BU.510.650 – Data Analytics

Assignment # 3

Please submit two documents: Your answers to each part of every question in .pdf or .doc format, and your R script, in .R format. Please submit your assignment through Canvas and name your files using the convention LastName_FirstName_AssignmentNumber.

Programming Part

1. In this question, you will use the K-means algorithm to cluster the US states, based on their 2018 US News & World Report rankings. US News and World Report ranks all 50 states, according to each of the following criteria: Healthcare, Education, Economy, Opportunity, Infrastructure, Crime and Corrections, Fiscal Stability, and Quality of Life. (Iowa was ranked as the best state overall, in case you are curious.) The file, `StateRankings.csv`, summarizes the data. (Please note that lower rank is better.) To begin your work on this question, please read the data in `StateRankings.csv` into a data frame called `Rankings`.

   Please include `set.seed(5)` once at the beginning of your code, so we all get the same results.

(a) Perform K-means clustering of all 50 states, using the Healthcare and Economy rankings only. Use five clusters and 20 different initializations. Answer the following questions: What are the sizes of the five clusters you obtained? What is the average Healthcare ranking for each cluster? What is the average Economy ranking for each cluster?

(b) Provide a plot of the states with Economy ranking on the x-axis, and Healthcare ranking on the y-axis, using a different shape and color for each cluster, and including the state names on the plot.

2. In this question, you will use the hierarchical clustering algorithm to cluster the US states, based on their 2018 US News & World Report rankings. Again, begin your work on this question by reading the data in `StateRankings.csv` into a data frame called `Rankings`.

   Please include `set.seed(5)` once at the beginning of your code, so we all get the same results.

(a) Apply hierarchical clustering with the method "complete," using only the Economy and Healthcare ranking data. Provide the resulting dendrogram.

(b) Based on the dendrogram you obtained, how many clusters do you think will be a good representation of the data? Why?

(c) Cut the dendrogram at a height that results in the number of clusters you suggested in part (b). Provide a plot of the states with Economy ranking on the x-axis, Healthcare ranking on the y-axis, using a different shape and color for each cluster, and including the state names on the plot.

Conceptual Part

For the conceptual section, there is no requirement to utilize any R coding. Prior to addressing this segment, kindly review Section 12.4 in the ISLR textbook (Pages 514 – 530). Some parts have already been completed for you. The sections that require your completion are highlighted.

Question 1 (K-means Clustering):

Let $C_1, \ldots, C_K$ denote clusters. If observation i is in cluster C_k, we write $i \in C_k$. Let $|C_k|$ denote the number of observations in cluster C_k.

The within-cluster variation for cluster k, $W(C_k)$ can be calculated using squared Euclidean distance:

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

K-means clustering aims to pick the clusters so that total within-cluster variation is minimized, that is, pick K clusters so that we minimize $\sum_{k=1}^{K} W(C_k)$

Suppose we have the following data with 5 observations. Each observation corresponds to the location of a city, with X1 representing the latitude (all of them North) and X2 representing the longitude (all of them East). Suppose the first two cities are in one cluster (Cluster 1), and the next three cities are in another cluster (Cluster 2).

Obs.	X1	X2
i = 1	35.7	51.4
i = 2	31.9	54.3
i = 3	39.9	116.4
i = 4	31.2	121.5
i = 5	30.6	104.1

For example, within-cluster variation for Cluster 1 is 11.4

$$d_{12} = (35.7 - 31.9)^2 + (51.4 - 54.3)^2 = 22.8 \qquad W(C_1) = \frac{1}{2}(22.8) = 11.4$$

• Compute the cluster variation for Cluster 2.

• Determine the total within-cluster variation for both Cluster 1 and Cluster 2.

Question 2 (K-means Clustering):

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.

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

Step 1: Initialization

We randomly assign observations 1, 3, and 5 to cluster 1, and observations 2, 4, and 6 to cluster 2.

Step 2:

a) We compute the centroid for each cluster:

Centroid of Cluster 1: X1 = (1+0+6)/3 = 2.33 X2 = (4+4+2)/3 = 3.33

Centroid of Cluster 2: X1 = (1+5+4)/3 = 3.33 X2 = (3+1+0)/3 = 1.33

b) We compute the centroid for each cluster:

Distance from obs. 1 to cluster 1: $\sqrt{(1-2.33)^2+(4-3.33)^2} \approx 1.5$

Distance from obs. 1 to cluster 2: $\sqrt{(1-3.33)^2+(4-1.33)^2} \approx 3.5$

So, observation 1 is assigned to Cluster 1.

The following table displays the complete calculations. The underlined values represent the shortest distances.

Cluster 1: Obs. 1, 2, 3 Cluster 2: Obs. 4, 5, 6

Obs.	Distance from C1	Distance from C2
1	1.5	3.5
2	1.4	2.9
3	2.4	4.3
4	3.5	1.7
5	3.9	2.7
6	3.7	1.5

- Repeat Step 2 iteratively until the obtained results cease to change.

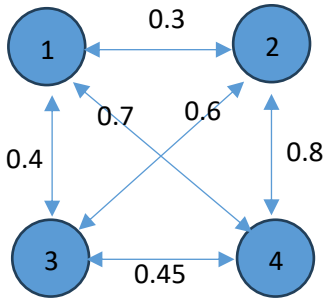
Question 3 (Hierarchical Clustering):

Suppose that we have four observations, for which we compute a dissimilarity matrix, given by

	C2	C3	C4
C1	0.30	0.40	0.70
C2	-	0.60	0.80
C3	-	-	0.45

For instance, the dissimilarity between the first and second observations is 0.3, and the dissimilarity between the second and fourth observations is 0.8.

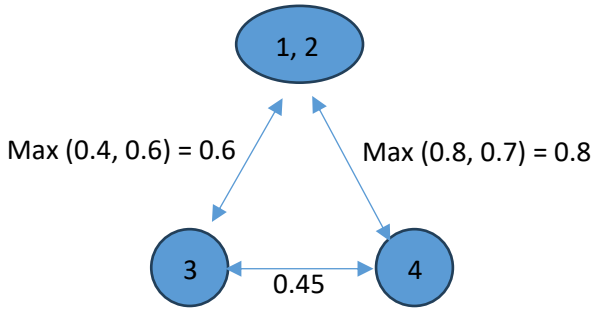
Based on this dissimilarity matrix, the dendrogram resulting from hierarchically clustering these four observations using "complete" linkage is depicted below:



Round 1:

The closest clusters are 1 and 2 with a dissimilarity of 0.30. Following the merger of Cluster 1 and 2, the updated dissimilarity matrix is displayed below:

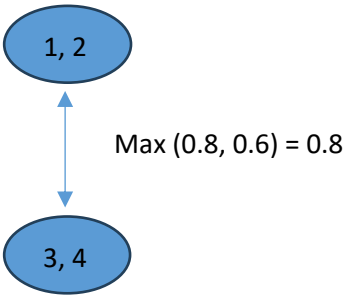
	C3	C4
C1, C2	0.60	0.80
C3	-	0.45



Round 2:

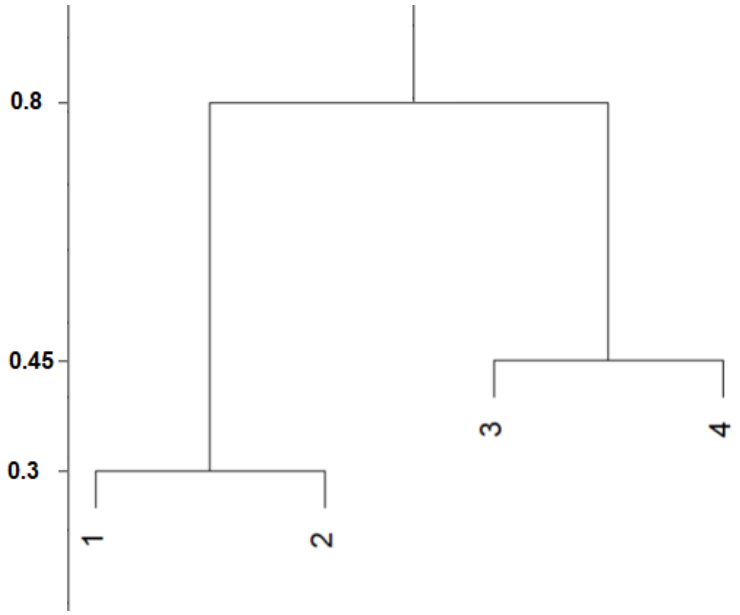
The closest clusters are 3 and 4 with the dissimilarity 0.45. Following the merger of Cluster 3 and 4, the updated dissimilarity matrix is displayed below:

	C3, C4
C1, C2	0.80



Round 3:

The only remaining clusters are (1, 2) and (3, 4) with the dissimilarity 0.8.



- Repeat the process above, this time using “single” linkage clustering.
- Repeat the process above, this time using “average” linkage clustering.