Name: Aryah Girish Kasliwal
SID: 55972222

Assignment 3 – Clustering

1. **Compare the hierarchical structures generated using single link, complete link and group average for the *Wine* data set. (30%)**

In hierarchical clustering, initially every single point is considered as a cluster on its own. Every single cluster is then merged into larger clusters until all data points form one single large cluster.

This process of hierarchical clustering can be represented by a dendrogram. The following are three dendrogram for single-link, complete-link, and average-link hierarchical clustering.

Complete link

In complete link hierarchical clustering, we merge clusters based on which ever merged cluster would have the smallest diameter. In other terms, between every cluster, the maximum pairwise distance is calculated, which would give us an estimate of the sizes of the possible merged clusters, then the smallest of these maximum pairwise distance is chosen, which would lead to the smallest merged diameter out of all the possible clusters.

Complete link clustering can also be understood in terms of cliques. Let d(n) be the diameter of the cluster formed at the nth step of clustering. If we then make a graph g(n) as a graph that links all data-points with a maximum distance of d(n). Then the cluster created after the nth step forms all cliques of g(n). Hence this process is called complete link clustering.

The dendrogram generated on performing complete link hierarchical clustering on the wine dataset in the UCI machine learning repository is attached below.

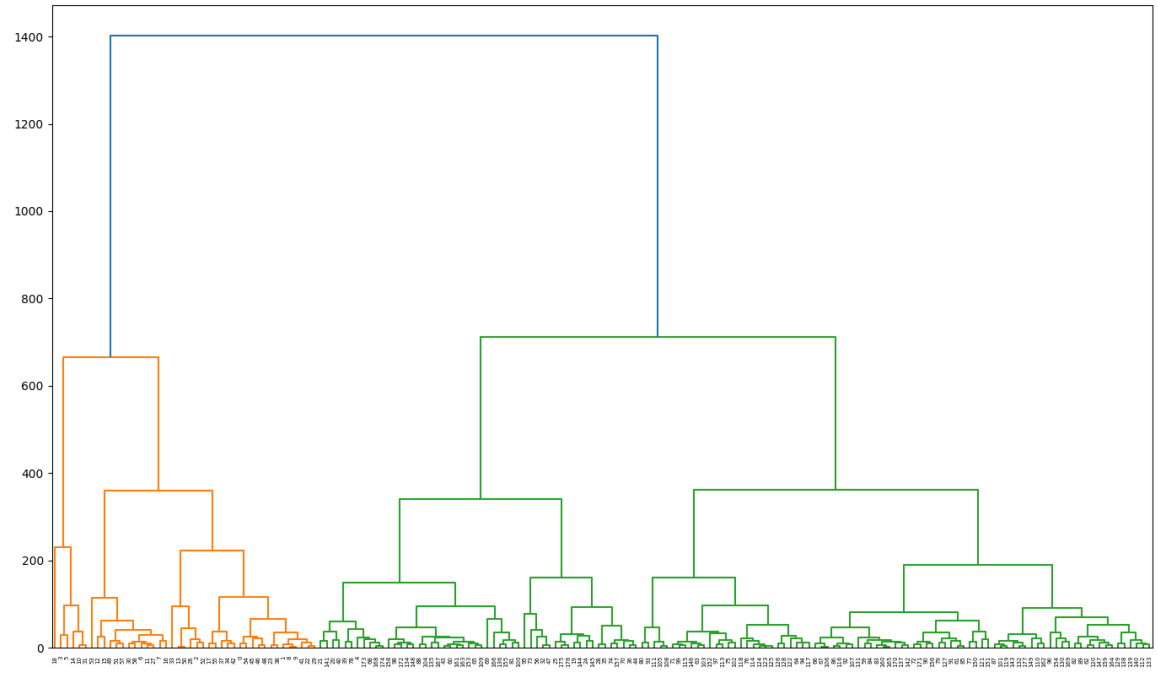


Figure 1

Single link

In single link hierarchical clustering, clusters are merged based on the smallest distance between the two closest points in the clusters. To be more exact, first for every pair of clusters, the two closest points in each are chosen and the distance between them is calculated. This same process is repeated for every pair of clusters. In the end the two clusters whose distance of closest points is smallest is chosen and merged.

Just like complete link, single link clusters can also be explained in theoretical terms. In step n, if d(n) is the distance between closest points of the two clusters merged, and g(n) is the graph of all data points with a maximum distance of d(n), then the clusters merged after step n are connected components of the graph.

The dendrogram generated on performing single link hierarchical clustering on the wine dataset in the UCI machine learning repository is attached below.

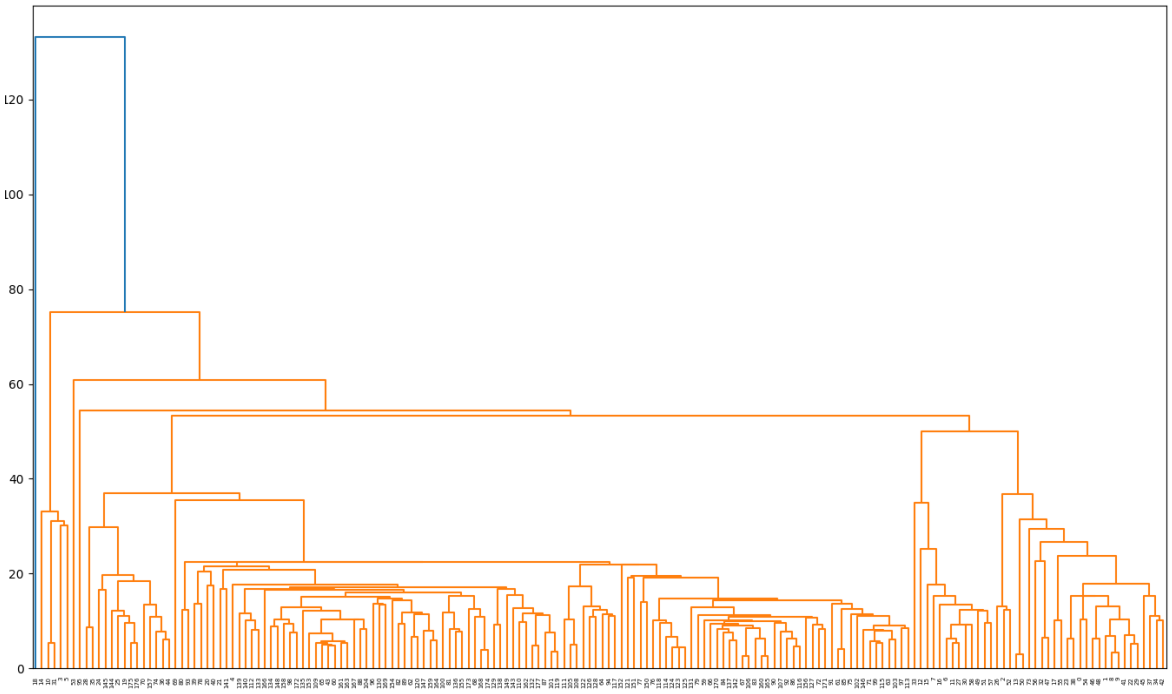


Figure 2

Average link

In average link hierarchical clustering, we find a balance between single link clustering and complete link clustering. Complete link clustering is sensitive to outliers, and single link clustering tends to form long chains of data points that go against the notion of clusters being spherical objects. In this form of clustering, in each iteration we pair clusters with highest cohesion. If our data points are represented as vectors, we can calculate the cohesion between them as the average dot product. Once we calculate the cohesion between all points in a cluster, we get the cohesion for that cluster. Repeating this process for all clusters gives us a clearer picture of which clusters are more cohesive and which are relatively less cohesive. We then merge clusters that lead to the highest cohesion out of all merges possible.

The dendrogram generated on performing average link hierarchical clustering on the wine dataset in the UCI machine learning repository is attached below.

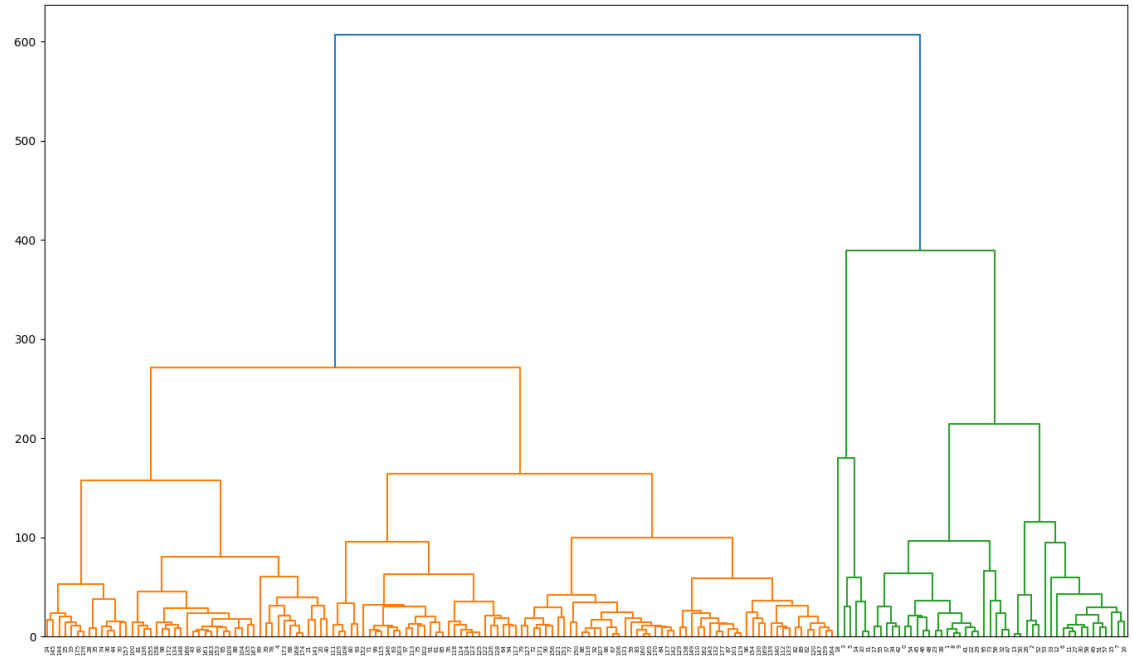


Figure 3

Structural comparison of the hierarchical structures

In figure 1 above we can see that the maximum height of the dendrogram is 1400 and two major clusters are formed around the value 700, these further forms 3 major clusters around the value 400, this tells us that average linkage forms clear and close clusters. Complete linkage forms two clear groups, shown as green and orange in the figure, with each sub-group further consisting of other clusters. Each sub cluster further forms multiple clusters almost containing equal number of data points; hence this is a good example of clustering.

In figure 2 above, we can see that the maximum height of the dendrogram is a little above 120 and just one major cluster is formed around the value 75, the other cluster is never formed, hence this is not a good example of clustering as all new points would be classified into the one cluster only. Even within this one cluster, the sub-clusters further formed are not equal in terms of the number of data points they contain, hence this is not a good example of clustering.

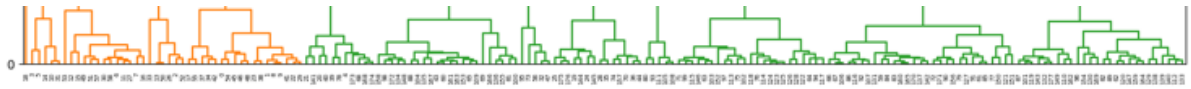
In figure 3 above, we can see that the maximum height of the dendrogram is 600 and two clusters green and orange are formed. Average linkage performs better than single linked. Though these clusters are formed at different values, 400 and 290 respectively, this means that values in the same cluster must be similar and values in different clusters must be different. Average linkage forms two clear groups, shown as green and orange in the figure, with each sub-group further consisting of other clusters. Each sub cluster further forms multiple clusters almost containing equal number of data points; hence this is a good example of clustering.

---

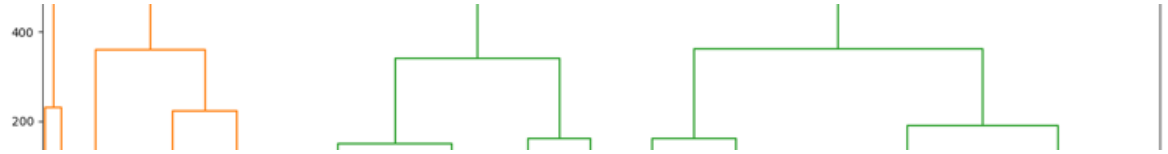
2. **For some of these hierarchical structures, observe the set of distance values at which cluster merge occurs and identify possible patterns from these values. (20%)**

Complete linked clustering

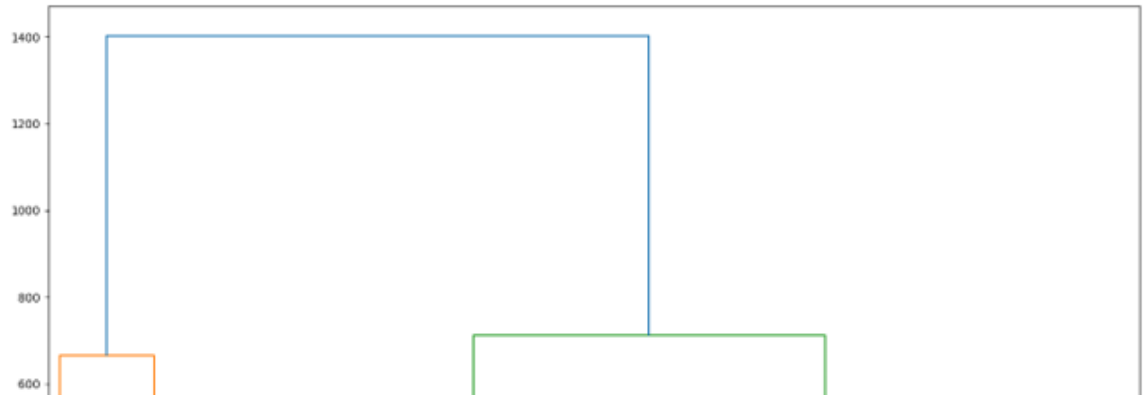
In figure 1 we can see that the first set of clustering occurs at a very low value, and only the data points that are close to each other are grouped together.



As our clusters grow larger, the clades (horizontal lines in the dendrogram where clusters are merged) become larger, however they are still consistent and equal to all other clusters clades too. The first notable clusters are formed around the value 200 which are further grouped together around the value 400. This indicates that complete clustering forms equal sized clusters and keeps all clusters equally balanced in terms of number of data points and merge values.

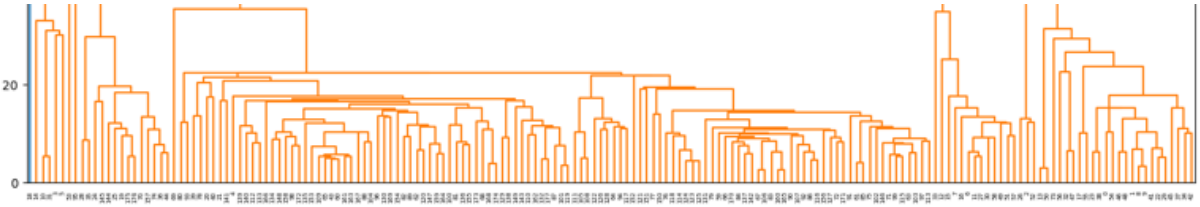


These clusters are further combined just above value 600 to form the two major groups, green and orange. These clusters formed are so different from each other than the combination of these two happens at 1400 which is way higher when compared to other methods of clustering namely single linked and average linked. A good balance between the cluster distribution helps us conclude that this is a good example of clustering the given data set and can effectively classify the data points into correct groups.



Single linked clustering

The dendrogram for single linked clustering (figure 2), we can notice that the bottom is very populated, with very long clades, this indicates that most of the data points are grouped into different clusters very soon. The long clades indicate that some data points that are not very similar are also grouped into clusters together. The narrow gap between one cluster merging once, and then merging again shows that some of the clusters formed are similar and are regrouped together very soon hence this methos is redundant.



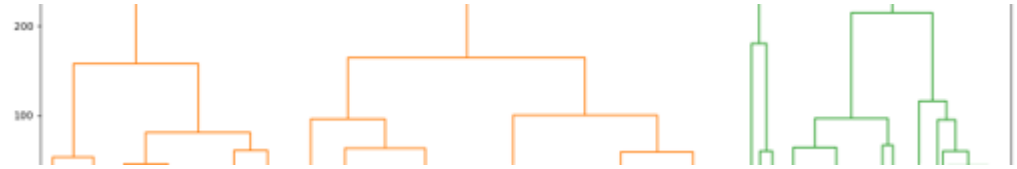
As we move further up this dendrogram, we notice that the clades become longer and longer, indicating that the clusters being grouped together are not very similar. If dis-similar clusters are grouped together, then the clustering is not efficient and may lead to wrong results.

In the end all clusters are merged into one cluster, with a clear imbalance as all data points are classified into orange colour and the other end of the blue line goes down all the way to the x axis. Hence this is not an ideal example of clustering and cannot effectively classify the data points into correct groups.

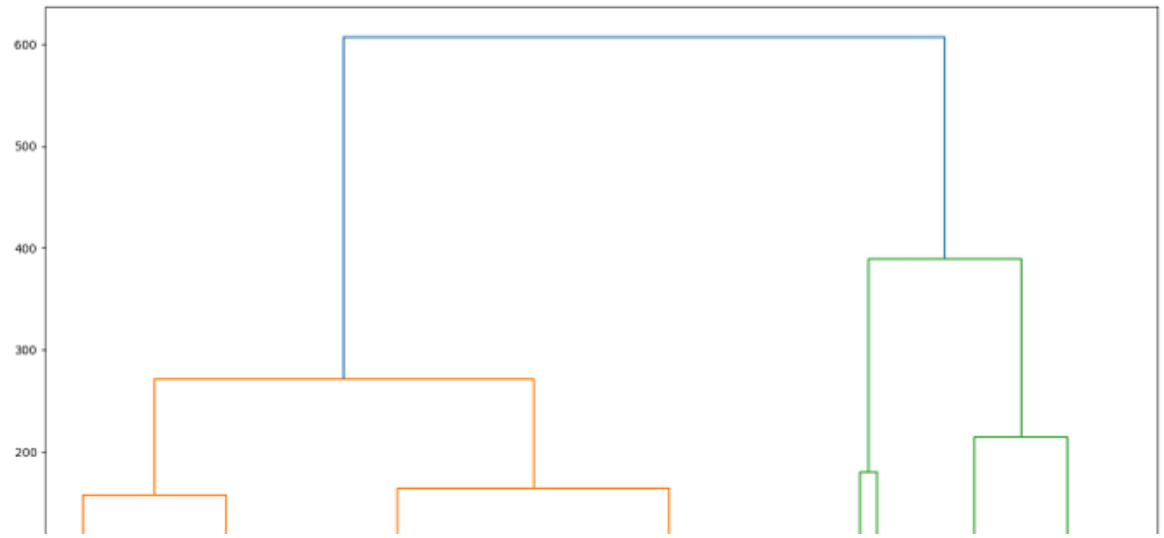
Average clustering

The dendrogram for average clustering is very balanced. A lot of data points that are similar to each other are group together early at a low value. This indicates that the points clustered are very similar to each other.

As we move further up the dendrogram, these clusters are grouped together to form larger clusters. It can be seen that the clades are of balanced length and all of them are formed around the same value, first around 100 and then around 170. This means that similar clusters are grouped together to form larger clusters and almost equal number of data points are in each cluster.



Finally, all these clusters form two large groups, green and orange. These groups merge finally at value 600, meaning these two groups are not as dis-similar from each other as the ones formed by complete linkage. A good balance between the cluster distribution helps us conclude that this is a good example of clustering the given data set and can effectively classify the data points into correct groups.



---

3. **Select different clustering solutions from the hierarchical structures, and compare the cluster groupings with the corresponding K-means clustering solutions (using the method KMeans from the module sklearn.cluster) based on, for example, the extent to which the clusters can capture the actual groupings of the data points according to their class labels. (30%)**

K-means result

On performing kmenas clustering on the wine dataset, I got the following result:

0 0

1 5

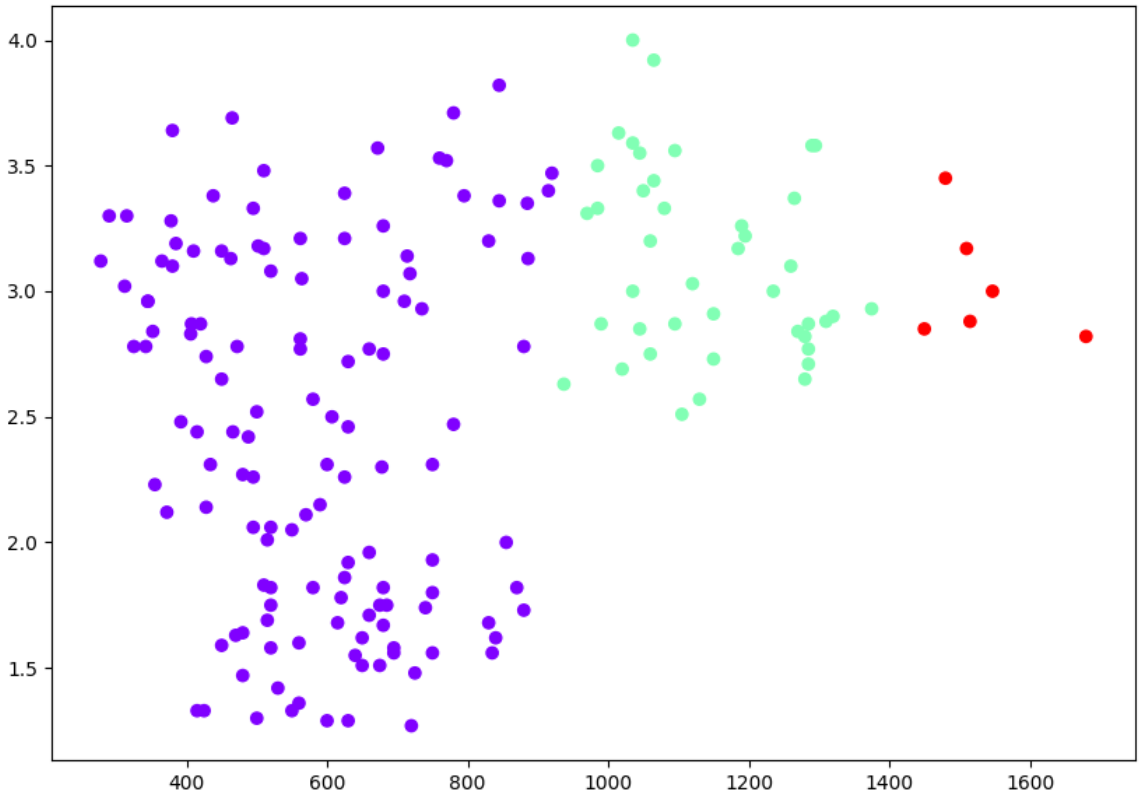
2 172

3 1

Which indicates that first cluster has 5 data point in it, second cluster has 172 data points in it, and the third cluster has 1 data point in it. We can now compare it with the graphs below formed by different forms of hierarchical clustering

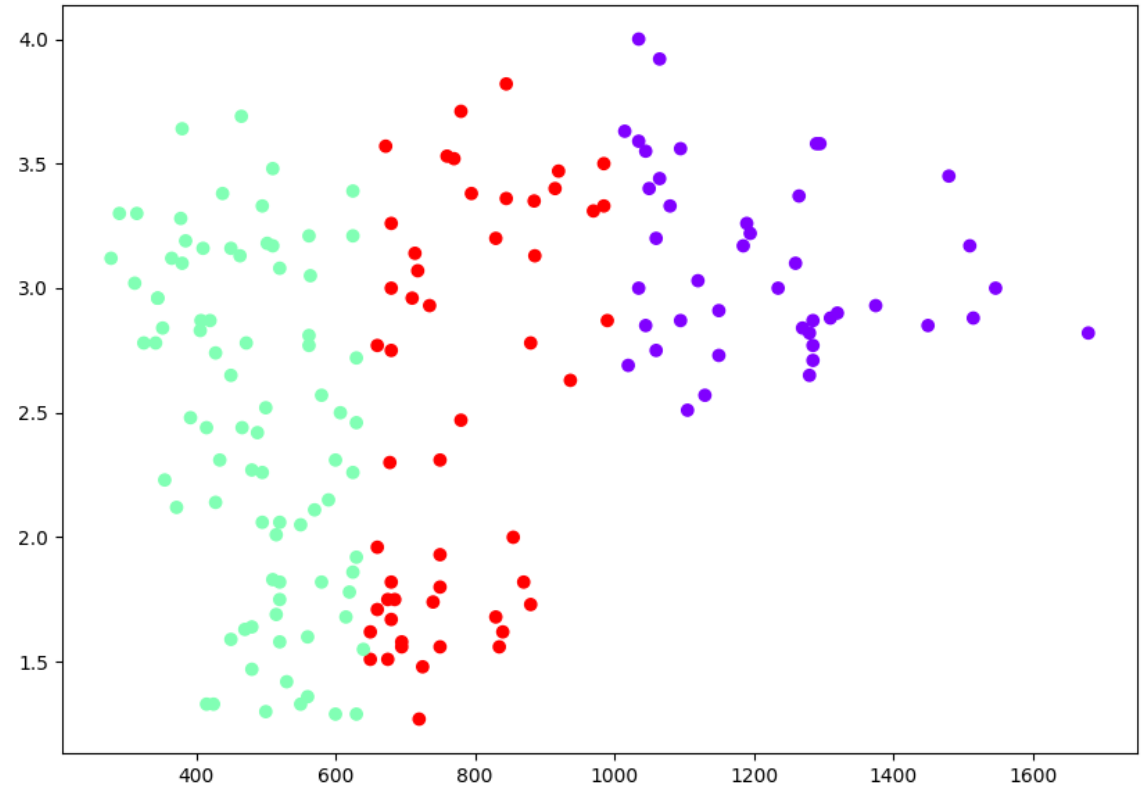
Average clustering

The following is the scatterplot for average linkage hierarchical clustering.



Complete linkage clustering

The following is the scatterplot for complete linkage hierarchical clustering.



On comparison, we can say that average linkage does a comparably better job than complete clustering because the results of average clustering are closer to k-means clustering than complete linkage is to k-means. K-means suggests that there are 5 data points in one cluster and 1 in another, but in complete linkage, the data points are equally divided in the clusters. However, the same is not the case with average linkage as it has lesser number of red, little teal, and most purple data points, which is closer to the result of k-means.

---

4. **Select different subsets of attributes from the data sets and re-perform hierarchical clustering. Compare the resulting hierarchical structures based on the selected attribute subsets with the original hierarchical structures. (20%)**

Note: In all the following subset groups of attributes, I have only performed average hierarchical linkage as it provided the best result with the complete dataset.

Group 1

In the first group of attributes selected, I've taken all except "alcohol", "ash", "proline", and "total_phenols" and perform average clustering. The dendrogram hence produced is shown in figure 4 below. This dendrogram is significantly different from figure 3 above. The resulting dendrogram is more like figure 2, with long clades and largely uneven distribution og clustered groups. Majority of data points are clustered in the green section and very less into the orange section. Further the maximum height of the dendrogram is also significantly less, with a max of just 50 when compared to 600.

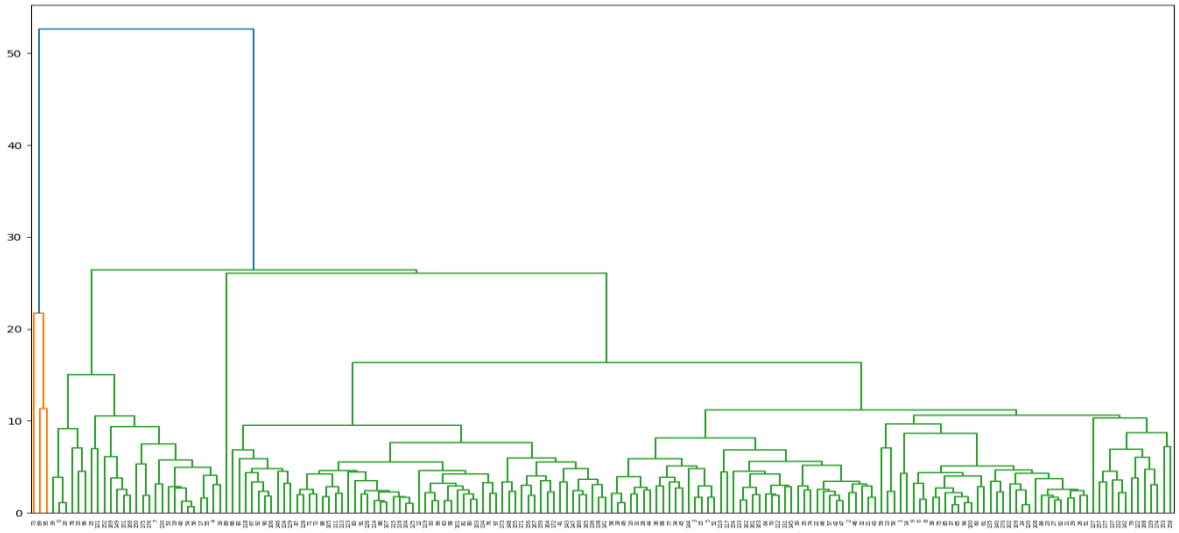
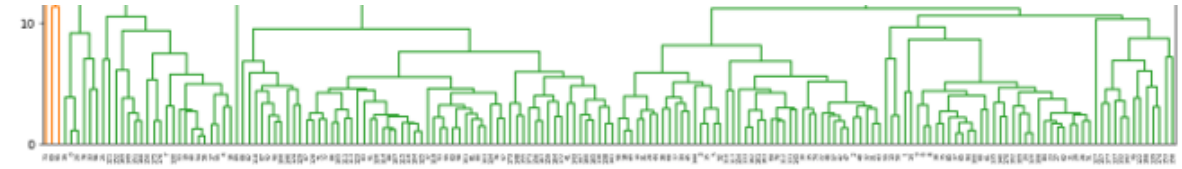


Figure 4

A lot of small clusters are formed very early and then merged together, which is unlike the result when all attributes are present. From this we can infer that the four attributes removed significantly affect the clustering.



Group 2

In the second group of attributes, I've taken all except "proline" and the resulting dendrogram is not much different from figure 4. From this we can say that proline is an important attribute as removing it makes a big difference in how clustering is performed. To test this, I also dropped other singular attributes such as ash and alcohol separately and the resulting dendrogram was more similar to figure 3, hence I can conclude that proline is a more important attribute than alcohol or ash.

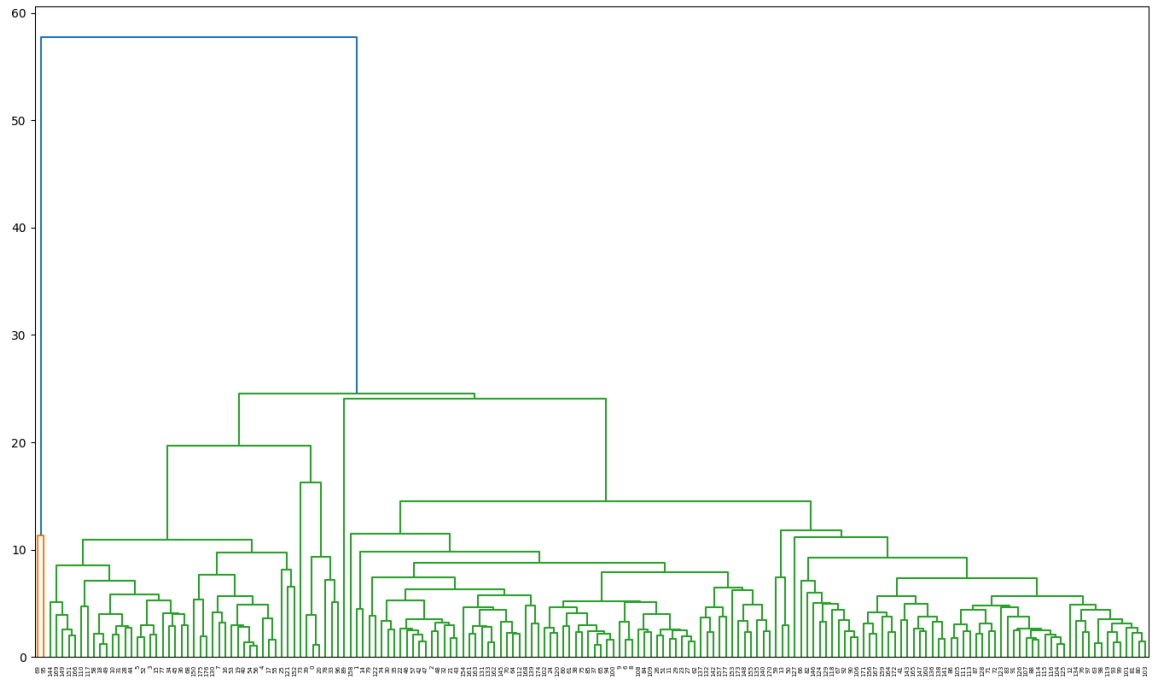


Figure 5

Group 3

To test my inference in group 2 that proline is one of the most important attributes, I formed a cluster using only attributes "proline" and "alcohol". The resulting dendrogram is shown below in figure 6. This dendrogram is similar to the one in figure 3 and hence it concludes our assumption in group 2 that proline is the most important attributes to decide the clustering.

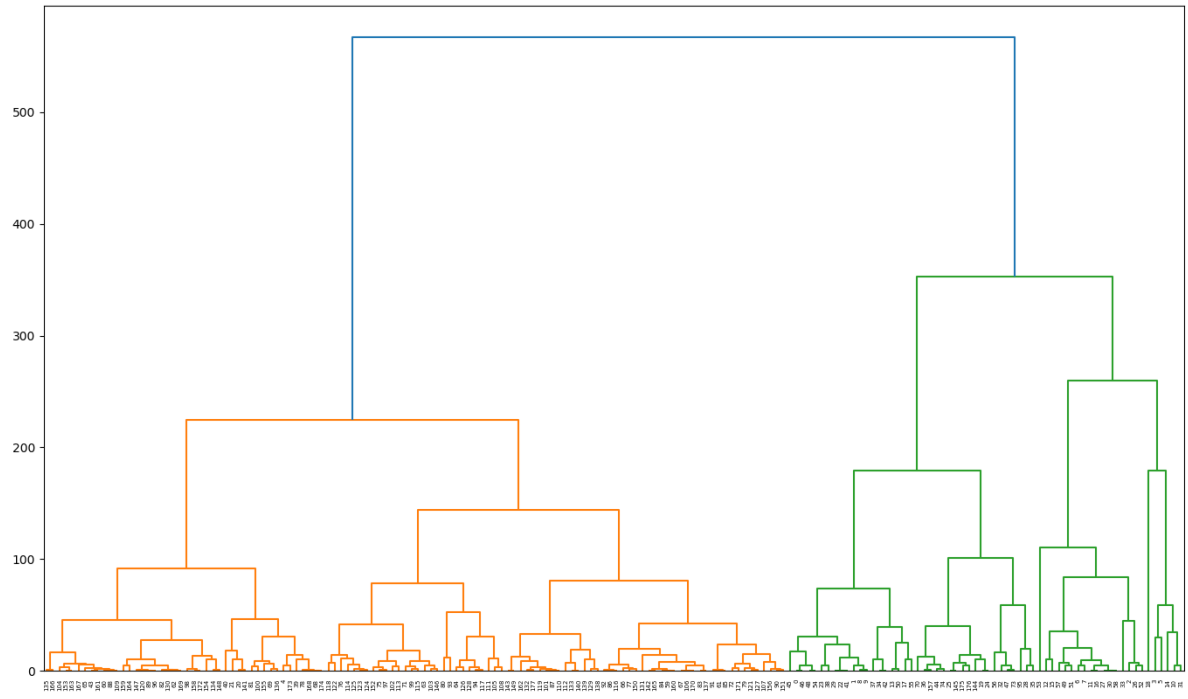


Figure 6

I tried dropping further attributes or choosing different subsets but none of them led to any substantial different is the dendrogram when compared to figure 3. Hence it can be said that proline is one of the most important attributes in this dataset.