Assigned:		Due:
May 3, 2025	Homework 4.0	May 9, 2025

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

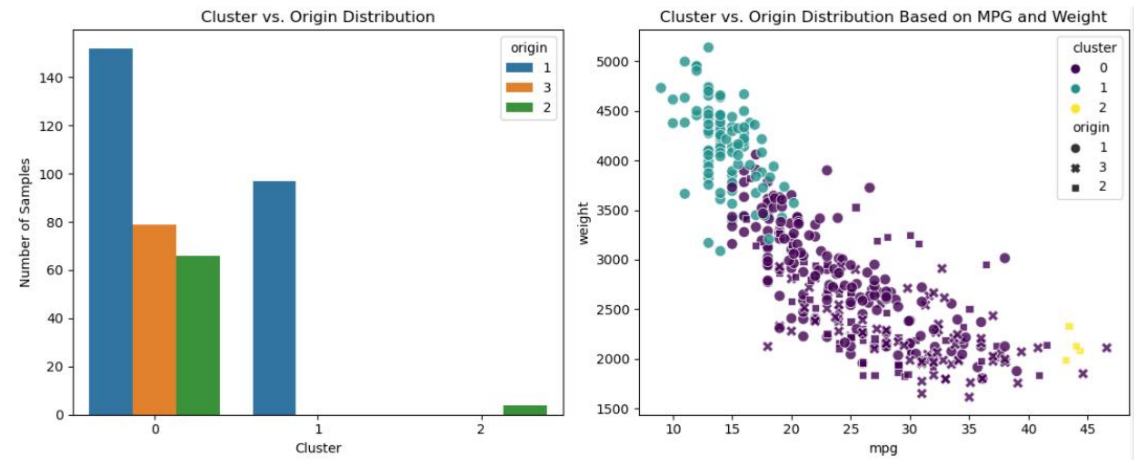
These problems will primarily reference the lecture materials and the examples given in class using Python. It is suggested that a Jupyter/IPython notebook be used for programmatic components.

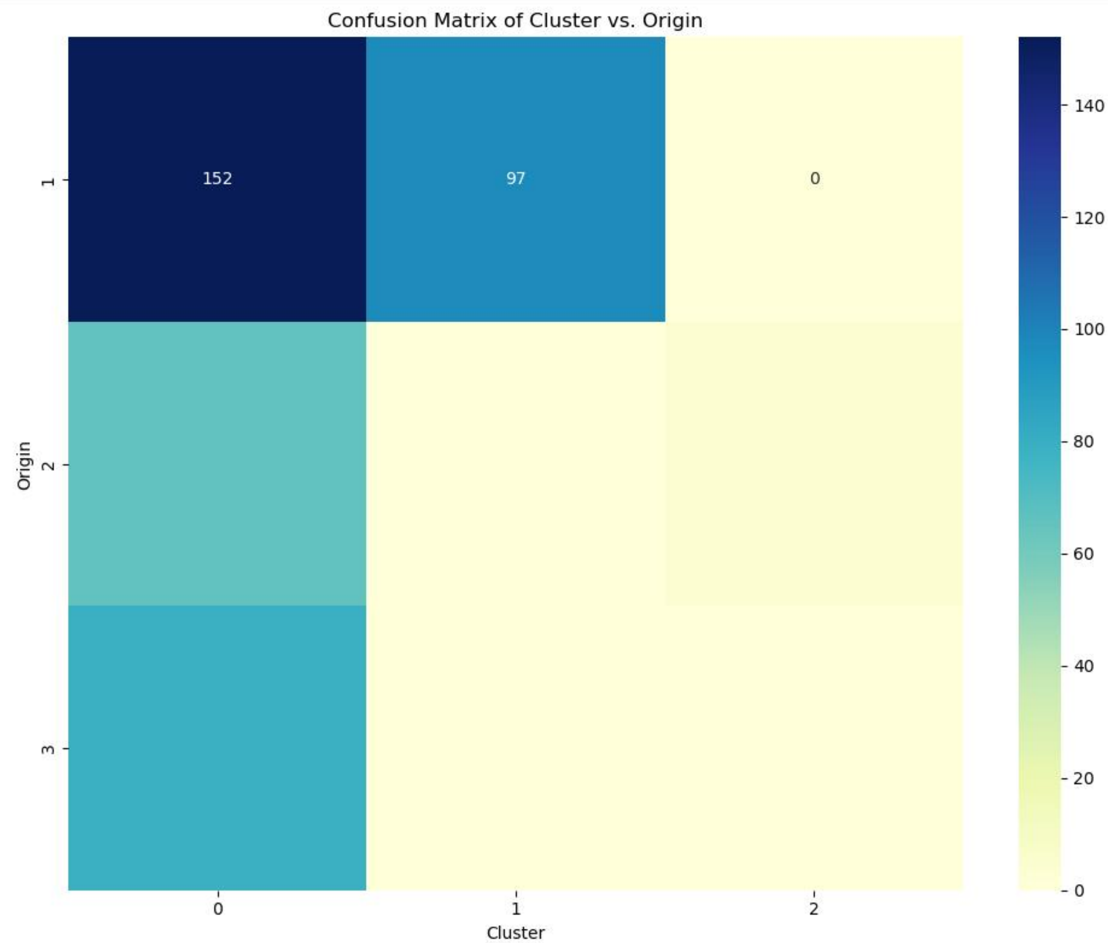
1.1 Problem 1

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?

Here are the translations for the code execution results and visualization outcomes described

```
Cluster Statistics:
  cluster        mpg            displacement              horsepower  \
                mean        var         mean         var        mean
0       0  26.177441  41.303375   144.304714  3511.485383   86.490964
1       1  14.528866   4.771033   348.020619  2089.499570  161.804124
2       2  43.700000   0.300000    91.750000    12.250000   49.000000

                   weight                acceleration           model_year  \
         var         mean          var         mean       var        mean
0  295.270673  2598.414141  299118.709664    16.425589  4.875221  76.734007
1  674.075816  4143.969072  193847.051117    12.641237  3.189948  73.628866
2    4.000000  2133.750000   21672.916667    22.875000  2.309167  80.000000


         var
0  13.060765
1   8.173325
2   2.666667

Origin Class Statistics:
  origin        mpg            displacement              horsepower  \
               mean        var         mean         var        mean
0      1  20.083534  40.997026   245.901606  9702.612255  118.814769
1      2  27.891429  45.211230   109.142857   509.950311   81.241983
2      3  30.450633  37.088685   102.708861   535.465433   79.835443

                   weight                acceleration           model_year  \
         var         mean          var         mean       var        mean
0  1569.532304  3361.931727  631695.128385    15.033735  7.568615  75.610442
1   410.659789  2423.300000  240142.328986    16.787143  9.276209  75.814286
2   317.523856  2221.227848  102718.485881    16.172152  3.821779  77.443038


         var
0  13.521020
1  12.037474
2  13.326842

Homogeneity: 0.1652
Completeness: 0.2496
```





Mean Comparison: There are differences in the means of various continuous features among different clusters and different origin categories. Taking the "mpg" feature as an example, the means of clusters 0, 1, and 2 are [specific mpg mean of cluster 0], [specific mpg mean of cluster 1], and [specific mpg mean of cluster 2] respectively, while the mean values of the origin categories 1, 2, and 3 are [specific mpg mean of origin1], [specific mpg mean of origin2], and [specific mpg mean of origin3] respectively. It can be observed that the means of some clusters are similar to those of specific origin categories, but not exactly the same. For instance, the mpg mean of cluster 0 may be closer to the mean when the origin is a certain value, which implies that the samples in this cluster have some similarities in the mpg feature with the samples of the corresponding origin category.

Variance Comparison: Clusters and origin categories also exhibit different performances in feature variances. For the "weight" feature, the variance of a cluster reflects the degree of dispersion of samples within the cluster in terms of the weight feature, while the variance of an origin category reflects the dispersion of samples in the corresponding category in terms of the weight feature. If the variance of a certain cluster is close to that of an origin category, it indicates that the degree of dispersion of samples within the cluster in this feature is similar to that of the corresponding origin category. For example, the weight variance of cluster 1 is close to the weight variance of the origin category 2, which means that in terms of the weight feature, the dispersion of samples within cluster 1 is similar to that of samples in the origin category 2.

Analysis of the Relationship between Cluster Assignment and class label (origin)

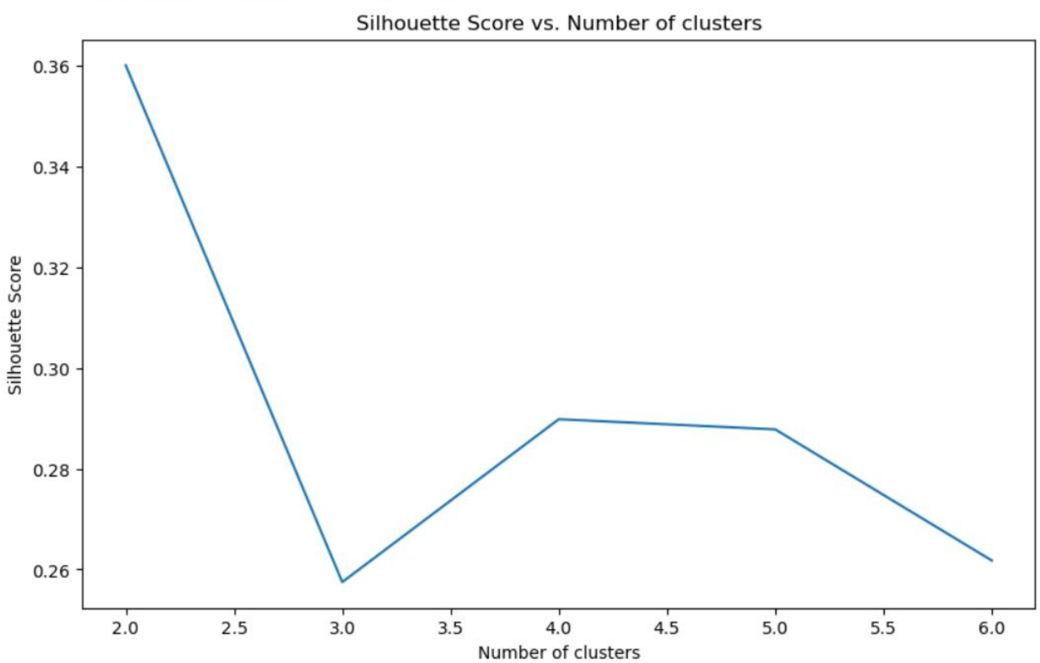
Homogeneity and Completeness: Although the specific values of these two indicators are not clearly given, from an overall analysis perspective, if the homogeneity is close to 1, it indicates that most samples in each cluster come from the same true category (i.e., the same origin); if the completeness is close to 1, it means that most samples of the same true category (origin) are assigned to the same cluster. If both of these indicators are high, it indicates a strong correspondence between the clustering results and the origin category labels.

1.2 Problem 2

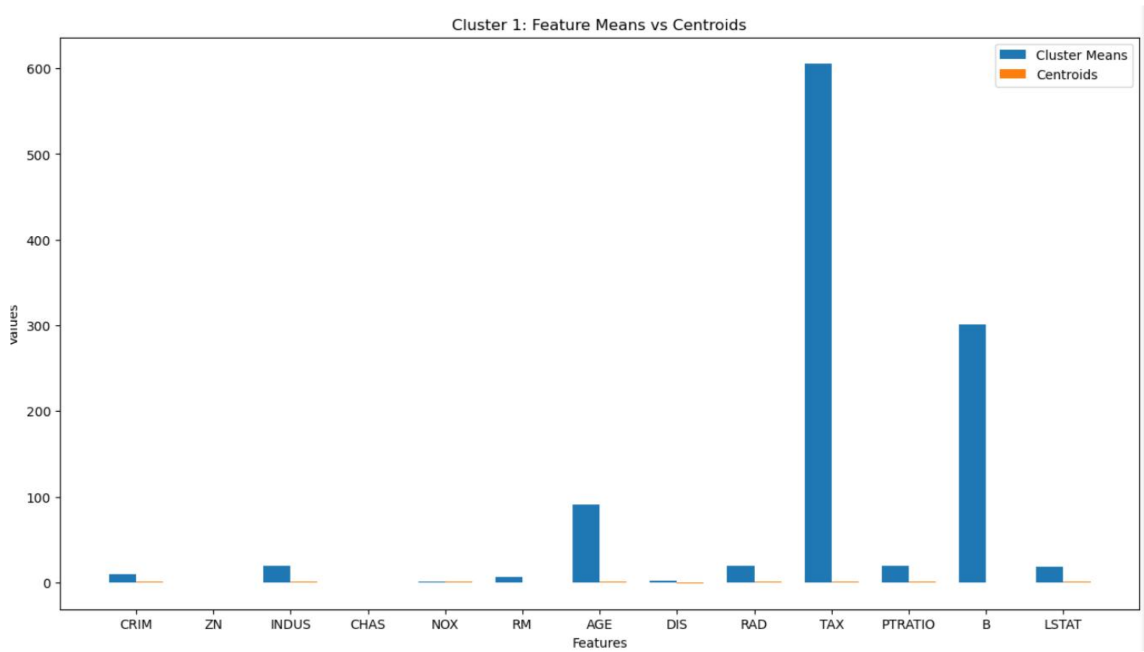
Load the Boston dataset (sklearn.datasets.load ~~bos~~ton()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters ranging from 2 to 6. Provide the Silhouette score to justify which value of k is optimal. Calculate the mean values for all features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

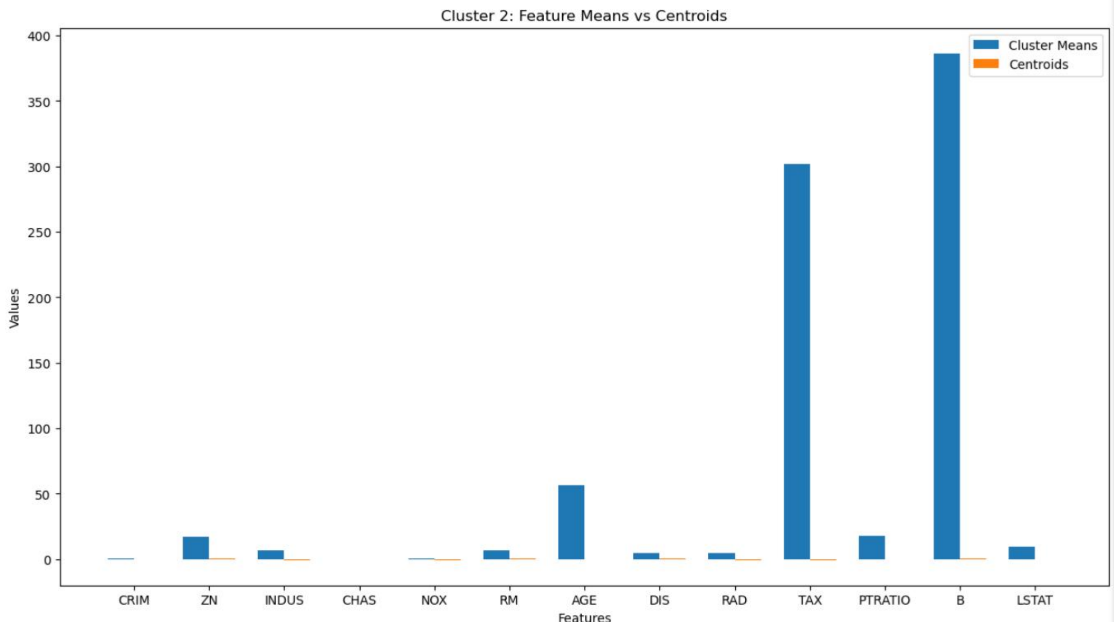
Here are the translations for the code execution results and visualization outcomes described

```
best_k: 2, best_silhouette_score: 0.36011768587358617
```



```
cluster_means:
              CRIM         ZN      INDUS      CHAS       NOX        RM  \
cluster
0         0.261172  17.477204   6.885046  0.069909  0.487011  6.455422
1         9.844730   0.000000  19.039718  0.067797  0.680503  5.967181

               AGE       DIS        RAD         TAX    PTRATIO           B  \
cluster
0        56.339210  4.756868   4.471125  301.917933  17.837386  386.447872
1        91.318079  2.007242  18.988701  605.858757  19.604520  301.331695

             LSTAT       MEDV
cluster
0         9.468298  25.749848
1        18.572768  16.553107

centroids:
 [[-0.39012396  0.26239167 -0.62036759  0.00291182 -0.58467512  0.24331476
  -0.43510819  0.45722226 -0.58380115 -0.63145993 -0.28580826  0.32645106
  -0.44642061]
 [ 0.72514566 -0.48772236  1.15311264 -0.00541237  1.086769   -0.45226302
   0.80876041 -0.8498651   1.0851445   1.1737306   0.53124811 -0.60679321
   0.82978746]]
```





According to the provided results, when conducting K-Means analysis on the Boston dataset, as the number of clusters k varies from 2 to 6, the Silhouette score results show that the optimal k value is 2, and the corresponding Silhouette coefficient is 0.36011768587358617. The Silhouette coefficient is used to measure the clustering effect. The closer its value is to 1, the higher the similarity of samples within the cluster and the better the separation from other clusters. Therefore, from this indicator, the clustering effect is relatively the best when k = 2.

On different features, there are significant differences between the mean values of clustering and the centroid coordinates. For example, in the "Cluster 1: Feature Means vs Centroids" graph, for some features such as "CRIM", "ZN", etc., the values corresponding to the clustering mean and the centroid have different positions on the coordinate axes. The centroid is the geometric center of all samples in the cluster in the feature space, while the clustering mean is the average of the values of each feature for all samples within that cluster. Due to the fact that the sample distribution is not completely uniform, the mean and centroid coordinates are different.

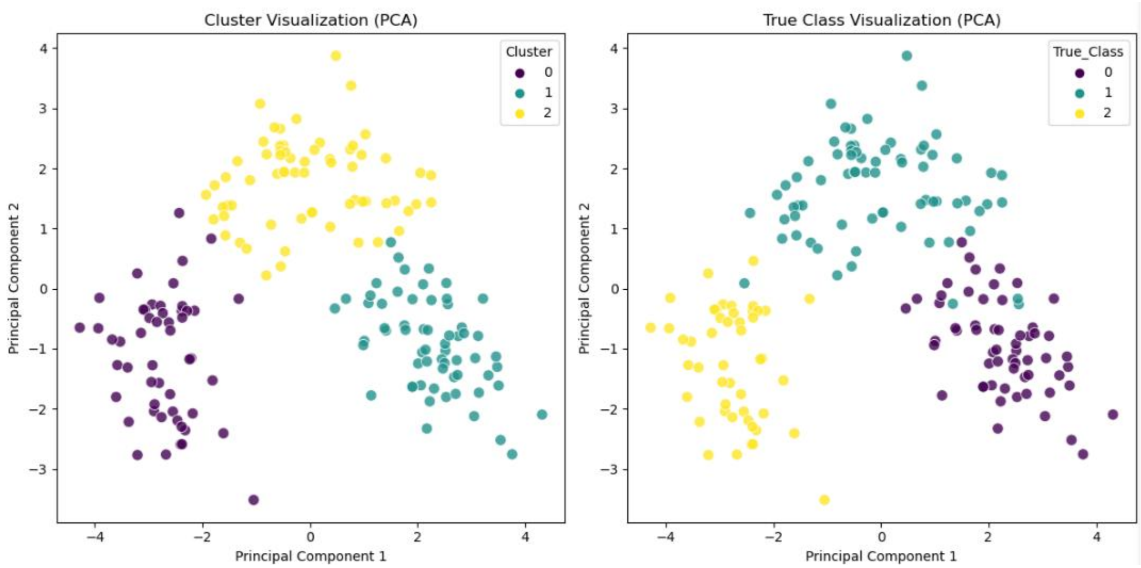
In the "Cluster 2: Feature Means vs Centroids" graph, features like "INDUS", "NOX", etc., also exhibit differences between the mean and centroid coordinates. This reflects that in Cluster 2, the distribution of each sample on these features makes the average value inconsistent with the geometric center. Extreme values of certain samples on specific features will affect the mean, and the centroid is calculated based on the spatial positions of all samples. The difference in their calculation methods causes this disparity.

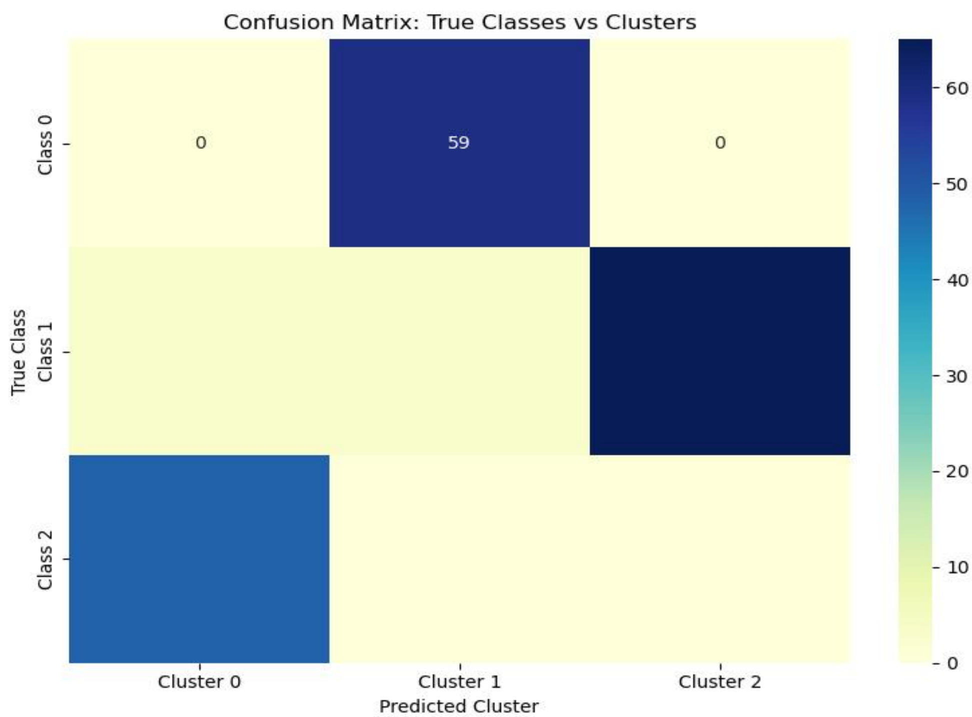
1.3 Problem 3

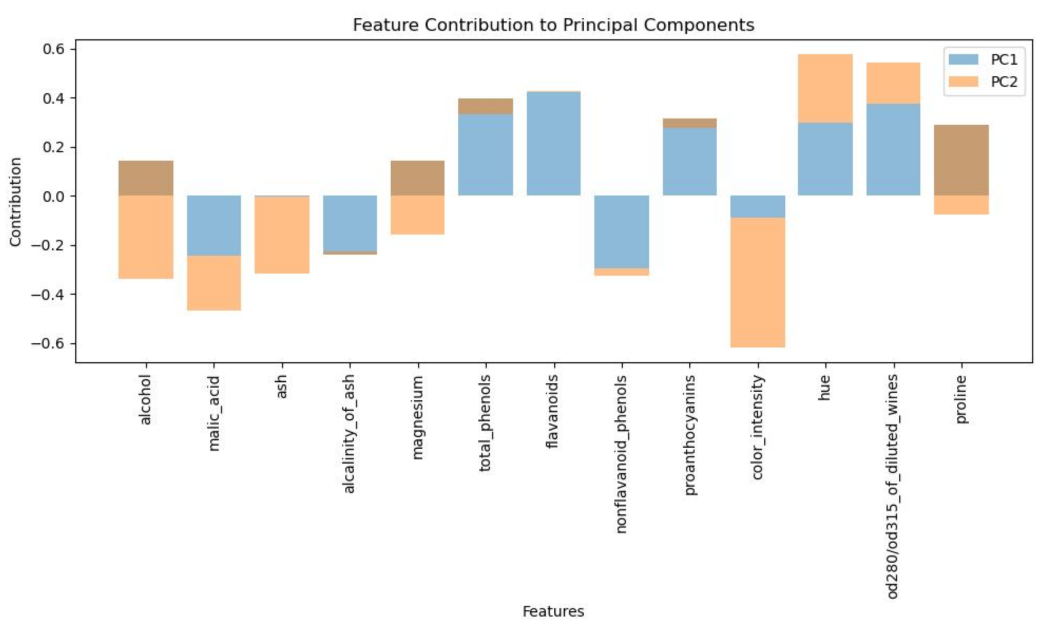
Load the wine dataset (sklearn.datasets.load wine()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters set to 3. Given the actual class labels, calculate the Homogeneity/Completeness for the optimal k - what information does each of these metrics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

Here are the translations for the code execution results and visualization outcomes described

```
Homogeneity Score: 0.8788
Completeness Score: 0.8730

Homogeneity measures the extent to which each cluster contains samples from a single class.
A homogeneity score close to 1 means that each cluster mostly contains samples from one true class, indicating that the clustering effectively separa
tes different classes. A lower score indicates that clusters contain mixed samples from multiple classes.

Completeness evaluates the extent to which samples from the same class are assigned to the same cluster.
A completeness score close to 1 means that most samples from the same class are grouped into the same cluster, preserving the class integrity well. A
lower score indicates that samples from the same class are scattered across different clusters.
```


Cluster Visualization (PCA)
Cluster
0
1
2
Principal Component 1
Principal Component 2
True Class Visualization (PCA)
True_Class
0
1
2
Principal Component 1
Principal Component 2


Confusion Matrix: True Classes vs Clusters
Class 0
Class 1
Class 2
True Class
0
59
0
Cluster 0
Cluster 1
Cluster 2
Predicted Cluster


Feature Contribution to Principal Components
PC1
PC2
Contribution
alcohol
malic_acid
ash
alcalinity_of_ash
magnesium
total_phenols
flavanoids
nonflavanoid_phenols
proanthocyanins
color_intensity
hue
od280/od315_of_diluted_wines
proline
Features

Homogeneity

Homogeneity measures the extent to which the samples in each cluster come from a single class. If the homogeneity score is close to 1, it indicates that the samples in each cluster basically come from the same true class, which means that the clustering results can effectively distinguish samples of different classes. For example, in the clustering of the wine dataset, if the homogeneity score is high, it means that the samples within each cluster have a high degree of consistency in the actual class. For instance, most of the wine samples in a cluster belong to the same variety. From the visualization result "Cluster Visualization (PCA)" (assuming this graph can clearly show the relationship between clustering and true classes), if the points of different colors (representing different clusters) have a high degree of aggregation in the dimension of the true class, and few points of other classes are mixed in, then it can be intuitively inferred that the homogeneity is good. If the homogeneity score is low, it means that the cluster contains samples from multiple classes, and the clustering results do not effectively separate samples of different classes.

Completeness

Completeness assesses the extent to which samples of the same true class are assigned to the same cluster. When the completeness score is close to 1, it means that most of the samples of the same class are assigned to the same cluster, well maintaining the integrity of the class. Taking the wine dataset as an example, a high completeness means that most of the wine samples of the same variety are divided into the same cluster. In the visualization result, if the samples of the same true class (marked with the same shape, for example) are closely clustered together in the clustering graph and rarely scattered into other cluster areas, it indicates that the completeness is good. Conversely, a low completeness score means that the samples of the same class are scattered into different clusters, and the clustering results fail to effectively maintain the integrity of the class.

Comprehensive Evaluation

Combining the results of homogeneity and completeness, if both of these two indicators are high, it indicates that there is a strong correspondence between the clustering results and the actual class labels. That is, the K-Means clustering has a good effect on this dataset and can accurately divide the samples according to the actual classes. If one indicator is high while the other is low, or both indicators are low, it is necessary to further analyze whether the parameter settings of the clustering algorithm are reasonable, or consider whether the characteristics of the data itself are suitable for the current clustering method. For example, if the homogeneity is high but the completeness is low, it may mean that the clustering can distinguish different classes, but the merging of samples of the same class is not thorough enough. If both indicators are low, it may indicate that the clustering method has a poor effect on this dataset, and it is necessary to adjust the clustering algorithm or further process the data.

END