Assignment 5

Abhiram Ravipati

2024-04-25

Chapter 12 1) a)

$$\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{ij}-x_{i'j}\right)^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2$$

where,

$$\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$$

is the mean of features j present in the cluster C_k

LHS:

$$= \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}\,x_{i'j}^2 + \frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{i'j}^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}\,x_{i'j}^2 \quad \text{—— Equation 1}$$

RHS:

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p}\left(x_{ij}-\bar{x}_{kj}\right)^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 4\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}\,\bar{x}_{kj} + 2\sum_{i\in C_k}\sum_{j=1}^{p}\bar{x}_{kj}^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 4|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2 + 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2$$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - 2|C_k|\sum_{j=1}^{p}\bar{x}_{kj}^2$$

Substituting the value of $\bar{x}_{kj}$

$$= 2\sum_{i\in C_k}\sum_{j=1}^{p} x_{ij}^2 - \frac{2}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p} x_{ij}\, x_{i'j} \text{—— Equation 2}$$

From Equation 1 and 2 we have, LHS = RHS

1)

b) The K-means algorithm as described in Algorithm 12.2, consistently reduces the objective function with each iteration. This is achieved by iteratively recalculating the centroids of clusters and assigning each point to the closest cluster center. As a result, the objective function decreases as the algorithm strives to minimize the overall distance between data points and their respective cluster centroids. This iterative process ensures the formation of well-defined clusters that are sufficiently separated from each other.

2)

a) We have,

Step 1:

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

Step 2: i = 4 We observe that the minimum dissimilarity is 0.3. Consequently, we merge observations 1 and 2 to create cluster (1,2) at a height of 0.3. Subsequently, we obtain the updated dissimilarity matrix.

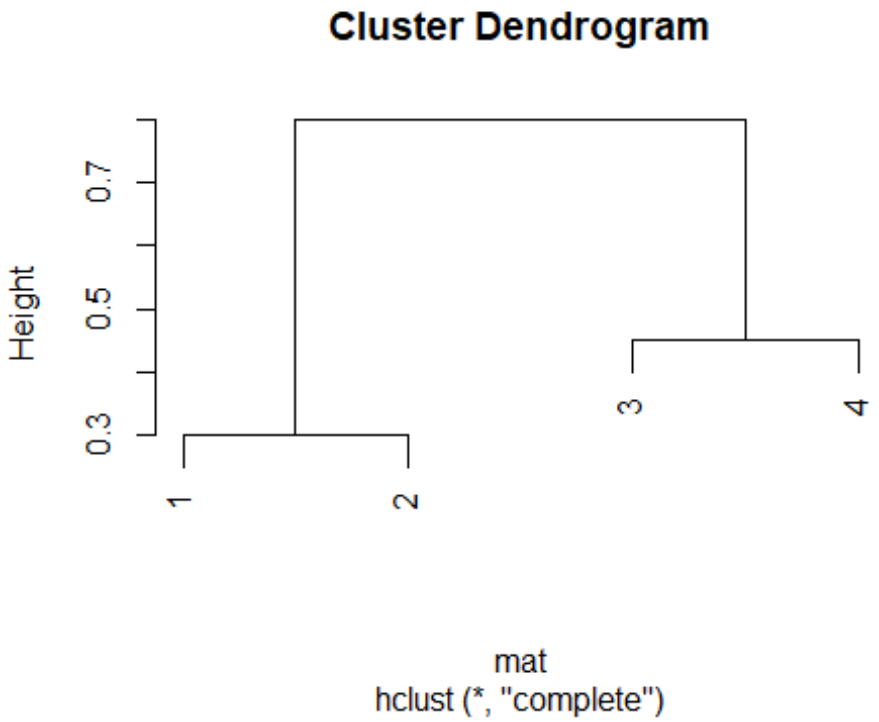
$$\begin{bmatrix} & 0.5 & 0.8 \\ 0.5 & & 0.45 \\ 0.8 & 0.45 & \end{bmatrix}$$

i = 3. After identifying that the maximum dissimilarity is 0.45 we proceed to merge observations 3 and 4 resulting in the formation of cluster (3,4) at a height of 0.45. Consequently, we obtain the updated dissimilarity matrix.

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

i = 4. The final step involves merging clusters (1,2) and (3,4) to create cluster ((1,2), (3,4)) at a height of 0.8.

```r
mat = as.dist(matrix(c(0, 0.3, 0.4, 0.7,
                       0.3, 0, 0.5, 0.8,
                       0.4, 0.5, 0.0, 0.45,
                       0.7, 0.8, 0.45, 0.0), nrow = 4))
plot(hclust(mat, method = "complete"))
```



b) We will utilize Algorithm 10.2 once more to illustrate the various steps leading to the formation of the dendrogram.

We have,

Step 1:

$$\begin{bmatrix} & 0.3 & 0.4 & 0.7 \\ 0.3 & & 0.5 & 0.8 \\ 0.4 & 0.5 & & 0.45 \\ 0.7 & 0.8 & 0.45 & \end{bmatrix}$$

Step 2: i = 4. Observing a minimum dissimilarity of 0.3, we proceed to merge observations 1 and 2 forming cluster (1,2) at a height of 0.3. Subsequently we obtain a new dissimilarity matrix.

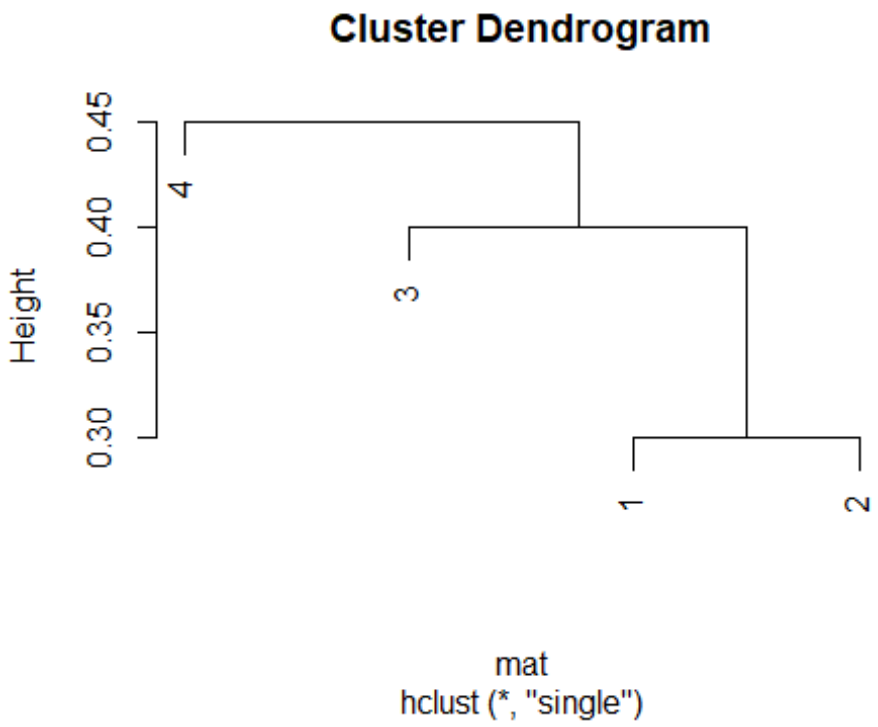
$$\begin{bmatrix} & 0.4 & 0.7 \\ 0.4 & & 0.45 \\ 0.7 & 0.45 & \end{bmatrix}$$

i = 3. Having identified a minimum dissimilarity of 0.4, we combine cluster (1,2) with observation 3 resulting in the formation of cluster ((1,2), 3) at a height of 0.4. Following this merge we update the dissimilarity matrix.

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

i = 4. The final step involves merging clusters ((1,2),3) with observation 4 to create cluster (((1,2),3),4) at a height of 0.45.

```
plot(hclust(mat, method = "single"))
```

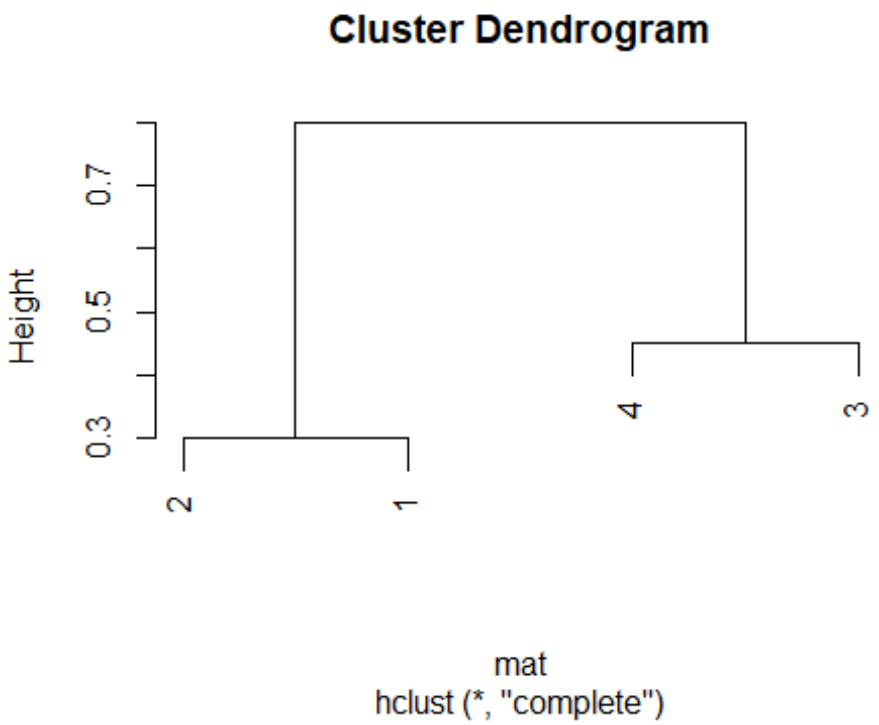


c) In this scenario we will have clusters (1,2) and (3,4).

d) In this scenario we will have clusters ((1,2),3) and (4).

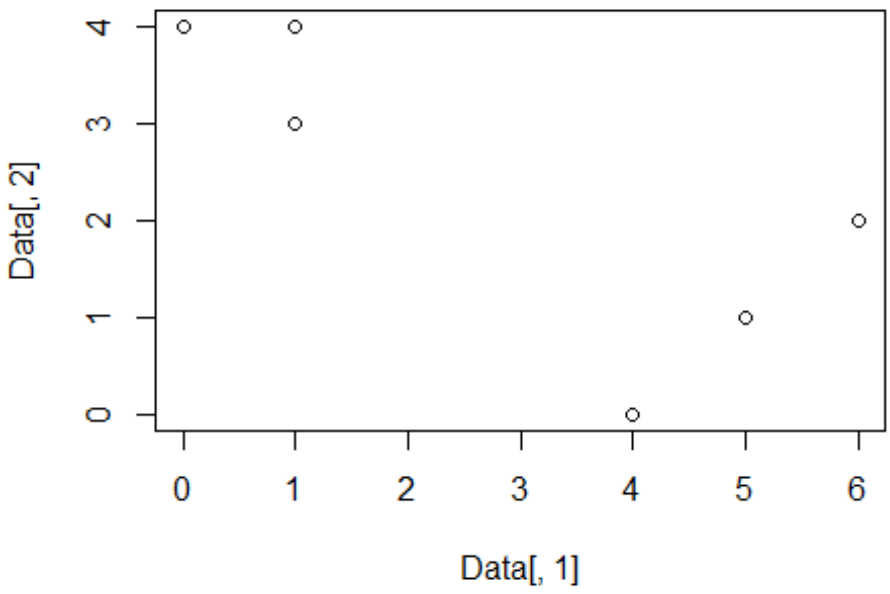
e) The interpretation of the dendrogram remains consistent even if the positions of the two clusters being merged are interchanged at each fusion point, as discussed in the chapter. Generate a dendrogram similar to the one in (a), ensuring that at least two leaf positions are altered while maintaining the intended meaning of the dendrogram.

```
plot(hclust(mat, method = "complete"), labels = c(2, 1, 4, 3))
```



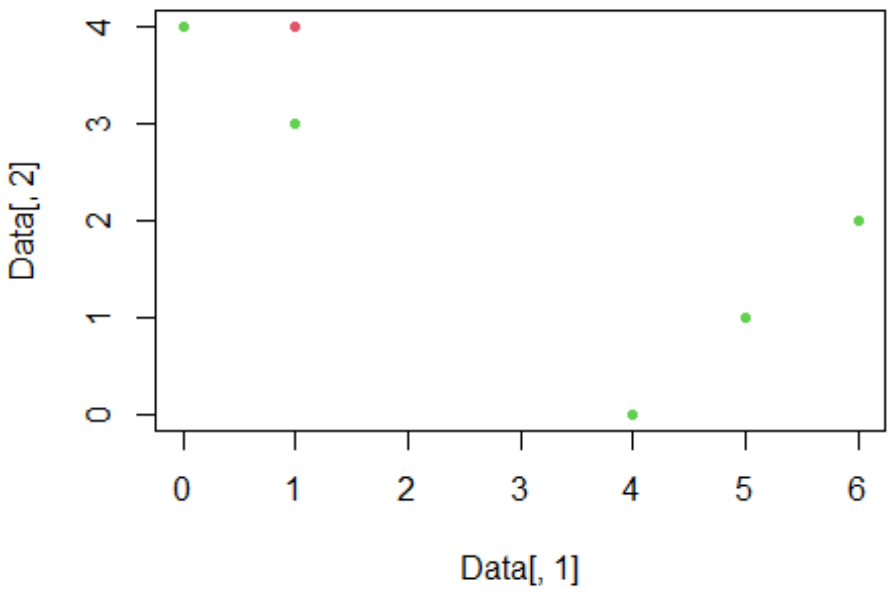
3) a)

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



b)

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

## [1] 1 2 2 2 2 2

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



c) The centroid for the green cluster can be computed using the following method:

$$\bar{x}_{11} = \frac{1}{3}(0 + 4 + 5) = 3$$

and

$$\bar{x}_{12} = \frac{1}{3}(4 + 0 + 1) = \frac{5}{3}$$

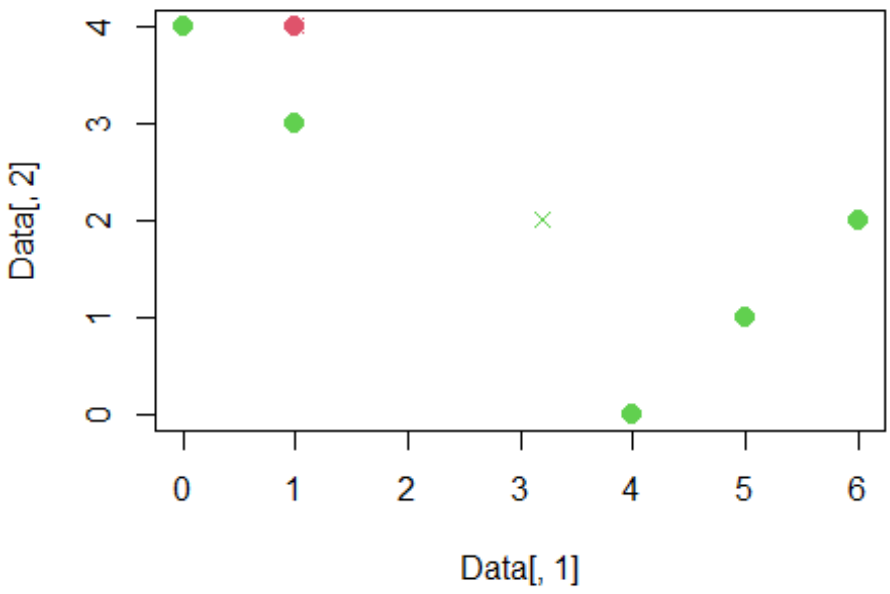
For red clusters:

$$\bar{x}_{21} = \frac{1}{3}(1 + 1 + 6) = \frac{8}{3}$$

and

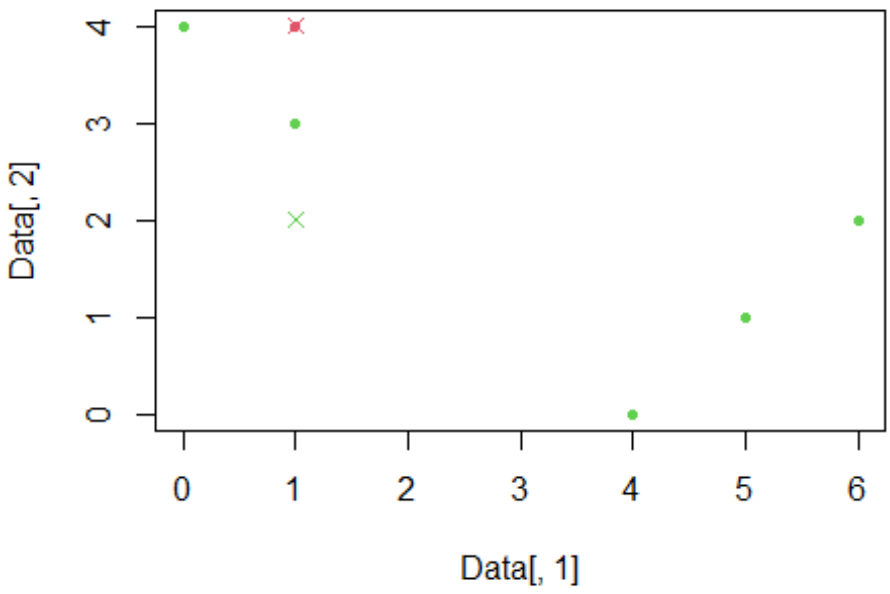
$$\bar{x}_{22} = \frac{1}{3}(2 + 4 + 3) = 3$$

```
c1 = c(mean(Data[clusterlabel == 1, 1]), mean(Data[clusterlabel == 1, 2]))
c2 = c(mean(Data[clusterlabel == 2, 1]), mean(Data[clusterlabel == 2, 2]))
plot(Data[,1], Data[,2], col = (clusterlabel + 1), pch = 20, cex = 2)
points(c1[1], c1[2], col = 2, pch = 4)
points(c2[1], c2[2], col = 3, pch = 4)
```



d) Assign each observation to the centroid that it is closest to in terms of Euclidean distance, and then report the cluster labels for each observation.

```
labels = c(1, 1, 1, 2, 2, 2)
plot(Data[, 1], Data[, 2], col = (clusterlabel + 1), pch = 20)
points(c1[1], c1[2], col = 2, pch = 4)
points(c1[1], c2[2], col = 3, pch = 4)
```



e) Continue steps c and d until the resulting cluster assignments no longer change. The centroid for the green cluster can be computed using the following method:

$$\bar{x}_{11} = \frac{1}{3}(4 + 5 + 6) = 5$$

and

$$\bar{x}_{12} = \frac{1}{3}(0 + 1 + 2) = 1$$

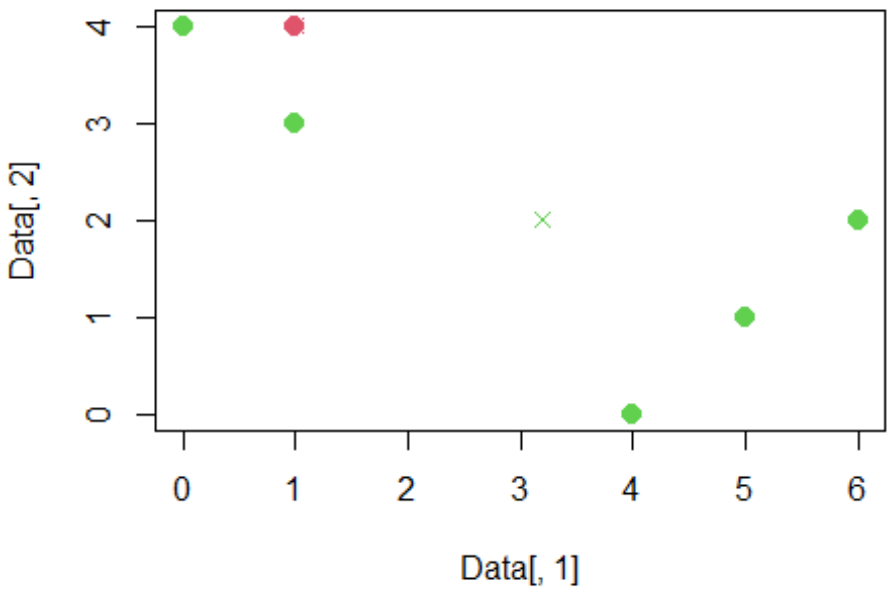
For red clusters:

$$\bar{x}_{21} = \frac{1}{3}(0 + 1 + 1) = \frac{2}{3}$$

and

$$\bar{x}_{22} = \frac{1}{3}(3 + 4 + 4) = \frac{11}{3}$$

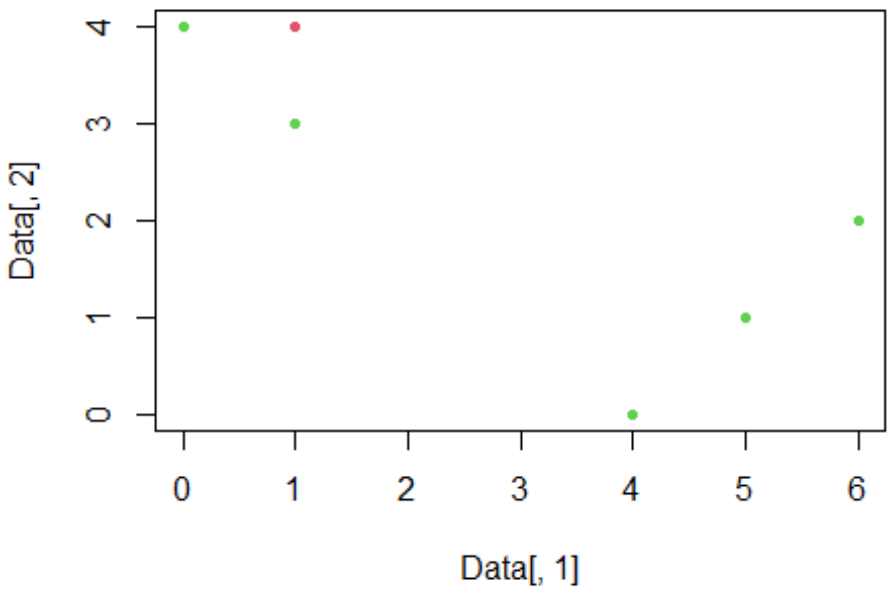
```
c1 = c(mean(Data[clusterlabel == 1, 1]), mean(Data[clusterlabel == 1, 2]))
c2 = c(mean(Data[clusterlabel == 2, 1]), mean(Data[clusterlabel == 2, 2]))
plot(Data[,1], Data[,2], col = (clusterlabel + 1), pch = 20, cex = 2)
points(c1[1], c1[2], col = 2, pch = 4)
points(c2[1], c2[2], col = 3, pch = 4)
```



As assigning each observation to the closest centroid yields no changes, the algorithm is concluded at this stage.As assigning each observation to the closest centroid yields no changes, the algorithm is concluded at this stage.

f)

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



4) a) In the cases where d(1,4)=2, d(1,5)=3, d(2,4)=1, d(2,5)=3, d(3,4)=4, and d(3,5)=1. The complete linkage dissimilarity in this instance is 4, while the single linkage dissimilarity between {1,2,3} and {4,5} is 1. As a result, they would fuse in the Single Linkage dendrogram at a height of 1 and in the Complete Linkage dendrogram at a height of 4. But in the case where all inter-observation distances are equivalent example 2, there would be no difference in the Single Linkage and Complete Linkage between {1,2,3} and {4,5}. They would merge on both dendrograms at the same height in such a scenario. Thus, it is impossible to say for sure whether the fusion will happen at the same height or higher on one dendrogram without precise distance information.

b) They would merge at the same height. For example, if the distance between 5 and 6, represented as d(5,6), is 2, then both the Single Linkage and Complete Linkage dissimilarities between {5} and {6} would be 2. This means that the fusion point for both single and complete linkages would occur at a height of 2.

Practicum Problems

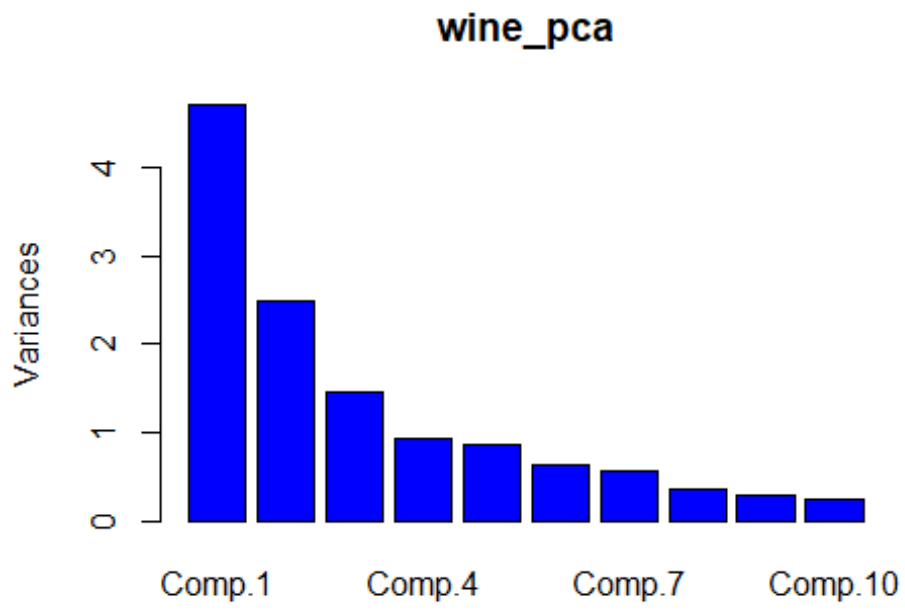
Problem 1:

```
dataset_wine = "https://archive.ics.uci.edu/ml/machine-learning-
databases/wine/wine.data"
dataset_wine_raw = read.csv(dataset_wine, header = FALSE)
colnames(dataset_wine_raw) = c('Type', 'Alcohol', 'Malic', 'Ash',
'Alcalinity', 'Magnesium', 'Phenols', 'Flavanoids', 'Nonflavanoids',
'Proanthocyanins', 'Color', 'Hue', 'Dilution', 'Proline')
```

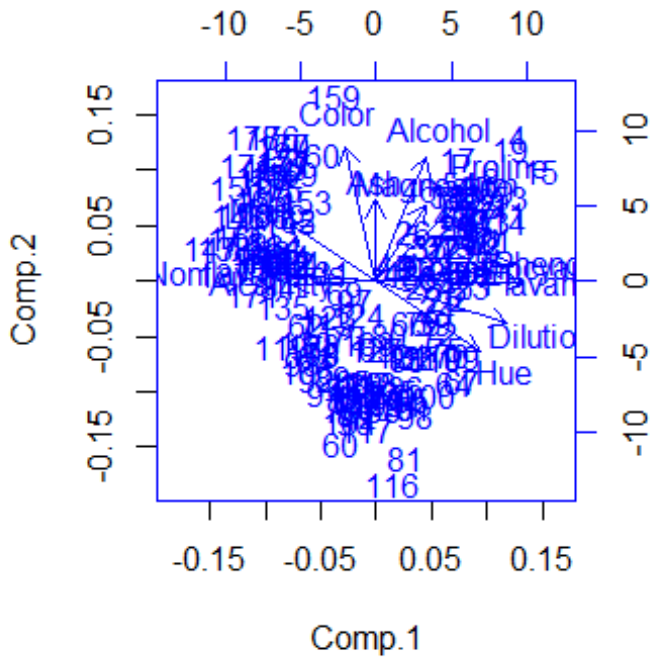
```r
data_wine = dataset_wine_raw
wine_pca = princomp(data_wine[,-1], cor = TRUE, scores = TRUE, covmat = NULL)
summary(wine_pca)
```

```
## 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(wine_pca, col = 'blue')
```

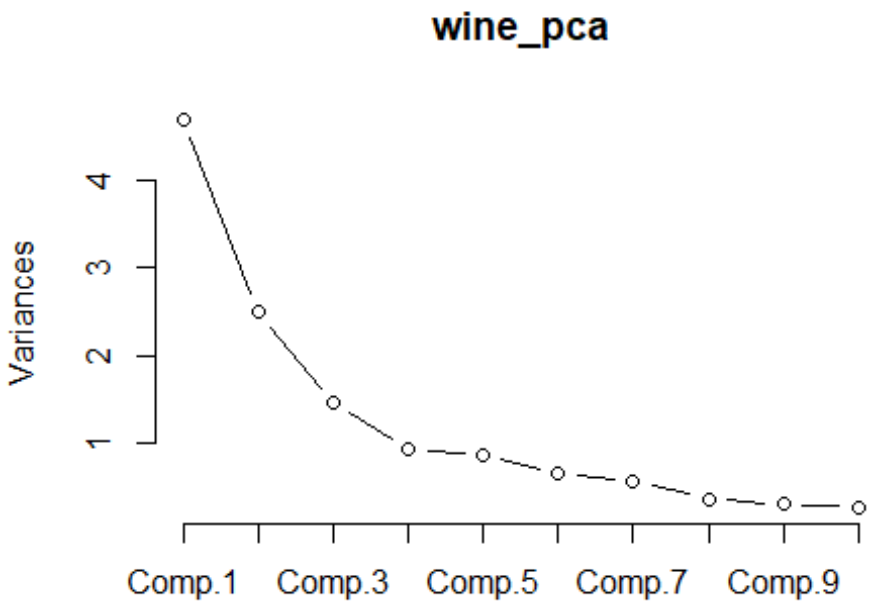



```r
biplot(wine_pca, col = 'blue')
```




```
wine_pca$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

screeplot(wine_pca, type = "lines", col = 'black')
```




```
summary(wine_pca)

## 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
```

The biplot analysis suggests that Malic and Hue exhibit opposing relationships. This opposition indicates that these features likely have different response profiles and underlying meanings within the dataset. This observation is further supported by the PCA component loading estimates. Specifically, the loading for Hue is 0.297, whereas for Malic, it is -0.309, indicating their statistical opposition.

The scree plot shows a change in slope at component 4. Furthermore, the summary indicates that PC1 explains 36.20% of the variance, while PC2 explains 19.21%.

Problem 2:

```
data("USArrests")
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
```

```
data_stats = 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))
data_stats = round(data_stats, 1)
head(data_stats)
```

```
##           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
```

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

```
##                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
```

```
kmeans_2 = kmeans(scaled_data, 2, nstart = 25)
kmeans_2
```

```
## 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"
```

```
kmeans_3 = kmeans(scaled_data, 3, nstart = 25)
kmeans_3
```

```
## K-means clustering with 3 clusters of sizes 17, 20, 13
##
## Cluster means:
##       Murder    Assault   UrbanPop       Rape
## 1 -0.4469795 -0.3465138  0.4788049 -0.2571398
## 2  1.0049340  1.0138274  0.1975853  0.8469650
## 3 -0.9615407 -1.1066010 -0.9301069 -0.9667633
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              2              2              2              1              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              1              1              2              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              1              3              2              1              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              3              2              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
```

```
##             1             2             3             2             2
##        Montana      Nebraska        Nevada New Hampshire    New Jersey
##             3             3             2             3             1
##     New Mexico      New York North Carolina  North Dakota          Ohio
##             2             2             2             3             1
##       Oklahoma        Oregon  Pennsylvania  Rhode Island South Carolina
##             1             1             1             1             2
##   South Dakota     Tennessee         Texas          Utah       Vermont
##             3             2             2             1             3
##       Virginia    Washington West Virginia     Wisconsin       Wyoming
##             1             1             3             3             1
##
## Within cluster sum of squares by cluster:
## [1] 19.62285 46.74796 11.95246
##  (between_SS / total_SS =  60.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans_4 = kmeans(scaled_data, 4, nstart = 25)
kmeans_4
```

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

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

```
kmeans_5 = kmeans(scaled_data, 5, nstart = 25)
kmeans_5
```

```
## K-means clustering with 5 clusters of sizes 13, 11, 12, 7, 7
##
## Cluster means:
##       Murder    Assault    UrbanPop        Rape
## 1 -0.2162425 -0.2611064 -0.04793489 -0.06172647
## 2 -1.1034717 -1.1654231 -0.99194587 -1.04874074
## 3  0.7298036  1.1188219  0.75717991  1.32135653
## 4  1.5803956  0.9662584 -0.77751086  0.04844071
## 5 -0.6958674 -0.5679476  1.12728218 -0.55096728
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              4              3              3              1              3
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              3              5              1              3              4
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              5              2              3              1              2
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              4              2              3
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              5              3              2              4              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              2              1              3              2              5
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              3              3              4              2              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              5              5              4
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              4              3              5              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              1              2              2              1
```

```
##
## Within cluster sum of squares by cluster:
## [1] 10.860162  8.499862 18.257332  6.128432  5.244931
##  (between_SS / total_SS =  75.0 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans_6 = kmeans(scaled_data, 6, nstart = 25)
kmeans_6
```

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

```
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans_7 = kmeans(scaled_data, 7, nstart = 25)
kmeans_7
```

```
## K-means clustering with 7 clusters of sizes 8, 1, 7, 3, 11, 10, 10
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1  0.8666035  1.2103171  0.8262657  0.84936722
## 2  0.5078625  1.1068225 -1.2117642  2.48420294
## 3  1.5803956  0.9662584 -0.7775109  0.04844071
## 4  0.4389842  0.8788344  1.2292659  2.19237920
## 5 -0.1642225 -0.3658283 -0.2822467 -0.11697538
## 6 -0.6286291 -0.4086988  0.9506200 -0.38883734
## 7 -1.1727674 -1.2078573 -1.0045069 -1.10202608
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              3              2              1              5              4
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              4              6              6              1              3
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              6              7              1              5              7
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              5              5              3              7              1
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              6              1              7              3              5
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              5              5              4              7              6
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              1              1              3              7              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              5              5              6              6              3
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              7              3              1              6              7
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              5              6              7              7              5
##
## Within cluster sum of squares by cluster:
## [1] 5.888384 0.000000 6.128432 1.682387 7.788275 9.326266 7.443899
##  (between_SS / total_SS =  80.5 %)
##
## Available components:
##
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans_8 = kmeans(scaled_data, 8, nstart = 25)
kmeans_8
```

```
## K-means clustering with 8 clusters of sizes 7, 10, 5, 7, 1, 3, 8, 9
##
## Cluster means:
##       Murder    Assault   UrbanPop        Rape
## 1  1.5803956  0.9662584 -0.7775109  0.04844071
## 2 -0.1028582 -0.1651114 -0.1547521 -0.08455771
## 3 -1.1176648 -1.2258563 -1.6124616 -1.23334676
## 4 -1.0500985 -1.0736357 -0.4419515 -0.83923219
## 5  0.5078625  1.1068225 -1.2117642  2.48420294
## 6  0.4389842  0.8788344  1.2292659  2.19237920
## 7  0.8666035  1.2103171  0.8262657  0.84936722
## 8 -0.6503130 -0.5437584  1.0066563 -0.36760301
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              5              7              2              6
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              6              8              2              7              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              8              4              7              2              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              2              2              1              3              7
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              8              7              4              1              2
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              4              4              6              4              8
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              7              7              1              3              8
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              2              2              8              8              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              7              8              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              2              8              3              4              2
##
## Within cluster sum of squares by cluster:
## [1] 6.128432 7.897361 2.196512 2.746293 0.000000 1.682387 5.888384
7.319063
##  (between_SS / total_SS =  82.7 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
```

```
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"

kmeans_9 = kmeans(scaled_data, 9, nstart = 25)
kmeans_9

## K-means clustering with 9 clusters of sizes 6, 5, 6, 8, 7, 3, 7, 7, 1
##
## Cluster means:
##        Murder     Assault   UrbanPop        Rape
## 1 -0.34546282 -0.06711651  0.3656939  0.24036311
## 2 -1.11766481 -1.22585634 -1.6124616 -1.23334676
## 3 -1.15669583 -1.12906137 -0.3712208 -0.89312299
## 4  0.86660350  1.21031715  0.8262657  0.84936722
## 5  1.58039562  0.96625839 -0.7775109  0.04844071
## 6  0.43898421  0.87883436  1.2292659  2.19237920
## 7 -0.69586737 -0.56794765  1.1272822 -0.55096728
## 8 -0.04972355 -0.41538414 -0.4912984 -0.32218561
## 9  0.50786248  1.10682252 -1.2117642  2.48420294
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              5              9              4              8              6
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              6              7              1              4              5
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              7              3              4              8              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              8              8              5              2              4
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              7              4              3              5              1
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              8              3              6              3              7
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              4              4              5              2              1
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              1              1              7              7              5
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              2              5              4              7              2
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              8              1              2              3              8
##
## Within cluster sum of squares by cluster:
## [1] 3.814022 2.196512 1.807927 5.888384 6.128432 1.682387 5.244931
3.183515
## [9] 0.000000
##  (between_SS / total_SS =  84.7 %)
##
## Available components:
##
```

```
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

```
kmeans_10 = kmeans(scaled_data, 10, nstart = 25)
kmeans_10
```

```
## K-means clustering with 10 clusters of sizes 7, 1, 7, 6, 5, 6, 4, 8, 3, 3
##
## Cluster means:
##         Murder     Assault   UrbanPop       Rape
## 1  -0.04972355 -0.41538414 -0.4912984 -0.3221856
## 2   0.50786248  1.10682252 -1.2117642  2.4842029
## 3  -0.69586737 -0.56794765  1.1272822 -0.5509673
## 4  -1.15669583 -1.12906137 -0.3712208 -0.8931230
## 5  -1.11766481 -1.22585634 -1.6124616 -1.2333468
## 6  -0.34546282 -0.06711651  0.3656939  0.2403631
## 7   1.60991488  0.60284869 -0.3309208  0.2981940
## 8   0.86660350  1.21031715  0.8262657  0.8493672
## 9   1.54103661  1.45080466 -1.3729643 -0.2845637
## 10  0.43898421  0.87883436  1.2292659  2.1923792
##
## Clustering vector:
##        Alabama         Alaska        Arizona       Arkansas     California
##              7              2              8              1             10
##       Colorado    Connecticut       Delaware        Florida        Georgia
##             10              3              6              8              7
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              4              8              1              4
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              1              1              7              5              8
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              8              4              9              6
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              1              4             10              4              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              8              8              9              5              6
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              6              6              3              3              9
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              5              7              8              3              5
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              1              6              5              4              1
##
## Within cluster sum of squares by cluster:
##  [1] 3.183515 0.000000 5.244931 1.807927 2.196512 3.814022 1.405705
5.888384
##  [9] 1.038324 1.682387
##  (between_SS / total_SS =  86.6 %)
##
```

```
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
"tot.withinss"
## [6] "betweenss"    "size"         "iter"         "ifault"
```

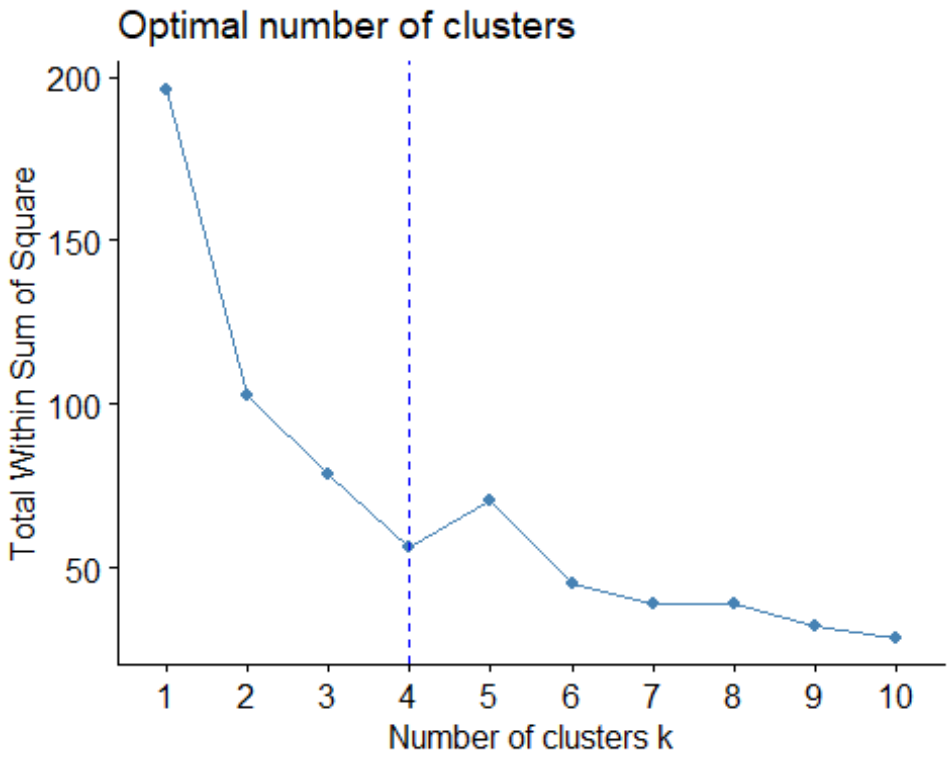
```
library("factoextra")
```

```
## Warning: package 'factoextra' was built under R version 4.3.3
```

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

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

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

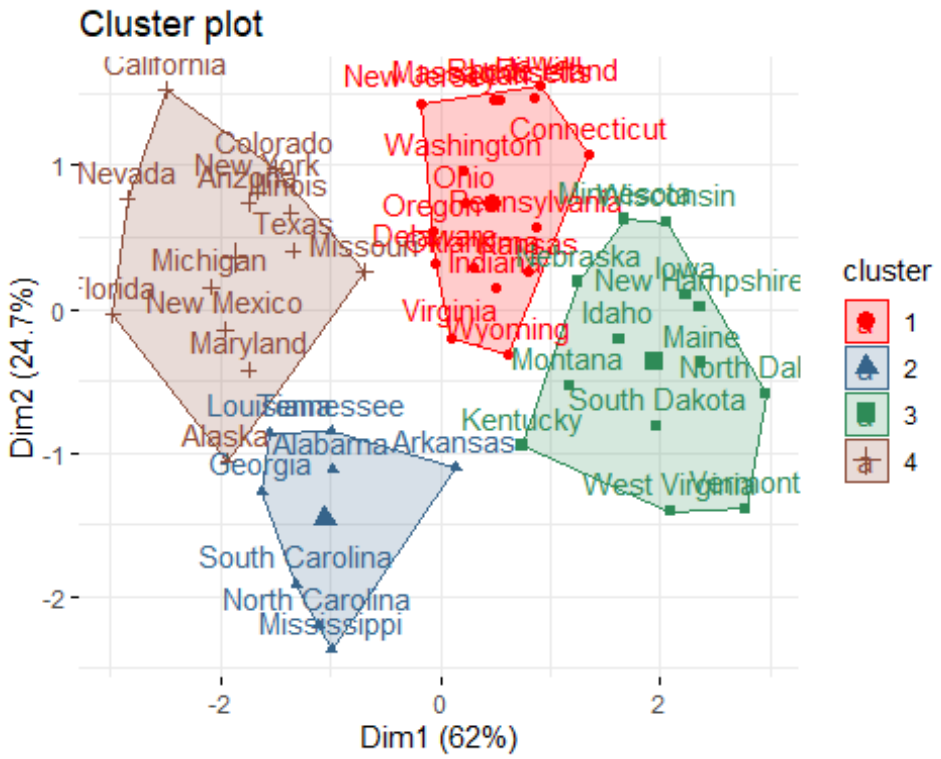



```
kmeans_4_result = kmeans(scaled_data, 4, nstart = 25)
kmeans_4_result
```

```
## 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(kmeans_4_result, data = scaled_data, palette = c("red",
"steelblue4", "seagreen4", "salmon4"), ggtheme = theme_minimal())
```

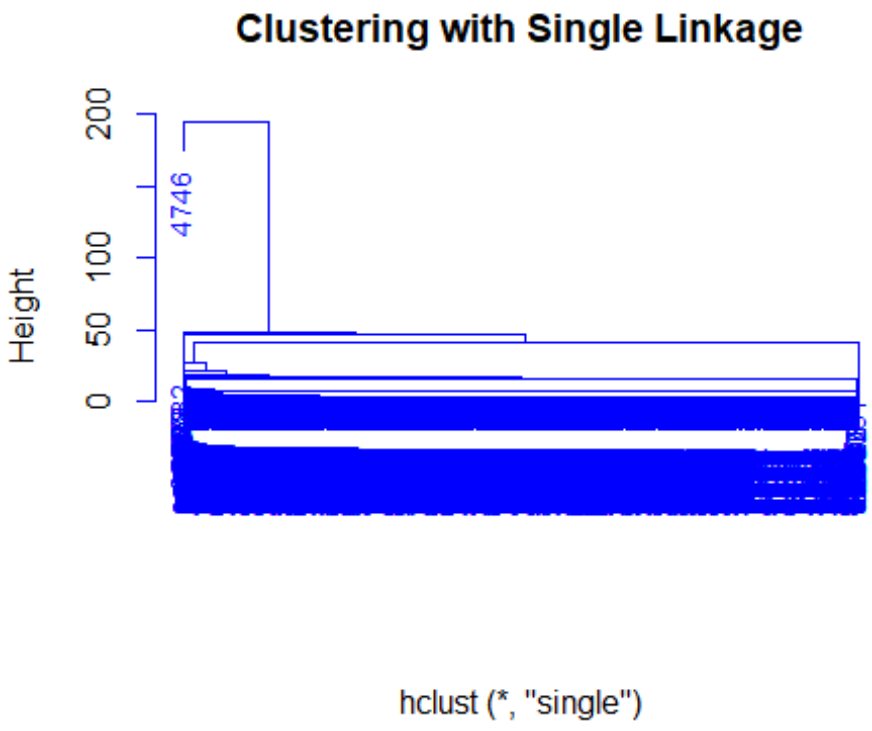


Problem 3:

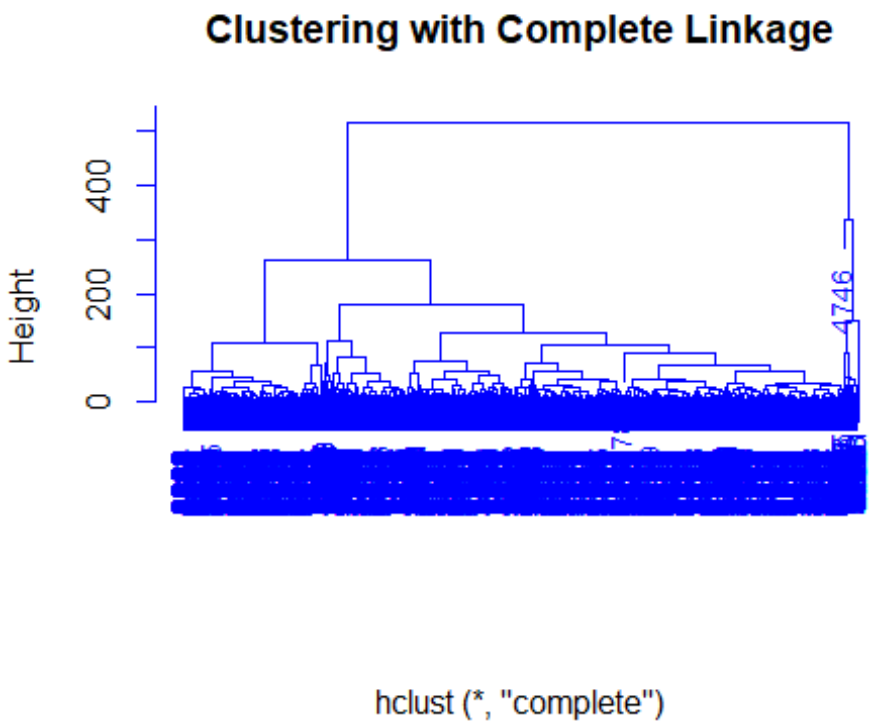
```
wine_white_url = "https://archive.ics.uci.edu/ml/machine-learning-
databases/wine-quality/winequality-white.csv"
wine_white_data = read.csv(wine_white_url, header = TRUE, sep = ";")
summary(wine_white_data)
```

```
##  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
```

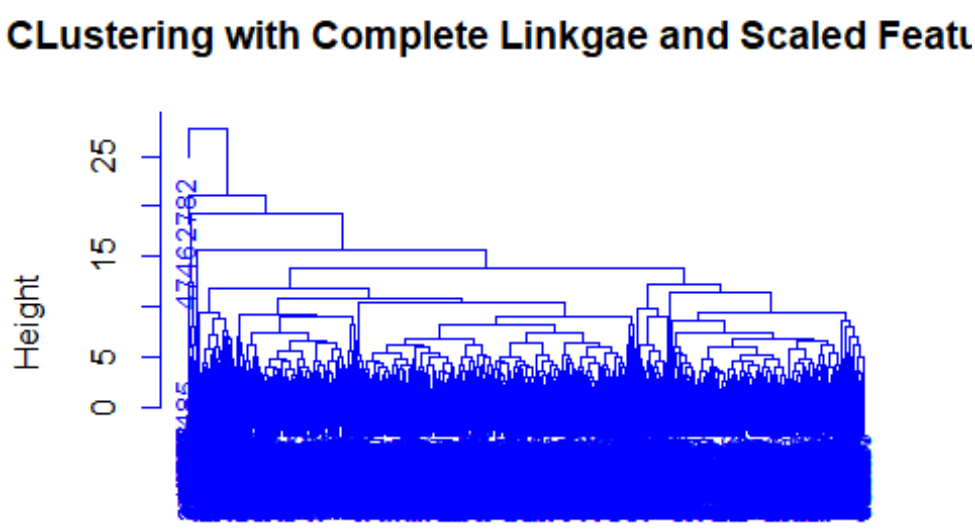
```
##  3rd Qu.:3.280   3rd Qu.:0.5500   3rd Qu.:11.40   3rd Qu.:6.000
##  Max.   :3.820   Max.   :1.0800   Max.   :14.20   Max.   :9.000

wine_white_data_filtered = subset(wine_white_data, select = -(quality))
single_linkage_hclust = hclust(dist(wine_white_data_filtered), method =
"single")
plot(single_linkage_hclust, main = "Clustering with Single Linkage", xlab =
"", cex = 0.9, col = "blue")
```



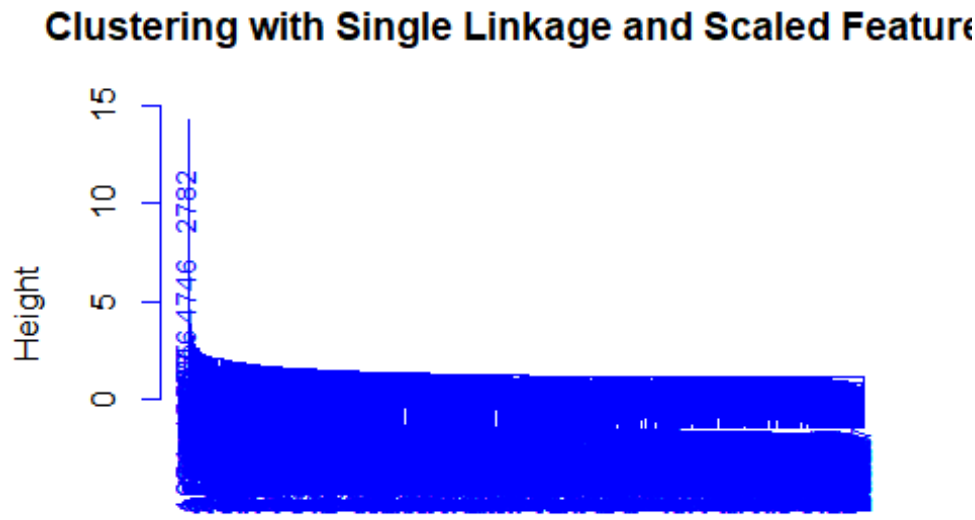

```
complete_linkage_hclust = hclust(dist(wine_white_data_filtered), method =
"complete")
plot(complete_linkage_hclust, main = "Clustering with Complete Linkage", xlab
= "", cex = 0.9, col = "blue")
```




```
wine_white_scaled_data = scale(wine_white_data_filtered)
complete_linkage_hclust = hclust(dist(wine_white_scaled_data), method =
"complete")
plot(complete_linkage_hclust, main = "CLustering with Complete Linkgae and
Scaled Features", xlab = "", sub = "", cex = 0.9, col = "blue")
```




```r
plot(hclust(dist(wine_white_scaled_data), method = "single"), main =
"Clustering with Single Linkage and Scaled Features", xlab = "", sub = "",
cex = 0.9, col = "blue")
```



```
complete_linkage_cutree = cutree(complete_linkage_hclust, k = 2)
complete_linkage_data = cbind(wine_white_data_filtered, cluster =
complete_linkage_cutree)
single_linkage_cutree = cutree(single_linkage_hclust, k = 2)
single_linkage_data = cbind(wine_white_data_filtered, cluster =
single_linkage_cutree)
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
```

```
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
```

```
complete_linkage_aggregate = aggregate(. ~ cluster, data =
complete_linkage_data, mean)
complete_linkage_aggregate
```

```
##   cluster fixed.acidity volatile.acidity citric.acid residual.sugar
chlorides
## 1       1      6.854595        0.2781009   0.3341372       6.379283
0.04576659
## 2       2      7.800000        0.9650000   0.6000000      65.800000
0.07400000
##   free.sulfur.dioxide total.sulfur.dioxide   density       pH sulphates
## 1            35.31366             138.3562 0.9940182 3.188225  0.489806
## 2             8.00000             160.0000 1.0389800 3.390000  0.690000
##    alcohol
## 1 10.51402
## 2 11.70000
```

```
single_linkage_aggregate = aggregate(. ~ cluster, data = single_linkage_data,
mean)
single_linkage_aggregate
```

```
##   cluster fixed.acidity volatile.acidity citric.acid residual.sugar
chlorides
## 1       1      6.854942        0.2782448   0.3342087       6.392128
0.04577211
## 2       2      6.100000        0.2600000   0.2500000       2.900000
0.04700000
##   free.sulfur.dioxide total.sulfur.dioxide   density       pH sulphates
## 1            35.25628             138.2991 0.9940276 3.188215 0.4898162
## 2           289.00000             440.0000 0.9931400 3.440000 0.6400000
##    alcohol
## 1 10.51427
## 2 10.50000
```