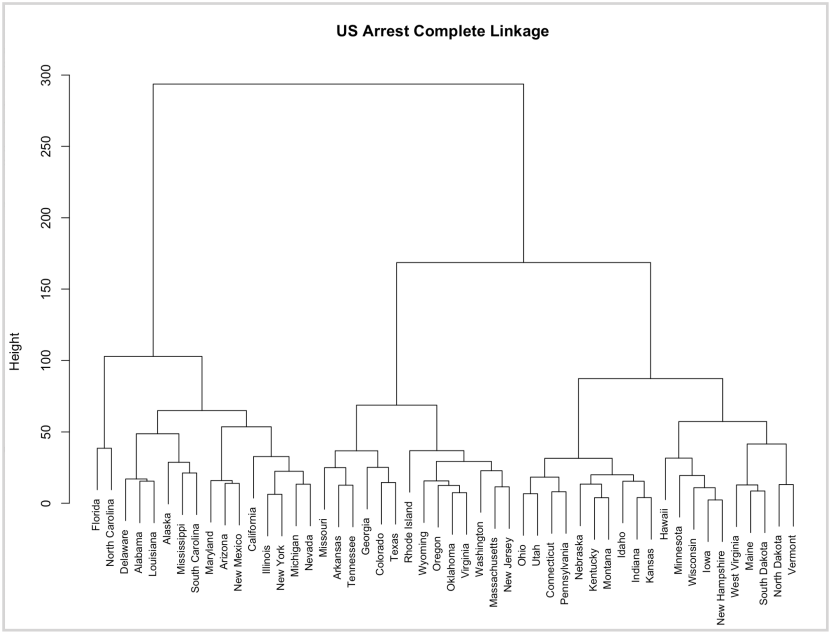
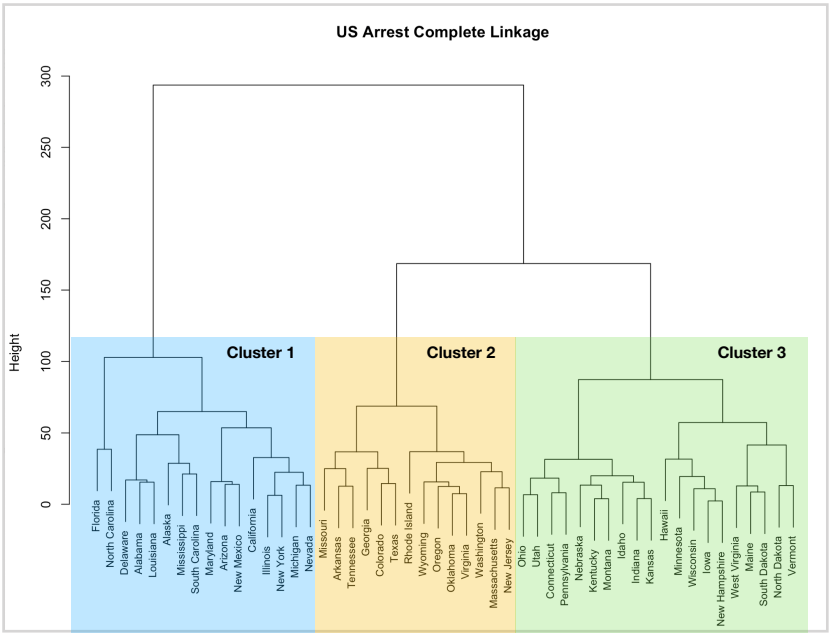
1) ISLR text: Chapter 10 Question 9

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.

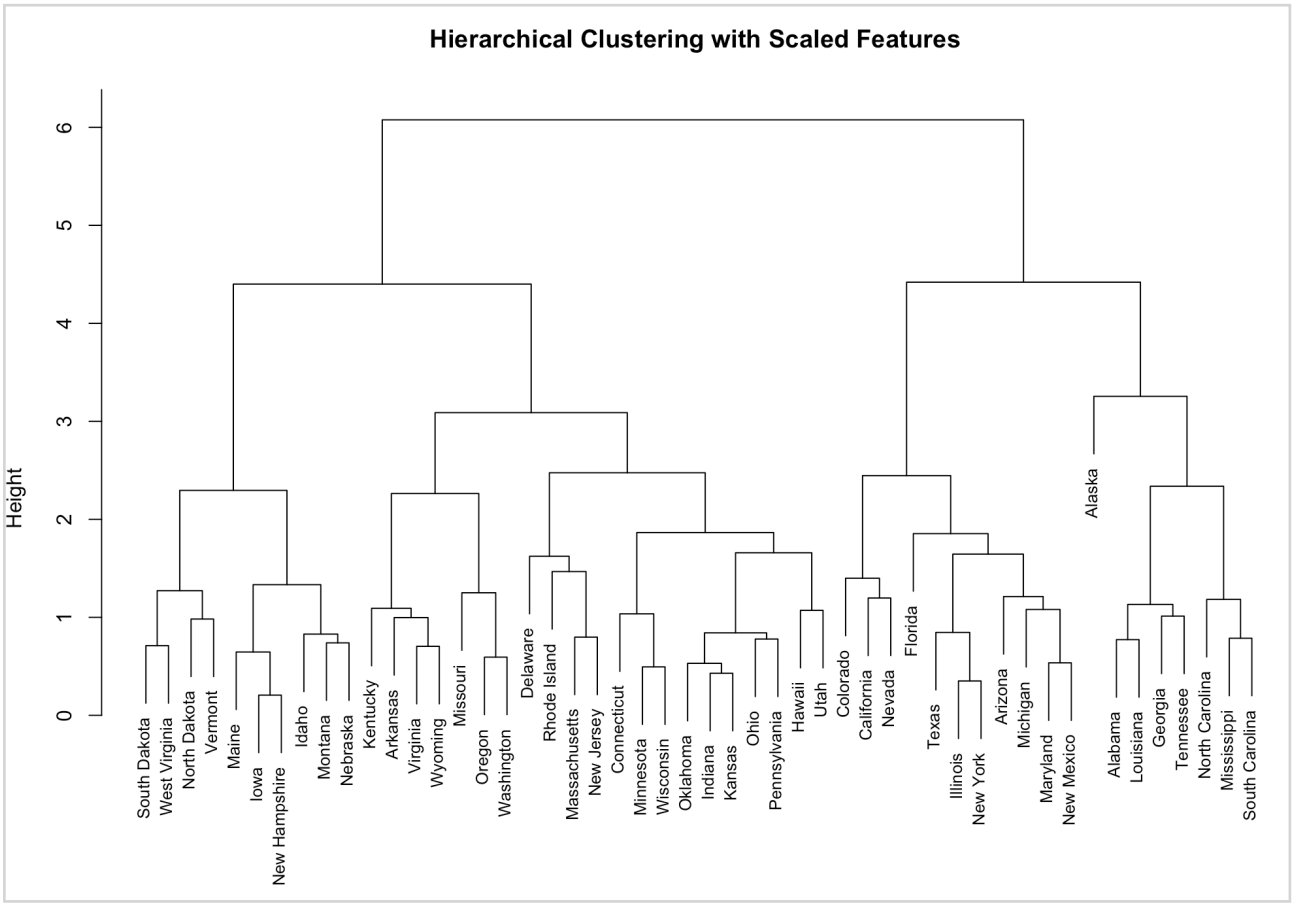


(b) Cut the dendrogram at a height that results in three distinct clusters. Which states belong to which clusters?



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

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



(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.

After scaling, we get different clusters. In Cluster 1, 9 of the 16 states moved to Cluster 2, and 1 moved to Cluster 3.

Original\ Scaled data	Cluster 1	Cluster 2	Cluster 3
Cluster 1	6	9	1
Cluster 2	2	2	10
Cluster 3	0	0	20

I think that the variable should not be scaled, because the branch is longer in the un-scaling dendrogram than the scaling dendrogram, also the cut point is clearer.

2) ISLR text: Chapter 10 Question 11

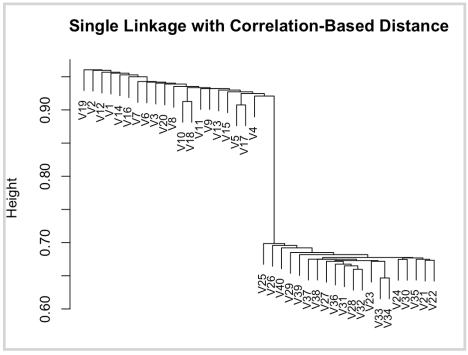
On the book website, www.StatLearning.com, there is a gene expression data set (Ch10Ex11.csv) that consists of 40 tissue samples with measurements on 1,000 genes. The first 20 samples are from healthy patients, while the second 20 are from a diseased group.

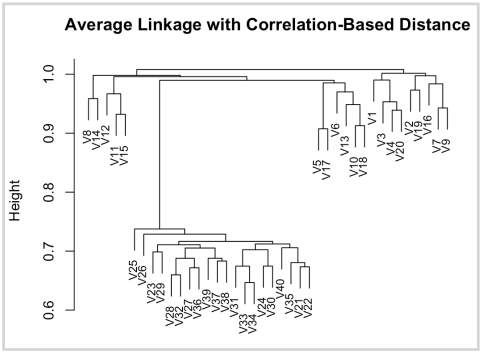
a) Load in the data using read.csv(). You will need to select header=F.

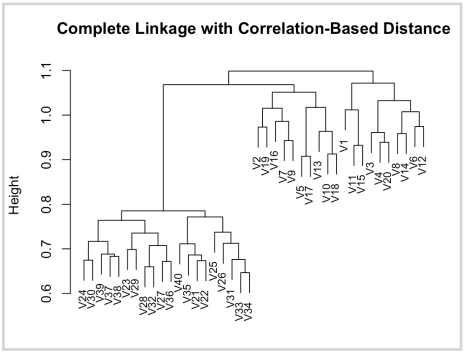
Ans: I used the read.csv with option header=False to store the Ch10Ex11.csv to the variable "gene".

b) Apply hierarchical clustering to the samples using correlation based distance, and plot the dendrogram. Do the genes separate the samples into the two groups? Do your results depend on the type of linkage used?

Ans: I've generated three types of linkage dendrogram, including single, average and complete. The genes with all three types of methods will present more than two clusters, but the complete linkage is more likely to separate into two groups. As a result, different linkages will affect the clustering results.

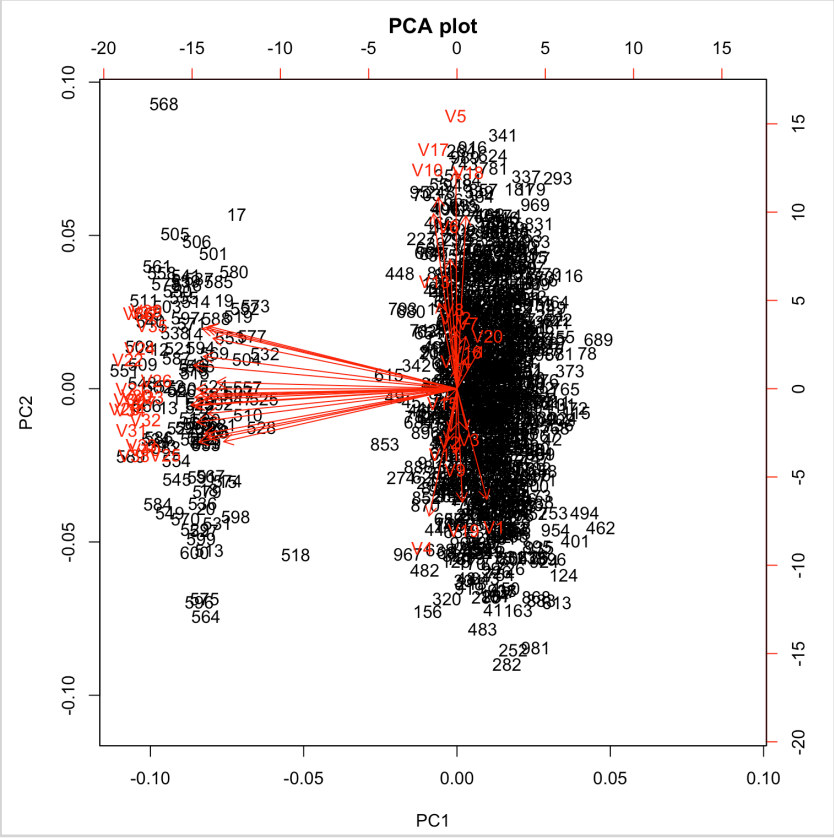


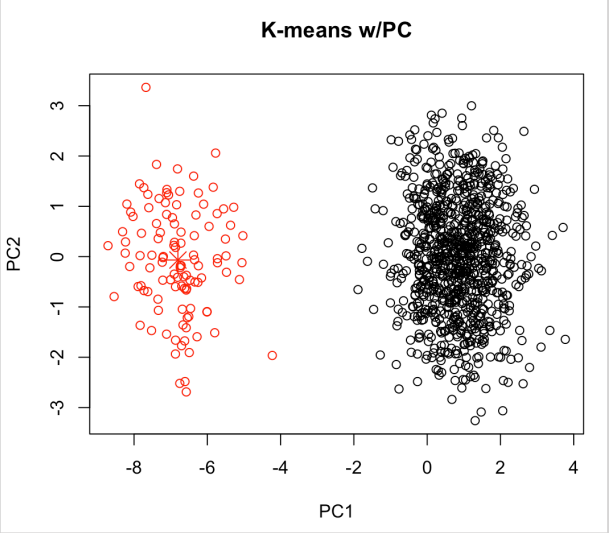




c) Your collaborator wants to know which genes differ the most across the two groups. Suggest a way to answer this question, and apply it here.

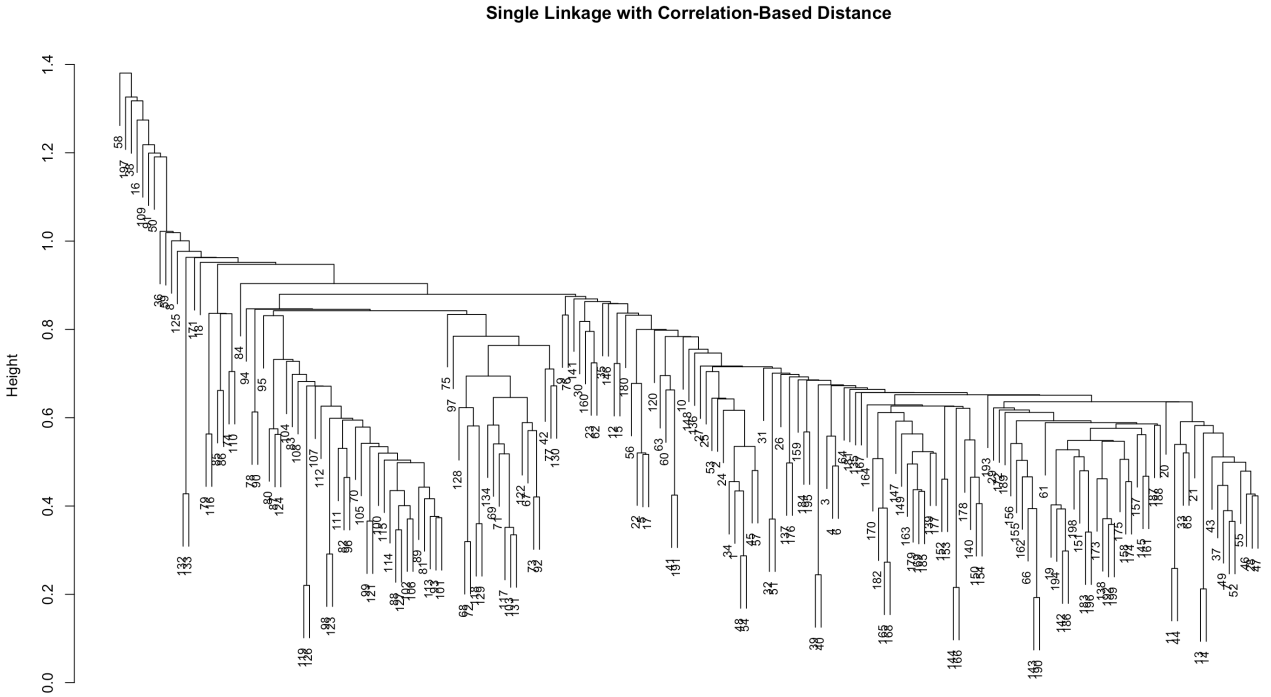
Ans: We can observe which genes differ the most across the two clusters through PCA, I've apply k-means to PC1 and PC2 then plot the group.

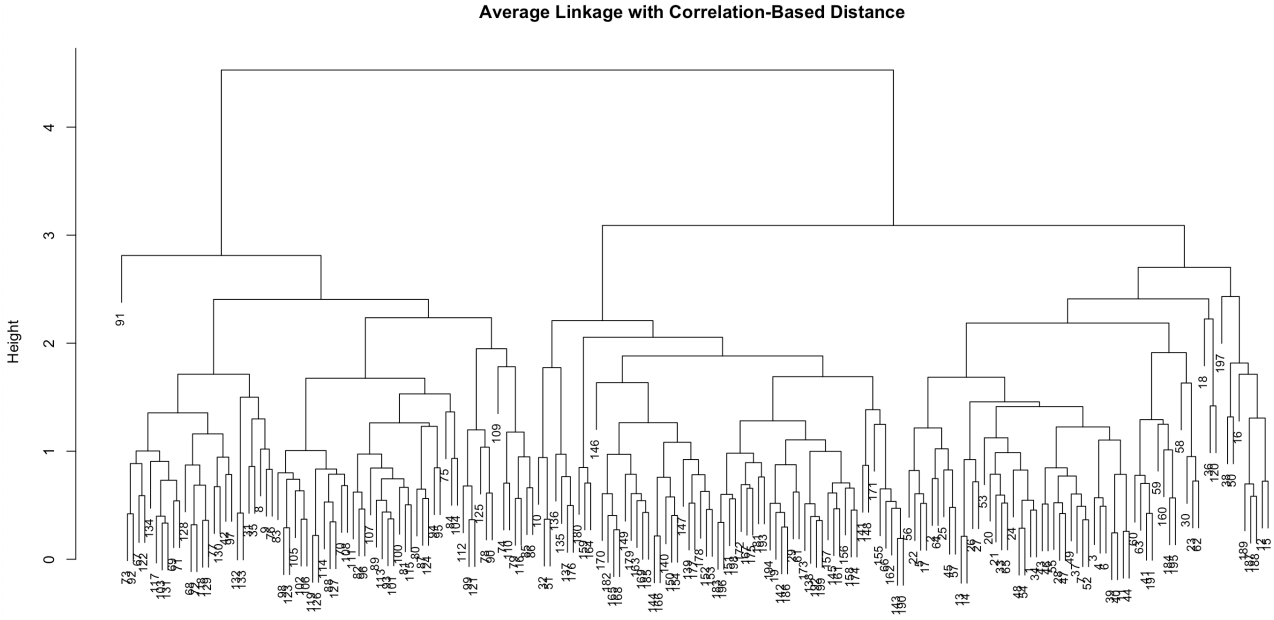


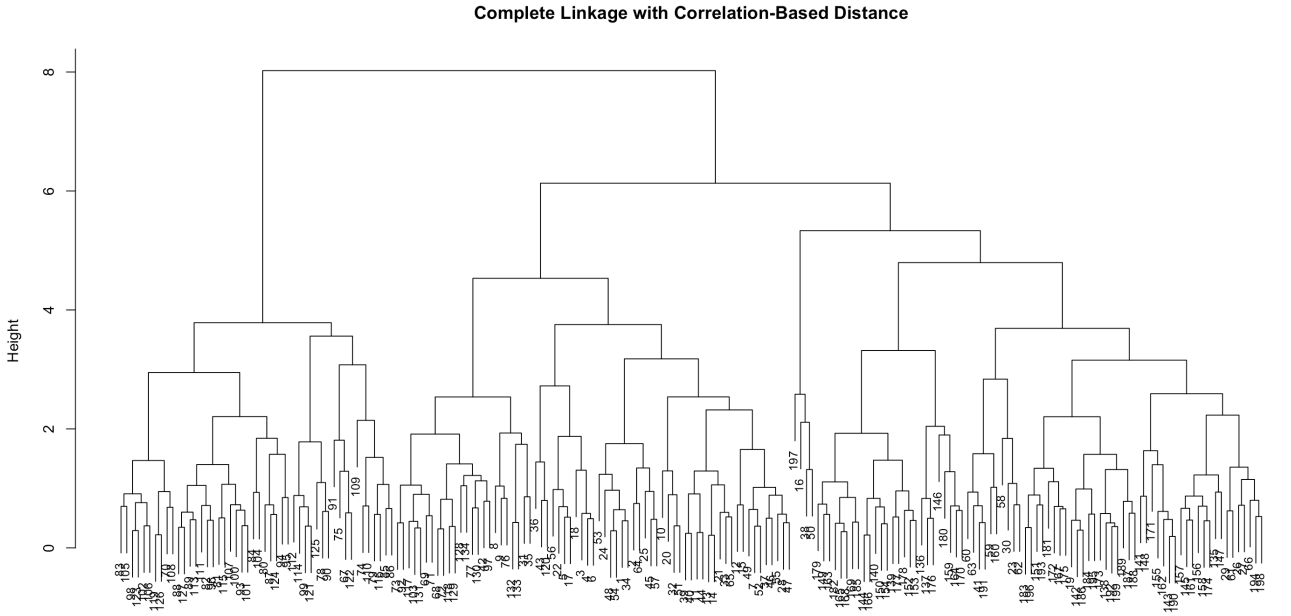


3) Access the data "seeds data" (on UB learns). This data contains the geometrical properties of kernels belonging to three different varieties of wheat (seed group). The original data can be found: https://archive.ics.uci.edu/ml/datasets/seeds, although I have modified the data slightly.

a) Cluster the data based on single-linkage, average linkage, and complete-linkage agglomerative hierarchical clustering. Do not use the "seed group" column to perform the clustering, but use it to help evaluate your results.







Decide on the groupings, and justify it, for all three methods. The justification should be based on a measure (you select which) that we learned in class.

Ans: I use silhouette to calculate the average silhouette width from k=2 to 6, and select the highest width to be the number of my cluster. According to the table below, all three methods have the highest average silhouette width at k = 2.

Linkage\ k	2	3	4	5	6
Single	0.066	0.002	-0.080	-0.091	-0.229
Average	0.462	0.393	0.306	0.279	0.255
Complete	0.421	0.346	0.334	0.272	0.252

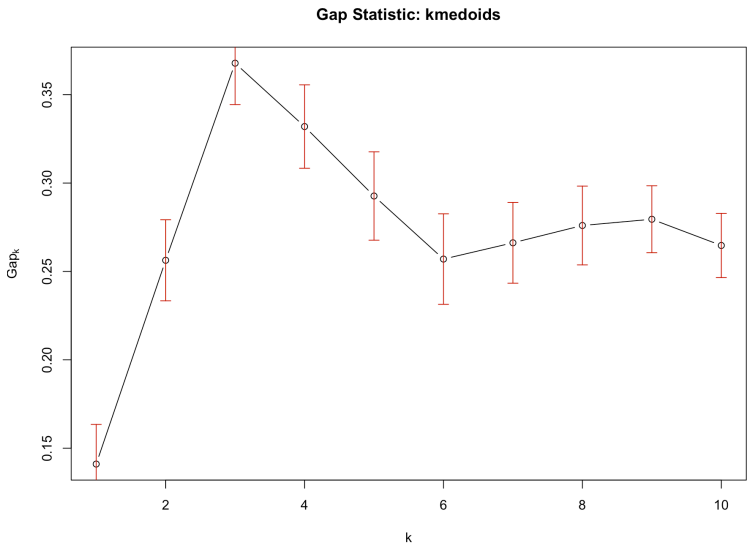
Which method "performed" the best and which method performed the worst? Was the result in line with your expectations?

Ans: The accuracy of the single, average and complete linkage agglomerative hierarchical clustering are 33%, 64%, and 57%. Hence, the cluster based on average linkage performed the best and the single linkage performed the worst.

My expectations are the same as the results. According to the dendrogram, the average-linkage dendrogram has longer branches than others, and has a broad shoulder for clean grouping, so average linkage will perform the best. On the contrary, the single-linkage dendrogram is not fitting well, we cannot do the good separation with the major arms, so it can be expected to be the worst result.

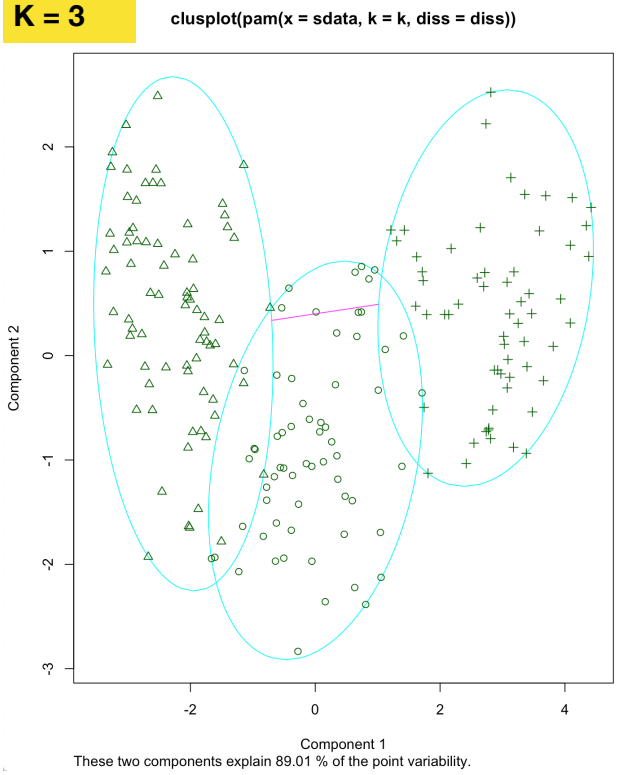
b) Cluster the data based on K-means or K-medoids. Use an analytical technique to justify your choice in "k". How did the performance compare to the hierarchical clustering of part a? Which did you feel was a better method for this data?

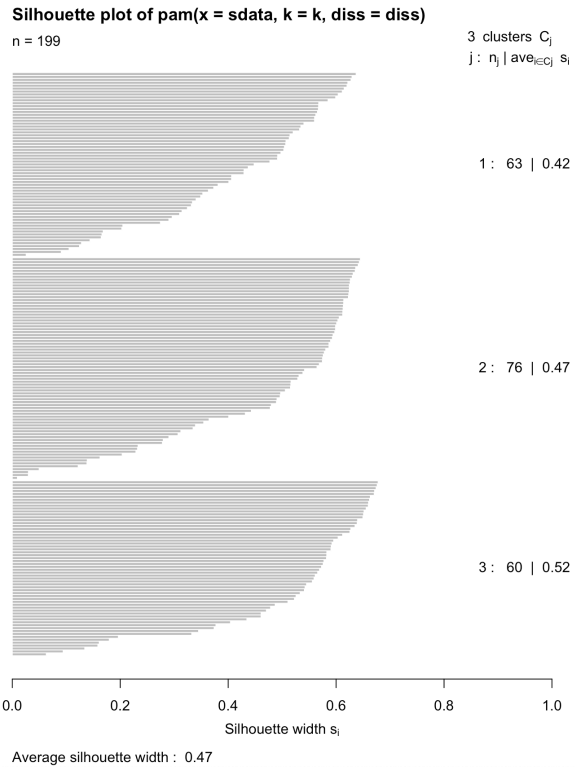
Ans: I clustered the data based on K-medoids, and used gap statistic to select k = 3. Since the accuracy of K-medoids with k= 3 is 27%, I've also calculated the accuracy of K-medoids with k = 2 for 60%, both are lower than the best case in average linkage hierarchical clustering (64%).



In my view, hierarchical clustering is a better method than k-medoids, because we can observe the dendrogram first then decide the value of k. It is easier to expect the result through hierarchical clustering.

K = 3





K = 2