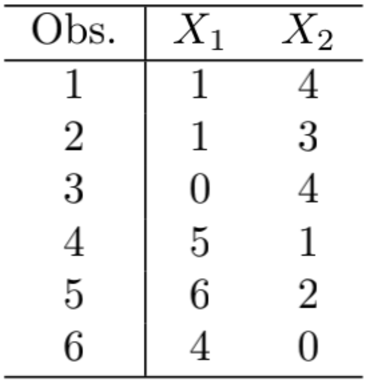
Chapter 12

Tong Sun

3/25/2022

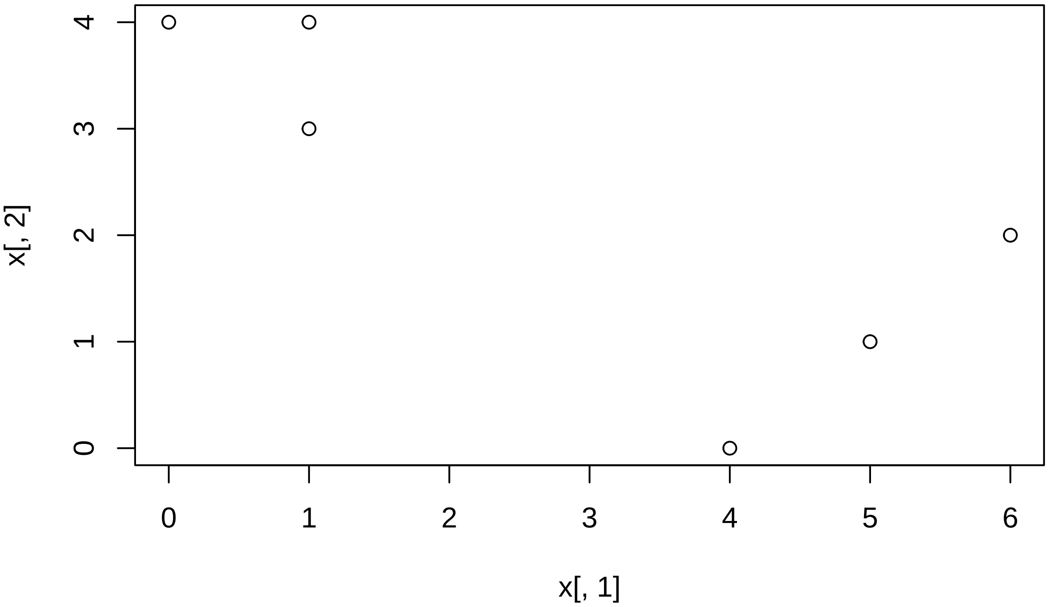
12.3

In this problem, you will perform K-means clustering manually, with K = 2, on a small example with n = 6 observations and p = 2 features. The observations are as follows.



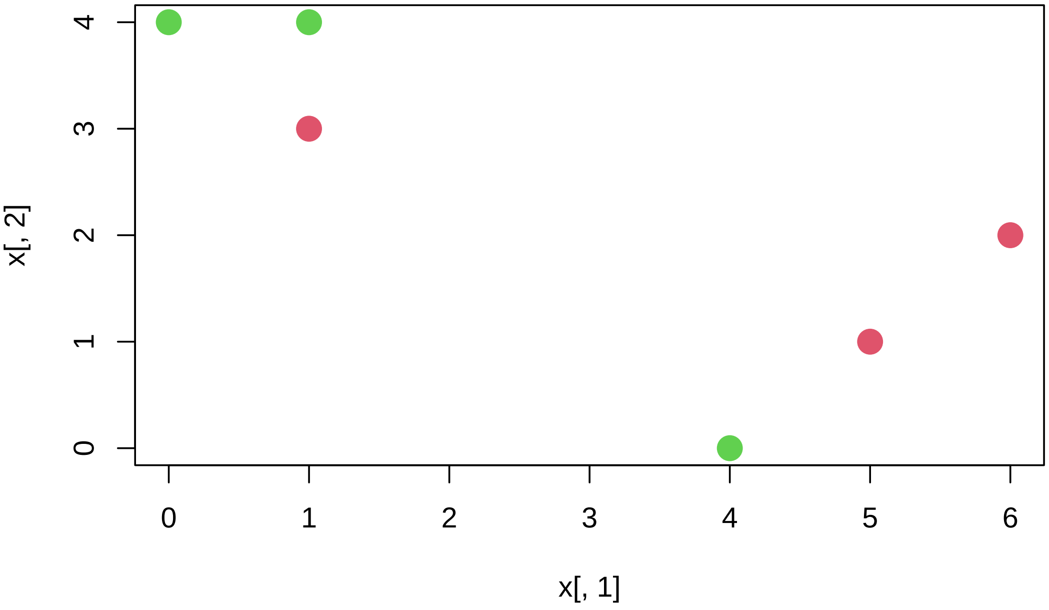
Obs.	X_1	X_2
1	1	4
2	1	3
3	0	4
4	5	1
5	6	2
6	4	0

##(a) Plot the observations.

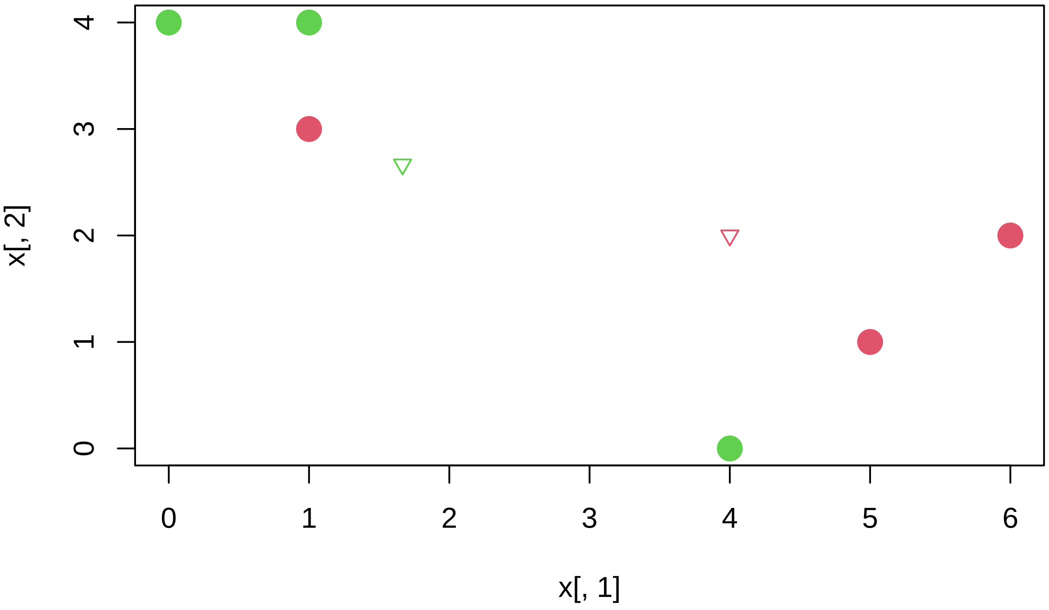


##(b) Randomly assign a cluster label to each observation. You can use the sample() command in R to do this. Report the cluster labels for each observation.

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

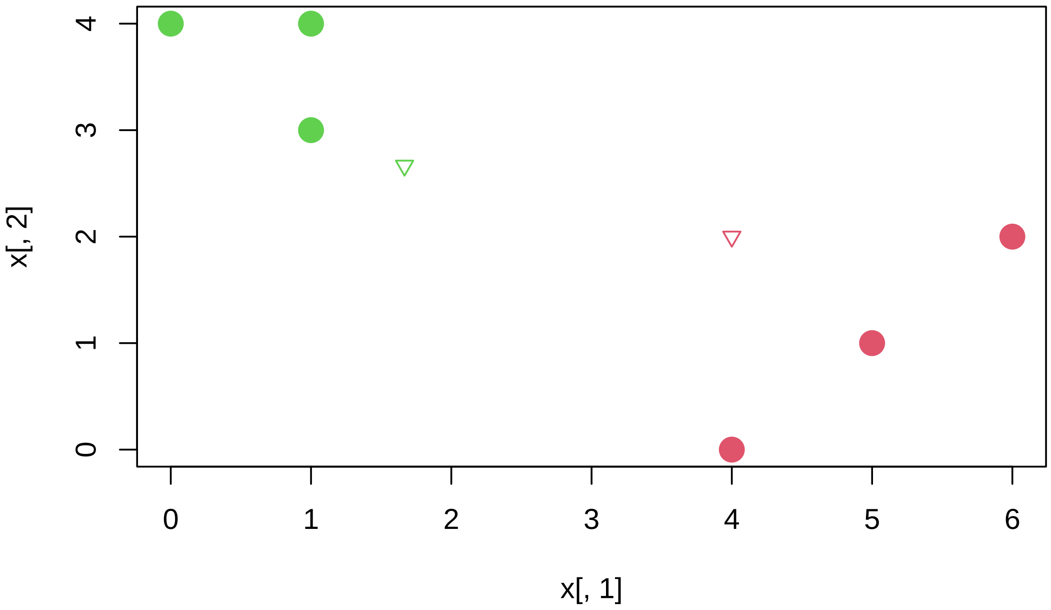


##(c) Compute the centroid for each cluster.

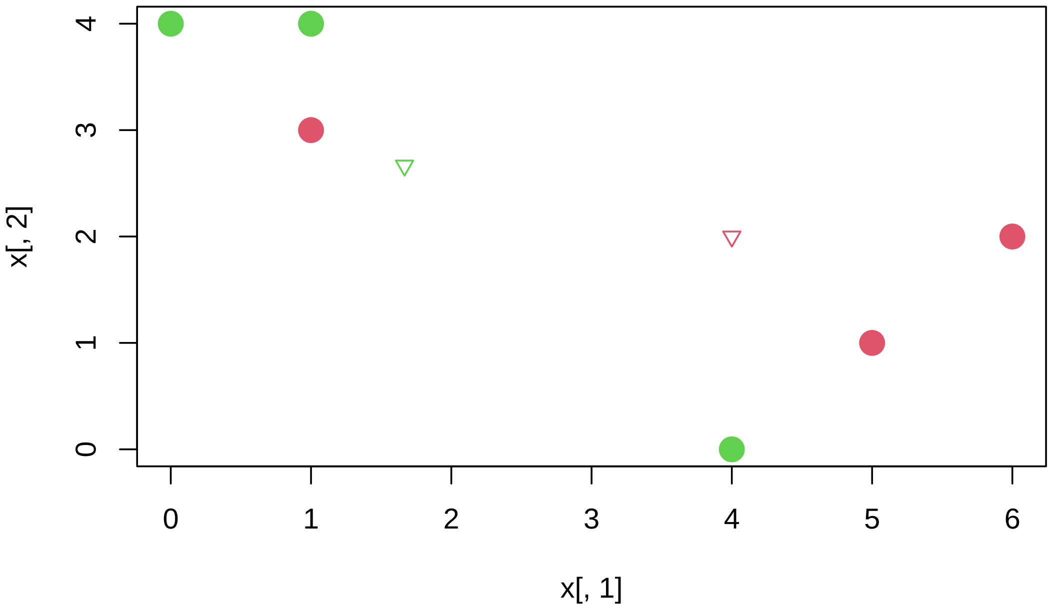


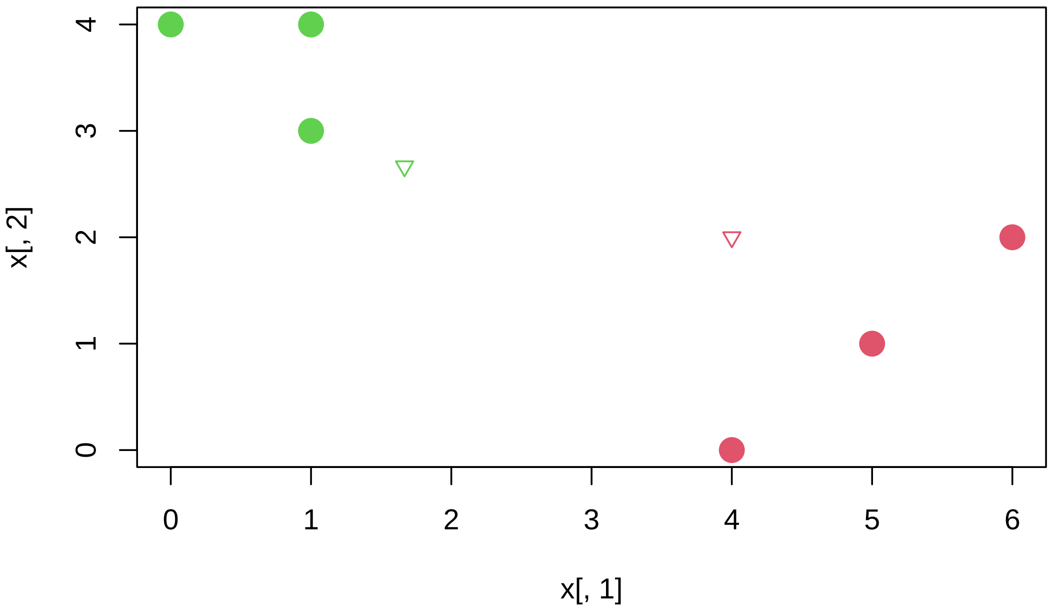
Manually, I compute the centroid for green cluster with $\bar{x}_{11} = \frac{1}{3}(0+1+4) = \frac{5}{3}$ and $\bar{x}_{12} = \frac{1}{3}(0+4+4) = \frac{3}{8}$.Also for the red cluster, $\bar{x}_{21} = \frac{1}{3}(1+5+6) = 4$ and $\bar{x}_{22} = \frac{1}{3}(3+1+2) = 2$.

##(d) Assign each observation to the centroid to which it is closest, in terms of Euclidean distance. Report the cluster labels for each observation.



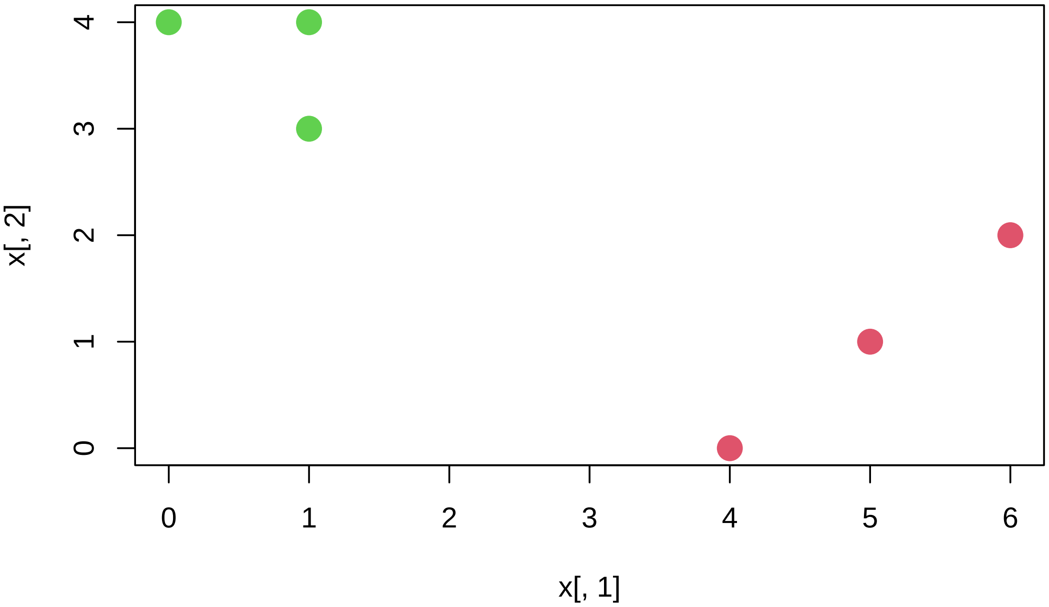
##(e) Repeat (c) and (d) until the answers obtained stop changing.





The algorithm is terminated at this step.

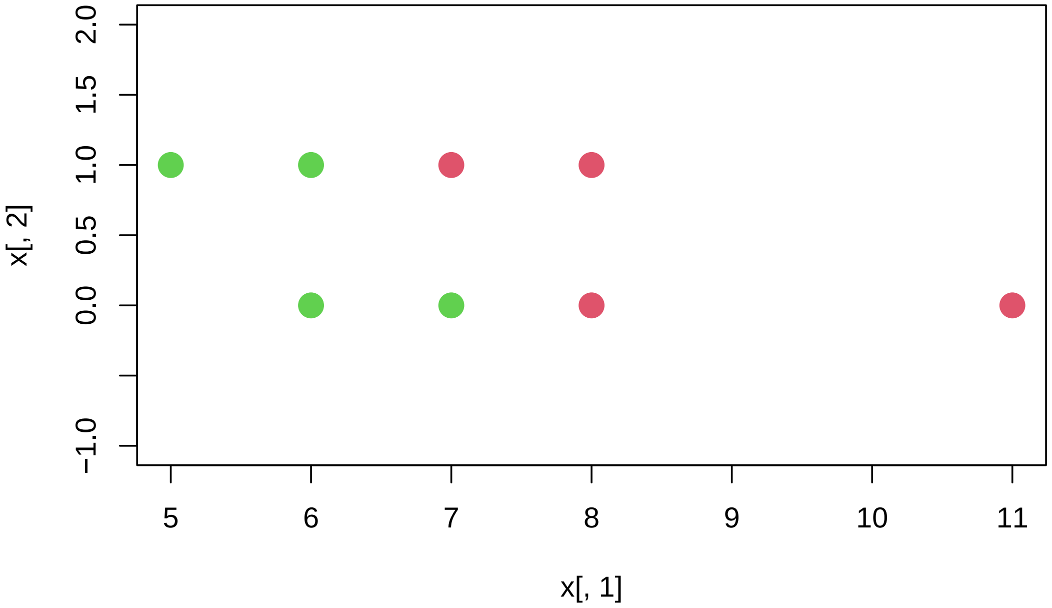
##(f) In your plot from (a), color the observations according to the cluster labels obtained.



12.5

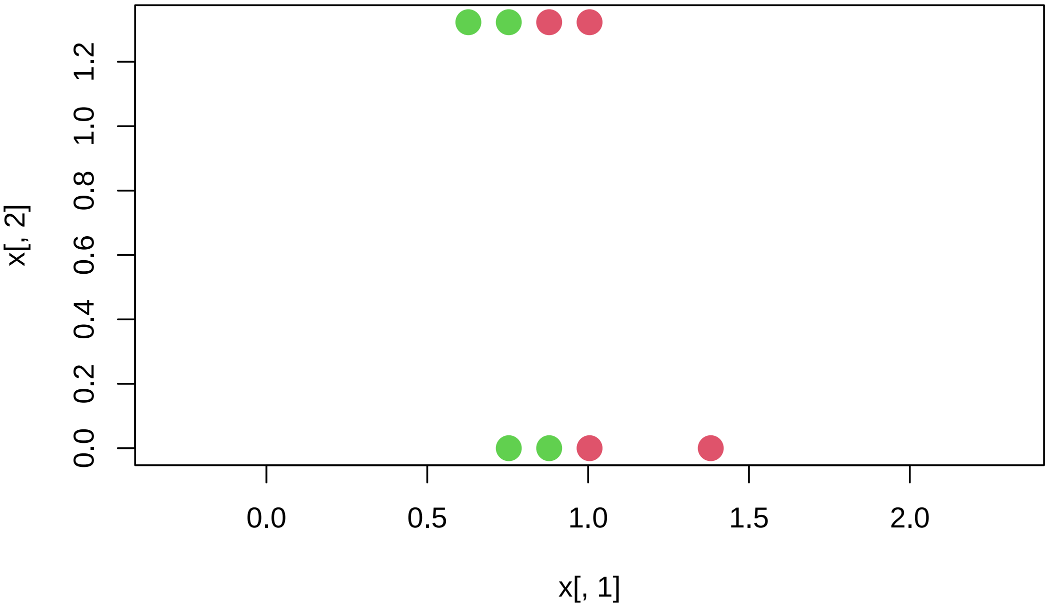
In words, describe the results that you would expect if you performed K-means clustering of the eight shoppers in Figure 12.16, on the basis of their sock and computer purchases, with K = 2. Give three answers, one for each of the variable scalings displayed. Explain.

If we take into consideration the unscaled variables, the number of socks plays a larger role than the number of computers, so we have the clusters{1,2,7,8}(least socks and computer) and {3,4,5,6}(more socks and computer).



If we take into consideration the scaled variables, the number of computers plays a much larger role than the number of socks, so we have the clusters{5, 6, 7, 8}(purchased computer) and {1, 2, 3, 4}(no computer purchased).

```
## [1] 0.5345225
```



If we take into consideration the variables measured by the number of dollars spent, here also the number of computers plays a much larger role than the number of socks, so we have the clusters{5, 6, 7, 8}(purchased computer) and {1, 2, 3, 4}(no computer purchased). ### 12.8 In Section 12.2.3, a formula for calculating PVE was given in Equation 12.10. We also saw that the PVE can be obtained using the "sdev" output of the "prcomp()" function. On the "USArrests" data, calculate PVE in two ways: ##(a) Using the "sdev" output of the "prcomp()" function, as was done in Section 12.2.3.

```
## [1] 4
```

From the Section 12.2.3 in the textbook, I found that the variables have vastly different variances : the UrbanPop variable measures the percentage of the population in each state living in an urban area, which is not a comparable number to the number of rapes om each state per 100,000 individuals. If we failed to scale the variables before performing PCA, then most of the principal components that we observed would be driven by the Assault variable, since it has by far the largest mean and variance. Thus, it's important to standardize the variables to have mean zero and standard deviation one before performing PCA. By default, the "prcomp()" function centers the variables to have mean zero. By using the option "scale = TRUE", I scale the variables to have standard deviation one. Then the variance explained by each principal component is obtained by squaring these. To compute the PVE by each principal component, I simply divide the variance explained by each principal component by the total variance explained by all four principal components.

```
## [1] 0.62006039 0.24744129 0.08914080 0.04335752
```

I find that the first principal component explains 62.0% of the variance in the data, the next principal component explains 24.7% of the variance, the third principal component explains 8.9% of the variance, and the last principal component explains 4.3% of the variance.

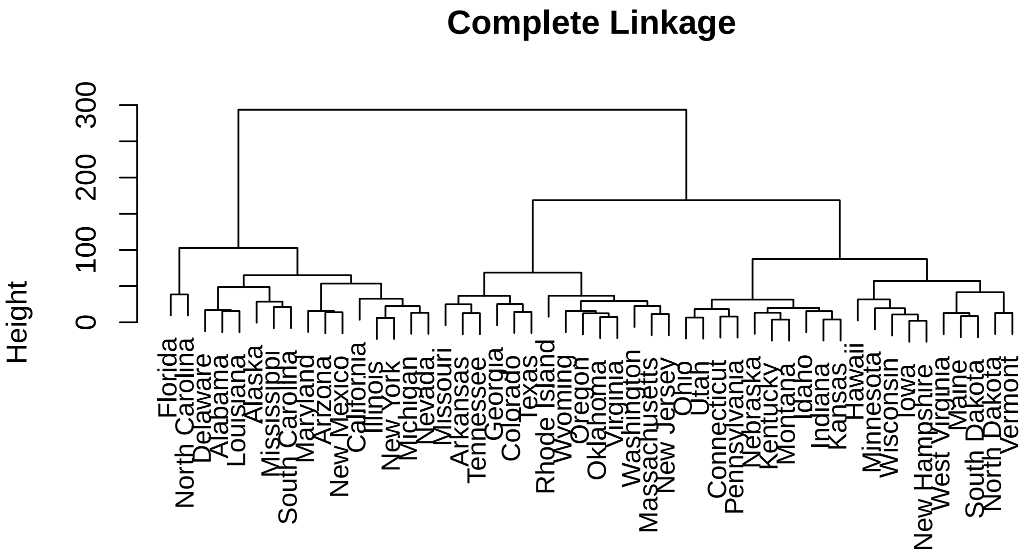
##(b) By applying Equation 12.10 directly. That is, use the "prcomp()" function to compute the principal component loadings. Then, use those loadings in Equation 12.10 to obtain the PVE.

```
##        PC1        PC2        PC3        PC4
## 0.62006039 0.24744129 0.08914080 0.04335752
```

The “rotation” matrix provides the principal component loadings; each column of “pr.out$rotation” contains the corresponding principal component loading vector.

12.9

Consider the “USArrests” data. We will now perform hierarchical clustering on the states. ##(a) Using hierarchical clustering with complete linkage and Euclidean distance, cluster the states.



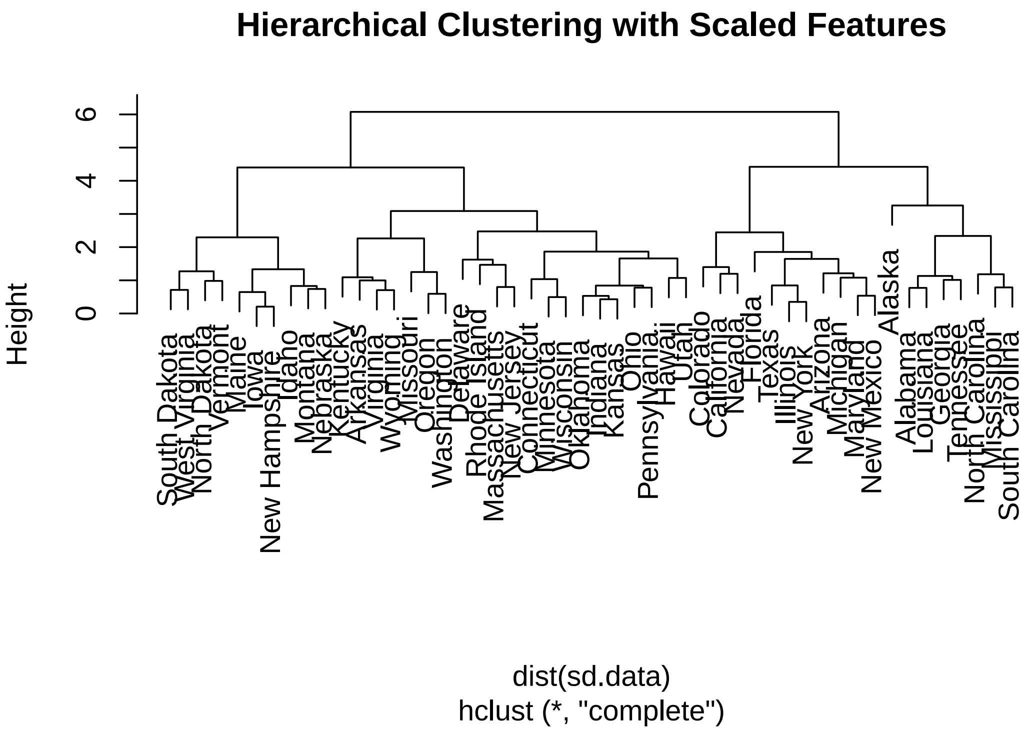
The ‘hclust()’ function implements hierarchical clustering in R. Here I use the data from “USArrests” data to plot the hierarchical clustering dendrogram using complete, single and average linkage clustering, with Euclidean distance as the dissimilarity measure. The ‘dist()’ function is used to compute the 50*50 inter-observation Euclidean distance matrix.

##(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?

```
##        Alabama         Alaska        Arizona       Arkansas     California
##              1              1              1              2              1
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              1              1              2
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              1              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              1
##   Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##               2              1              3              1              2
##         Montana       Nebraska         Nevada  New Hampshire     New Jersey
##               3              3              1              3              2
##      New Mexico       New York North Carolina   North Dakota           Ohio
##               1              1              1              3              3
##        Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##               2              2              3              2              1
##    South Dakota      Tennessee          Texas           Utah        Vermont
##               3              2              2              3              3
##        Virginia     Washington  West Virginia      Wisconsin        Wyoming
##               2              2              3              3              2
```

To determine the cluster labels for each observation associated with a given cut of the dendrogram, I use the 'cutree()' function. The second argument to 'cutree()' is the number of clusters we wish to obtain. For this question, I set it to 3. And for this data, complete linkage generally separate the observations into their correct groups.

##(c) Hierarchically cluster the states using complete linkage and Euclidean distance, after scaling the variables to have standard deviation one.



To scale the variables before performing hierarchical clusering of the observations, i use the 'scale()' function.

##(d) What effect does scaling the variables have on the hierarchical clustering obtained? In your opinion, should the variables be scaled before the inter-observation dissimilarities are computed? Provide a justification for your answer.

```
##         Alabama         Alaska        Arizona       Arkansas     California
##               1              1              2              3              2
##       Colorado    Connecticut       Delaware        Florida        Georgia
##              2              3              3              2              1
##         Hawaii          Idaho       Illinois        Indiana           Iowa
##              3              3              2              3              3
##         Kansas       Kentucky      Louisiana          Maine       Maryland
##              3              3              1              3              2
##  Massachusetts       Michigan      Minnesota    Mississippi       Missouri
##              3              2              3              1              3
##        Montana       Nebraska         Nevada  New Hampshire     New Jersey
##              3              3              2              3              3
##     New Mexico       New York North Carolina   North Dakota           Ohio
##              2              2              1              3              3
##       Oklahoma         Oregon   Pennsylvania   Rhode Island South Carolina
##              3              3              3              3              1
##   South Dakota      Tennessee          Texas           Utah        Vermont
##              3              1              2              3              3
##       Virginia     Washington  West Virginia      Wisconsin        Wyoming
##              3              3              3              3              3
```

```
table(cutree(hc.complete, 3), cutree(hc.complete.sd, 3))
```

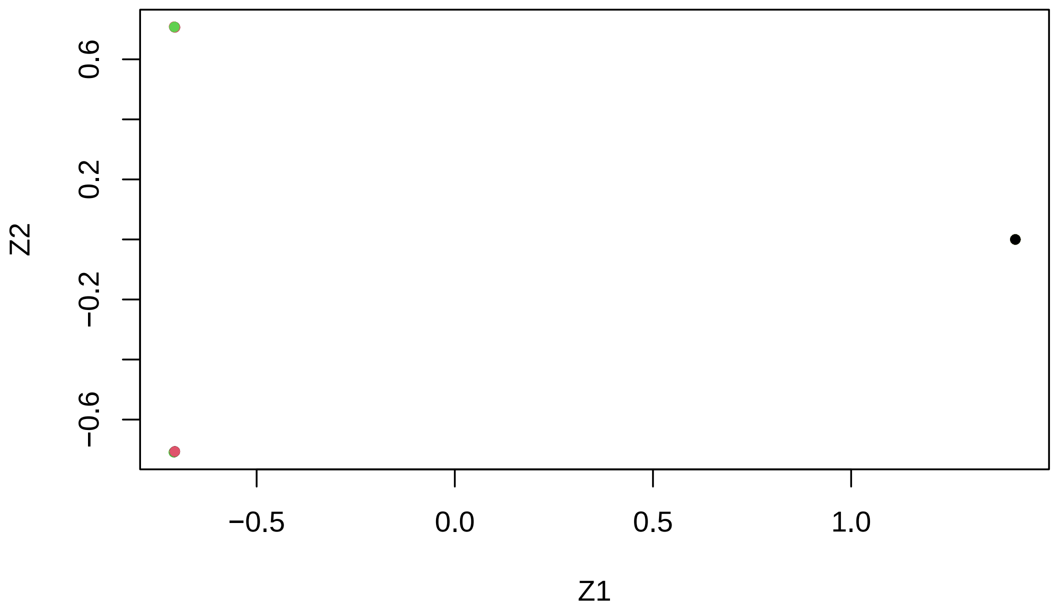
```
##
##       1  2  3
##   1   6  9  1
##   2   2  2 10
##   3   0  0 20
```

Scaling the variables affect the clusters obtained although the trees are similar. The variables should be scaled beforehand because the data measures have different units.

12.10

In this problem, you will generate simulated data, and then perform PCA and K-means clustering on the data. ##(a) Generate a simulated data set with 20 observations in each of three classes (i.e. 60 observations total), and 50 variables.

##(b) Perform PCA on the 60 observations and plot the first two principal component score vectors. Use a different color to indicate the observations in each of the three classes. If the three classes appear separated in this plot, then continue on to part (c). If not, then return to part (a) and modify the simulation so that there is greater separation between the three classes. Do not continue to part (c) until the three classes show at least some separation in the first two principal component score vectors.



##(c) Perform K-means clustering of the observations with K = 3.How well do the clusters that you obtained in K-means clustering compare to the true class labels?

```
##
## true.labels  1  2  3
##           1  0 20  0
##           2  0  0 20
##           3 20  0  0
```

The observations are perfectly clustered.

##(d) Perform K-means clustering with K = 2. Describe your results.

```
##
## true.labels  1  2
##           1 20  0
##           2  0 20
##           3 20  0
```

All observations of one of the three clusters is now absorbed in one of the two clusters.

##(e) Now perform K-means clustering with K = 4, and describe your results.

```
##
## true.labels  1  2  3  4
##           1  0  0  0 20
##           2 11  0  9  0
##           3  0 20  0  0
```

The second cluster is splitted into two clusters.

##(f) Now perform K-means clustering with K = 3 on the first two principal component score vectors, rather than on the raw data.That is, perform K-means clustering on the 60 × 2 matrix of which the first column is the first principal component score vector, and the second column is the second principal component score vector. Comment on the results.

```
##
## true.labels  1  2  3
##           1 20  0  0
##           2  0 20  0
##           3  0  0 20
```

All observations are perfectly clustered once again.

##(g) Using the scale() function, perform K-means clustering with K = 3 on the data after scaling each variable to have standard deviation one. How do these results compare to those obtained in (b)? Explain.

```
##
## true.labels  1  2  3
##           1  6  7  7
##           2 11  6  3
##           3  7  6  7
```

We may see that we have worse results than with unscaled data, as scaling affects the distance between the observations.