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

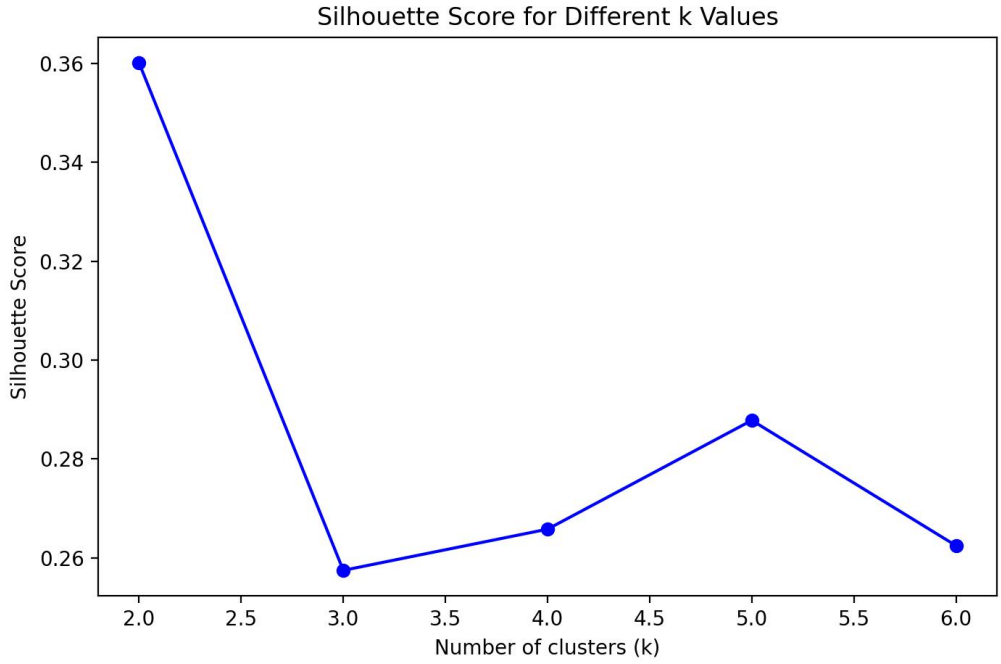
Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into Python using a Pandas dataframe. Using only the continuous fields as features, impute any missing values with the mean, and perform Hierarchical Clustering (Use sklearn.cluster.AgglomerativeClustering) with 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?

```
The mean and variance of each cluster:
              mpg             ... acceleration
             mean        var  ...         mean        var
cluster                       ...
0       27.365414  41.976309  ...    16.298120   5.718298
1       13.889062   3.359085  ...    13.025000   3.591429
2       17.510294   8.829892  ...    15.105882  10.556980

[3 rows x 10 columns]

The mean and variance for each category when using origin as the class label:
              mpg            displacement  ...         weight acceleration
             mean        var         mean  ...            var         mean       var
origin                                     ...
1       20.083534  40.997026   245.901606  ...  631695.128385    15.033735  7.568615
2       27.891429  45.211230   109.142857  ...  240142.328986    16.787143  9.276209
3       30.450633  37.088685   102.708861  ...  102718.485881    16.172152  3.821779

[3 rows x 10 columns]
```

The mean value of mpg for cluster 0 is 27.365414, while the mean value of mpg for the category labeled "origin" as 2 is 27.891429, which is relatively close; however, the mean value of acceleration for cluster 1, 13.889062, differs significantly from the mean values of mpg for each category label. The mean value of acceleration for cluster 0, 16.298120, does not show a clear and consistent correspondence compared to the values of 15.033735 for "origin" as 1, 16.787143 for "origin" as 2, and 16.172152 for "origin" as 3. There is no regular matching pattern in the mean values of this feature across different clusters and different category labels. Through the observation of these mean values and variances of the features, no very clear and stable correspondence between the cluster allocation and the category labels based on "origin" was found. There is no obvious relationship between the two.

1.2 Problem 2

Load the Boston dataset (sklearn.datasets.load ~~bos~~ston()) 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?




```
For k=2, the silhouette score is 0.3601
For k=3, the silhouette score is 0.2575
For k=4, the silhouette score is 0.2658
For k=5, the silhouette score is 0.2878
For k=6, the silhouette score is 0.2625

Optimal number of clusters: 2

Cluster means (original scale):
       CRIM            ZN      INDUS  ...    PTRATIO           B      LSTAT
0  0.261172  1.747720e+01   6.885046  ...  17.837386  386.447872   9.468298
1  9.844730  1.243450e-14  19.039718  ...  19.604520  301.331695  18.572768

[2 rows x 13 columns]

Centroid coordinates (scaled):
       CRIM        ZN     INDUS  ...   PTRATIO         B     LSTAT
0 -0.390124  0.262392 -0.620368  ... -0.285808  0.326451 -0.446421
1  0.725146 -0.487722  1.153113  ...  0.531248 -0.606793  0.829787

[2 rows x 13 columns]
```

The difference in numerical scale: The cluster mean is the statistical result of the data in the original scale, while the centroid coordinate is based on the standardized data. For instance, in the "CRIM" feature, the mean of cluster 0 is 0.261172, and the centroid coordinate is -0.390124. Standardization transforms the data to a distribution with a mean of 0 and a variance of 1, which leads to the difference in their values.

The difference in calculation essence: The cluster mean is the arithmetic average of the feature values of all sample points in the cluster, reflecting the average feature of the data within the cluster. The centroid coordinate is the position of the cluster center

determined by the K-means clustering algorithm during the iterative process, obtained by continuously optimizing the objective function (such as minimizing the sum of the squared distances from the sample points to the centroid). It focuses more on representing the central position of the cluster. The calculation logic and meaning of the two are different. Although they are related in measuring the characteristics of the cluster, their specific values are generally not equal.

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?

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

The homogeneity score is 0.8788432003662366, which measures whether each cluster contains only samples of a single category. The maximum score is 1. In the K-means clustering results of this wine dataset, most clusters have relatively pure sample categories within them, and the mixing degree of samples from different categories is relatively low.

The completeness score is 0.8729636016078731, which measures whether all samples of the same category are classified into the same cluster. The full score is 1. This result indicates that most of the wine samples of the same category have been successfully classified into the same cluster, but there are still a small number of samples of the same category that are scattered into different clusters.

Overall, these two indicators indicate that the clustering effect of the wine dataset based on the K-means algorithm (with k = 3) is relatively good, but there is still room for further optimization.

END