Please complete the assigned problems to the best of your abilities. Ensure that your work is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

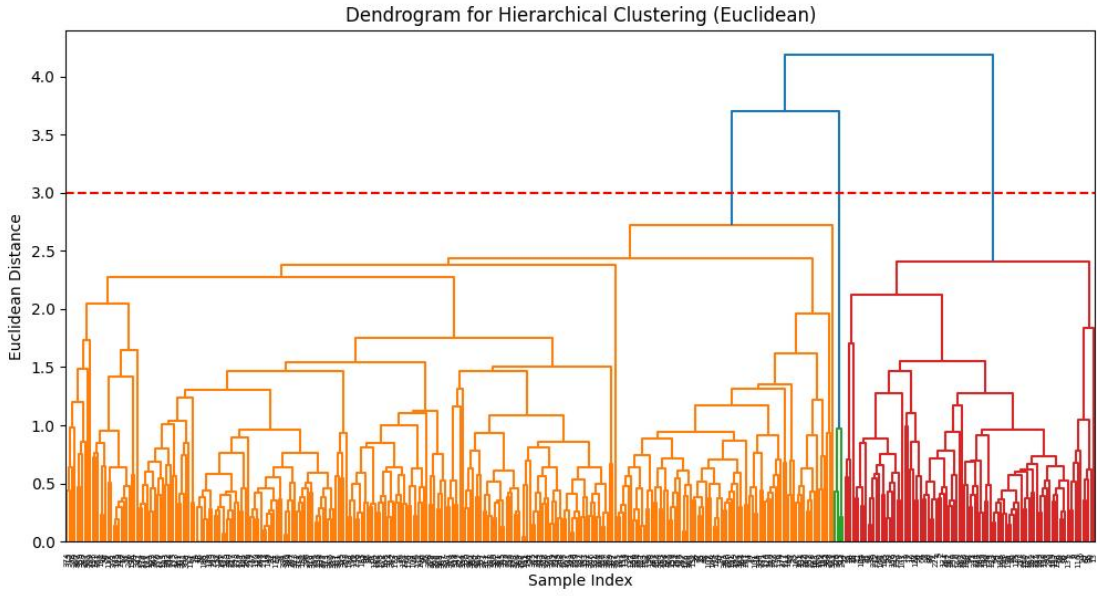
1. Practicum Problems

These problems will primarily reference the lecture materials and the examples given in class using Python. It is suggested that a Jupyter/IPython notebook be used for programmatic components.

1.1 Problem 1

Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into Python using a Pandas dataframe. Using only the continuous fields as features, impute any missing values with the mean,and perform Hierarchical Clustering (Use sklearn.cluster.AgglomerativeClustering) with linkage set to average and the default affinity set to a euclidean. Set the remaining parameters to obtain a shallow tree with 3 clusters as the target. Obtain the mean and variance values for each cluster and compare these values to the values obtained for each class if we used origin as a class label. Is there a Clear relationship between cluster assignment and class label?

The dendrogram after hierarchical clustering is as follows:



The mean and variance values of each cluster are as follows:

```
Hierarchical Cluster Stats:
                            mpg            displacement               \
                           mean        var         mean          var
hierarchical_cluster
0                     26.177441  41.303375   144.304714  3511.485383
1                     14.528866   4.771033   348.020619  2089.499570
2                     43.700000   0.300000    91.750000    12.250000

                      horsepower                   weight                \
                            mean        var         mean           var
hierarchical_cluster
0                      86.120275  294.554450  2598.414141  299118.709664
1                     161.804124  674.075816  4143.969072  193847.051117
2                      49.000000    4.000000  2133.750000   21672.916667

                     acceleration
                             mean       var
hierarchical_cluster
0                       16.425589  4.875221
1                       12.641237  3.189948
2                       22.875000  2.309167

Origin Class Stats:
              mpg            displacement               horsepower  \
             mean        var         mean          var        mean
origin
1       20.083534  40.997026   245.901606  9702.612255  119.048980
2       27.891429  45.211230   109.142857   509.950311   80.558824
3       30.450633  37.088685   102.708861   535.465433   79.835443

                         weight                acceleration
                var         mean           var         mean       var
origin
1       1591.833657  3361.931727  631695.128385    15.033735  7.568615
2        406.339772  2423.300000  240142.328986    16.787143  9.276209
3        317.523856  2221.227848  102718.485881    16.172152  3.821779
```

The contingency table constructed from the "origin" column and the "hierarchical_cluster" column is as follows:

```
Hierarchical vs Origin:
 hierarchical_cluster    0   1  2
origin
1                      152  97  0
2                       66   0  4
3                       79   0  0
```

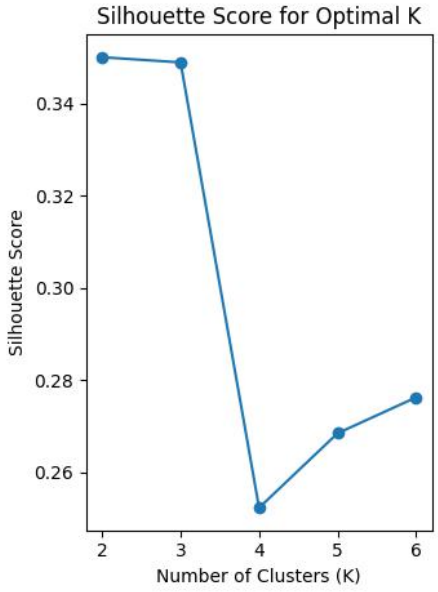
From the cross-distribution table of the clustering results and the "origin" categories, it can be seen that there is **a certain relationship between the cluster assignment and the category labels, but this relationship is not obvious.** Specifically, the distribution of different "origin" categories in each cluster is not uniform: the samples of Class 1 are mainly concentrated in Cluster 0 and Cluster 1, the samples of Class 2 are mostly concentrated in Cluster 2, and almost all samples of Class 3 are distributed in Cluster 0. If there is no relationship, each category should be evenly distributed in each cluster. The non-uniform mapping between this clustering structure and the labels indicates that there is a certain correspondence between them.

To further quantitatively analyze this relationship, I calculated the F-measure value between the hierarchical clustering results and the "origin" categories, and the result is 0.497. Considering that the F-value ranges from 0 (no association) to 1 (complete agreement), 0.497 indicates a moderate degree of correspondence between the two, but it is not sufficient to show that the clustering results can clearly distinguish the "origin" categories. Therefore, it can be considered that there is a certain correlation between the clustering and the class labels, but this relationship is not significant.

1.2 Problem 2

Load the Boston dataset (sklearn.datasets.load boston()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters ranging from 2 to 6. Provide the Silhouette score to justify which value of k is optimal. Calculate the mean values for all features in each cluster for the optimal clustering - how do these values differ from the centroid coordinates?

After performing the K-means analysis, the silhouette coefficient is calculated, and the graph of the change in the silhouette coefficient with the change in the number of clusters is as follows.



It can be seen that when the cluster is **2**, the contour coefficient is the largest, so the optimal k value is 2. At this point, the contour coefficient is 0.35.Calculate the mean of all features in each cluster under the optimal clustering, as shown below:

```
The mean value of all features in each cluster:
                   0                   1                   2   \
0 -0.3898012191570719  0.2623916673149328 -0.6152940245726287
1  0.7245457689416755 -0.4877223646701291  1.1436821134711581

                   3                   4                   5   \
0  0.0029118214285368 -0.5829159437747091  0.2449126250915631
1 -0.0054123686440034  1.0834991271292609 -0.4552330714978773

                   6                   7                   8   \
0 -0.4335841611084604  0.4544914112686059 -0.5834517228200965
1  0.8059276214953864 -0.8447891203806294  1.0844950102136257

                   9                   10                  11  \
0 -0.6297268870156213 -0.2946620052887642  0.3286002735406785
1  1.1705092984640646  0.5477050832768544 -0.6107880790671389

                   12                  13
0 -0.4534974696857152  0.3536413215392679
1  0.8429416244440686 -0.6573333038780759
Centroid coordinates:
[[-0.3898012191570719  0.2623916673149329 -0.6152940245726283
   0.0029118214285365 -0.582915943774709   0.244912625091563
  -0.4335841611084604  0.4544914112686062 -0.5834517228200966
  -0.6297268870156213 -0.2946620052887644  0.3286002735406782
  -0.4534974696857151  0.3536413215392679]
 [ 0.7245457689416759 -0.4877223646701286  1.1436821134711592
  -0.0054123686440036  1.0834991271292618 -0.4552330714978776
   0.8059276214953867 -0.8447891203806296  1.0844950102136266
   1.1705092984640684  0.5477050832768559 -0.610788079067139
   0.8429416244440692 -0.6573333038780758]]
```

The mean values of the features within the cluster are almost the same as the centroid coordinates because the objective of the K-Means algorithm is to minimize the sum of the squared distances from the samples to the centroids of the clusters they belong to. In each iteration of the algorithm, the centroid of each cluster is updated to the arithmetic mean of all samples in each feature dimension within that cluster. That is to say, the centroid is essentially the mean point of the samples. Therefore, when the algorithm converges, the calculation methods for the feature means within the cluster and the centroids are completely consistent, and the numerical values are naturally almost the same, with only minor numerical errors or differences in the index order.

1.3 Problem 3

Load the wine dataset (sklearn.datasets.load_wine()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters set to 3. Given the actual class labels, calculate the Homogeneity/Completeness for the optimal k - what information does each of these metrics provide?

The results of the index calculation are as follows:

```
Homogeneity: 0.8788432003662366
Completeness: 0.8729636016078731
```

Homogeneity measures whether the samples within each clustering cluster belong to the same true category. The closer the value is to 1, the higher the purity within the cluster, and the better the clustering effect in distinguishing different categories. **The homogeneity value of this clustering is 0.8788432003662366, indicating that most of the samples within the clusters in the clustering result come from the same actual category, that is, the clustering algorithm has well distinguished the samples of different categories.**

Completeness measures whether the samples of the same true category are all classified into the same clustering cluster. The closer the value is to 1, the more concentrated the samples of the same category are in one cluster. **The completeness value of this clustering is 0.8729636016078731, indicating that during the clustering process, most of the samples of the same actual category have been grouped into the same cluster, but there are also some cases where samples of the same category are divided into different clusters.**

Overall, these two scores are relatively high, indicating that there is a good corresponding relationship between the clustering structure and the true category labels.

END