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?

There is no clear correspondence because vehicles of the same origin are dispersed into different clusters, and vehicles of different origins are heavily mixed in the same cluster (origin=2 and 3 in cluster 0)

The clustering results reflect more on the similarity of technical parameters (such as mpg/displacement) rather than the origin attribute. Only a small number of extreme cases (Cluster 2) show perfect correspondence, but the sample size is too small, indicating a non-linear relationship between the technical parameter features of vehicles and the origin label. Hierarchical clustering based solely on continuous features cannot effectively reconstruct the classification structure of origin

```
Cluster statistical information (mean and variance):
               mpg                 displacement                horsepower  \
              mean        var           mean          var          mean
cluster
0        26.177441  41.303375     144.304714  3511.485383     86.120275
1        14.528866   4.771033     348.020619  2089.499570    161.804124
2        43.700000   0.300000      91.750000    12.250000     49.000000

                        weight                   acceleration
                var        mean            var         mean        var
cluster
0        294.554450  2598.414141  299118.709664   16.425589  4.875221
1        674.075816  4143.969072  193847.051117   12.641237  3.189948
2          4.000000  2133.750000   21672.916667   22.875000  2.309167

Statistical information grouped by origin (mean and variance):
              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

Cross tabulation of clustering and origin:
origin     1   2   3
cluster
0        152  66  79
1         97   0   0
2          0   4   0
```

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?

optimal k=2

The centroid coordinates display the position of the center point in the standardized space, while the clustering mean is a statistical measure of the original scale

In terms of scope, the centroid coordinates are within the range of [-1,1] (standardized), and the clustering mean maintains the original variable units.

In terms of calculation method,

Cluster mean is obtained by taking the arithmetic mean of the corresponding feature values of all samples within each cluster. For example, for the feature Crimea, in cluster 0, the sum of the Crimea values of all samples in the cluster is divided by the number of samples to obtain the mean 0.261172. It is a simple statistic that reflects the average level of samples within a cluster on that feature.

The centroid coordinates are the center positions of each cluster found by the K-Means algorithm during the iterative optimization process. The algorithm continuously adjusts the centroid position to minimize the sum of distances between samples within the cluster and the centroid. For example, for the feature crim, the centroid coordinates of cluster 0 are -0.390124, which is the result of algorithm optimization and may not necessarily equal the arithmetic mean.

```
k = 2 ,  silhouette_score: 0.3601176858735861
k = 3 ,  silhouette_score: 0.2574894522739463
k = 4 ,  silhouette_score: 0.2898322145974091
k = 5 ,  silhouette_score: 0.2878157430985233
k = 6 ,  silhouette_score: 0.2982352318859569
The optimal k value is: 2
Cluster feature 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
cluster
0         9.468298
1        18.572768

Centroid coordinates:
       crim        zn     indus      chas       nox        rm       age  \
0 -0.390124  0.262392 -0.620368  0.002912 -0.584675  0.243315 -0.435108
1  0.725146 -0.487722  1.153113 -0.005412  1.086769 -0.452263  0.808760

        dis       rad       tax   ptratio         b     lstat
0  0.457222 -0.583801 -0.631460 -0.285808  0.326451 -0.446421
1 -0.849865  1.085145  1.173731  0.531248 -0.606793  0.829787
```

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?

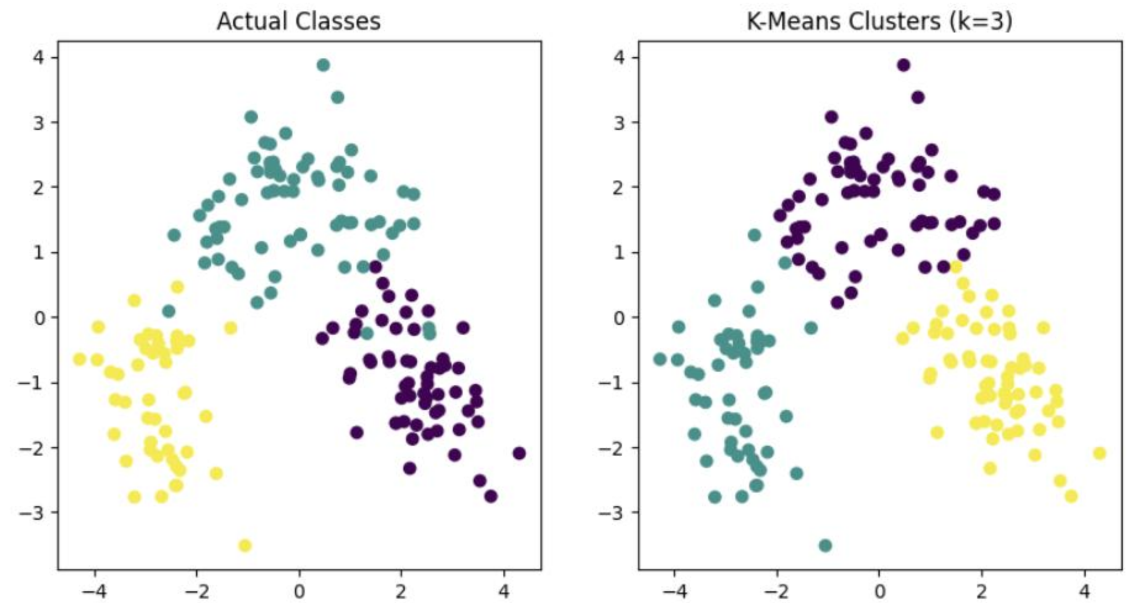
Homogeneity is a measure of the degree to which each cluster contains only samples from a single category. The higher the value, the better the consistency between the clustering results and the true categories. 0.805 indicates that most clusters have effectively concentrated a single category of wine

Completeness measures the degree to which samples of the same category are assigned to the same cluster, with higher values indicating the degree to which wines of the same category are clustered together. 0.814 indicates that most wines in the same category are correctly classified into the same cluster

V-measure is the harmonic mean of homogeneity and completeness, which is a single indicator for comprehensively evaluating the quality of clustering.

Both indicators exceed 0.8, indicating that the K-Means clustering results are highly consistent with the actual categories
The chemical characteristics of wine can indeed reflect its actual category differences.
The PCA dimensionality reduction graph shows that the clustering results are generally consistent with the actual category distribution. There are a few overlapping areas, which is consistent with the situation where the index is slightly lower than 1.0. K-Means, an unsupervised method, can discover structures similar to known categories. The indicators show that there is still about 20% room for improvement, which may require more complex models or feature engineering

```
Homogeneity score: 0.8788
Completeness score: 0.8730
V-measure score: 0.8759
```




```
Cluster means:
           alcohol  malic_acid       ash  alcalinity_of_ash   magnesium  \
cluster
0        12.250923    1.897385  2.231231          20.063077   92.738462
1        13.134118    3.307255  2.417647          21.241176   98.666667
2        13.676774    1.997903  2.466290          17.462903  107.967742

         total_phenols  flavanoids  nonflavanoid_phenols  proanthocyanins  \
cluster
0             2.247692    2.050000              0.357692         1.624154
1             1.683922    0.818824              0.451961         1.145882
2             2.847581    3.003226              0.292097         1.922097

         color_intensity       hue  od280/od315_of_diluted_wines      proline  \
cluster
0               2.973077  1.062708                      2.803385   510.169231
1               7.234706  0.691961                      1.696667   619.058824
2               5.453548  1.065484                      3.163387  1100.225806

           target
cluster
0        1.000000
1        1.941176
2        0.048387
```

END