Question 1

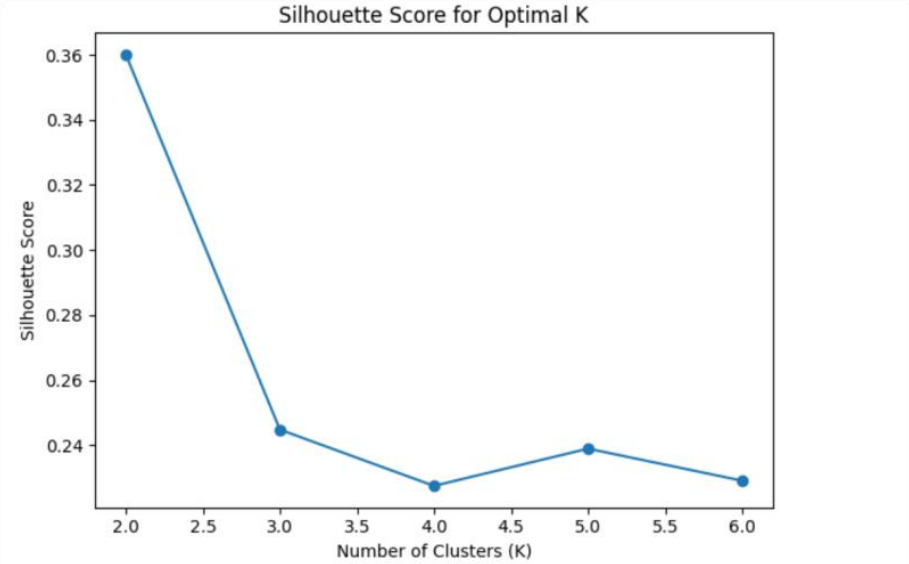
From the final output of the code in jupyter notebook(as follow), we can find that even through every sample in origin 3 is in Cluster 0, and the vast majority of samples in origin 2 are in Cluster 0, which means origin 2 and origin 3 have a strong correlation with Cluster 0. However, because Cluster 0 contains samples of origin 1, origin 2, and origin 3, and the samples in origin 1 are scattered, so there is no Clear relationship between cluster assignment and class label.

Hierarchical vs Origin:

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

Question 2

From the final output of the code in jupyter notebook(as follow), we can obviously find that when k=2 we will have the highest Silhouette Score, so 2 is the optimal value of k. The mean values for all features in each cluster and the centroid coordinates are identical. However, there are minor differences, which may be caused by floating-point precision issues



```
 the mean values for all features in each cluster:
              CRIM         ZN      INDUS      CHAS       NOX        RM  \
Cluster
0         0.388774  15.582656   8.420894  0.073171  0.511847  6.388005
1        12.299162   0.000000  18.451825  0.058394  0.670102  6.006212

               AGE       DIS        RAD         TAX    PTRATIO           B  \
Cluster
0        60.632249  4.441272   4.455285  311.926829  17.809214  381.042575
1        89.967883  2.054470  23.270073  667.642336  20.196350  291.039051

             LSTAT
Cluster
0        10.417453
1        18.674526

centroid coordinates:
        CRIM            ZN      INDUS      CHAS       NOX        RM  \
0   0.388774  1.558266e+01   8.420894  0.073171  0.511847  6.388005
1  12.299162  3.019807e-14  18.451825  0.058394  0.670102  6.006212

         AGE       DIS        RAD         TAX    PTRATIO           B  \
0  60.632249  4.441272   4.455285  311.926829  17.809214  381.042575
1  89.967883  2.054470  23.270073  667.642336  20.196350  291.039051

       LSTAT
0  10.417453
1  18.674526
```

Question 3

Homogeneity is used to determine whether the cluster contains only sample points of the same category, and Completeness is used to determine whether sample points of the same class are grouped into the same cluster. From the final output of the code in jupyter notebook(as follow),we can know that both Homogeneity and Completeness are close to 1, which means the clustering results are highly consistent with the real categories.