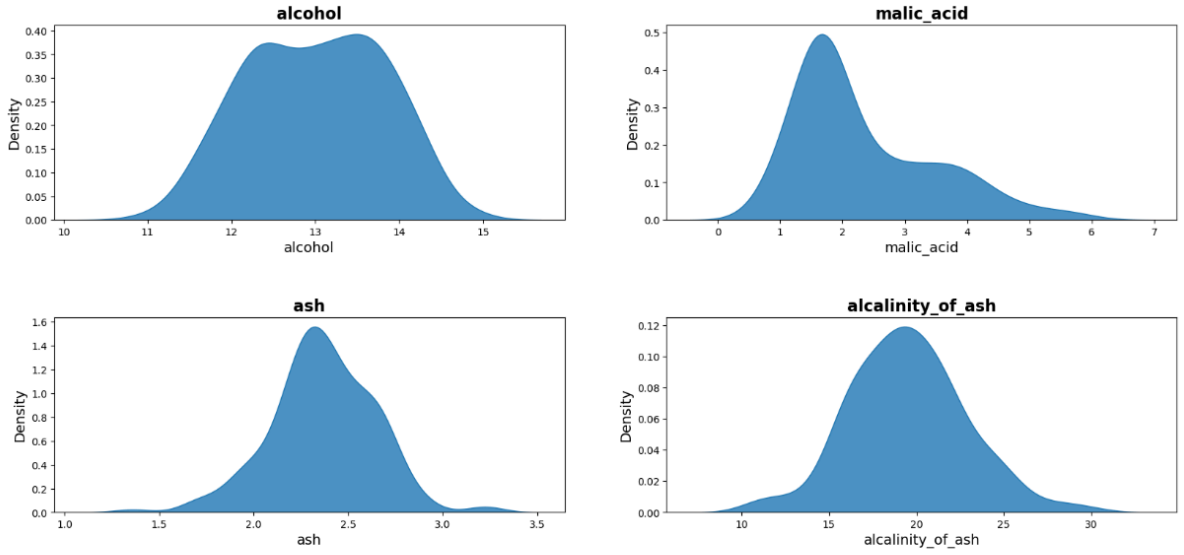
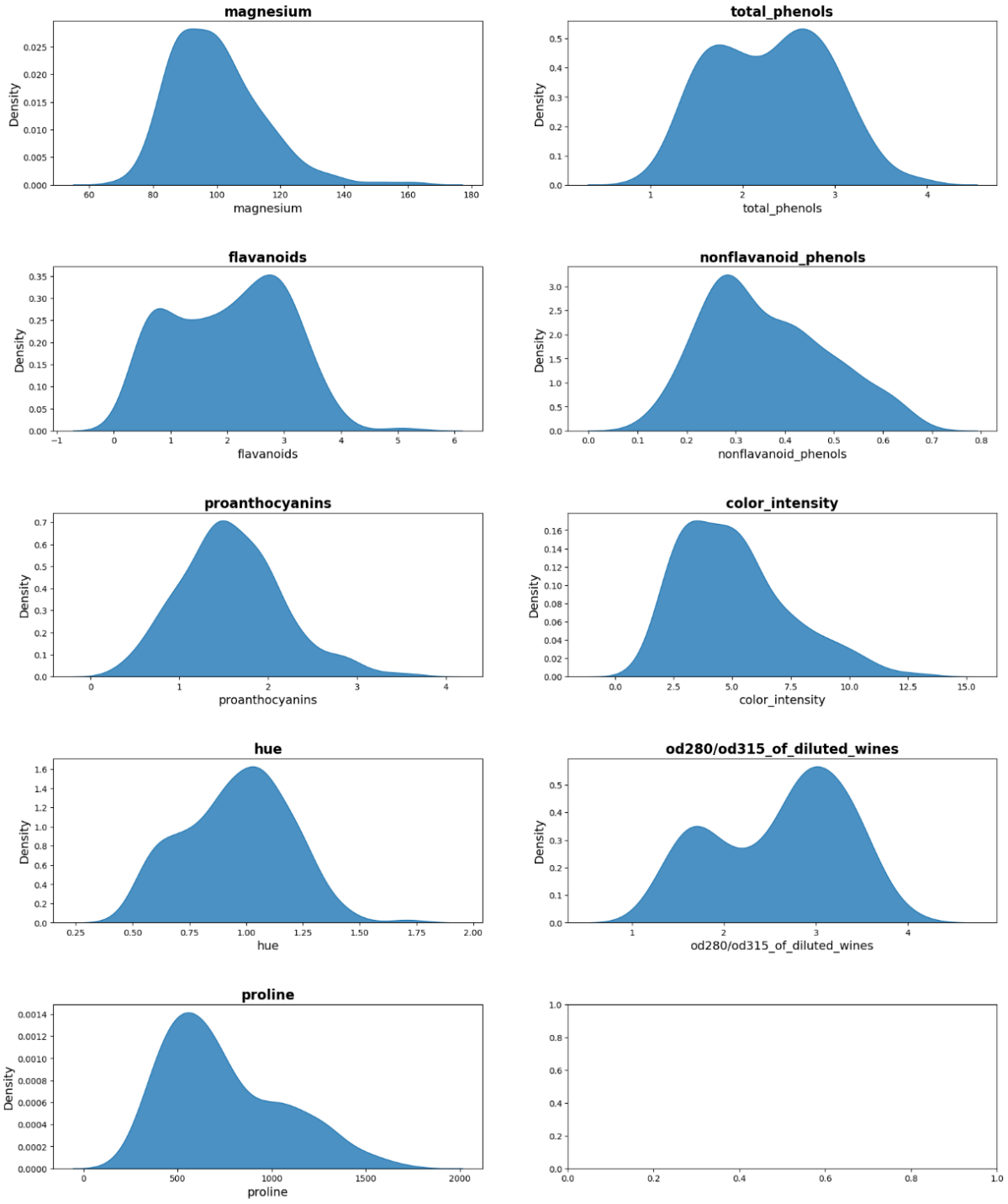
Task 1

To perform some preprocessing steps before clustering the data, I examined the **attributes table** using the describe() function:

	count	mean	std	min	25%	50%	75%	max
alcohol	178.0	13.000618	0.811827	11.03	12.3625	13.050	13.6775	14.83
malic_acid	178.0	2.336348	1.117146	0.74	1.6025	1.865	3.0825	5.80
ash	178.0	2.366517	0.274344	1.36	2.2100	2.360	2.5575	3.23
alcalinity_of_ash	178.0	19.494944	3.339564	10.60	17.2000	19.500	21.5000	30.00
magnesium	178.0	99.741573	14.282484	70.00	88.0000	98.000	107.0000	162.00
total_phenols	178.0	2.295112	0.625851	0.98	1.7425	2.355	2.8000	3.88
flavanoids	178.0	2.029270	0.998859	0.34	1.2050	2.135	2.8750	5.08
nonflavanoid_phenols	178.0	0.361854	0.124453	0.13	0.2700	0.340	0.4375	0.66
proanthocyanins	178.0	1.590899	0.572359	0.41	1.2500	1.555	1.9500	3.58
color_intensity	178.0	5.058090	2.318286	1.28	3.2200	4.690	6.2000	13.00
hue	178.0	0.957449	0.228572	0.48	0.7825	0.965	1.1200	1.71
od280/od315_of_diluted_wines	178.0	2.611685	0.709990	1.27	1.9375	2.780	3.1700	4.00
proline	178.0	746.893258	314.907474	278.00	500.5000	673.500	985.0000	1680.00

I plotted the graphs using the sns.kdeplot() function to visualize the **distribution of data** for each attribute. It can be seen that some attributes followed a roughly normal distribution, while others exhibited bimodal distributions with two distinct peaks. I suppose that the means of these attributes may have different values for different target class values:

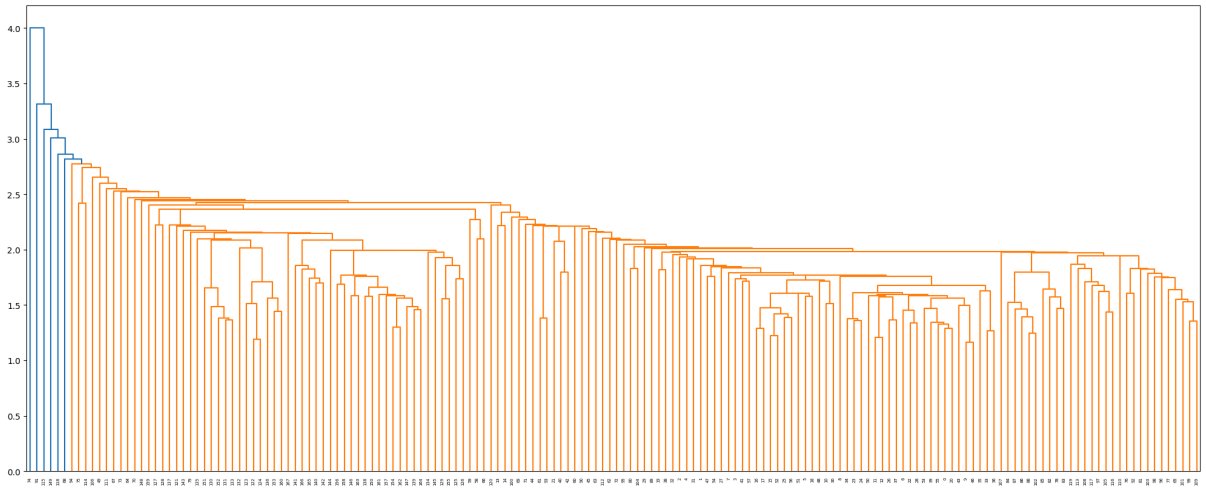




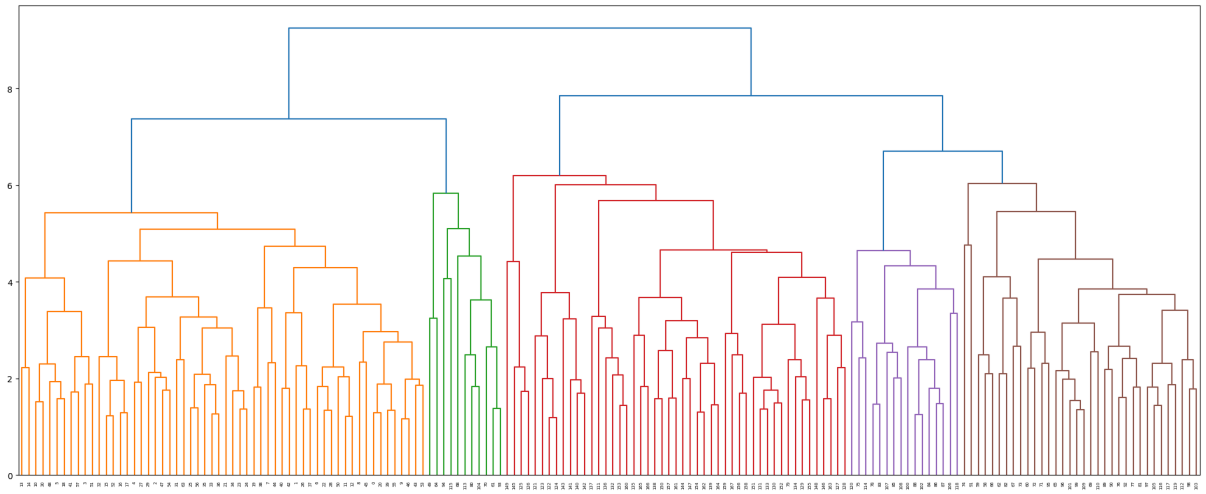
The Attributes table revealed that attributes had different ranges. Some attributes had relatively large values ("proline" with a mean of 747), while others had values less than one ("nonflavanoid_phenols" with a mean of 0.36). To address the varying scales of different attributes, **data normalization** was performed using the Standard Scaler fit_transform() function. This technique transformed the data to have zero mean and unit variance. To enhance accuracy, the zscore() function was applied to **identify outliers** with a z-value greater than 3 (3 standard deviations). These data points were considered potentially anomalous and were dropped from the dataset. After outlier removal, the dataset consisted of 168 data rows.

Hierarchical clustering was performed using the linkage() and dendrogram() methods with three different linkage methods: single, average, and complete. The dendrograms generated from complete and average linkage methods exhibited clearer cluster structures compared to the dendrogram from the single linkage method. Notably, there were similarities between the patterns observed on the right side of the complete linkage dendrogram and the left side of the average linkage dendrogram(and vice versa), middle parts are also similar, indicating potential cluster associations.

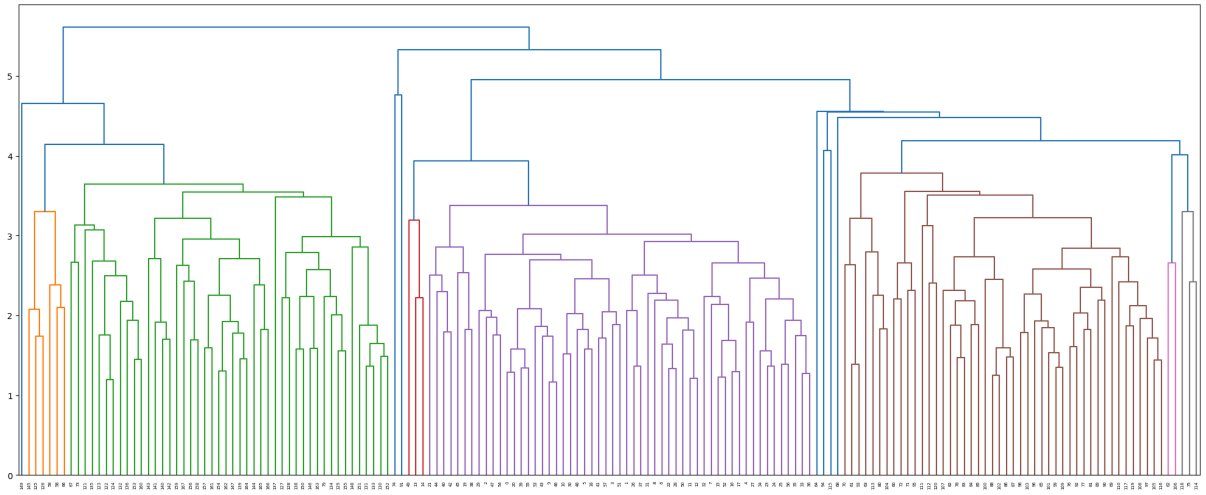
Single:



Complete:

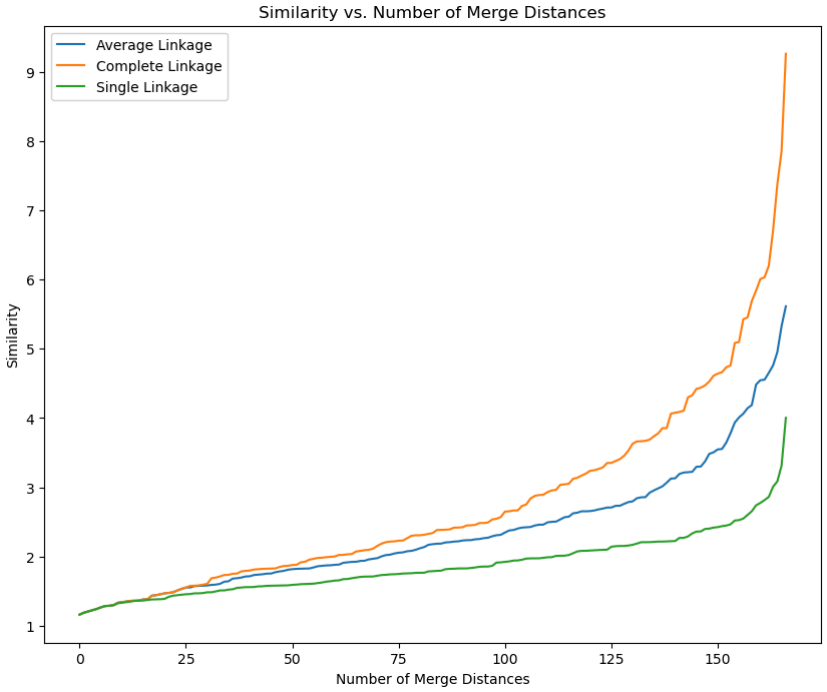


Average:

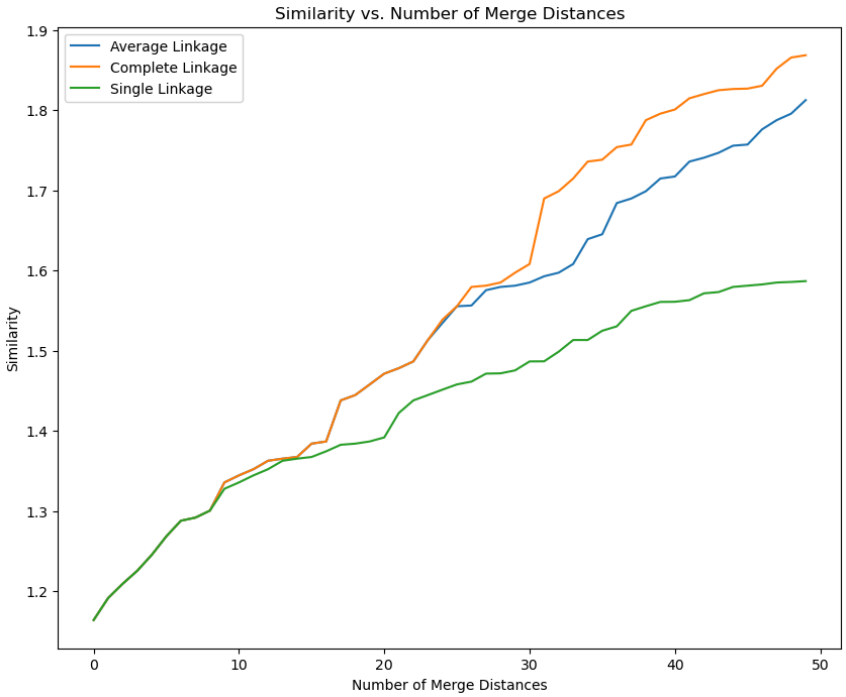


Task 2

I got the list of **merge distance values**, using Z_average[:, 2], and plotted a graph based on it. We see that for each linkage method, the distance values increase, following a similar growth rate at low merge distances. As expected, the complete linkage method exhibited the fastest growth rate, while the single linkage method showed the slowest growth rate:



To see more details, I zoomed in on the graph to focus on the first 50 merge distances. In this range, we see that all the curves had nearly identical values until around merge distance 15. Additionally, the curves for the complete and average linkage methods displayed the same values until approximately merge distance 25:



Task 3

I created a **K-Means** clustering solution using the KMeans() method and compared it with the single, average, and complete hierarchical clustering solutions. The number of clusters was set to 3 using the fcluster() method. To evaluate the extent to which the clusters captured the class structure of the dataset, I calculated the **accuracy** of each solution.

To overcome the label inconsistency between different clustering solutions and the original target values in the dataset, I implemented a loop that iterated through each cluster assignment permutation. This allowed me to identify the permutation that yielded the highest accuracy, indicating the best correspondence between cluster labels and original class labels.

Accuracy results:

Clustering Solution	Best Accuracy	Best Permutation
single linkage	36.31%	1, 0, 2
complete linkage	92.26%	0, 2, 1
average linkage	63.69%	2, 1, 0
K-Means	98.21%	1, 0, 2

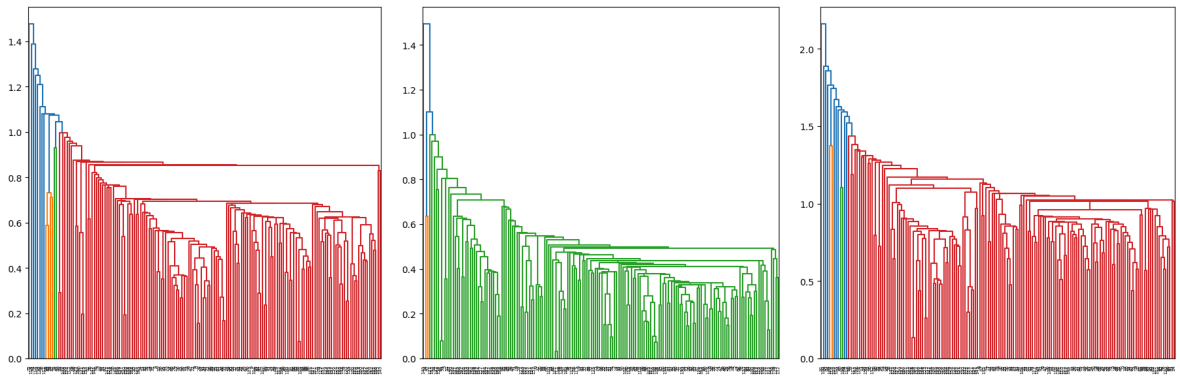
From the results, K-Means clustering solution achieved the highest accuracy of 98.21%. The next best solution was the complete linkage hierarchical clustering with an accuracy of 92.26%. The average linkage hierarchical clustering had a moderate accuracy of 63.69%, while the single linkage clustering performed the worst with an accuracy of 36.31%.

Task 4:

I created **three subsets** from the original attribute list and generated dendrograms using the single, average, and complete linkage methods. Here are the attributes included in each subset:
Subset 1: malic_acid, alcalinity_of_ash, hue, proline
Subset 2: magnesium, total_phenols, color_intensity
Subset 3: alcohol, ash, flavonoids, proanthocyanidins, nonflavanoid_phenols, od280/od315_of_diluted_wines

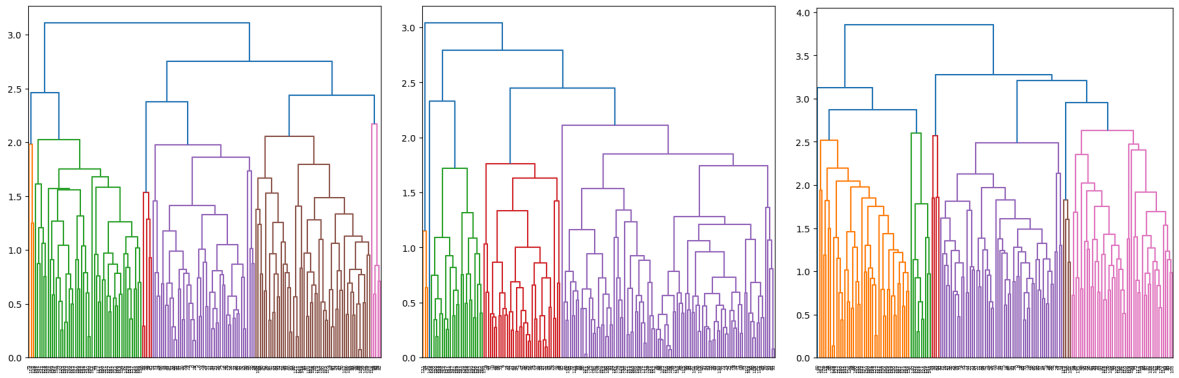
Single link dendrograms:

Subset 1 | Subset 2 | Subset 3



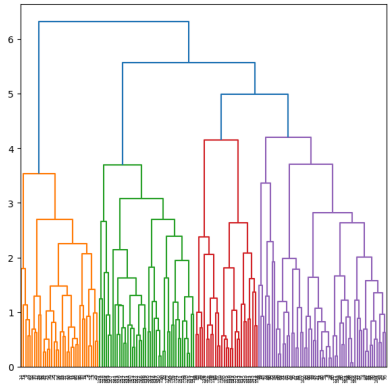
Average link dendrograms:

Subset 1 | Subset 2 | Subset 3

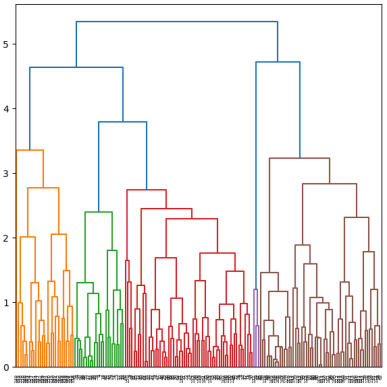


Complete link dendrograms:

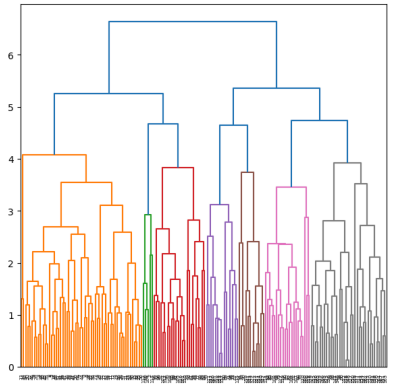
Subset 1



Subset 2



Subset 3



Solution	Single Link Accuracy	Average Link Accuracy	Complete Link Accuracy
Original	36.31%	63.69%	92.26%
Subset 1	36.31%	87.5%	74.4%
Subset 2	36.9%	51.79%	58.93%
Subset 3	36.31%	63.1%	59.52%

We see that the best accuracy is still achieved using the complete linkage method when considering the original whole set of attributes. However, when applying the clustering methods to the individual subsets, the accuracy scores decreased significantly for the complete linkage method. It is worth noting that subset 1 achieved a relatively higher accuracy of 74.4% with the complete linkage method, which is not a bad result.

On the other hand, the average linkage method showed an improvement in performance for subset 1, achieving an accuracy of 87.5%. Worth noting that both the average and complete linkage methods are known to be sensitive to outliers. To reduce the influence of outliers and improve the accuracy, we may consider a more strict outlier reduction strategy during the data preprocessing stage (e.g., $z > 2$ instead of 3).

The single linkage method consistently performed poorly across all subsets and the original dataset, showing accuracy scores of around 36%. The results obtained using the single linkage method were almost identical for all subsets and the original dataset.

Overall, subset 1 generally demonstrated higher accuracy scores across the different linkage methods, indicating that the attributes included in this subset may contribute more significantly to the classification of wines.