CS3481 Fundamentals of Data Science

Assignment 2

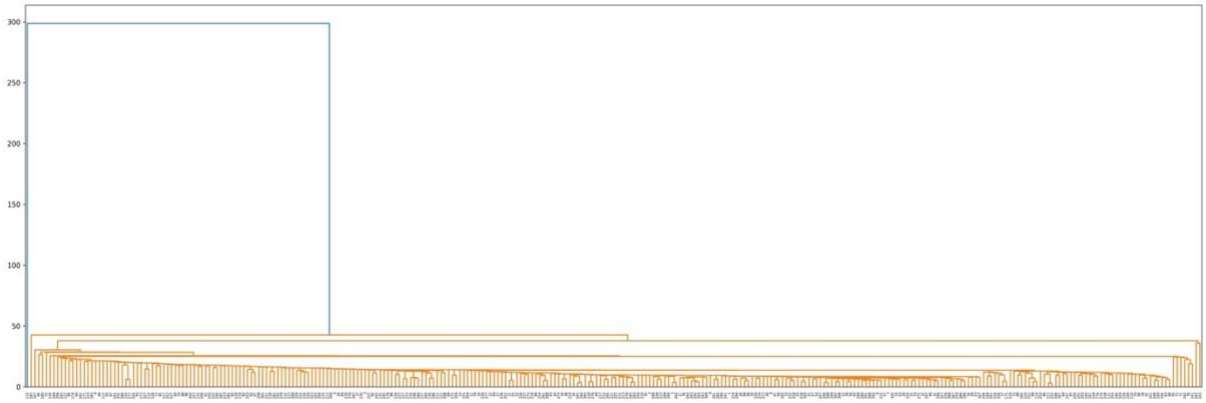
Clustering Assignment

Student Name: LUO Peiyuan
SID: 56642728

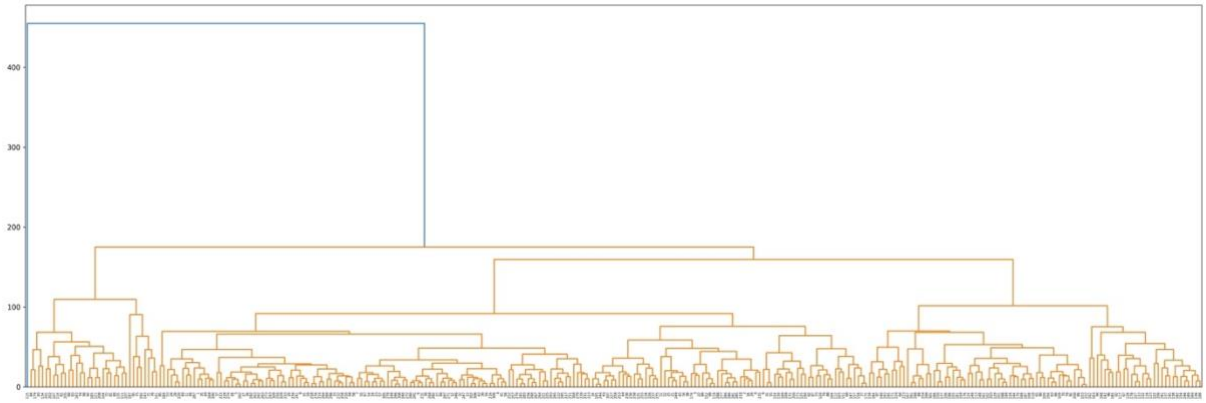
Table of Contents

Hierarchical Clustering for Original Dataset (Raw Data)

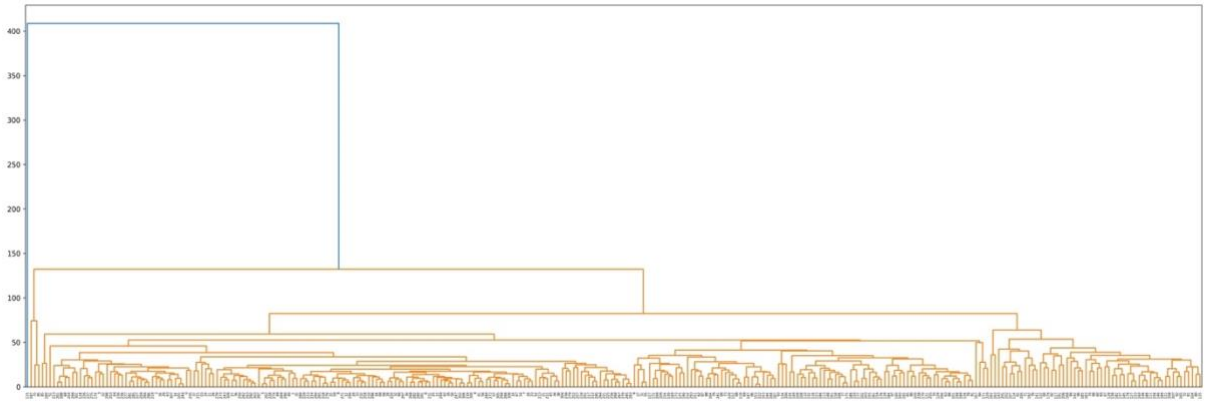
Single Linkage



Complete Linkage



Group Average Linkage



Conclusion

The generated hierarchical clustering result is too bad. Hence, we need further preprocess the data to get a more appropriate result.

Data Preprocessing

Four Major Steps

Standardize the data

Use StandardScaler() to make the data fit the Normal Distribution.

```
scaler = StandardScaler()
X_scaled = scaler.fit_transform(X)
```

Remove the Outliers

Use stats.zscore library to remove the data points with z-score greater than 3 This is to handle the outliers, also known as the 3-sigma rule or 3 times standard deviation rule. The basic principle is that for a normally distributed random variable, the vast majority of data (about 99.7%) is within ±3 standard deviations from the mean. Therefore, if a data point has a z-score greater than 3, it can be considered an outlier and needs to be removed.

```
z_scores = np.abs(stats.zscore(X_scaled))
X_new = X_scaled[(abs(z_scores) < 3).all(axis=1)]
```

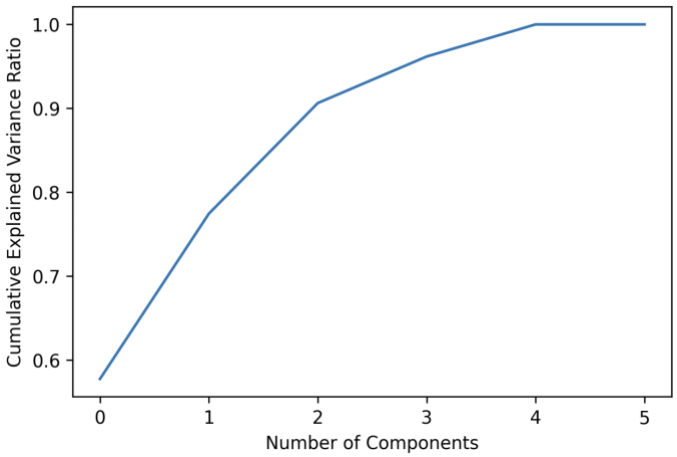
Normalize the Data

Normalize the data in order to map the value range of a dataset to [0, 1].

```
normalized_data = (X_new - X_new.min()) / (X_new.max() - X_new.min())
```

Principal Component Analysis

To reduce the dimension, but still retain most of the information in the original data. Here, as the below figure shows, we choose n_components to be 2, as it retain the Cumulative Explained Variance Ratio to around 0.90.



Question 1 Analysis

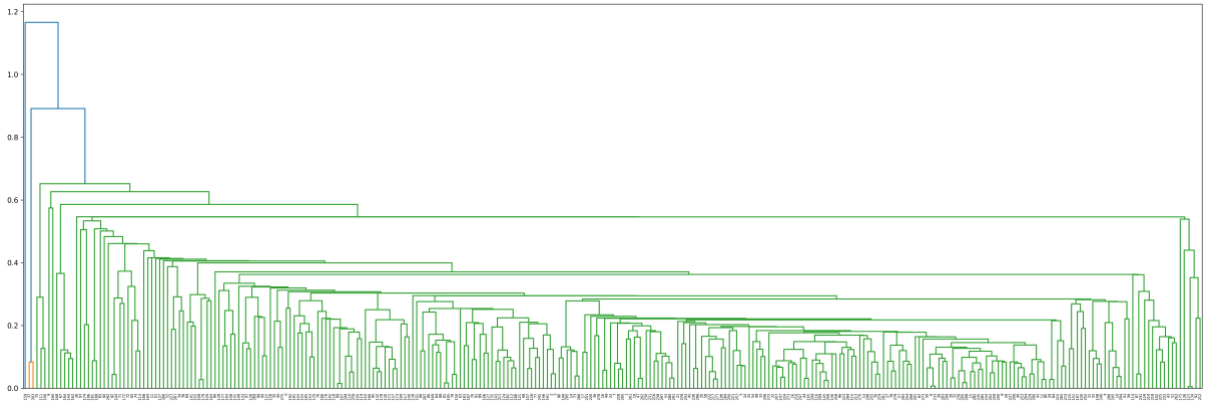
Single Linkage

Structures Overview
The dendrogram of the single linkage presents a chain-like structure, where clusters are connected by chains. This is because Single linkage method determines which clusters should be merged by calculating the minimum distance between different clusters. As a result, the newly formed clusters are usually connected to the original clusters.

Size of Clusters
The cluster size is uneven. And it's quite unbalanced, and present as a chain-like structure.

Type of Merge Steps
The merge steps of the single linkage are shown as below, the merging of the clusters is determined by the minimum distance between any two points, one from each cluster. And as we see from the low-level merging steps, it's easily seen that the merging step is more like **point-to-point**.

Dendrogram
The generated dendrogram of the hierarchical clustering dendrogram with single linkage is shown as below. As, you may see, the clustering result is not so good.



Conclusion
As the figure above shown, the clustering performance of the single linkage is a bit poor. And we can easily see that it has several weaknesses as below:

- Sensitivity to outliers: it is usually sensitive to the outliers and noise data. For example, we can easily see that the leftmost points cause the entire cluster to be merged with another cluster, resulting in a poor clustering result.
- Chain effect: it tends to produce long, chain-like clusters. This is because it only considers the minimum distance between any two points in the clusters but may ignore the overall structure of the data.
- Bad for handling non-convex clusters: it would meet difficulties to handle the clusters, where the points within a cluster are not all connected by a straight line.

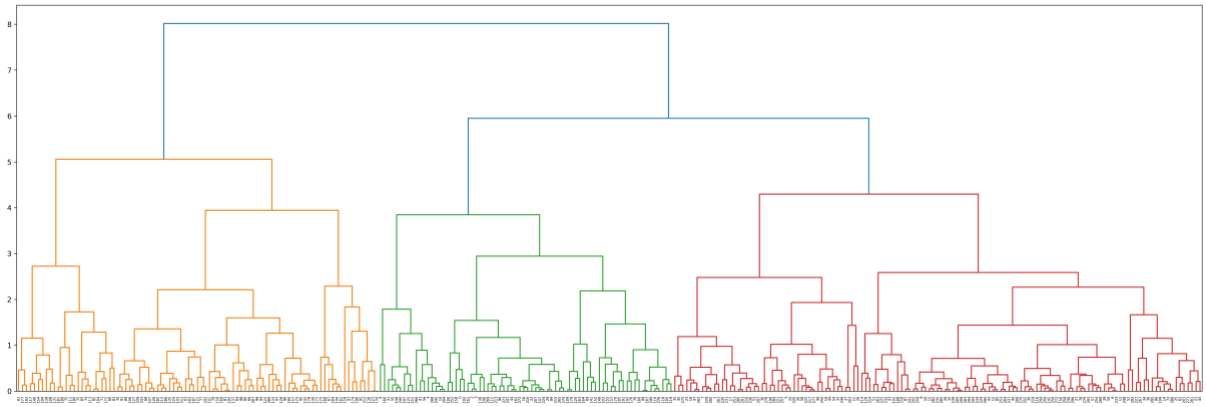
Complete Linkage

Structure Overview
As shown below, you may see that the cluster size of the complete linkage is quite balanced. And it tends to produce compact, spherical clusters that are well-separated from each other.

Size of Clusters
The cluster size is even and balanced to some extent.

Type of Merge Steps
As below figure shows, the merging step of the complete linkage is more like **cluster-to-cluster**. The merge step is performed on clusters, not individual data points. The clustering starts with each data point as a separate cluster and iteratively merges the closest clusters together until all data points are in a single cluster. This means that the merge step is more like cluster-to-cluster than point-to-point or cluster-to-point.

Dendrogram
The generated dendrogram of the hierarchical clustering dendrogram with complete linkage is shown as below. As you may see, the clustering result is quite good. The clusters are well-separated.



Group Average Linkage

Structure Overview
Group average linkage hierarchical clustering tends to produce clusters that are more evenly sized and well-separated than single linkage clustering, but less compact and spherical than complete linkage clustering, which is because group average linkage clustering calculates the distance between clusters based on the average distance between all pairs of points from different clusters. This approach tends to produce clusters with a moderate diameter and an even distribution of points, as it considers the distances between all pairs of points from different clusters. In addition, group average linkage clustering is less sensitive to noise and outliers than single linkage clustering, but more sensitive than complete linkage clustering. This can result in clusters that are less elongated than those produced by single linkage clustering, but more elongated than those produced by complete linkage clustering.
In a group average linkage dendrogram, the branches tend to be longer than those in a complete linkage dendrogram, indicating that the clusters are less compact. However, the

clusters tend to be more evenly sized and well-separated than those produced by single linkage clustering, with a moderate diameter and an even distribution of points.

Size of Clusters

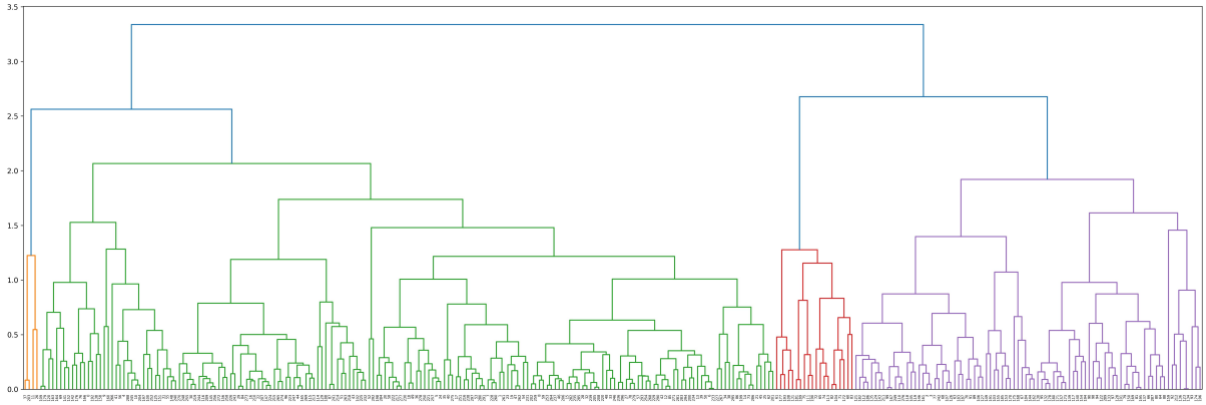
As shown below, you may see that the cluster size of the group average linkage is quite balanced, but less compact than that of linkage. It is less elongated than those produced by single linkage clustering, but more elongated than those produced by complete linkage clustering.

Type of Merge Steps

As below figure shows, the merge step is performed on clusters, not individual data points. The algorithm starts with each data point as a separate cluster and iteratively merges the closest clusters together until all data points are in a single cluster, which means that the merge step is more like **cluster-to-cluster** than point-to-point or cluster-to-point.

Dendrogram

The generated dendrogram of the hierarchical clustering dendrogram with group average linkage is shown as below. As you may see, the clusters are also well-sepatated.



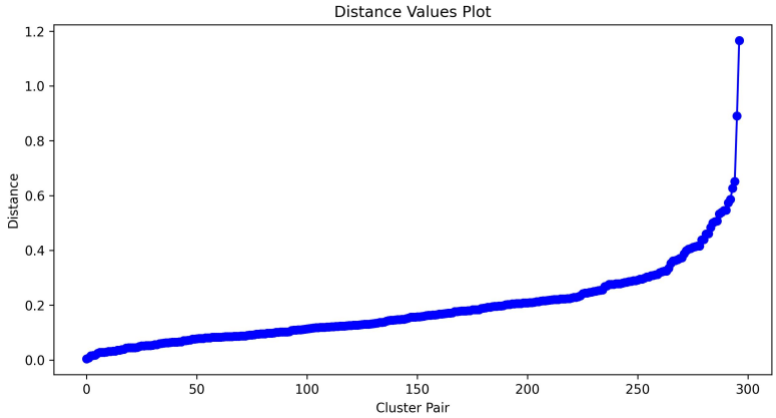
Overall Comparisons

Features / Linkage	Sizes of Cluster	Type of Merge Steps	Sensitivity to Outliers and Noise Data	Aggregation	Structure Characteristic
Single Linkage	Not balanced	Point-to-Point	Very sensitive	Tend to form larger clusters	Chain-like (Most elongated)
Complete Linkage	Relatively more balanced	Cluster-to-Cluster	Relatively more Robustness	Tend to form compact, spherical clusters	Have Clear Boundary and compact
Group Average Linkage	Relatively more balanced	Cluster-to-Cluster	Relatively more Robustness	Tend to form compact, spherical clusters	Tightly packed but with clear boundary

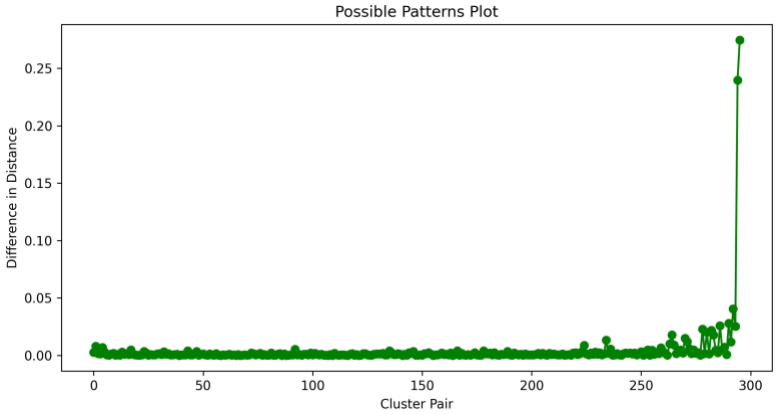
Question 2 Analysis

Single Linkage

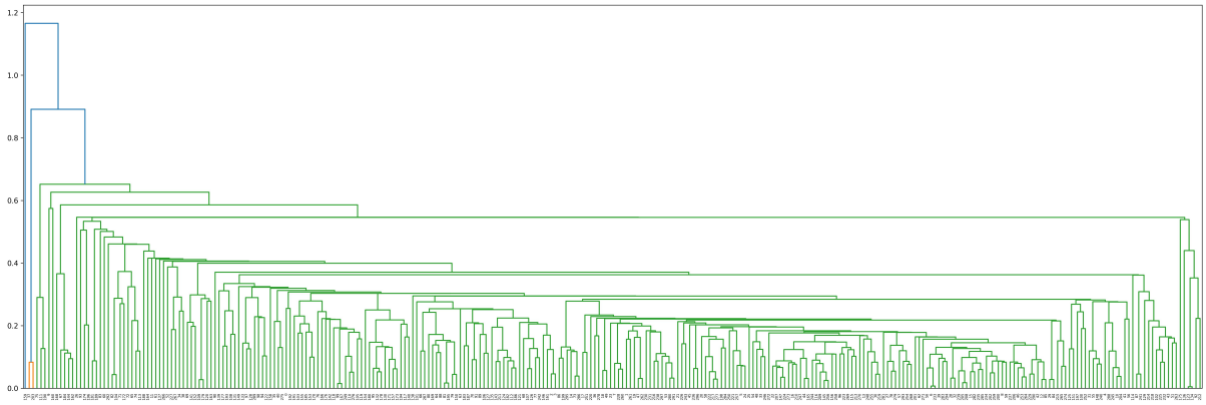
Distance Values Plot



Possible Values Plot



Dendrogram

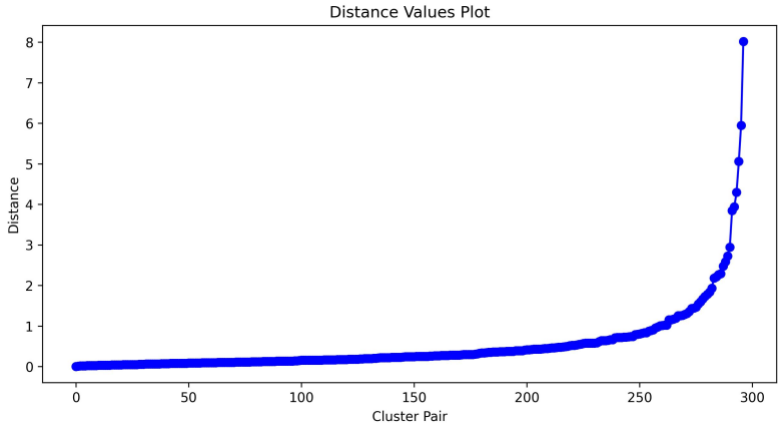


Conclusion and Observation

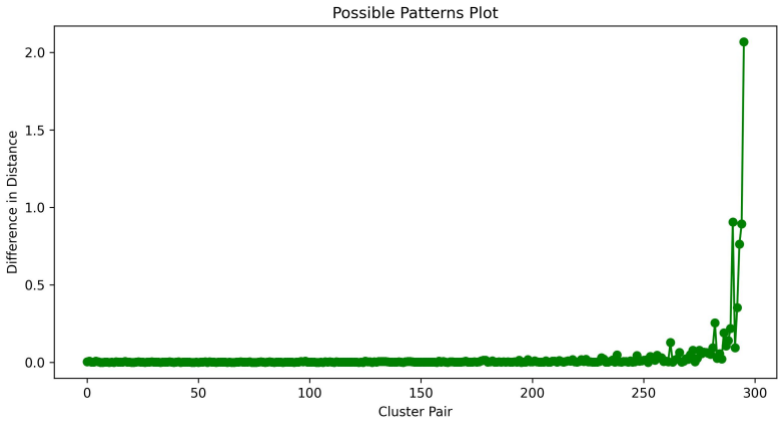
Observed from the above three figures, we can know that, at the start of the merging, the distance increased smoothly for a very long time, and the difference distance keeps at a level of around 0.00-0.05. However, it increases sharply when the distance value reach to 0.5 and simultaneously the difference in distance increased sharply. It also can be observed from the dendrogram that when the difference in distance value (showing as the step of the clusters) or when the distance value at a relatively large value, this kind of merging step is not appropriate, as it means that it's combining two relatively distant cluster together, which would lower the performance of the clustering.

Complete Linkage

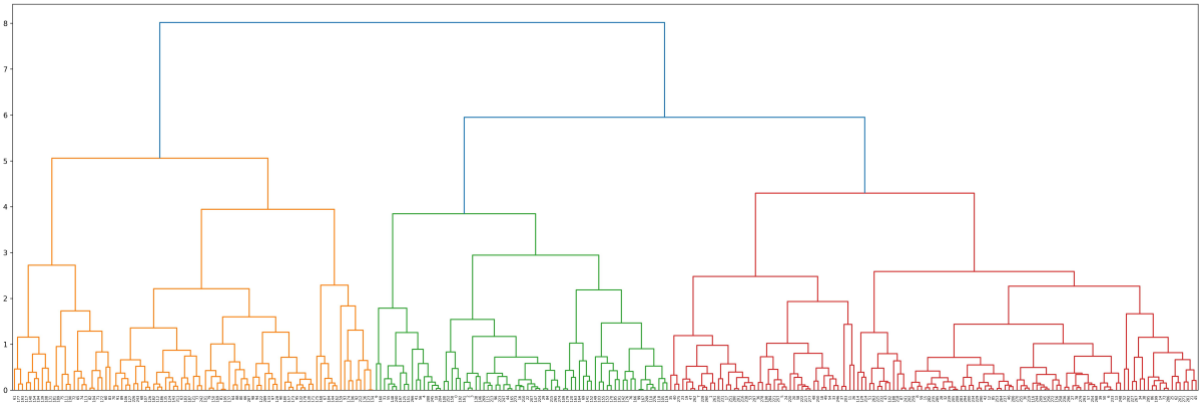
Distance Values Plot



Possible Values Plot



Dendrogram

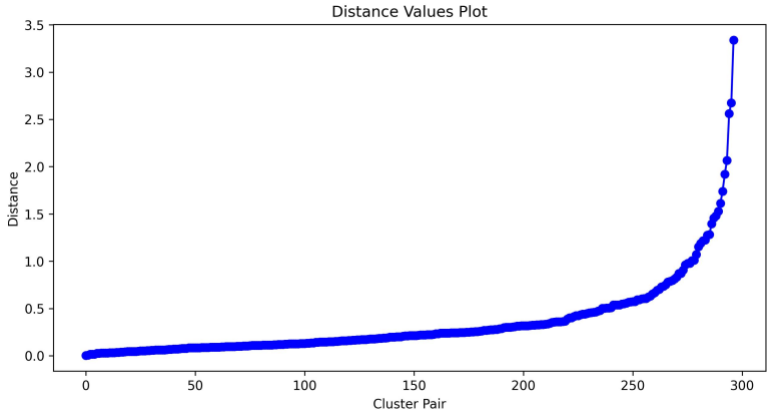


Conclusion and Observation

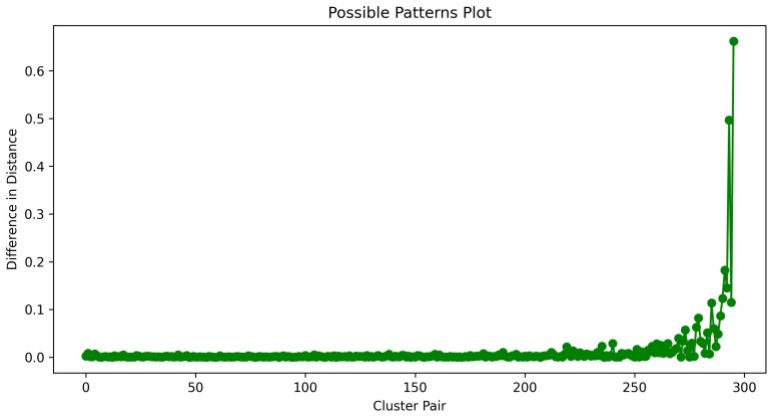
Observed from the above three figures, we can know that, at the start of the merging, the distance increased smoothly for a very long time, and the difference distance keeps at a level of around 0.00-0.1. However, it increases sharply when the distance value reach to 3 and simultaneously the difference in distance increased sharply. It also can be observed from the dendrogram that when the difference in distance value (showing as the step of the clusters) or when the distance value at a relatively large value, and it's easily seen that it's combining two relatively distant cluster together, which is more appropriate than the single linkage.

Group Average Linkage

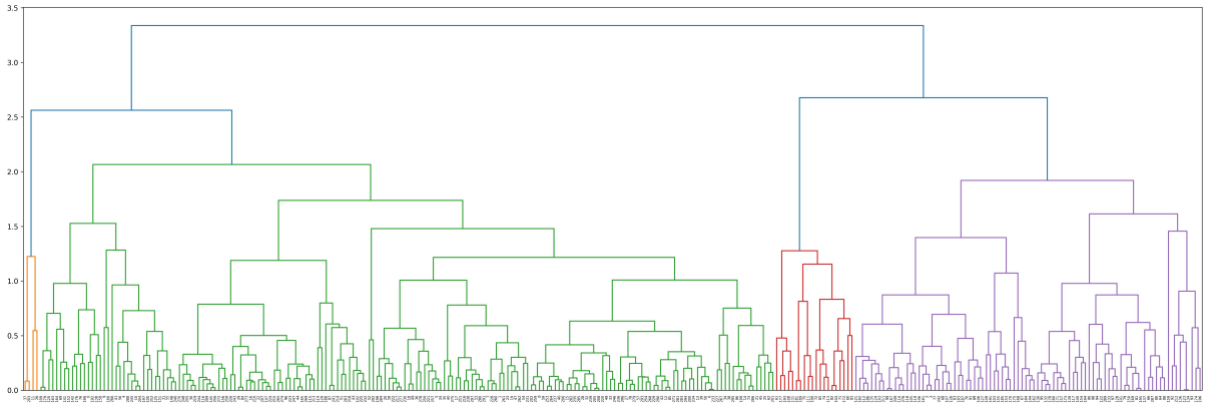
Distance Values Plot



Possible Values Plot



Dendrogram



Conclusion and Observation

Observed from the above three figures, we can know that, at the start of the merging, the distance increased smoothly for a very long time, and the difference distance keeps at a level of around 0.00-0.05. However, it increases sharply when the distance value reach to 1.5 and simultaneously the difference in distance increased sharply. It also can be observed from the dendrogram that when the difference in distance value (showing as the step of the clusters) or when the distance value at a relatively large value, and it's easily seen that it's combining two relatively distant cluster together.

Overall Conclusion and Patterns Observations

Sharply increased distance value

When the distance value increases sharply at a certain stage or when the difference in distance increase suddenly, it may be caused by the clustering merges two distinct clusters together.If there are clear threshold distance values at which many clusters merge, it may suggest that the data is naturally organized into distinct groups. This can be useful for determining the appropriate number of clusters to use when applying other clustering algorithms.

Gradual merging

If clusters merge at a relatively consistent rate over a range of distance values, it may suggest that the data has a continuous structure, rather than being organized into distinct groups. This can be useful for understanding the general structure of the data and may suggest that other clustering algorithms, such as density-based clustering, may be more appropriate.

Multiple merging events

If there are multiple sets of distance values at which clusters merge, it may suggest that the data has multiple levels of organization or that there are multiple underlying structures in the data. This can be useful for identifying subgroups within larger groups or for identifying different types of patterns in the data.

Outliers

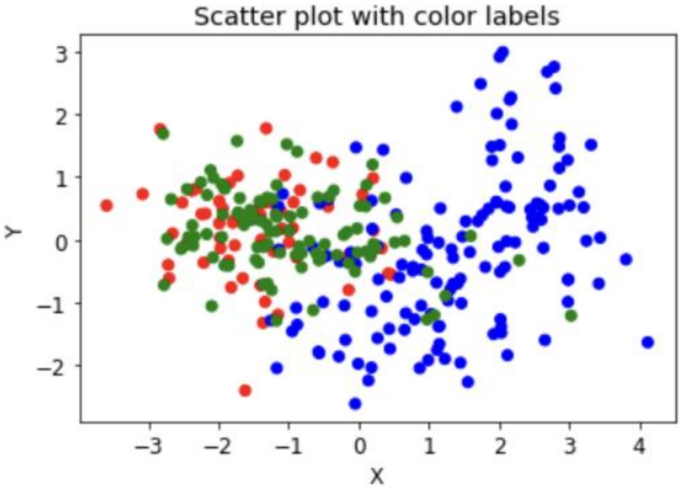
If there are clusters that merge at much larger or smaller distances than other clusters, it may suggest that there are outliers in the data that are not well-represented by the other clusters. This can be useful for identifying potential anomalies or for determining whether certain data points should be excluded from the analysis.

Question 3 Analysis

Background

I generate the scatter plot, Silhouette score, Calinski-Harabasz score, and Davies-Bouldin score for the clustering result with n_components equaling to 3, and I chose several result displayed as below.

Original Class Structures



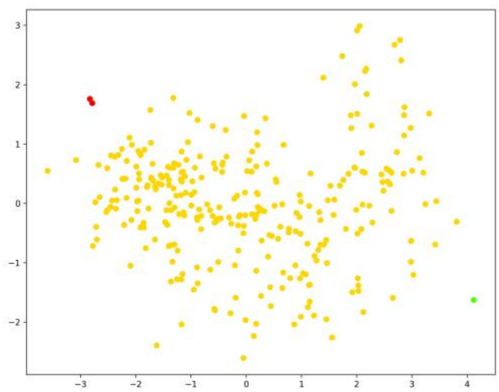
Labels

[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2,
2, 2, 2, 2, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1]

number of clusters=3(Best Scores)

Single Linkage

Scatter Plot



Evaluation Scores

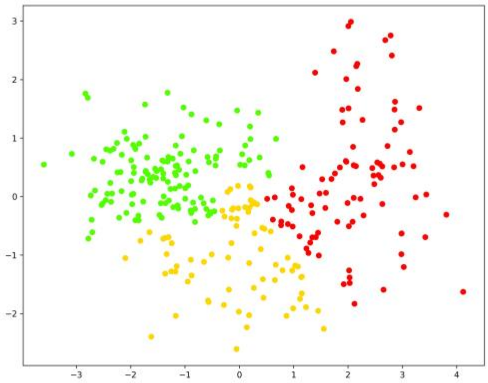
Purity=0.4664429530201342
Silhouette score=0.0891
Calinski-Harabasz score=5.5548
Davies-Bouldin score=0.4891

Predicted Labels

[2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 3 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 2 2]

Complete Linkage

Scatter Plot



Evaluation Scores

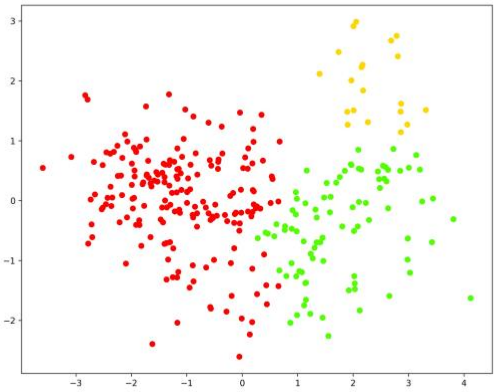
Purity = 0.6677852348993288
Silhouette score=0.3625
Calinski-Harabasz score=282.3412
Davies-Bouldin score=0.9525

Predicted Labels

[2 3 2 2 2 3 3 3 3 2 2 3 3 3 3 3 2 3 3 3 3 3 2 3 3 3 3 3 3 2 3 2 3 3 3 3 3
 3 3 3 3 2 3 3 2 3 3 3 3 3 3 3 3 3 3 2 2 3 3 3 1 1 3 1 1 1 1 1 1 2 1 1 1 3
 1 1 2 1 3 1 1 1 1 1 2 3 3 3 1 1 1 1 1 1 1 1 2 1 1 2 2 1 3 2 1 2 1 1 1 2 2
 1 1 1 3 2 3 1 2 2 2 1 1 1 1 2 1 1 1 3 1 1 1 1 1 1 1 1 2 1 1 2 2 1 1 3 2 2
 2 2 1 2 2 2 1 2 3 1 2 1 2 2 1 2 2 2 2 2 1 1 2 1 1 1 1 1 2 1 1 2 1 1 1 1 1
 2 1 1 3 1 1 2 2 1 1 1 1 2 3 3 3 3 3 3 2 3 3 2 3 3 2 1 1 2 3 2 3 3 3 3 3 3
 3 2 2 3 1 3 3 3 3 3 3 3 3 3 3 3 3 3 2 3 1 1 3 2 3 2 3 2 3 3 2 3 3 3 3 3 3
 3 3 3 3 3 3 2 2 3 3 2 3 3 3 2 2 3 2 3 3 3 3 3 3 3 3 3 3 1 2 1 3 3 3 3 3 3
 3 3]

Group Average Linkage

Scatter Plot



Evaluation Scores

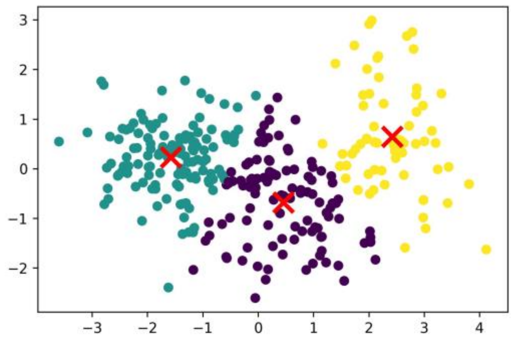
Purity = 0.6442953020134228
Silhouette score=0.4322
Calinski-Harabasz score=244.8449
Davies-Bouldin score=0.7424

Predicted Labels

[1 1 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 1 3 3 2 3 3 3 1 3 2 2 1
 2 2 1 3 1 3 3 3 2 2 1 1 1 1 3 3 3 3 3 3 3 3 1 3 3 3 1 3 1 3 2 3 3 3 2 1 3
 2 3 2 1 1 1 3 3 3 1 3 3 3 3 1 3 3 3 1 3 2 3 3 3 3 3 3 1 3 3 1 1 3 3 1 3 1
 1 3 3 1 1 3 3 3 1 3 1 3 1 3 3 1 1 1 1 3 3 3 3 3 2 3 3 3 1 2 3 1 3 2 3 3 3
 3 3 3 1 2 2 3 1 2 3 3 3 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 3 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1
 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 3 1 1 1 1 1 1 1 1 1 1 3 1 3 1 1 1 1 1 1
 1 1]

K-Means Clustering

Scatter Plot



Evaluation Scores

Purity=0.6711409395973155

Silhouette score=0.4246
Calinski-Harabasz score=344.3625
Davies-Bouldin score=0.9525

Predicted Labels

[0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1
 1 1 1 1 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 2 2 1 2 2 2 2 0 0 0 0 2 2 1
 2 2 0 0 1 2 2 2 2 2 0 1 1 0 2 0 2 0 2 2 2 2 0 2 2 0 0 2 0 0 2 0 0 0 2 0 0
 2 0 2 1 0 0 2 0 0 0 0 2 2 0 0 0 0 2 0 2 2 2 2 2 0 2 2 0 2 2 0 0 0 0 0 0 0
 1 0 2 0 0 0 2 0 1 2 0 2 0 0 0 0 0 0 0 0 0 2 0 2 2 2 2 2 0 2 2 0 2 2 0 0 0
 0 0 0 0 2 2 0 0 2 2 2 2 0 1 1 1 1 1 1 1 1 0 1 0 0 0 2 0 0 0 1 1 0 1 1 1
 1 0 0 0 0 1 1 1 1 1 0 1 1 1 1 1 1 1 0 0 0 1 1 1 0 1 0 1 1 0 1 1 1 0 1 1
 1 1 1 1 0 1 0 1 1 1 0 1 1 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 2 1 2 1 1 1 1 1 1
 1 1]

Observation and Conclusions

- In terms of the figure: Single is obviously worse than the other.
- Purity Score: K-Means>Complete>Group Average > Single
- Silhouette score: K-Means > Group Average > Complete > Single
- Calinski-Harabasz score: K-Means > Group Average > Complete <>Single
- Davies-Bouldin score: Group Average > Complete > K-Means > Single

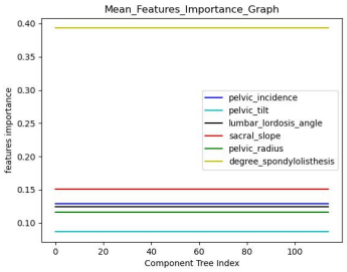
Overall Comparisons

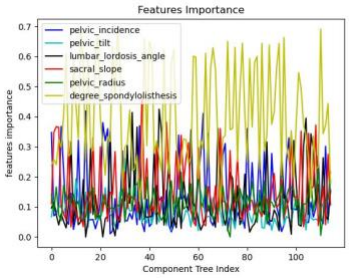
The Single Linkage is much worse than the other, and the others (complete, group average, k-means) are quite close to each other.

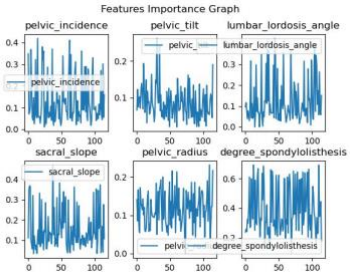
Question 4 Analysis

Background

According to the previous assignment, we get the following information shown as below. I try to use degree spondylolisthesis, sacral slope, and pelvic incidence to generate the dendrogram.







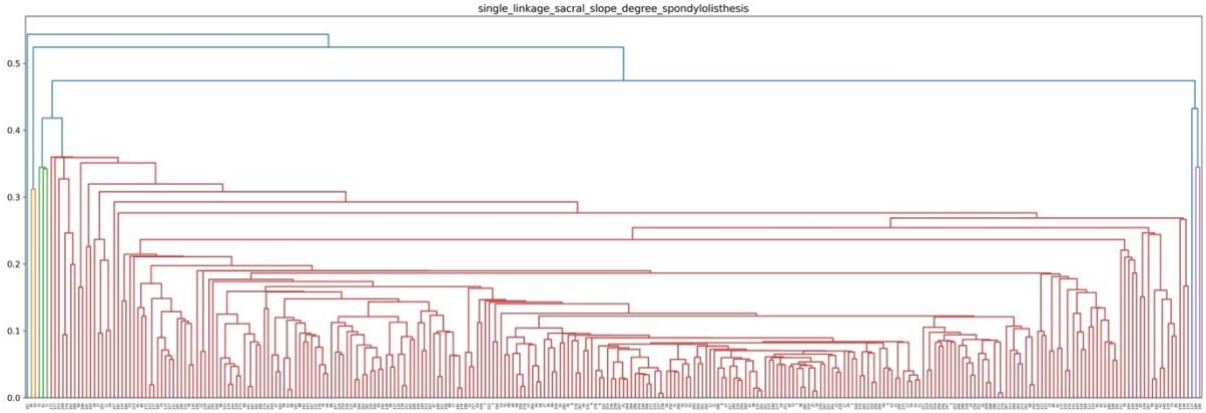
Degree spondylolisthesis and sacral slope

Single Linkage

New Structure

Observation

Compared to the original dendrogram, this one is more even, and looks more appropriate, and the merging steps is better. And it tends to less chain-like, and has more sub-clusters, as we can see from the different-colored clusters.



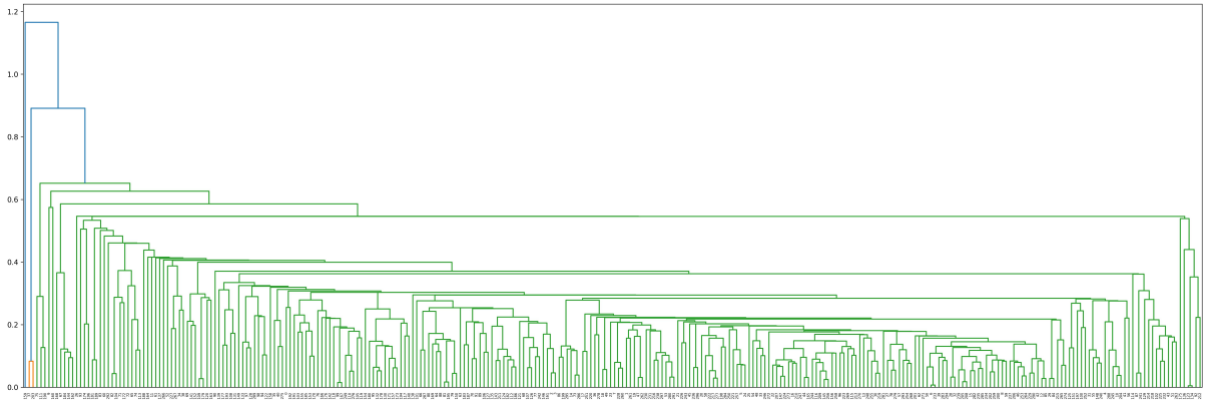
Evaluation Scores

Silhouette score=0.3821
Calinski-Harabasz score=8.8486
Davies-Bouldin score=0.4059

Original Structure

Evaluation Scores

Silhouette score=0.0891
Calinski-Harabasz score=5.5548
Davies-Bouldin score=0.4891



Complete Linkage

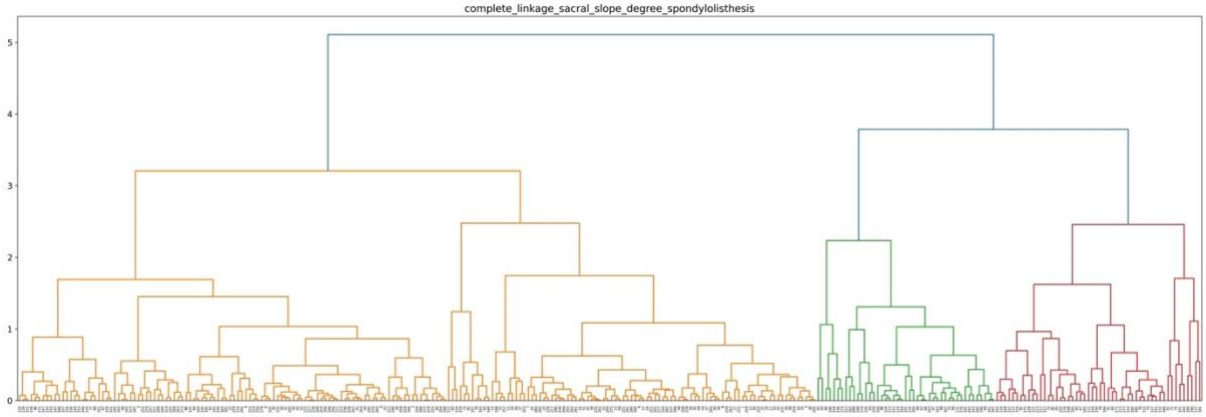
New Structure

Observation

Compared to the original dendrogram, they are quite similar to each other. However, one of the large cluster has a much larger size, which makes the new generated structure less balanced.

Evaluation Scores

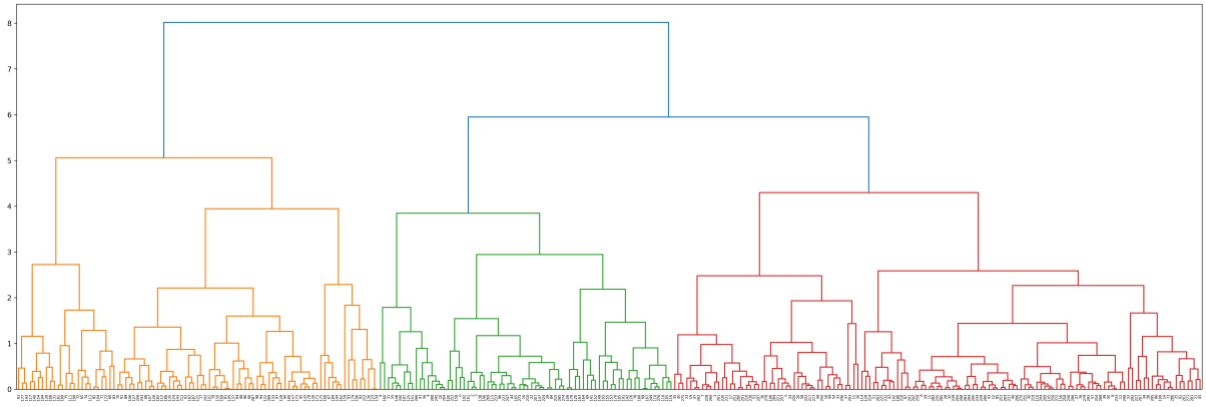
Silhouette score=0.4219
Calinski-Harabasz score=227.5041
Davies-Bouldin score=0.8762



Original Structure

Evaluation Scores

Silhouette score=0.3625
Calinski-Harabasz score=282.3412
Davies-Bouldin score=0.9525



Group Average Linkage

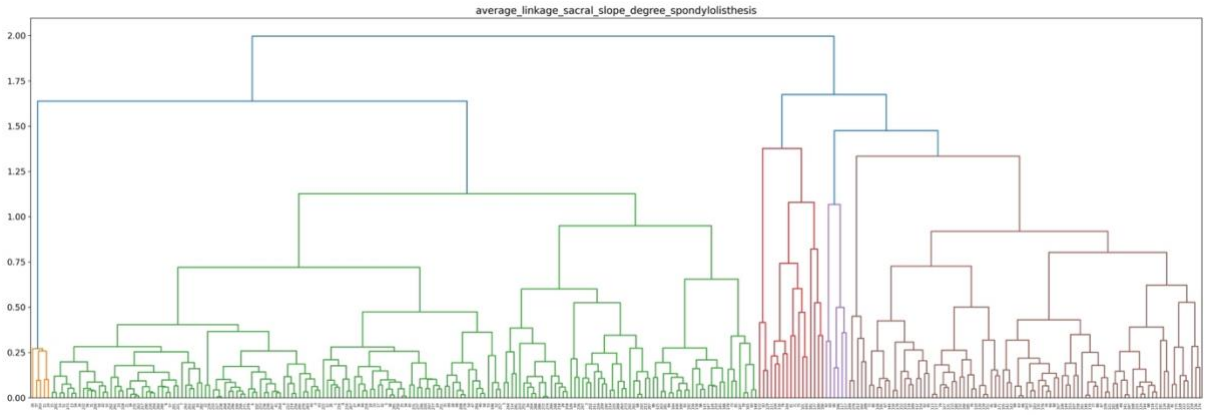
New Structure

Observation

Compared to the original dendrogram, they are quite similar to each other.

Evaluation Scores

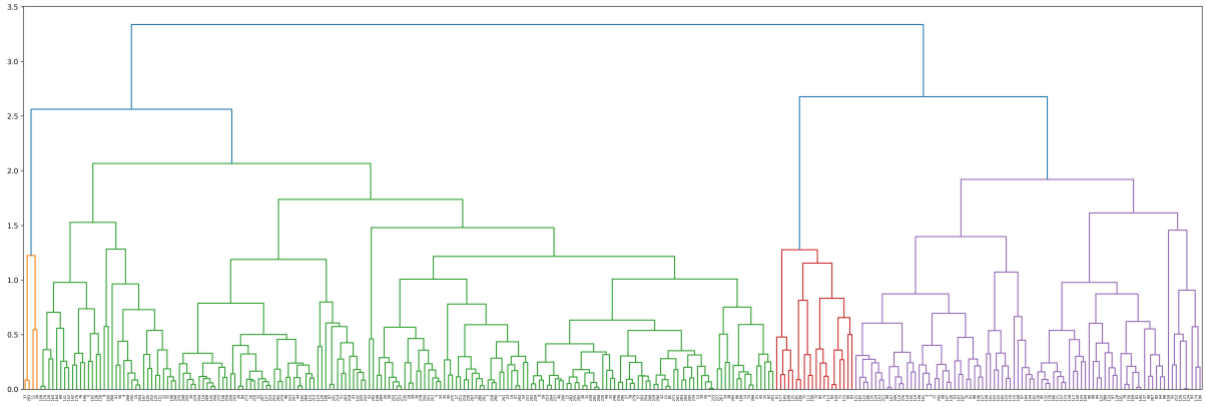
Silhouette score=0.4663
Calinski-Harabasz score=251.9780
Davies-Bouldin score=0.8190



Original Structure

Evaluation Scores

Silhouette score=0.4322
Calinski-Harabasz score=244.8449
Davies-Bouldin score=0.7424



Degree spondylolisthesis and pelvic incidence

Single Linkage

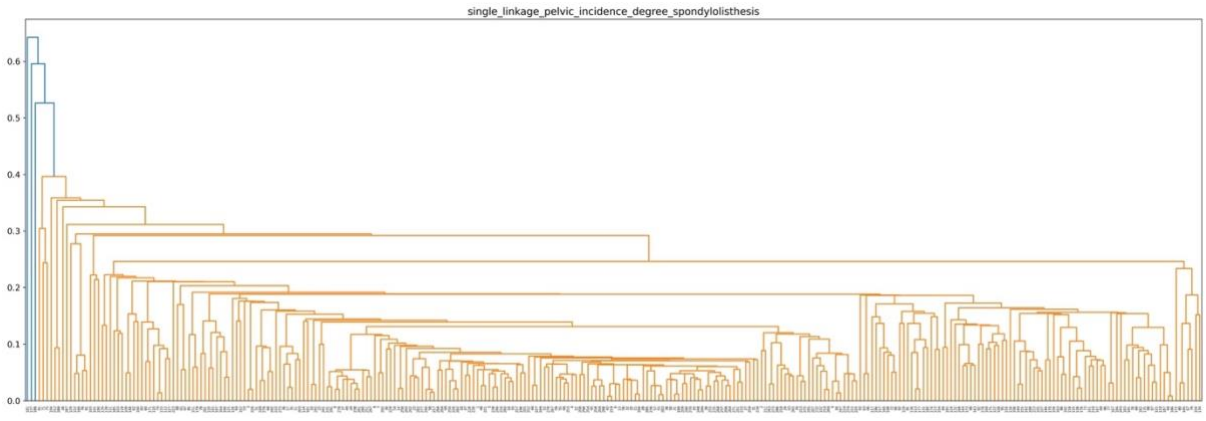
New Structure

Observation

Compared to the original dendrogram, this one didn't improve the structure.

Evaluation Scores

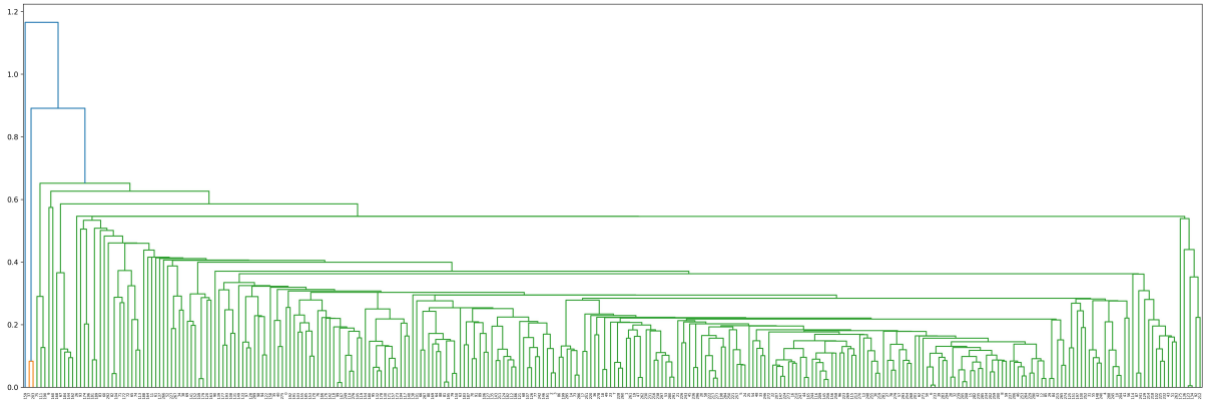
Silhouette score=0.3586
Calinski-Harabasz score=4.6576
Davies-Bouldin score=0.4415



Original Structures

Evaluation Scores

Silhouette score=0.0891
Calinski-Harabasz score=5.5548
Davies-Bouldin score=0.4891



Complete Linkage

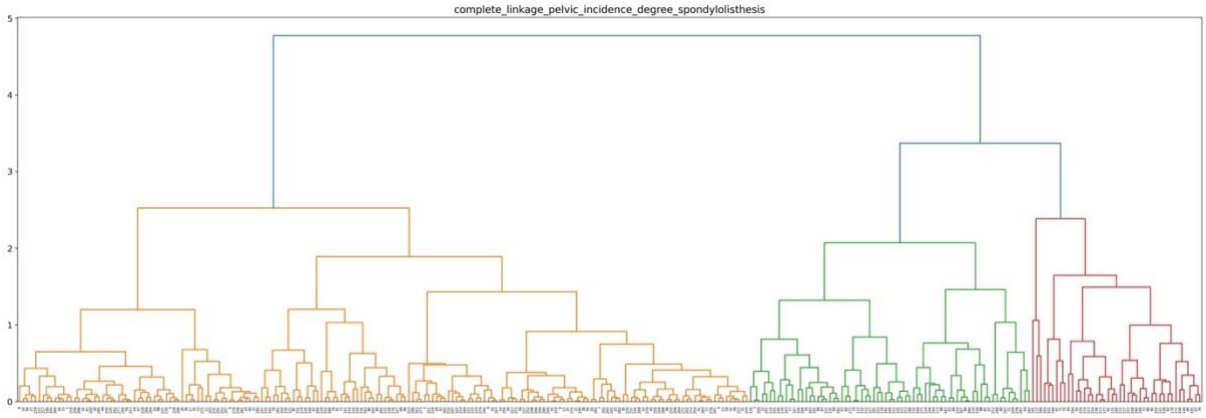
New Structure

Observation

Compared to the original dendrogram, this new structure has a less balanced structure. However, in general, it still has three major clusters.

Evaluation Scores

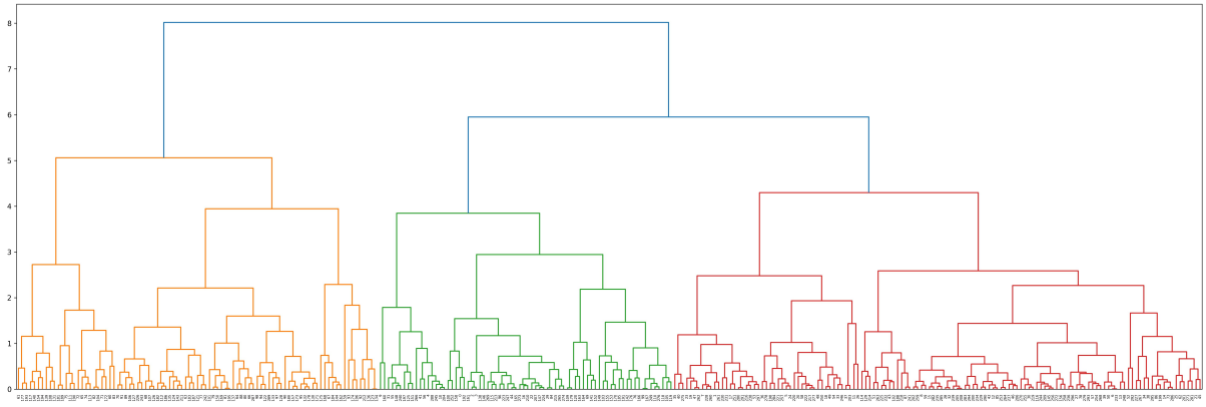
Silhouette score=0.4340
Calinski-Harabasz score=369.0485
Davies-Bouldin score=0.8148



Original Structures

Evaluation Scores

Silhouette score=0.3625
Calinski-Harabasz score=282.3412
Davies-Bouldin score=0.9525



Group Average Linkage

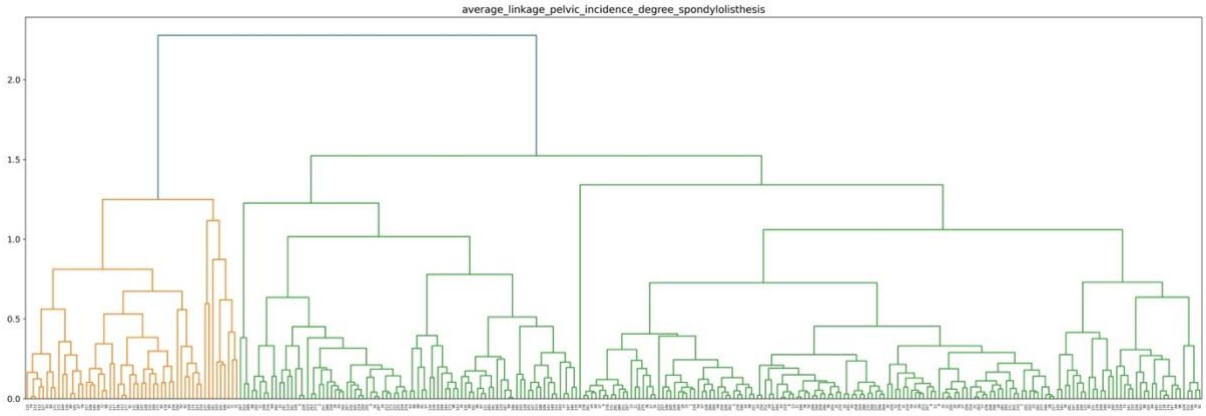
New Structure

Observation

Compared to the original dendrogram, the performance of this structure is worse, as its structures are less balanced.

Evaluation Scores

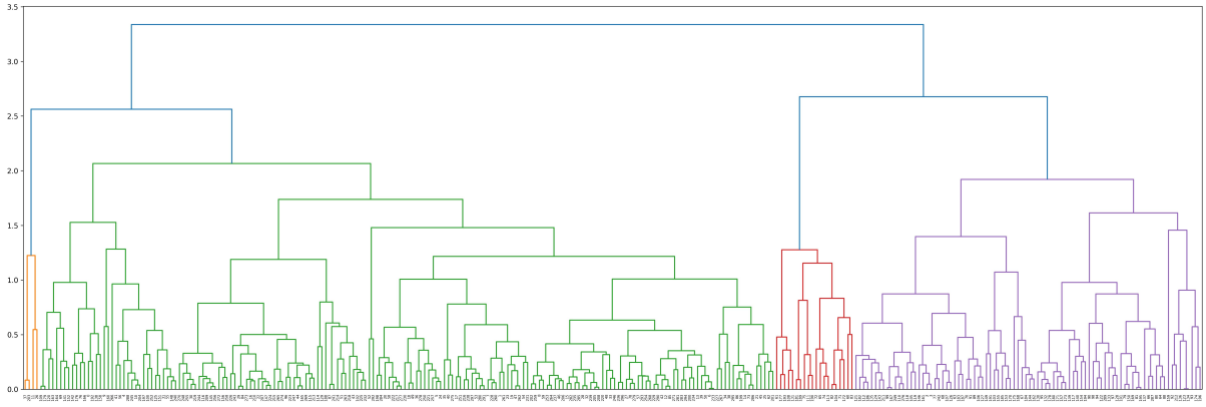
Silhouette score=0.4412
Calinski-Harabasz score=414.8504
Davies-Bouldin score=0.7860



Original Structure

Evaluation Scores

Silhouette score=0.4322
Calinski-Harabasz score=244.8449
Davies-Bouldin score=0.7424



General Conclusion

As the original clustering for single, complete, and group average linkage is generated by the data produced by data preprocessing (remove outliers, standard scaler, normalization, and pca), so it already has a good structure. However, compared with the Degree spondylolisthesis and sacral slope subset and Degree spondylolisthesis and pelvic incidence subset using both view and evaluation scores (under the assumptions of 3 clusters) of Silhouette score, Calinski-Harabasz score, and Davies-Bouldin score, we found that the following comparisons.

Single Linkage Performance: Degree spondylolisthesis and sacral slope subset > Degree spondylolisthesis and pelvic incidence subset > original

Complete Linkage Performance: Degree spondylolisthesis and pelvic incidence subset > Degree spondylolisthesis and sacral slope subset > original

Group Average Linkage Performance: Degree spondylolisthesis and pelvic incidence subset close to Degree spondylolisthesis and sacral slope subset close to original

Hence, in general, they all achieve a good structure, but the two subsets has a slightly better structures.

Appendix

My Original Code is already pushed to my github repository:
https://github.com/alfreddLUO/Hierachical-Clustering-and-K-Means-Clustering-Analysis