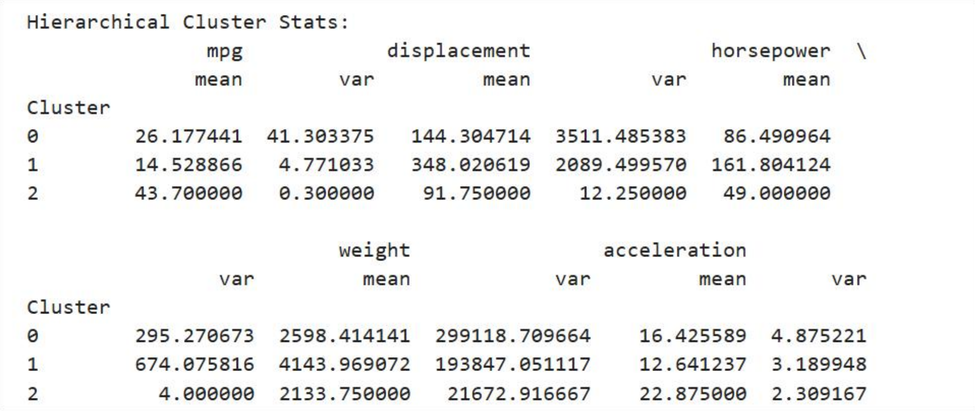
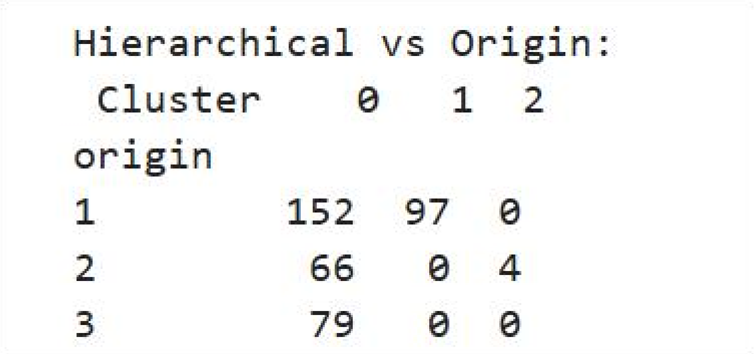
Question 1

From the final output of the code in jupyter notebook(as follow), we can find that every samples in Cluster 1 is comes from origin 1 and every samples in Cluster 2 comes from origin 2. So there is a clear relationship between Cluster 1 and origin 1, and a clear relationship between Cluster 2 and origin 2. But Cluster 0 contains samples from origin 1, origin 2, and origin 3, Cluster 0 is impure and has no clear relationship to class label.



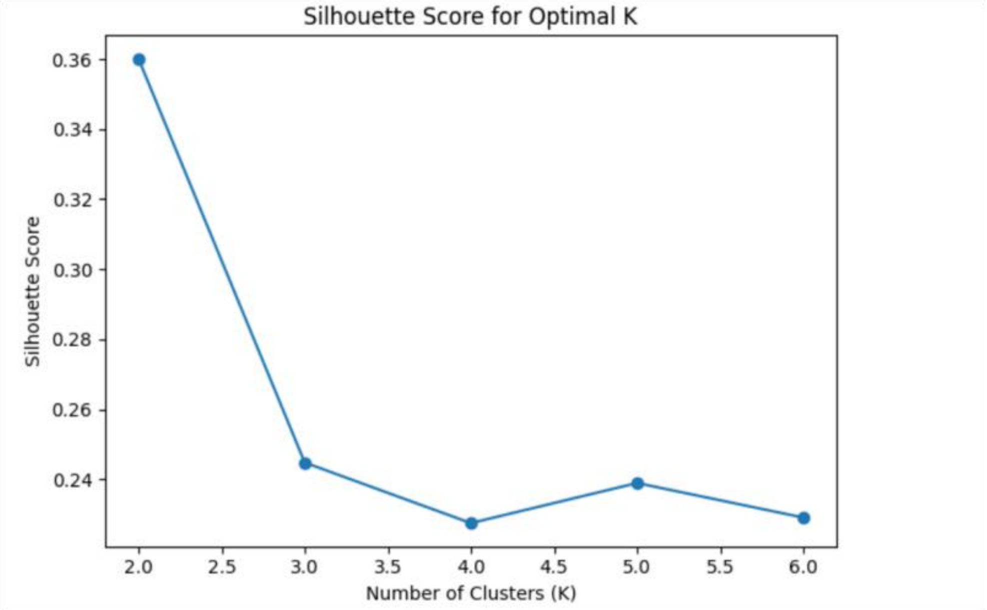
```
Hierarchical Cluster Stats:
               mpg           displacement               horsepower  \
              mean       var         mean         var         mean
Cluster
0        26.177441  41.303375  144.304714  3511.485383   86.490964
1        14.528866   4.771033  348.020619  2089.499570  161.804124
2        43.700000   0.300000   91.750000    12.250000   49.000000

                         weight               acceleration
                var        mean          var         mean       var
Cluster
0        295.270673  2598.414141  299118.709664    16.425589  4.875221
1        674.075816  4143.969072  193847.051117    12.641237  3.189948
2          4.000000  2133.750000   21672.916667    22.875000  2.309167
```

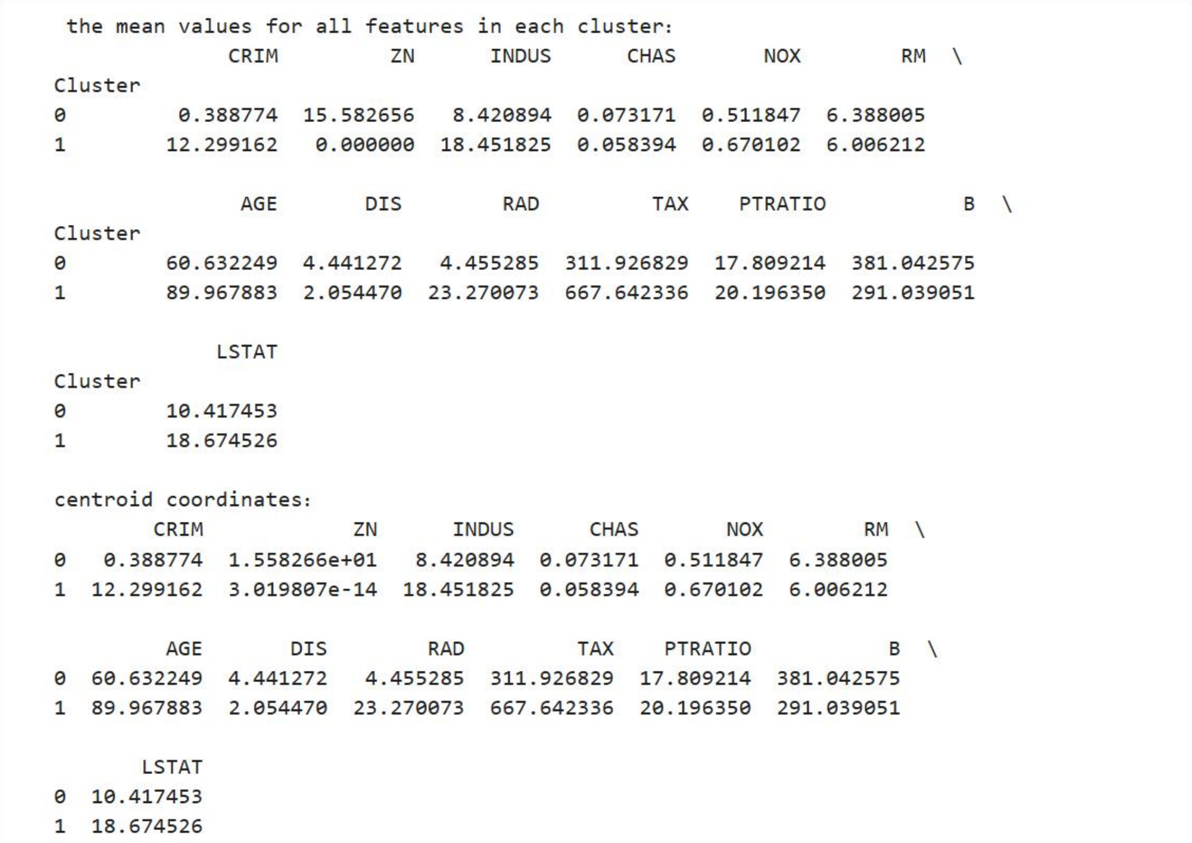


```
Hierarchical vs Origin:
 Cluster    0   1  2
origin
1         152  97  0
2          66   0  4
3          79   0  0
```

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 .

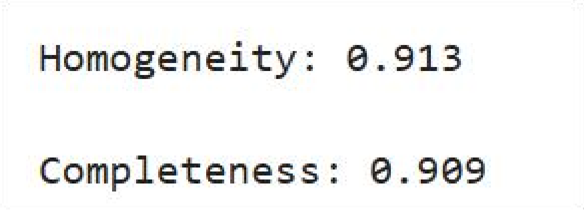




```
 the mean values for all features in each cluster:
               CRIM         ZN      INDUS      CHAS       NOX        RM  \
Cluster
0          0.388774  15.582656   8.420894  0.073171  0.511847  6.388005
1         12.299162   0.000000  18.451825  0.058394  0.670102  6.006212

                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.



```
Homogeneity: 0.913

Completeness: 0.909
```

END