P1：

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?

Cluster Statistics (Mean and Variance)

This section provides the mean and variance for each feature within the clusters identified by hierarchical clustering.

plaintext

Cluster statistics (mean and variance):

	mpg		cylinders		displacement	
	mean	var	mean	var	mean	var
cluster						
0	27.365414	41.976309	4.443609	0.851525	131.934211	2828.083391
1	13.889062	3.359085	8.000000	0.000000	358.093750	2138.213294
2	17.510294	8.829892	7.014706	1.059482	278.985294	2882.492318

	horsepower		weight		acceleration
	mean	var	mean	var	mean
cluster					
0	84.300061	369.143491	2459.511278	182632.099872	16.298120
1	167.046875	756.521577	4398.593750	74312.340278	13.025000
2	124.470588	713.088674	3624.838235	37775.809263	15.105882

	model_year	
	var	mean	var
cluster			
0	5.718298	76.751880	14.141978
1	3.591429	73.375000	5.984127
2	10.556980	75.588235	10.454785

Explanation of Cluster Statistics

Cluster 0:

MPG: Mean = 27.37, Variance = 41.98

Cylinders: Mean = 4.44, Variance = 0.85

Displacement: Mean = 131.93, Variance = 2828.08

Horsepower: Mean = 84.30, Variance = 369.14

Weight: Mean = 2459.51, Variance = 182632.10

Acceleration: Mean = 16.30, Variance = 16.30
Model Year: Mean = 76.75, Variance = 14.14

Cluster 1:
MPG: Mean = 13.89, Variance = 3.36
Cylinders: Mean = 8.00, Variance = 0.00
Displacement: Mean = 358.09, Variance = 2138.21
Horsepower: Mean = 167.05, Variance = 756.52
Weight: Mean = 4398.59, Variance = 74312.34
Acceleration: Mean = 13.03, Variance = 13.03
Model Year: Mean = 73.38, Variance = 5.98

Cluster 2:
MPG: Mean = 17.51, Variance = 8.83
Cylinders: Mean = 7.01, Variance = 1.06
Displacement: Mean = 278.99, Variance = 2882.49
Horsepower: Mean = 124.47, Variance = 713.09
Weight: Mean = 3624.84, Variance = 37775.81
Acceleration: Mean = 15.11, Variance = 15.11
Model Year: Mean = 75.59, Variance = 10.45

Class Statistics (Mean and Variance)
This section provides the mean and variance for each feature within the classes defined by the `origin` field.

plaintext
Class statistics (mean and variance):

	mpg		cylinders		displacement	\
	mean	var	mean	var	mean	var
origin						
1	20.083534	40.997026	6.248996	2.760332	245.901606	9702.612255
2	27.891429	45.211230	4.157143	0.250311	109.142857	509.950311
3	30.450633	37.088685	4.101266	0.348588	102.708861	535.465433

	horsepower		weight		acceleration	\
	mean	var	mean	var	mean	
origin						
1	119.048980	1591.833657	3361.931727	631695.128385	15.033735	
2	80.558824	406.339772	2423.300000	240142.328986	16.787143	
3	79.835443	317.523856	2221.227848	102718.485881	16.172152	

	model_year	
	var	mean	var
origin			

1	7.568615	75.610442	13.521020
2	9.276209	75.814286	12.037474
3	3.821779	77.443038	13.326842

Explanation of Class Statistics

Class 1:

MPG: Mean = 20.08, Variance = 40.99

Cylinders: Mean = 6.25, Variance = 2.76

Displacement: Mean = 245.90, Variance = 9702.61

Horsepower: Mean = 119.05, Variance = 1591.83

Weight: Mean = 3361.93, Variance = 631695.13

Acceleration: Mean = 15.03, Variance = 15.03

Model Year: Mean = 75.61, Variance = 13.52

Class 2:

MPG: Mean = 27.89, Variance = 45.21

Cylinders: Mean = 4.16, Variance = 0.25

Displacement: Mean = 109.14, Variance = 509.95

Horsepower: Mean = 80.56, Variance = 406.34

Weight: Mean = 2423.30, Variance = 240142.33

Acceleration: Mean = 16.79, Variance = 16.79

Model Year: Mean = 75.81, Variance = 12.04

Class 3:

MPG: Mean = 30.45, Variance = 37.09

Cylinders: Mean = 4.10, Variance = 0.35

Displacement: Mean = 102.71, Variance = 535.47

Horsepower: Mean = 79.84, Variance = 317.52

Weight: Mean = 2221.23, Variance = 102718.49

Acceleration: Mean = 16.17, Variance = 16.17

Model Year: Mean = 77.44, Variance = 13.33

Crosstab of Class Labels and Clusters

This section shows the distribution of clusters within each class label.

plaintext

Crosstab of class labels and clusters:

col_0	0	1	2
origin			
1	120	64	65
2	67	0	3
3	79	0	0

Explanation of Crosstab

Class 1: 120 data points in Cluster 0, 64 in Cluster 1, and 65 in Cluster 2.
Class 2: 67 data points in Cluster 0, 0 in Cluster 1, and 3 in Cluster 2.
Class 3: 79 data points in Cluster 0, 0 in Cluster 1, and 0 in Cluster 2.

Conclusion
The crosstab indicates that there is some overlap between the clusters and the class labels, but Class 3 is predominantly in Cluster 0. This suggests that while there is some relationship between the cluster assignments and the class labels, it is not perfectly aligned.

P2：

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?

K-Means Clustering Analysis on the Boston Housing Dataset
Silhouette Scores
The silhouette scores for different numbers of clusters (k) are as follows:
Silhouette score for k=2: 0.382
Silhouette score for k=3: 0.268
Silhouette score for k=4: 0.287
Silhouette score for k=5: 0.272
Silhouette score for k=6: 0.264

Based on the silhouette scores, the optimal number of clusters is determined to be 2.

Mean Values for Each Cluster
The mean values of the features for each cluster are presented in the table below:

Feature	Cluster 0 Mean	Cluster 1 Mean
CRIM	0.261172	9.844730
ZN	17.477204	0.000000
INDUS	6.885046	19.039718
NOX	0.487011	0.680503
RM	6.455422	5.967181
AGE	56.339210	91.318079
DIS	4.756868	2.007242
RAD	4.471125	18.988701
TAX	301.917933	605.858757
PTRATIO	17.837386	19.604520
B	386.447872	301.331695

| LSTAT | 9.468298 | 18.572768 |

Cluster Center Coordinates

The coordinates of the cluster centers are presented in the table below:

Feature	Cluster 0 Center	Cluster 1 Center
CRIM	-0.390124	0.725146
ZN	0.262392	-0.487722
INDUS	-0.620368	1.153113
NOX	-0.584675	1.086769
RM	0.243315	-0.452263
AGE	-0.435108	0.808760
DIS	0.457222	-0.849865
RAD	-0.583801	1.085145
TAX	-0.631460	1.173731
PTRATIO	-0.285808	0.531248
B	0.326451	-0.606793
LSTAT	-0.446421	0.829787

Comparison of Mean Values and Cluster Center Coordinates

Cluster 0

Mean Values:

CRIM: 0.261172

ZN: 17.477204

INDUS: 6.885046

NOX: 0.487011

RM: 6.455422

AGE: 56.339210

DIS: 4.756868

RAD: 4.471125

TAX: 301.917933

PTRATIO: 17.837386

B: 386.447872

LSTAT: 9.468298

Cluster Center Coordinates:

CRIM: -0.390124

ZN: 0.262392

INDUS: -0.620368

NOX: -0.584675

RM: 0.243315

AGE: -0.435108

DIS: 0.457222
RAD: -0.583801
TAX: -0.631460
PTRATIO: -0.285808
B: 0.326451
LSTAT: -0.446421

Cluster 1
Mean Values:
CRIM: 9.844730
ZN: 0.000000
INDUS: 19.039718
NOX: 0.680503
RM: 5.967181
AGE: 91.318079
DIS: 2.007242
RAD: 18.988701
TAX: 605.858757
PTRATIO: 19.604520
B: 301.331695
LSTAT: 18.572768

Cluster Center Coordinates:
CRIM: 0.725146
ZN: -0.487722
INDUS: 1.153113
NOX: 1.086769
RM: -0.452263
AGE: 0.808760
DIS: -0.849865
RAD: 1.085145
TAX: 1.173731
PTRATIO: 0.531248
B: -0.606793
LSTAT: 0.829787

Interpretation
Cluster 0:

This cluster represents neighborhoods with lower crime rates (CRIM), higher proportions of residential land zoned for lots over 25,000 sq.ft （ZN), and lower proportions of non-retail business acres (INDUS).

These neighborhoods have fewer nitric oxides concentration (NOX), more rooms per dwelling (RM), and older housing units (AGE).

They are located closer to employment centers (DIS), have lower accessibility to radial highways (RAD), lower property-tax rates (TAX), and lower pupil-teacher ratios (PTRATIO).

The proportion of blacks (`B`) is higher, and the percentage of lower status of the population (`LSTAT`) is lower.

Cluster 1:

This cluster represents neighborhoods with higher crime rates (`CRIM`), no residential land zoned for lots over 25,000 sq.ft (`ZN`), and higher proportions of non-retail business acres (`INDUS`).

These neighborhoods have higher nitric oxides concentration (`NOX`), fewer rooms per dwelling (`RM`), and older housing units (`AGE`).

They are farther from employment centers (`DIS`), have higher accessibility to radial highways (`RAD`), higher property-tax rates (`TAX`), and higher pupil-teacher ratios (`PTRATIO`).

The proportion of blacks (`B`) is lower, and the percentage of lower status of the population (`LSTAT`) is higher.

Conclusion

The optimal number of clusters for the Boston housing dataset is 2. The two clusters represent distinct types of neighborhoods based on the housing characteristics. Cluster 0 consists of neighborhoods with lower crime rates and better housing conditions, while Cluster 1 consists of neighborhoods with higher crime rates and older, less desirable housing conditions. The mean values and cluster center coordinates provide insights into the characteristics of each cluster.

P3

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: 0.8788
Completeness Score: 0.8730
```

the Homogeneity Score is 0.8788 and the Completeness Score is 0.8730. Both scores are quite high, indicating that K-Means clustering model performs well on the wine dataset. Here is a detailed analysis of these two scores:

Homogeneity Score

Score: 0.8788

Explanation: A homogeneity score close to 1 indicates that each cluster mainly contains data points from a single class. This means that the clustering result has high "purity," with most data points in each cluster belonging to the same class and very few data points from other classes being mixed in.

Analysis:

Strengths: The clustering result is very clear, with each cluster almost only containing one type of wine. This helps to distinguish between different types of wines effectively.

Potential for Improvement**: Although the score is already high, there may still be some data points that are incorrectly assigned to other clusters. You could try adjusting the parameters of the clustering algorithm (such as the random seed, initialization method, etc.), or experiment with other clustering algorithms (such as hierarchical clustering, DBSCAN, etc.) to further improve homogeneity.

Completeness Score

Score: 0.8730

Explanation: A completeness score close to 1 indicates that most data points from the same class are assigned to the same cluster. This means that the clustering result is good in terms of "completeness," with data points from each class being correctly grouped together.

Analysis:

Strengths: The clustering result can effectively reflect the true class distribution of the data. For the wine dataset, this means that most data points of the same type of wine are correctly clustered together.

Potential for Improvement: Despite the high score, there may still be a few data points that are not correctly clustered. You could further analyze the feature distribution of the data or try different feature selection methods to improve the completeness score.

Comprehensive Analysis

Clustering Effect: Both the homogeneity and completeness scores are high, indicating that the K-Means clustering performs well on the wine dataset. The clustering result has both high purity and high completeness, meaning that it is highly consistent with the actual class labels.

Practical Application: In practical applications, such a clustering result can be used to classify wines, for example, in wine quality assessment, variety identification, etc. The high homogeneity and completeness scores indicate that the clustering result is highly credible.

Further Optimization: Although the scores are already high, there is still some room for improvement. You could try the following methods:

Adjust Clustering Parameters: Adjust parameters such as `n_clusters` (number of cluster centers) and `random_state` (random seed) to see if further improvements in scores can be achieved.

Feature Engineering: Conduct more in-depth feature engineering on the data, such as selecting more relevant features or performing feature dimensionality reduction.

Try Other Algorithms: In addition to K-Means, you could also try other clustering algorithms (such as DBSCAN, hierarchical clustering, etc.) to verify whether better clustering results can be obtained.

Summary

K-Means clustering model performs excellently on the wine dataset, with both homogeneity and completeness scores close to 0.9. This indicates that the clustering result has high purity and completeness. Such a clustering result is highly valuable in practical applications. However, there is still some room for improvement.