Assignment # 3

Beiming Zhang

Programming Part

Question 1 (a):

Output:

K-means clustering with 5 clusters of sizes 9, 12, 11, 11, 7

Cluster means:

	Healthcare	Economy
1	10.11111	6.666667
2	11.66667	28.416667
3	35.54545	14.363636
4	29.45455	37.818182
5	47.00000	42.857143

Clustering vector:

Alabama	Alaska	Arizona	Arkansas	California
5	4	3	5	1
Colorado	Connecticut	Delaware	Florida	Georgia
1	2	3	3	3
Hawaii	Idaho	Illinois	Indiana	Iowa
2	1	4	3	1
Kansas	Kentucky	Louisiana	Maine	Maryland
4	5	5	4	2
Massachusetts	Michigan	Minnesota	Mississippi	Missouri
1	3	2	5	4
Montana	Nebraska	Nevada	New Hampshire	New Jersey
4	2	3	1	2
New Mexico	New York	North Carolina	North Dakota	Ohio
4	2	3	2	4
Oklahoma	Oregon	Pennsylvania	Rhode Island	South Carolina
5	1	4	2	3
South Dakota	Tennessee	Texas	Utah	Vermont
2	3	3	1	2
Virginia	Washington	West Virginia	Wisconsin	Wyoming
4	1	5	2	4

Within cluster sum of squares by cluster:

[1] 518.8889 1131.5833 757.2727 858.3636 176.8571

(between_SS / total_SS = 83.5 %)

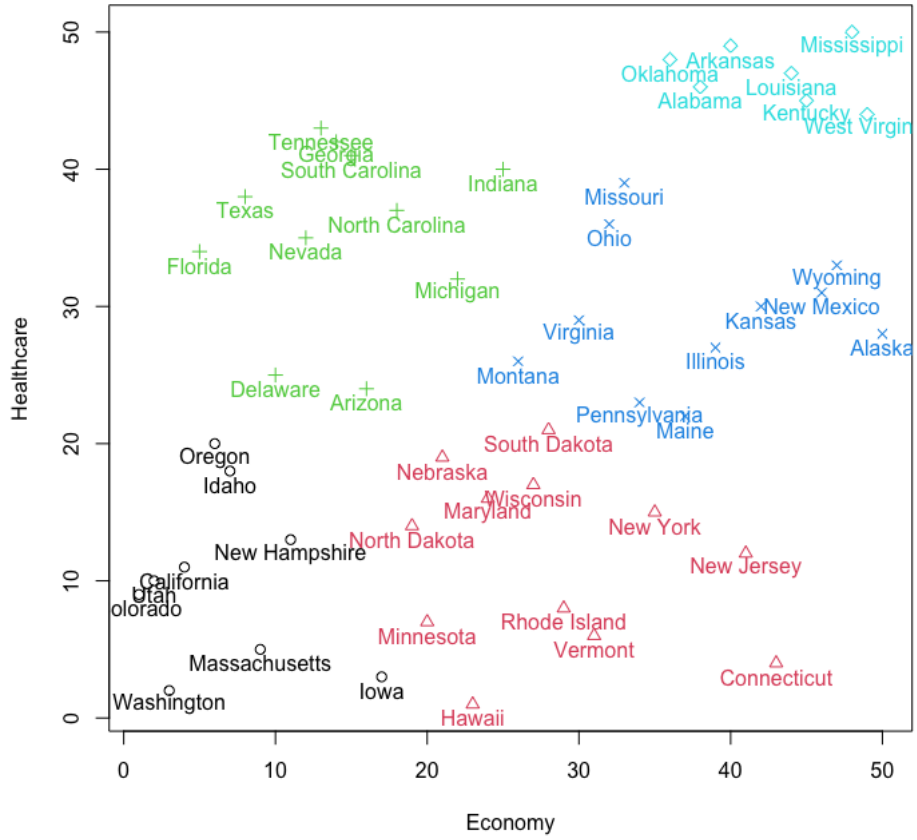
Available components:

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

- The sizes of the five clusters are 9, 12, 11, 11 and 7. This means that there are 9, 12, 11, 11 and 7 states respectively distributed in cluster 1-5.
- The average Healthcare ranking for cluster 1-5 are 10.11, 11.67, 35.55, 29.45 and 47.00.
- The average Economy ranking for cluster 1-5 are 6.67, 28.42, 14.36, 37.82 and 42.86.

Question 1 (b)

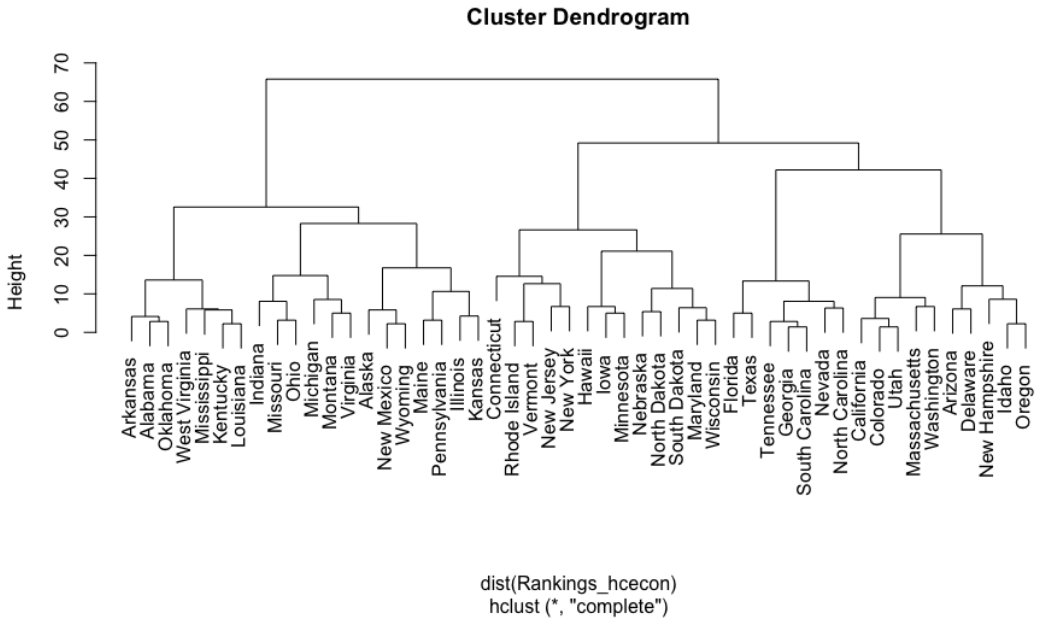
Output:



- The K-means clustering plot is shown above.

Question 2 (a):

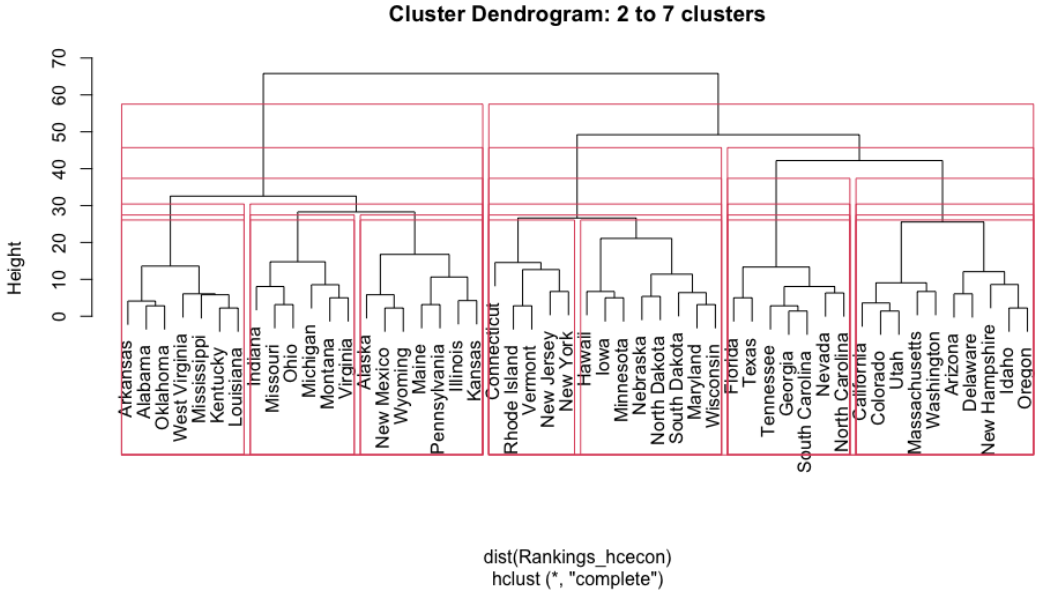
Output:

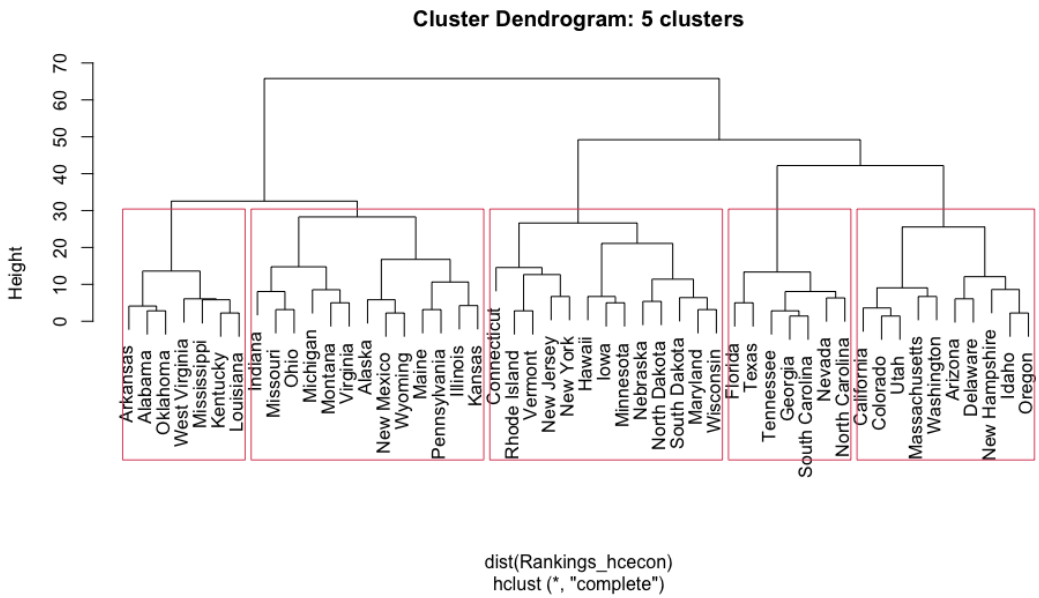


♦ The resulting dendrogram is shown above.

Question 2 (b):

Output:

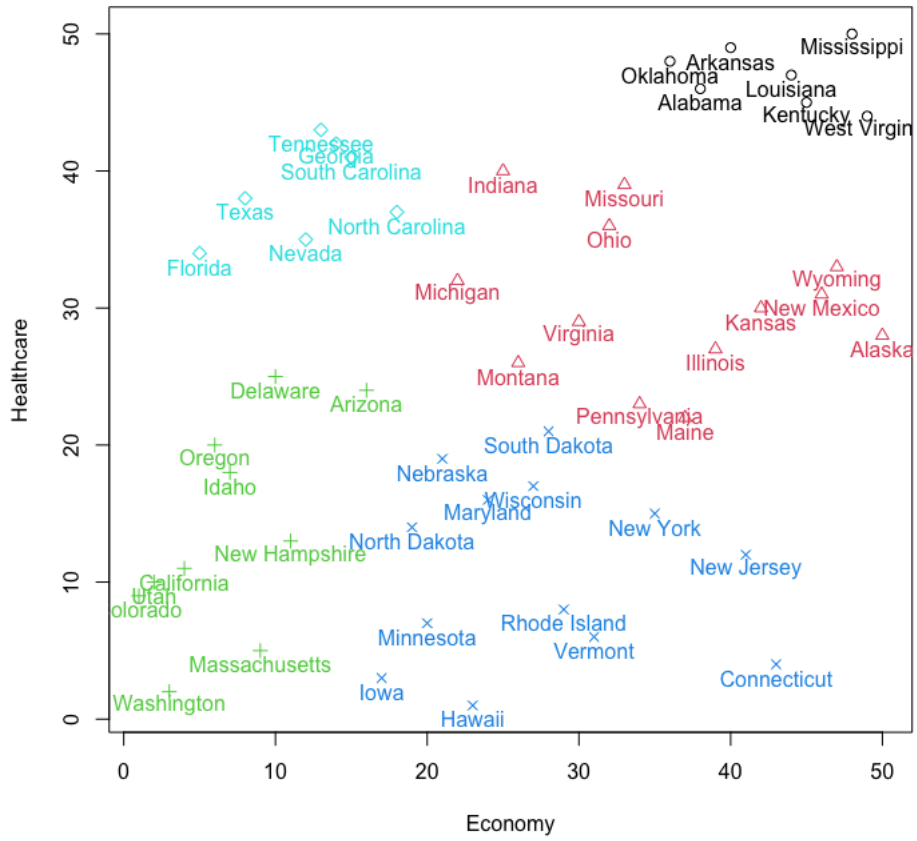




- By plotting images for 2 clusters to 7 clusters, we can observe that the intergroup dissimilarity between the clusters begins to significantly decrease after dividing into 5 clusters. On the other hand, the result of 5 clusters is relatively close to that of the K-means algorithm (compare the conclusions of Question 1 (b) and Question 2 (c)), and the number of clusters is moderate, making it suitable for subsequent analysis.

Question 2 (c):

Output:



♦ The hierarchical clustering plot of 5 clusters is shown above.

Conceptual Part

Question 1:

- Cluster variation for Cluster 2:

$$d_{34} = (39.9 - 31.2)^2 + (116.4 - 121.5)^2 = 101.7$$
$$d_{35} = (39.9 - 30.6)^2 + (116.4 - 104.1)^2 = 237.8$$
$$d_{45} = (31.2 - 30.6)^2 + (121.5 - 104.1)^2 = 303.1$$
$$d_{34} + d_{35} + d_{45} = 642.6$$
$$W(C_2) = \frac{1}{3}(642.6) = 214.2$$

- The total within-cluster variation for both Cluster 1 and Cluster 2:

$$W(C_1) + W(C_2) = 225.6$$

Question 2:

- Centroid of Cluster 1:

$$X_1 = (1 + 1 + 0) \div 3 = 0.67, X_2 = (4 + 3 + 4) \div 3 = 3.67$$

- Centroid of Cluster 2:

$$X_1 = (5 + 6 + 4) \div 3 = 5.00, X_2 = (1 + 2 + 0) \div 3 = 1.00$$

- It is easy to derive from the following formula the distances of the observations to the centroids Cluster 1 and Cluster 2:

$$\sqrt{(X_{1i} - 0.67)^2 + (X_{2i} - 3.67)^2}, \sqrt{(X_{1i} - 5.00)^2 + (X_{2i} - 1.00)^2}$$

- The following table displays the complete calculations:

Obs.	Distance from C1	Distance from C2
1	<u>0.47</u>	5.00
2	<u>0.75</u>	4.47
3	<u>0.75</u>	5.83
4	5.09	<u>0.00</u>
5	5.59	<u>1.41</u>
6	4.96	<u>1.41</u>

- Therefore, observations 1, 2, and 3 are in cluster 1, and observations 3, 4, and 5 are in cluster 2. There is no change in the obtained results, so the clustering results are reasonable.

Question 3:

1) Using "single" linkage clustering:

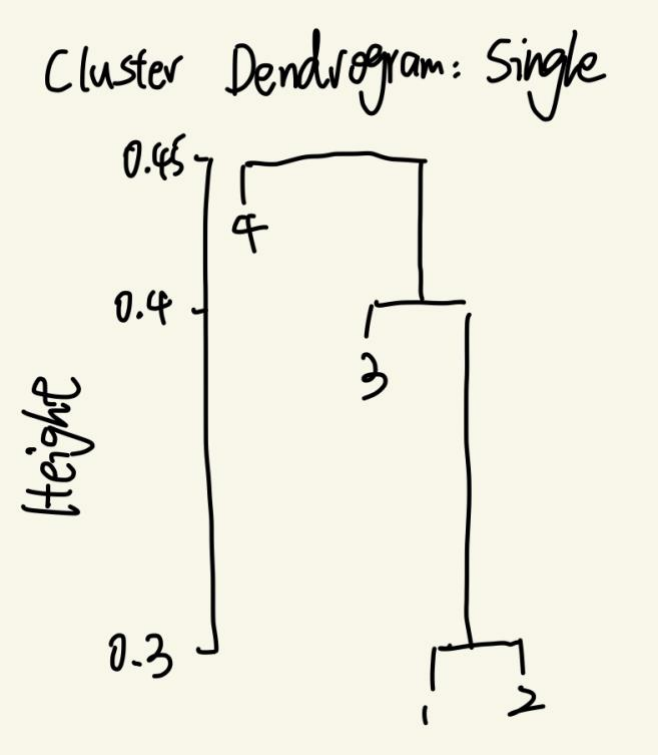
- The closest clusters are 1 and 2 with a dissimilarity of 0.30.
- Following the merger of Cluster 1 and 2, the updated dissimilarity matrix is displayed below:

	C3	C4
C1, C2	Min (0.40, 0.60) = 0.40	Min (0.80, 0.70) = 0.70
C3		0.45

- The closest clusters are 1, 2 and 3 with a dissimilarity of 0.40.
- Following the merger of Cluster 1, 2 and 3, the updated dissimilarity matrix is displayed below:

	C4
C1, C2, C3	Min (0.70, 0.80, 0.45) = 0.45

- The only remaining clusters are (1, 2, 3) and (4) with the dissimilarity 0.45.



- The resulting dendrogram is shown above.

2) Using "average" linkage clustering:

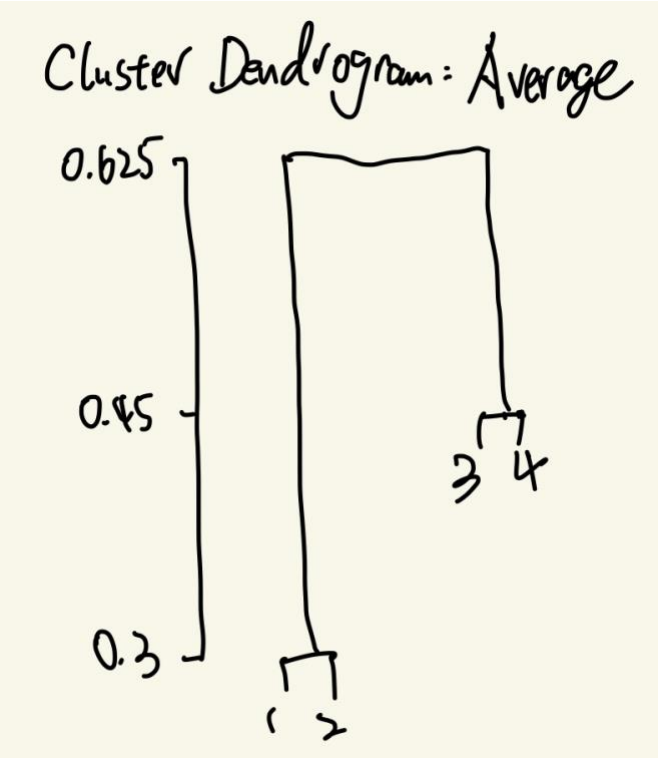
- The closest clusters are 1 and 2 with a dissimilarity of 0.30.
- Following the merger of Cluster 1 and 2, the updated dissimilarity matrix is displayed below:

	C3	C4
C1, C2	Avg (0.40, 0.60) = 0.50	Avg (0.80, 0.70) = 0.75
C3		0.45

- The closest clusters are 3 and 4 with a dissimilarity of 0.45.
- Following the merger of Cluster 1, 2 and 3, the updated dissimilarity matrix is displayed below:

	C3, C4
C1, C2	Avg (0.40, 0.70, 0.60, 0.80) = 0.625

- The only remaining clusters are (1, 2) and (3, 4) with the dissimilarity 0.625.



- The resulting dendrogram is shown above.