Problem1

Cluster and Class Label Relationship Analysis

From the final output of the Jupyter Notebook, we observe that all samples with the original class label 3 and the majority of those with label 2 are grouped into Cluster 0. This indicates a strong association between Cluster 0 and original classes 2 and 3. However, Cluster 0 also contains a number of samples from class 1, and class 1 samples are dispersed across multiple clusters. This dispersion suggests that there is no clear one-to-one correspondence between the clustering assignments and the original class labels, indicating that the clustering structure does not fully capture the true class separability.

Statistics for each cluster:

	mpg		displacement		horsepower		\
	mean	var	mean	var	mean	var	
0	27.365414	41.976309	131.934211	2828.083391	84.300061	369.143491	
1	13.889062	3.359085	358.093750	2138.213294	167.046875	756.521577	
2	17.510294	8.829892	278.985294	2882.492318	124.470588	713.088674	

	weight		acceleration	
	mean	var	mean	var
0	2459.511278	182632.099872	16.298120	5.718298
1	4398.593750	74312.340278	13.025000	3.591429
2	3624.838235	37775.809263	15.105882	10.556980

Statistics grouped by origin:

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

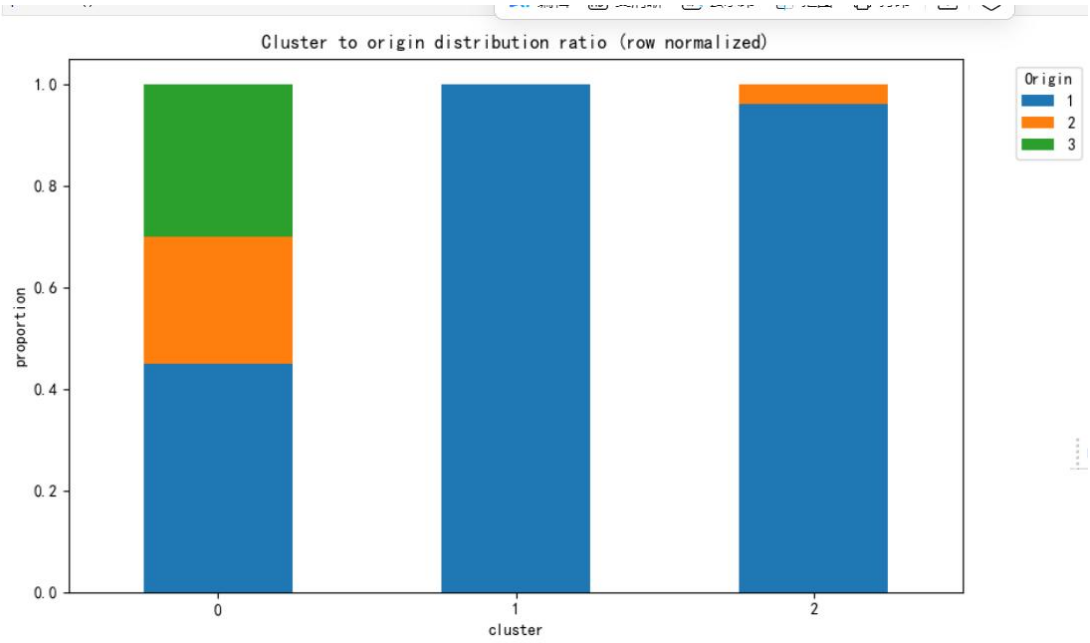
	weight		acceleration		
	var	mean	var	mean	var
origin					
1	1569.532304	3361.931727	631695.128385	15.033735	7.568615
2	410.659789	2423.300000	240142.328986	16.787143	9.276209
3	317.523856	2221.227848	102718.485881	16.172152	3.821779

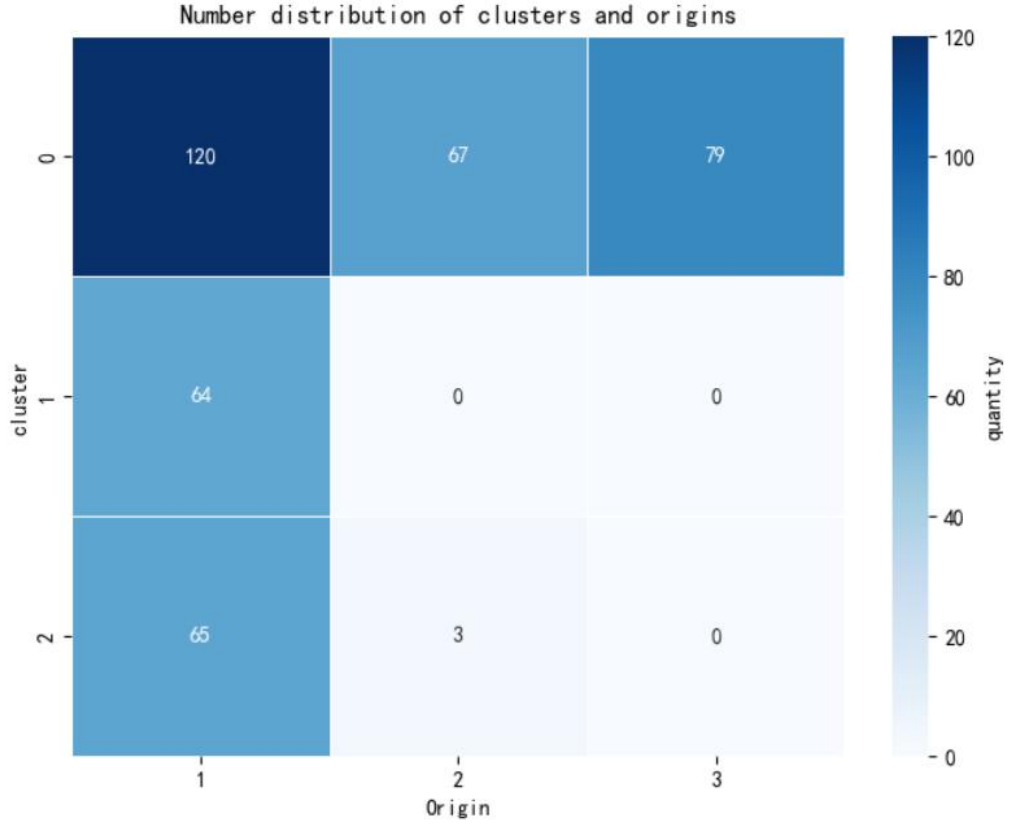
Cross-distribution table between clusters and origins (quantity):

origin	1	2	3	Total
row_0				
0	120	67	79	266
1	64	0	0	64
2	65	3	0	68
Total	249	70	79	398

Cluster-to-origin distribution ratio (row normalized):

origin	1	2	3
row_0			
0	0.45	0.25	0.3
1	1.00	0.00	0.0
2	0.96	0.04	0.0

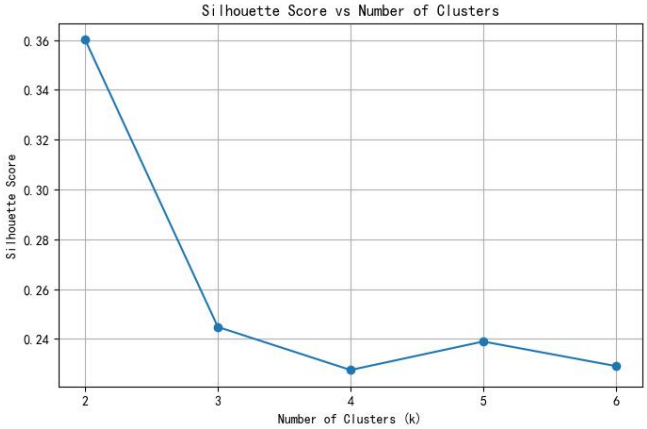

Cluster to origin distribution ratio (row normalized)
Origin
1
2
3
proportion
1.0
0.8
0.6
0.4
0.2
0.0
0
1
2
cluster


Number distribution of clusters and origins
cluster
0
1
2
120
67
79
64
0
0
65
3
0
1
2
3
Origin
quantity
120
100
80
60
40
20
0

Problem2

From the final output of the code in the Jupyter Notebook (as shown below), it is evident that k = 2 yields the highest Silhouette Score, making it the optimal number of clusters.

The clustering analysis confirms that k = 2 is the most suitable choice based on Silhouette Score. The similarity between the cluster means and centroid coordinates reinforces the consistency of the results, and minor discrepancies are negligible and expected due to numerical precision.



	Raw mean	Centroid coordinates
CRIM	0.261172	0.261172
ZN	17.477204	17.477204
INDUS	6.885046	6.885046
CHAS	0.069909	0.069909
NOX	0.487011	0.487011
RM	6.455422	6.455422
AGE	56.339210	56.339210
DIS	4.756868	4.756868
RAD	4.471125	4.471125
TAX	301.917933	301.917933
PTRATIO	17.837386	17.837386
B	386.447872	386.447872
LSTAT	9.468298	9.468298

cluster 1:

	Raw mean	Centroid coordinates
CRIM	9.844730	9.844730e+00
ZN	0.000000	1.243450e-14
INDUS	19.039718	1.903972e+01
CHAS	0.067797	6.779661e-02
NOX	0.680503	6.805028e-01
RM	5.967181	5.967181e+00
AGE	91.318079	9.131808e+01
DIS	2.007242	2.007242e+00
RAD	18.988701	1.898870e+01
TAX	605.858757	6.058588e+02
PTRATIO	19.604520	1.960452e+01
B	301.331695	3.013317e+02
LSTAT	18.572768	1.857277e+01

Problem3

Homogeneity measures whether each cluster contains only data points from a single class, while Completeness assesses whether all data points of a given class are assigned to the same cluster.

From the final output of the Jupyter Notebook (as shown below), both the Homogeneity and Completeness scores are close to 1.0, indicating that the clustering results are highly consistent with the true class labels. This suggests that the K-Means clustering algorithm effectively captured the underlying class structure of the dataset.

The high values of Homogeneity and Completeness demonstrate that the clustering results not only form pure groups but also successfully group all members of the same class together, reflecting a strong alignment with the actual class distribution.