CS3481
Introduction to Data Science

Assignment 3
Clustering

Table of Contents

***Question-1*...2**

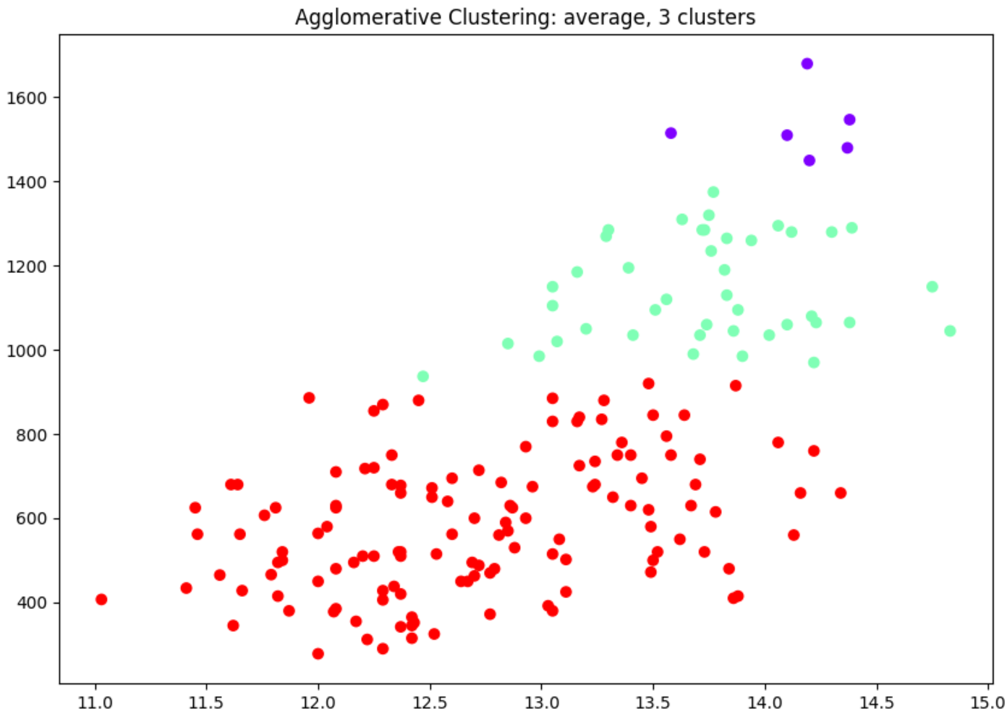
***Question-2*...7**

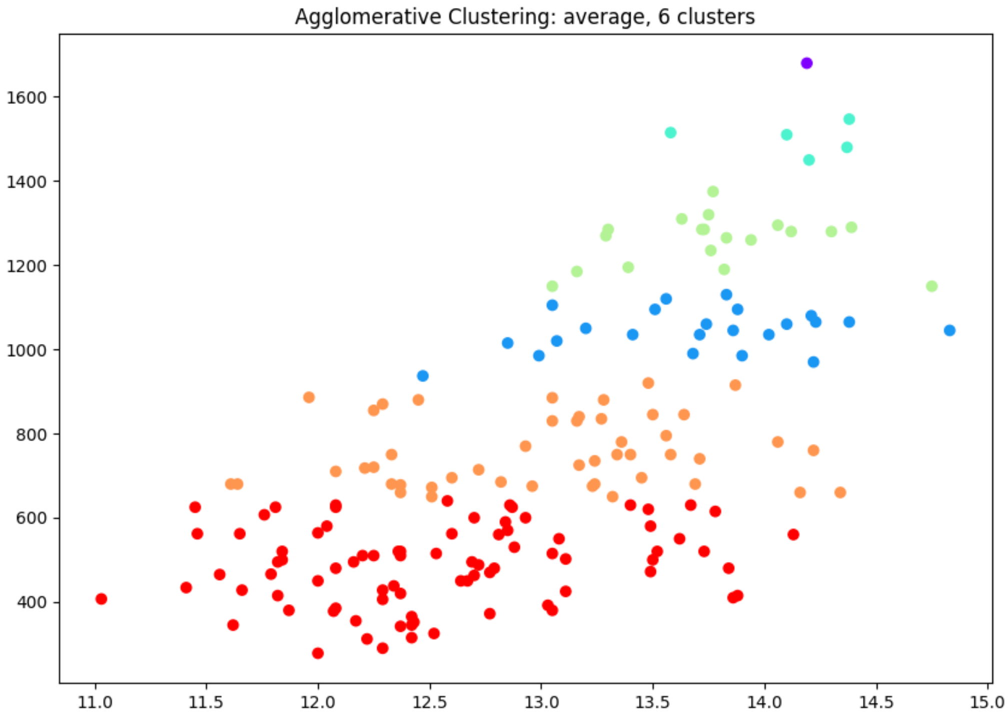
***Question-3*...8**

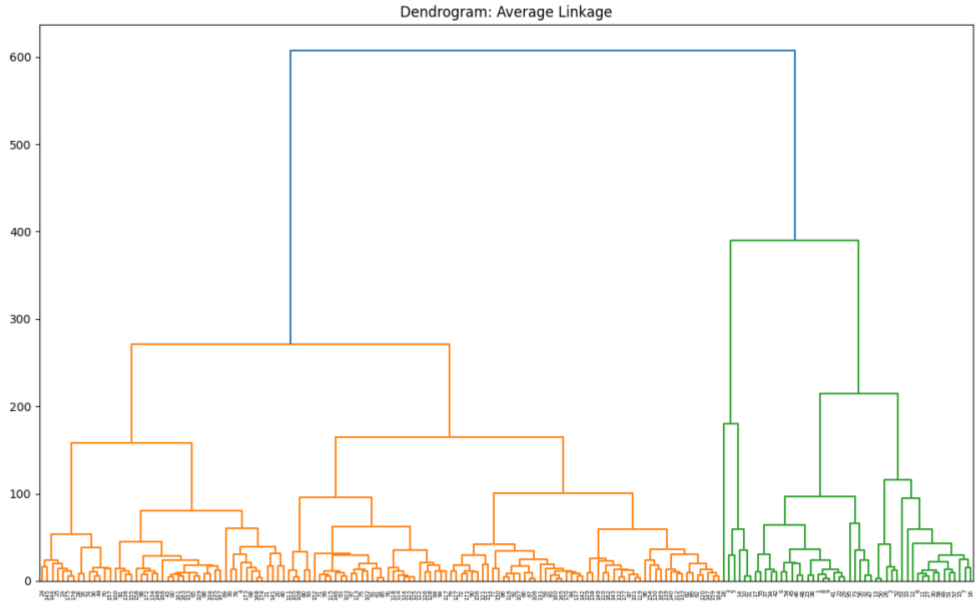
***Question-4*...10**

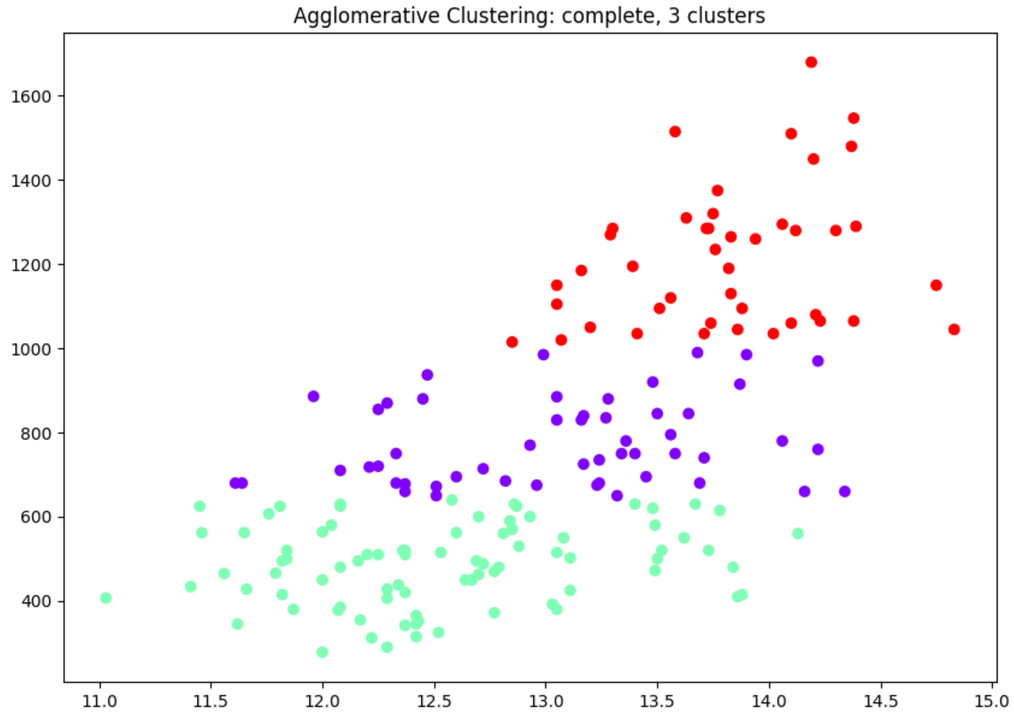
Name: Avi Malhotra
Student ID: 55773896

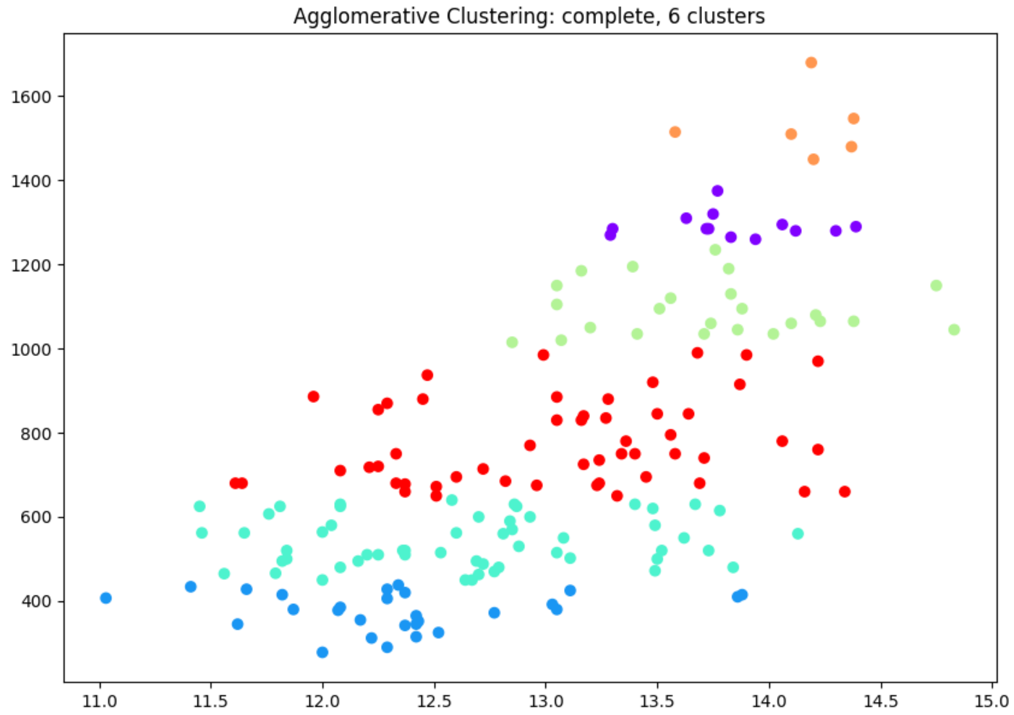
Question-1

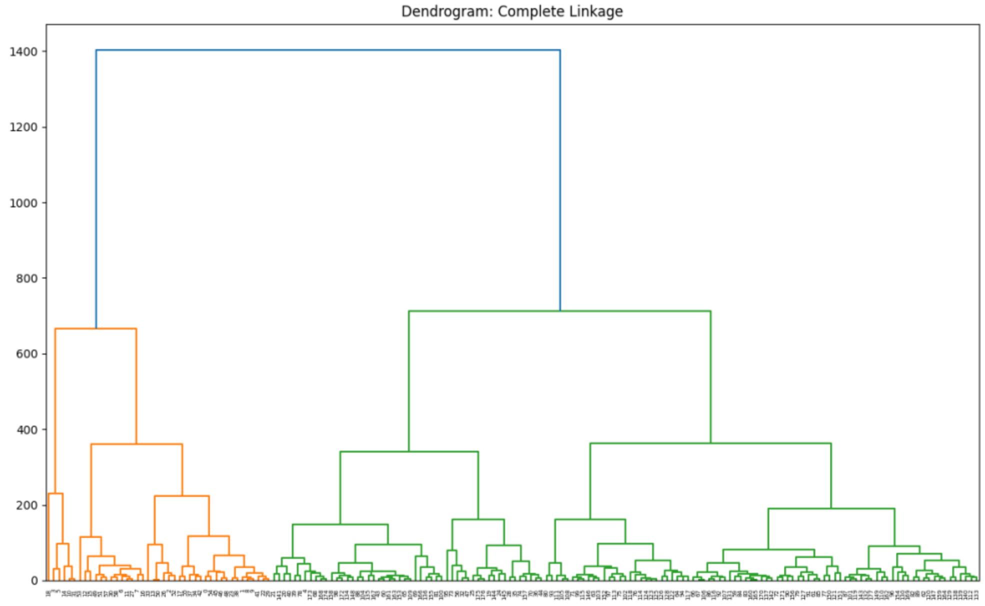

Agglomerative Clustering: average, 3 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0

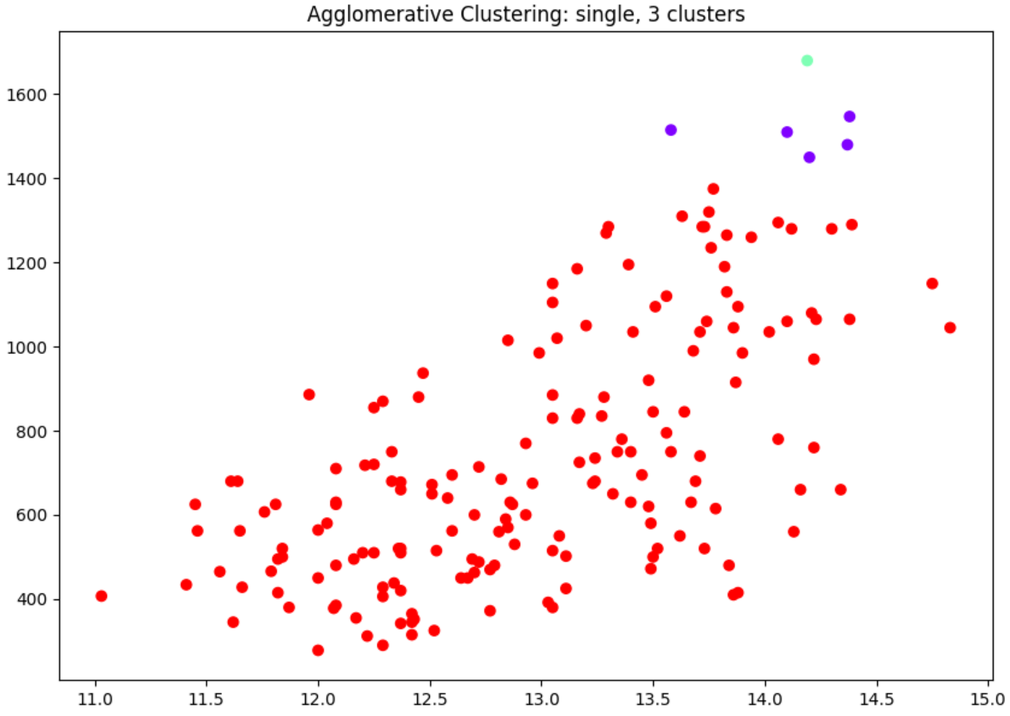

Agglomerative Clustering: average, 6 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0

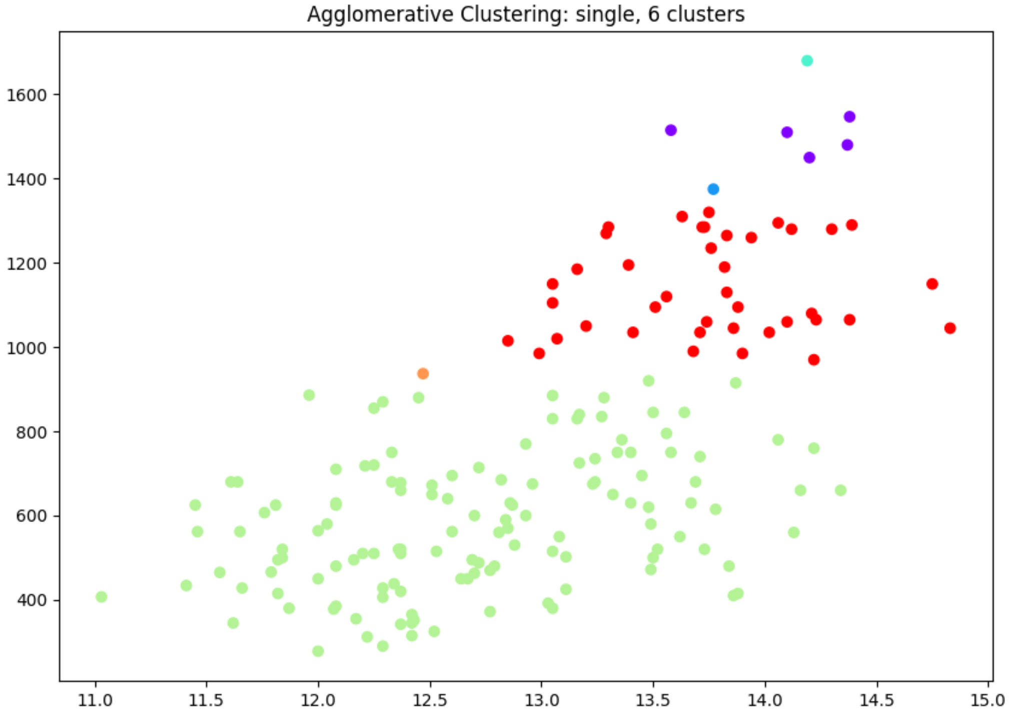

Dendrogram: Average Linkage
600
500
400
300
200
100
0

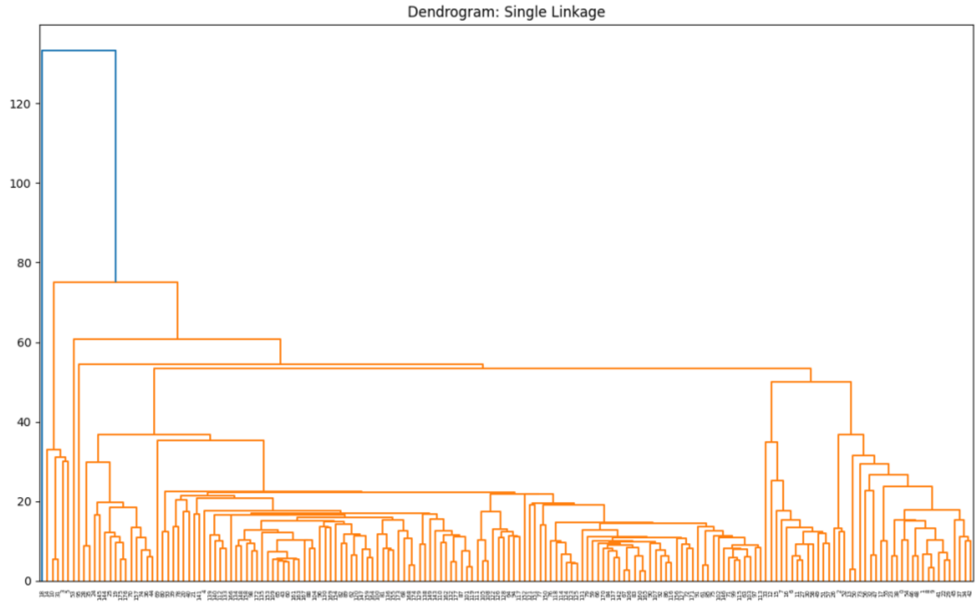

Agglomerative Clustering: complete, 3 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0


Agglomerative Clustering: complete, 6 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0


Dendrogram: Complete Linkage
1400
1200
1000
800
600
400
200
0


Agglomerative Clustering: single, 3 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0


Agglomerative Clustering: single, 6 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0



Analyzing the dendrograms, it can be seen that the Single link Clustering is a clear exception given the reasonably consistent shape of the other two clustering linkages. The nodes in Single Link Clustering are tightly intervened, perhaps because the dendrogram uses the MIN version of hierarchal clustering.

Furthermore, the clusters were visualized using Agglomerative clustering, wherein we see that for the cases of single linkage with 3 and 6 clusters, and for average linkage with 6 clusters, there is a cluster that contains only a single entry. These findings are largely in tandem with the dendrograms formed.

Given that complete clustering uses the MAX version of hierarchical clustering, the distances between nodes is largest in the group. The clusters predominant in the Agglomerative diagrams for complete linkage with clusters = 3 and 6 are purple and red respectively.

Average linkage is the intermediate of single and complete, as it takes the average distance computed across every pair. When comparing the distances of average linkage's dendrogram, they clearly lie between single linkage and complete linkage.

Question-2

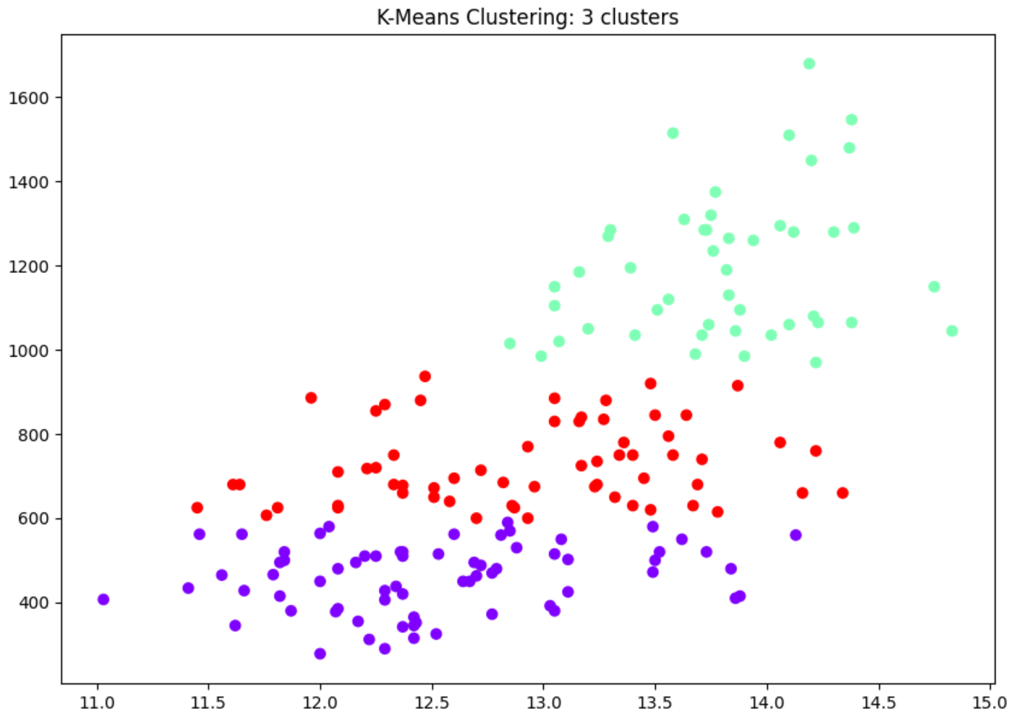
The dendrograms are particularly useful in deciphering the following patterns:

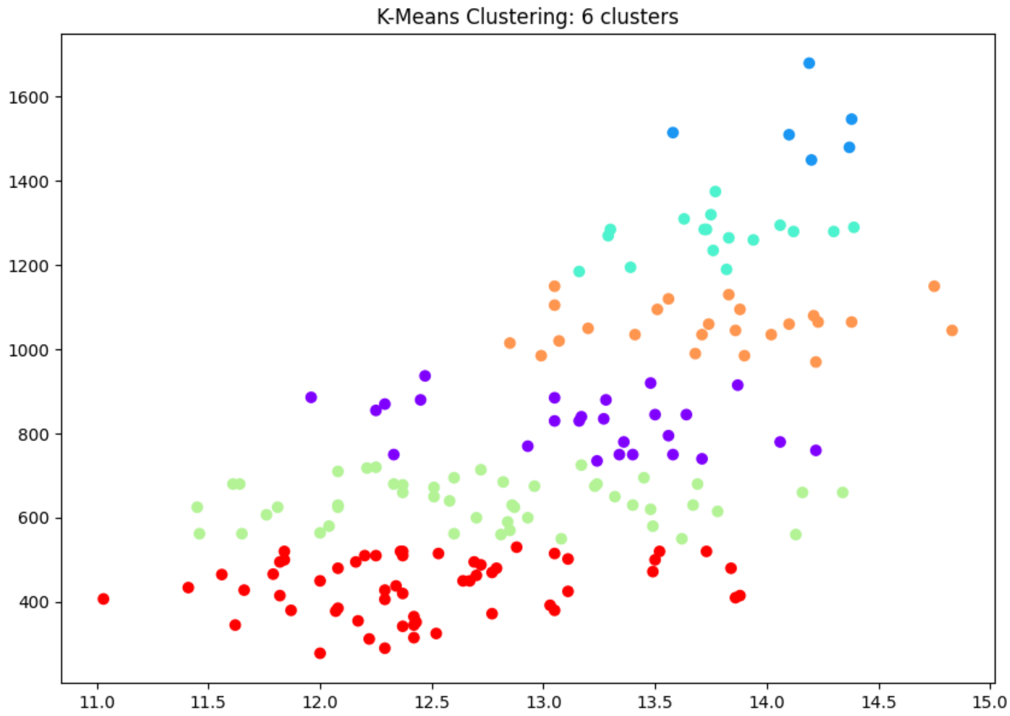
- In the case of single linkage, the cluster merging distance is the lowest of the set, given that the clusters use the minimum distance for merging. This is further reflected by the densely populated node distances.

- In the case of average linkage, the merging distance is larger than single linkage, and the nodes are also comparatively more spaced out.

- In the case of complete linkage, the cluster merging distance is the largest, given that the clusters use the maximum distance for merging. This is further reflected by the highly spread-apart node distances.

Question-3


K-Means Clustering: 3 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0

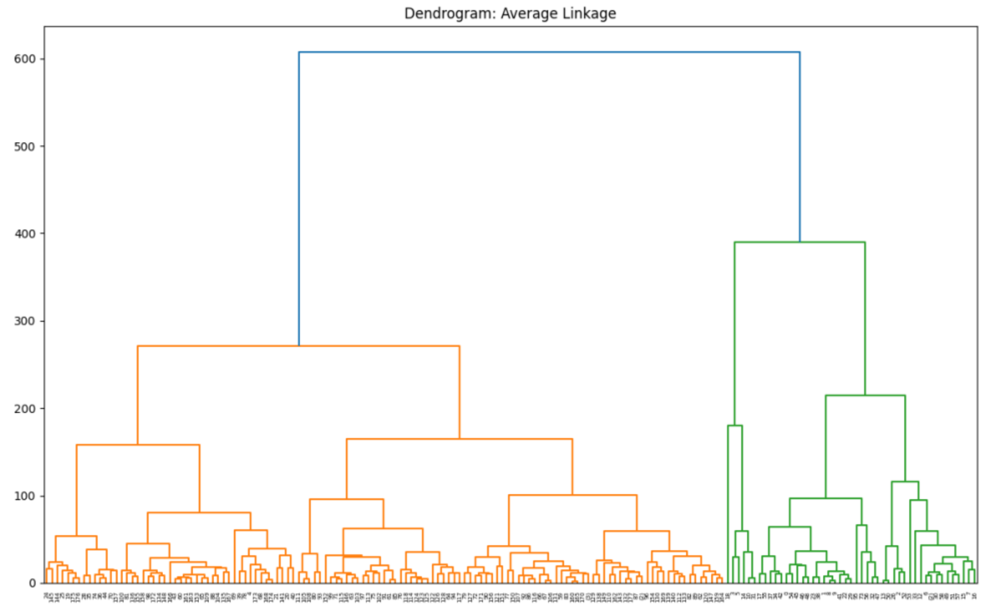

K-Means Clustering: 6 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0

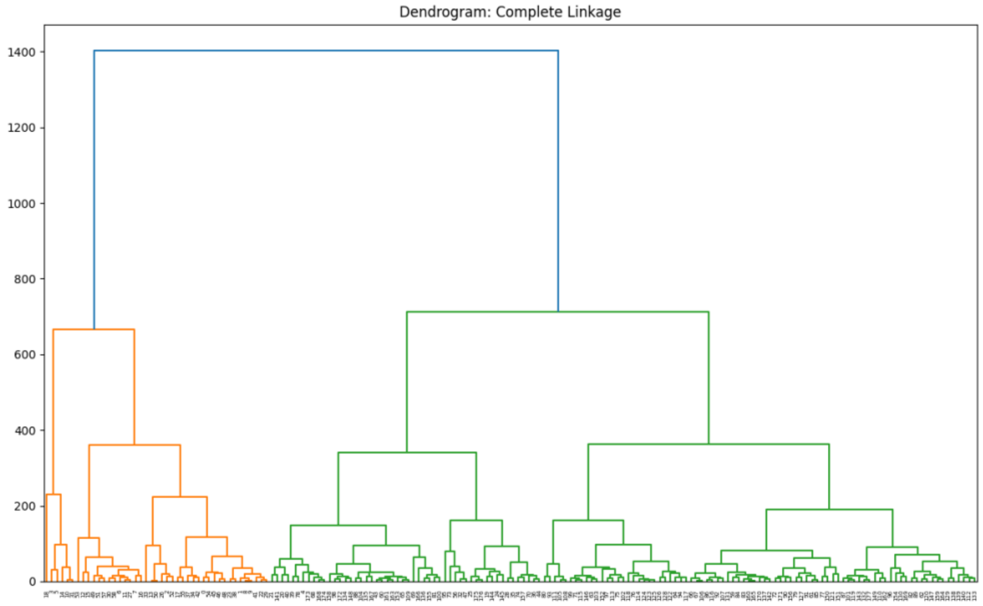
As seen from the above diagrams and the ones shown in question-1, the borders of Agglomerative clusters are the same as the ones for KMeans clusters in both cases wherein the number of clusters equals 3 or 6. Moreover, the number of nodes each cluster for KMeans clustering, complete linkage in Agglomerative clustering, and average linkage in Agglomerative clustering are the same. However, single linkage again continues to be the outlier here, bearing no resemblance whatsoever to any of the other 3 scatterplots.

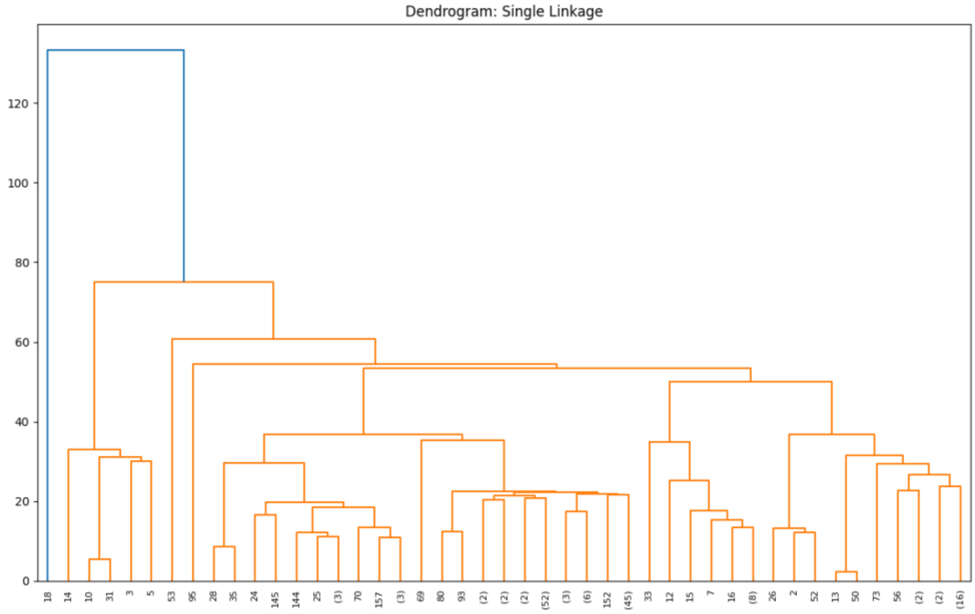
Question-4

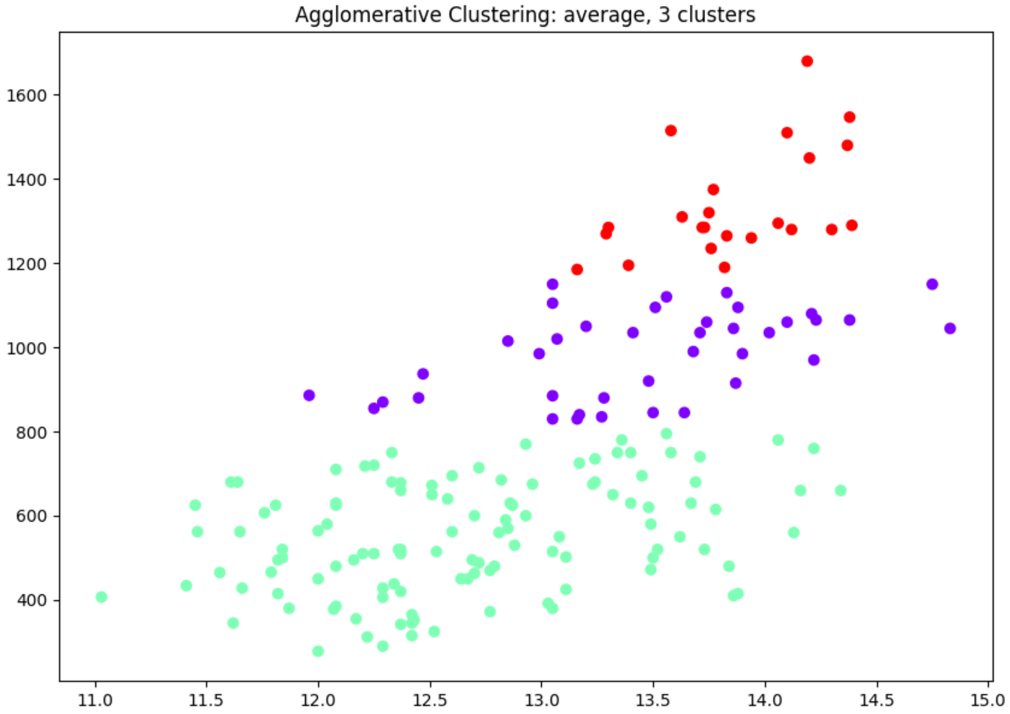
The following sets of attributes were removed from the original data:

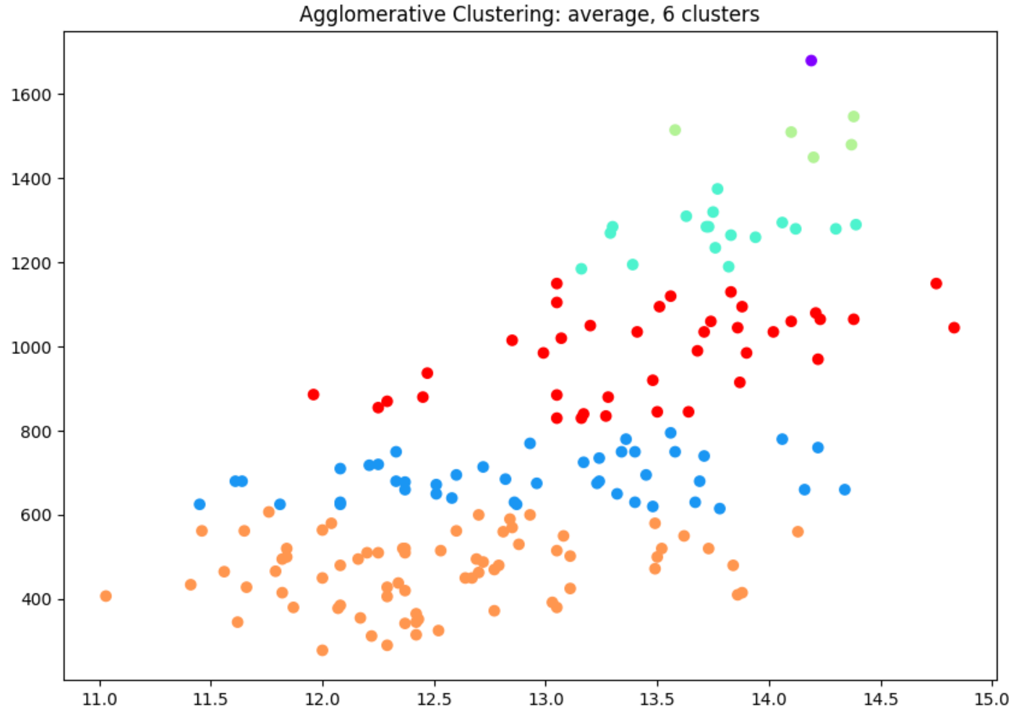
- Alcohol,
- Ash,
- Hue

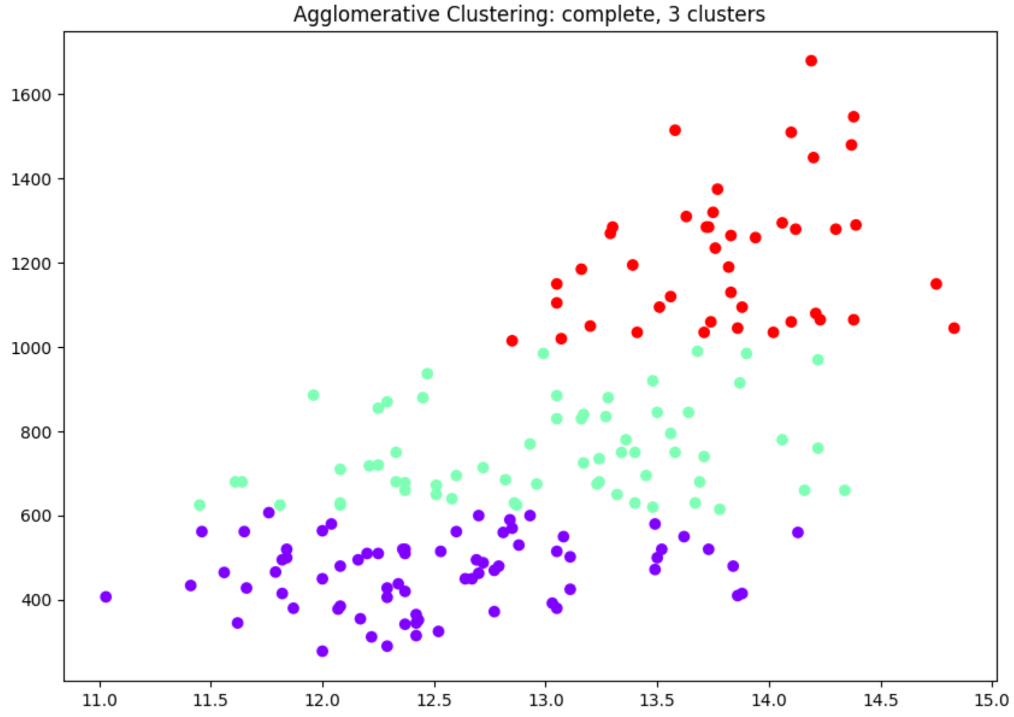


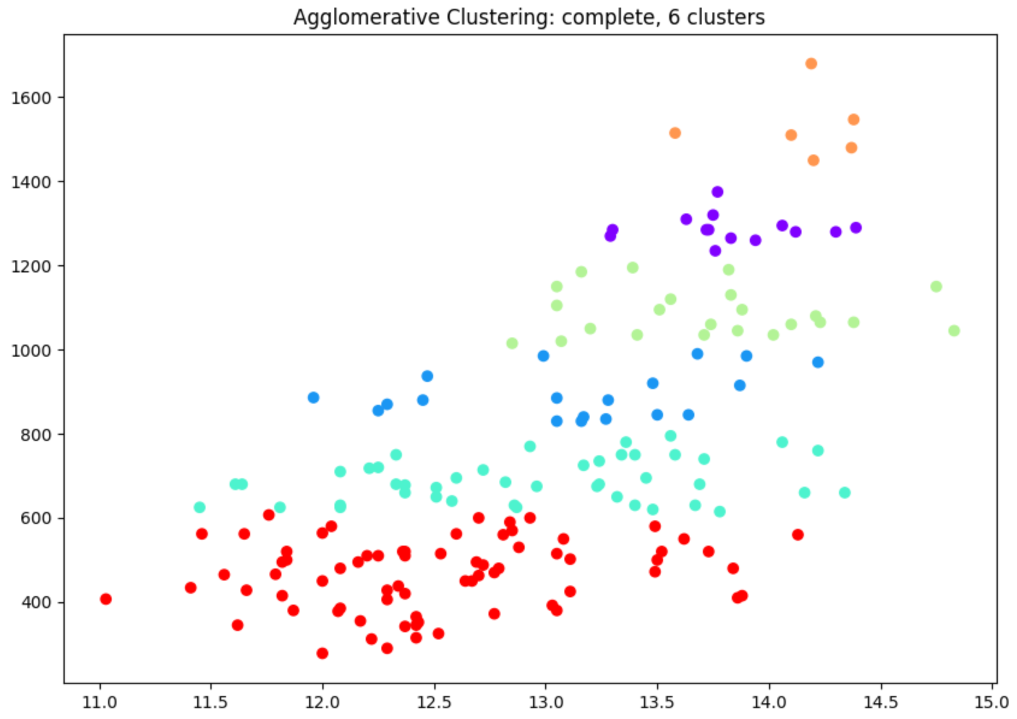


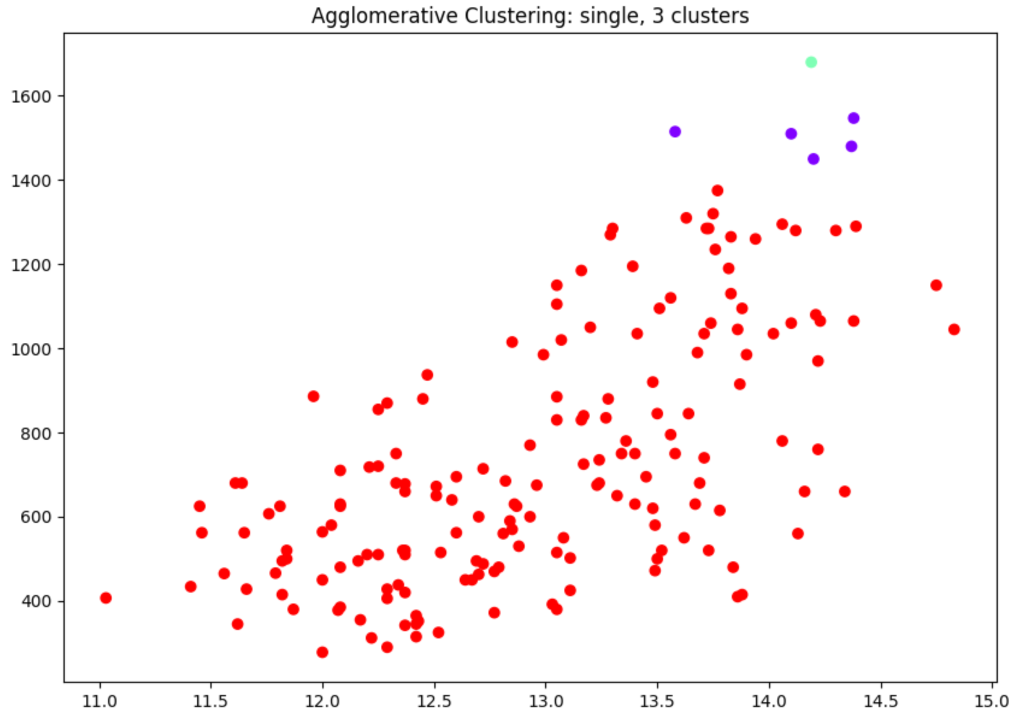

Dendrogram: Single Linkage

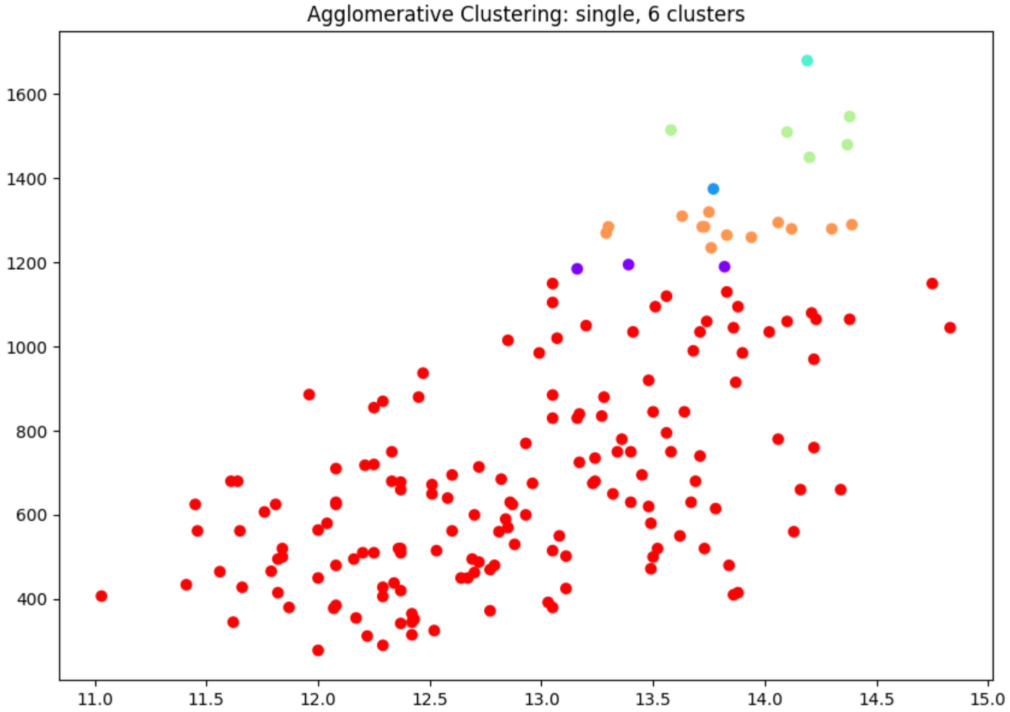

Agglomerative Clustering: average, 3 clusters


Agglomerative Clustering: average, 6 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0


Agglomerative Clustering: complete, 3 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0


Agglomerative Clustering: complete, 6 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0


Agglomerative Clustering: single, 3 clusters
1600
1400
1200
1000
800
600
400
11.0
11.5
12.0
12.5
13.0
13.5
14.0
14.5
15.0



As seen from the above set of diagrams, the Agglomerative Clustering for the filtered set of attributes does resemble the diagrams shown in question 1 wherein the number of clusters equals 3. Even though the same statement is not entirely true when the number of clusters increases to 6, the same characteristics are still prevalent in the Agglomerative clustering diagrams. That is, particularly the proximity of different sets of nodes. As for the dendrograms, it was very surprising to see a varied dissimilarity index in the case of single linkage. Examining the x axis more carefully reveals that removing the alcohol, ash, and hue attributes has reduces the horizontal distance. While this may not be prevalent in the case when the clusters are 3, as the diagrams are essentially the same, the difference is quite noticeable in the case when the number of clusters equals 6 for both the above agglomerative cluster in single linkage and the one depicted in question 1.