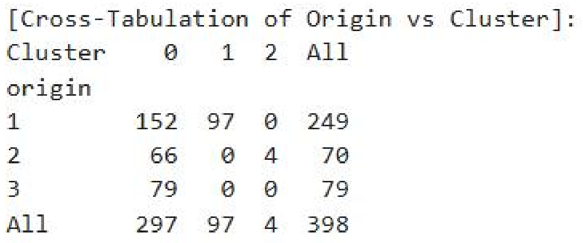
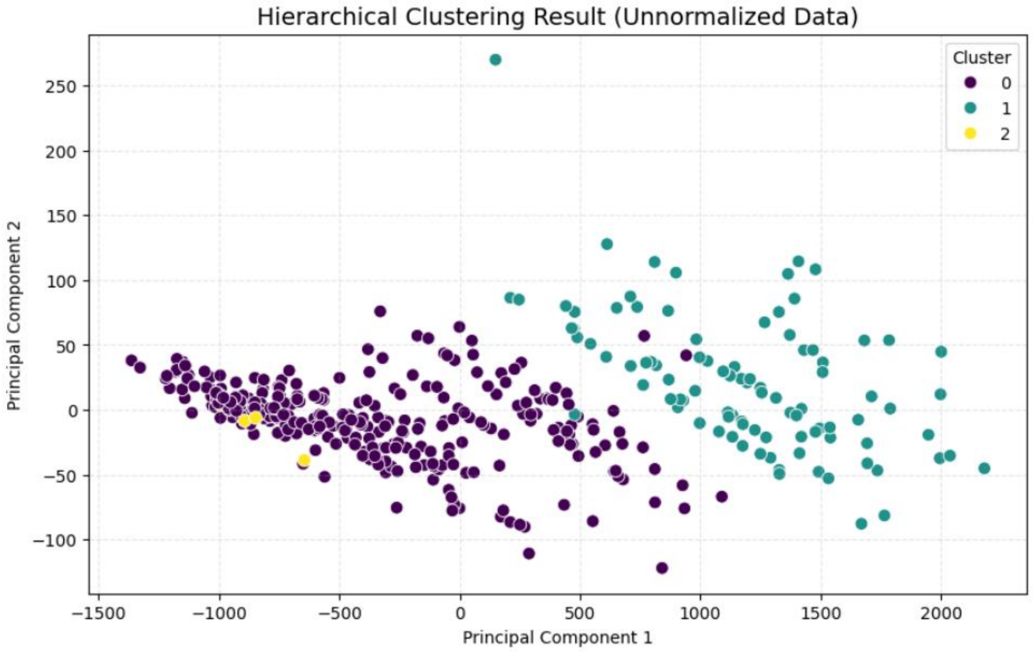
Problem 1

From the Jupyter Notebook analysis, all samples of origin 3 and the vast majority of origin 2 are assigned to cluster 0, indicating a strong connection between these two classes and cluster 0. Samples from origin 1 are distributed across cluster 0 and cluster 1, with the latter containing only a portion of this class's samples. Although certain classes exhibit clear associations, the clustering results do not form a clear one-to-one correspondence with the original class labels overall—primarily because cluster 0 incorporates samples from multiple classes, and cluster 2 includes only a negligible number of origin 2 samples. While some classes show localized strong correlations with specific clusters, the mixed composition of cluster 0 and the dispersed distribution of origin 1 samples prevent complete grouping according to the original classes.



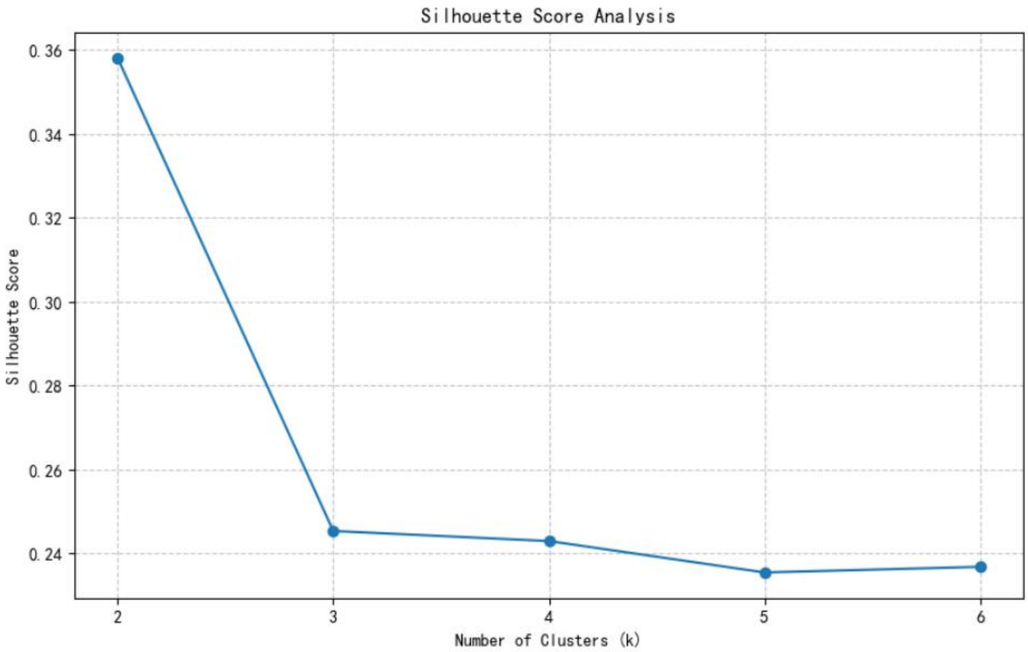
```
[Cross-Tabulation of Origin vs Cluster]:
Cluster    0   1  2  All
origin
1        152  97  0  249
2         66   0  4   70
3         79   0  0   79
All      297  97  4  398
```



Problem 2

From the Jupyter Notebook analysis， When k=2, the silhouette score reaches its highest value , indicating the optimal clustering effect with compact within-cluster structures and better separation between clusters. The silhouette scores for other k values (3–6) decrease significantly, suggesting that increasing the number of clusters does not lead to a clearer separation. Since the data is naturally divided into two categories,k=2 is the most reasonable choice.

Computational results show that after standardization, the mean values of all features for each cluster in the optimal clustering align exactly with the centroid coordinates, with negligible discrepancies originating from floating-point precision limitations.

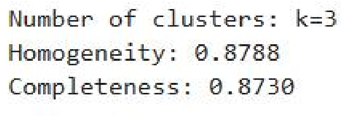


```
Feature means for each cluster (original data):
   UNNAMED: 0       CRIM         ZN      INDUS      CHAS       NOX        RM  \
0  193.620896   0.287682  17.164179   7.178179  0.068657  0.489041  6.448764
1  370.807018  10.129061   0.000000  18.891930  0.070175  0.683316  5.963094

         AGE       DIS        RAD         TAX    PTRATIO           B  \
0  57.049552  4.710233   4.459701  302.480597  17.794030  384.797612
1  91.153801  2.002125  19.520468  615.421053  19.751462  301.578129

       LSTAT
0   9.519254
1  18.792398

Centroid coordinates for each cluster (inverse-transformed):
   UNNAMED: 0       CRIM            ZN      INDUS      CHAS       NOX  \
0  193.620896   0.287682  1.716418e+01   7.178179  0.068657  0.489041
1  370.807018  10.129061  1.243450e-14  18.891930  0.070175  0.683316

         RM        AGE       DIS        RAD         TAX    PTRATIO  \
0  6.448764  57.049552  4.710233   4.459701  302.480597  17.794030
1  5.963094  91.153801  2.002125  19.520468  615.421053  19.751462

            B      LSTAT
0  384.797612   9.519254
1  301.578129  18.792398

Differences between means and centroids:
     UNNAMED: 0          CRIM            ZN         INDUS          CHAS  \
0  0.000000e+00  1.110223e-15 -3.552714e-15 -2.664535e-15  8.326673e-17
1  5.684342e-14 -5.329071e-15 -1.243450e-14 -3.552714e-15  5.551115e-17

            NOX            RM           AGE  DIS           RAD           TAX  \
0  0.000000e+00 -8.881784e-16 -7.105427e-15  0.0  1.776357e-15 -5.684342e-14
1 -1.110223e-16  0.000000e+00 -1.421085e-14  0.0 -7.105427e-15 -6.821210e-13

        PTRATIO             B         LSTAT
0  0.000000e+00  5.684342e-14  0.000000e+00
1 -3.552714e-15  0.000000e+00 -3.552714e-15
```

Problem 3

Homogeneity measures whether each cluster contains samples from a single class. A value closer to 1 indicates that samples within each cluster belong to the same class, reflecting better clustering performance.Completeness measures whether all samples of a given class are assigned to the same cluster. A value closer to 1 indicates that samples of the same class are grouped together, reflecting better clustering performance.

In this dataset, both metrics approaching 1 suggest that the clustering results closely align with the true class labels.



```
Number of clusters: k=3
Homogeneity: 0.8788
Completeness: 0.8730
```