            8.5     0.058
## 6           8.1             0.28        0.40            6.9     0.050
##   free.sulfur.dioxide total.sulfur.dioxide density   pH sulphates alcohol
## 1                  45                  170  1.0010 3.00      0.45     8.8
## 2                  14                  132  0.9940 3.30      0.49     9.5
## 3                  30                   97  0.9951 3.26      0.44    10.1
## 4                  47                  186  0.9956 3.19      0.40     9.9
## 5                  47                  186  0.9956 3.19      0.40     9.9
## 6                  30                   97  0.9951 3.26      0.44    10.1
##   cluster
## 1       1
## 2       1
## 3       1
## 4       1
## 5       1
## 6       1
```

```
# Compute mean values for each cluster for complete linkage data
aggregate_complete = aggregate(complete_linkage_data, by=list(cluster=complete_linkage_data$cluster), me

# Compute mean values for each cluster for single linkage data
aggregate_single = aggregate(single_linkage_data, by=list(cluster=single_linkage_data$cluster), mean)
```