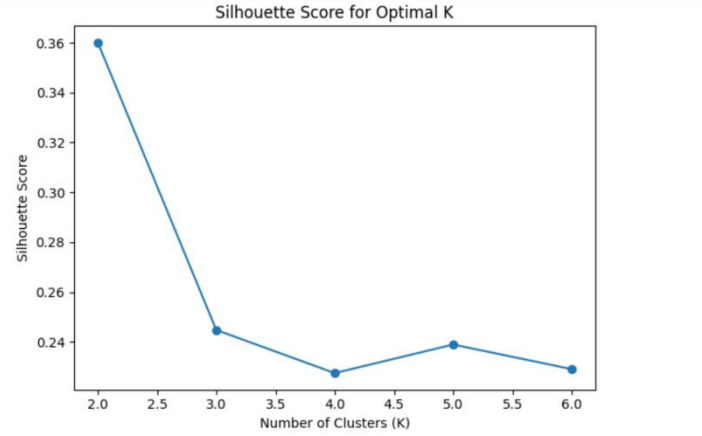
Question 1

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. Theoretically, the mean values for all features in each cluster and the centroid coordinates should be identical, but the mean values for all features in each cluster comes from the original data, while the centroid coordinates comes from scaled data, so they looks completely different.




```
 the mean values for all features in each cluster:
             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
```

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.