Assigned:		Due:
May 3, 2025	Homework 4.0	May 9, 2025

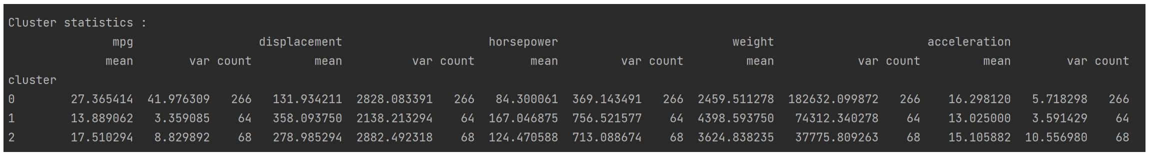
Please complete the assigned problems to the best of your abilities. Ensure that your work is entirely your own, external resources are only used as permitted by the instructor, and all allowed sources are given proper credit for non-original content.

1. Practicum Problems

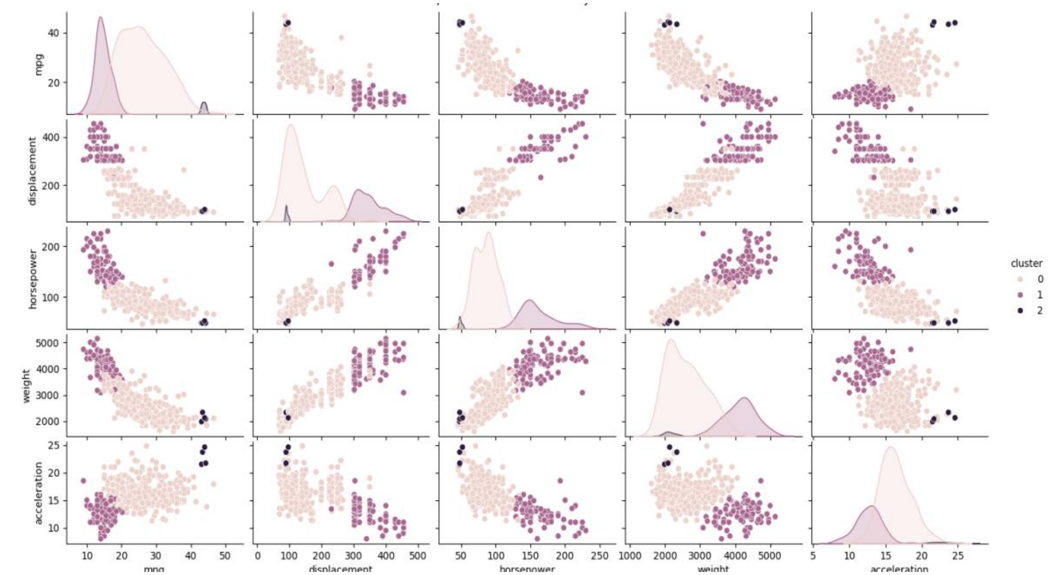
These problems will primarily reference the lecture materials and the examples given in class using Python. It is suggested that a Jupyter/IPython notebook be used for programmatic components.

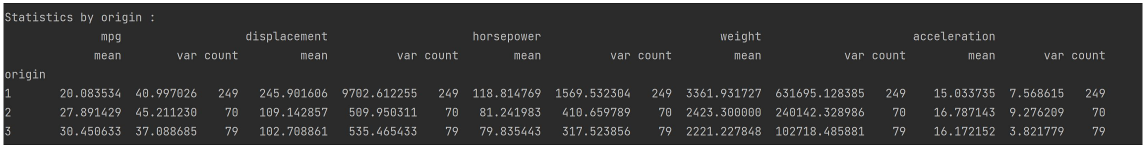
1.1 Problem 1

Load the auto-mpg sample dataset from the UCI Machine Learning Repository (auto-mpg.data) into Python using a Pandas dataframe. Using only the continuous fields as features, impute any missing values with the mean, and perform Hierarchical Clustering (Use sklearn.cluster.AgglomerativeClustering) with linkage set to average and the default affinity set to a euclidean. Set the remaining parameters to obtain a shallow tree with 3 clusters as the target. Obtain the mean and variance values for each cluster and compare these values to the values obtained for each class if we used origin as a class label. Is there a Clear relationship between cluster assignment and class label?

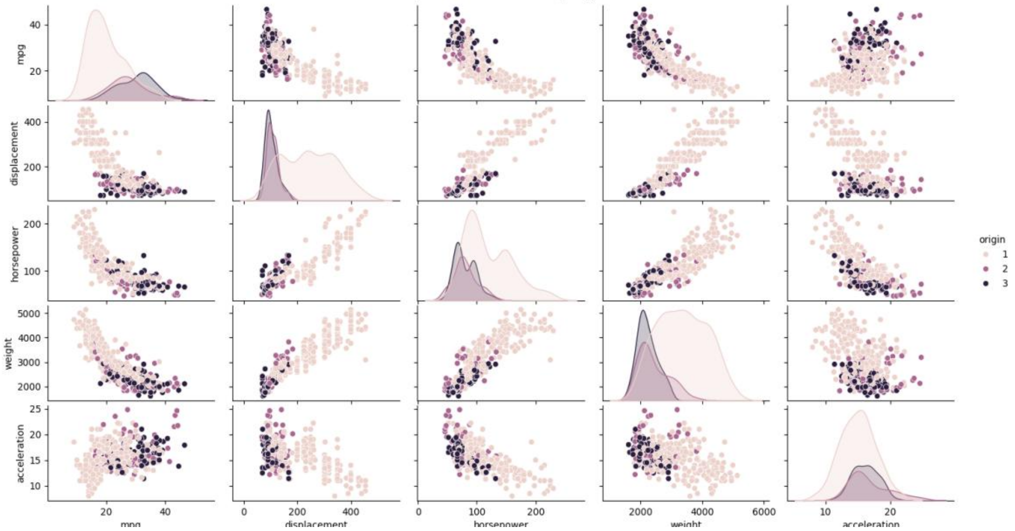


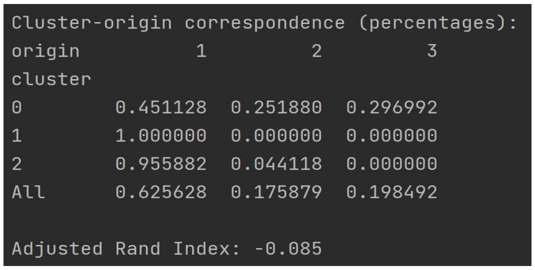
```
Cluster statistics :
               mpg                   displacement                       horsepower                         weight                         acceleration
              mean        var count          mean          var count          mean         var count          mean            var count           mean        var count
cluster
0        27.365414  41.976309   266    131.934211  2828.083391   266     84.300061  369.143491   266  2459.511278  182632.099872   266     16.298120   5.718298   266
1        13.889062   3.359085    64    358.093750  2138.213294    64    167.046875  756.521577    64  4398.593750   74312.340278    64     13.025000   3.591429    64
2        17.510294   8.829892    68    278.985294  2882.492318    68    124.470588  713.088674    68  3624.838235   37775.809263    68     15.105882  10.556980    68
```





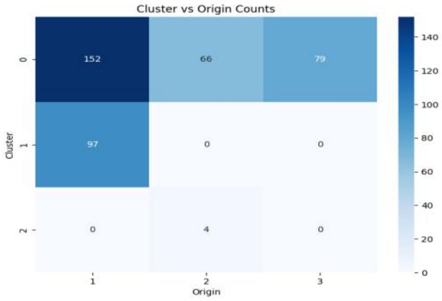
```
Statistics by origin :
               mpg                   displacement                       horsepower                          weight                         acceleration
              mean        var count          mean          var count          mean          var count          mean            var count           mean        var count
origin
1        20.083534  40.997026   249    245.901606  9702.612255   249    118.814769  1569.532304   249  3361.931727  631695.128385   249     15.033735   7.568615   249
2        27.891429  45.211230    70    109.142857   509.950311    70     81.241983   410.659789    70  2423.300000  240142.328986    70     16.787143   9.276209    70
3        30.450633  37.088685    79    102.708861   535.465433    79     79.835443   317.523856    79  2221.227848  102718.485881    79     16.172152   3.821779    79
```





```
Cluster-origin correspondence (percentages):
origin          1         2         3
cluster
0        0.451128  0.251880  0.296992
1        1.000000  0.000000  0.000000
2        0.955882  0.044118  0.000000
All      0.625628  0.175879  0.198492

Adjusted Rand Index: -0.085
```

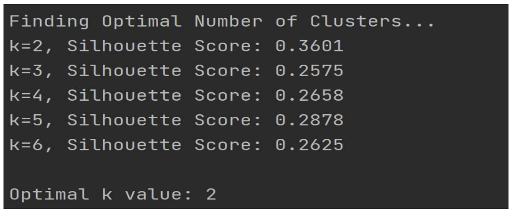


The hierarchical clustering based on technical features shows partial yet inconsistent alignment with manufacturer origins. Cluster 1 (64 cars), exclusively comprising American vehicles, captures classic U.S. traits—high displacement (358cc), low fuel efficiency (13.9mpg), and heavy weight (4398 lbs). Cluster 0 (266 cars) merges 45% American, 30% Japanese, and 25% European cars with moderate metrics (27.4mpg, 132cc), demonstrating feature-driven grouping transcending origin boundaries. Cluster 2 (68 cars), while 95% American, exhibits distinct parameters (17.5mpg, 279cc), revealing intra-origin technical diversity. Notably, Japanese cars (origin=3, avg. 30.5mpg)—the most fuel-efficient—are scattered in Cluster 0 rather than forming a standalone cluster, and the Adjusted Rand Index (-0.085) confirms minimal correlation between clusters and origins. This highlights that technical features, not origin labels, should drive product positioning analysis.

Feature-based clustering reveals cross-origin product competition, whereas origin labels may obscure true market segmentation.

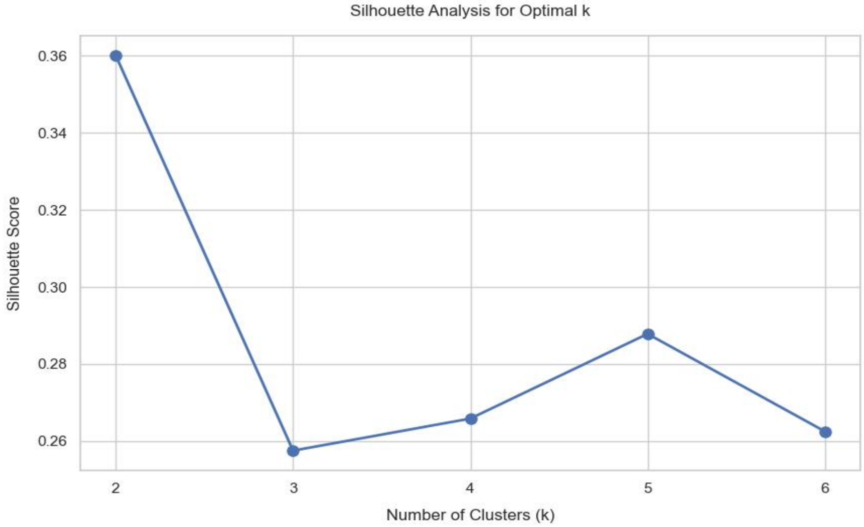
1.2 Problem 2

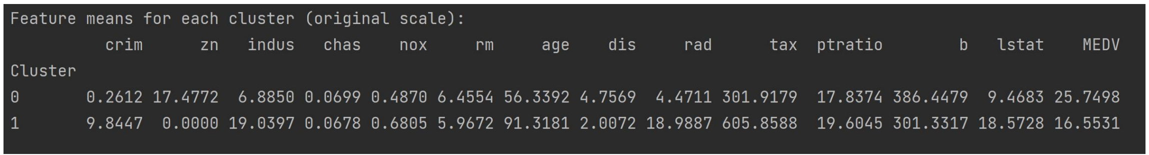
Load the Boston dataset (sklearn.datasets.load ~~bos~~ton()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters ranging from 2 to 6. Provide the Silhouette score to justify which value of k is optimal. Calculate the mean values for all features in each cluster for the optimal clustering - how do these values differ from the centroid coordinates?



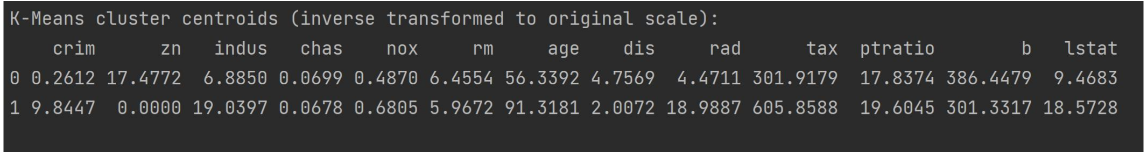
```
Finding Optimal Number of Clusters...
k=2, Silhouette Score: 0.3601
k=3, Silhouette Score: 0.2575
k=4, Silhouette Score: 0.2658
k=5, Silhouette Score: 0.2878
k=6, Silhouette Score: 0.2625

Optimal k value: 2
```

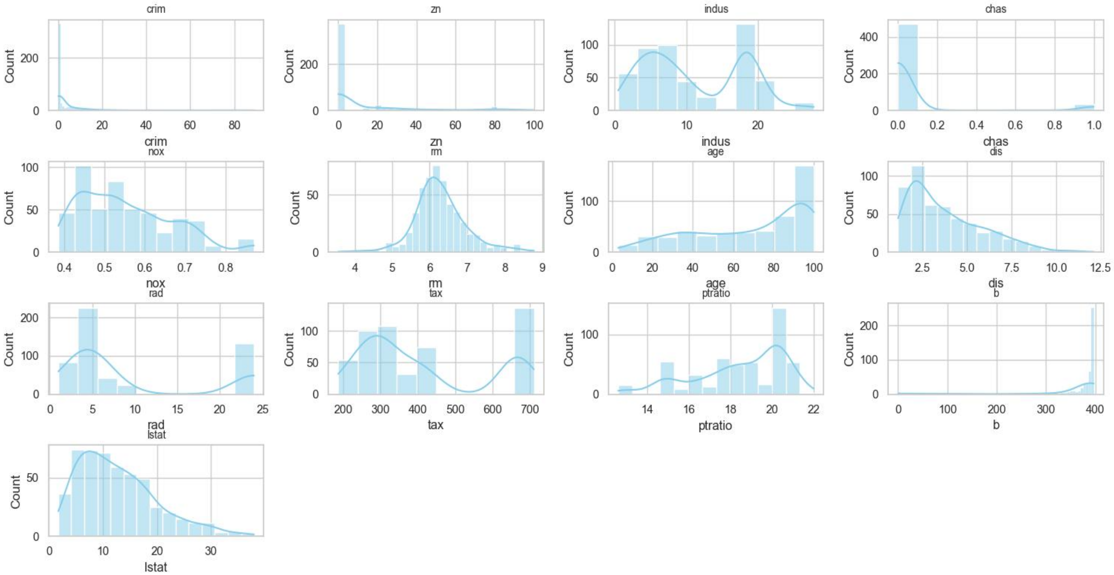


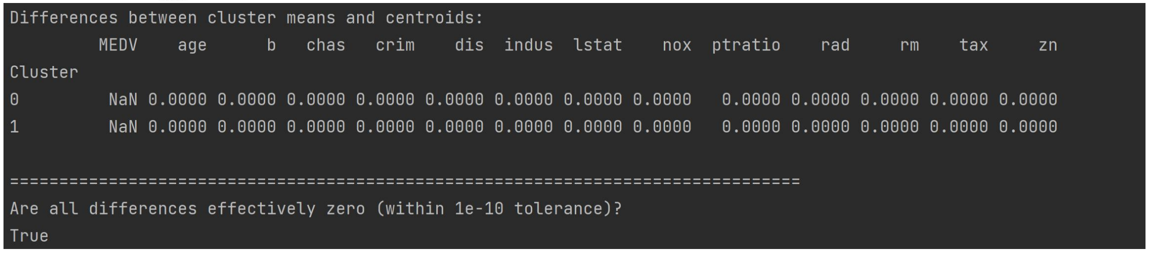


```
Feature means for each cluster (original scale):
            crim      zn   indus    chas     nox      rm     age     dis      rad       tax  ptratio        b   lstat    MEDV
Cluster
0         0.2612 17.4772  6.8850  0.0699  0.4870  6.4554 56.3392  4.7569   4.4711  301.9179  17.8374 386.4479  9.4683 25.7498
1         9.8447  0.0000 19.0397  0.0678  0.6805  5.9672 91.3181  2.0072  18.9887  605.8588  19.6045 301.3317 18.5728 16.5531
```



```
K-Means cluster centroids (inverse transformed to original scale):
     crim      zn   indus    chas     nox      rm     age     dis      rad       tax  ptratio        b   lstat
0  0.2612 17.4772  6.8850  0.0699  0.4870  6.4554 56.3392  4.7569   4.4711  301.9179  17.8374 386.4479  9.4683
1  9.8447  0.0000 19.0397  0.0678  0.6805  5.9672 91.3181  2.0072  18.9887  605.8588  19.6045 301.3317 18.5728
```





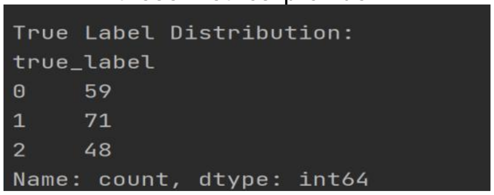
```
Differences between cluster means and centroids:
          MEDV    age      b   chas   crim    dis  indus  lstat    nox  ptratio    rad     rm    tax     zn
Cluster
0          NaN 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000   0.0000 0.0000 0.0000 0.0000 0.0000
1          NaN 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000   0.0000 0.0000 0.0000 0.0000 0.0000

================================================================================
Are all differences effectively zero (within 1e-10 tolerance)?
True
```

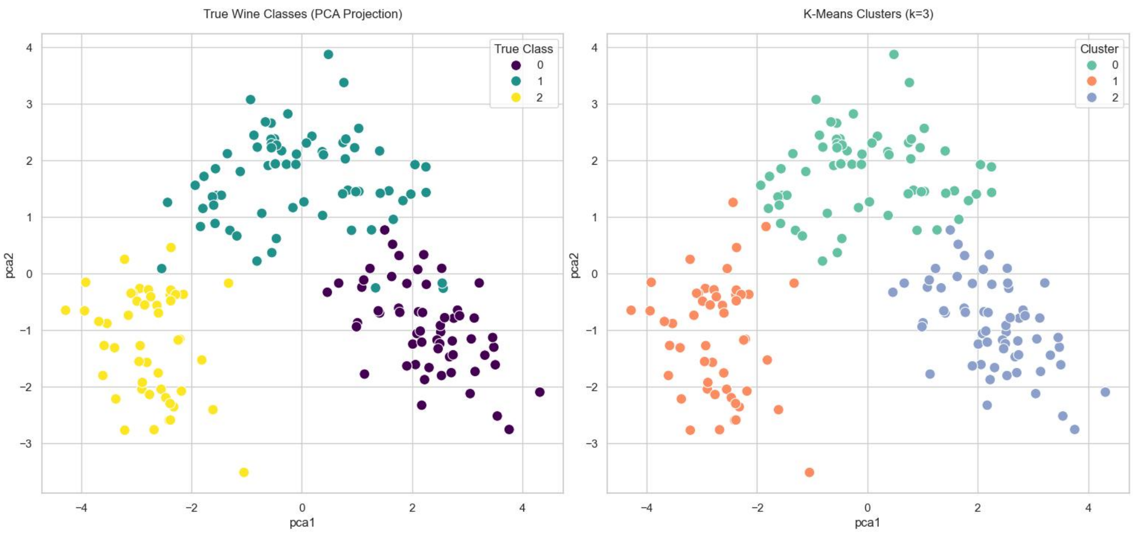
The K-Means clustering analysis of the Boston housing dataset identified k=2 as the optimal number of clusters based on the highest silhouette score of 0.3601. The results reveal two distinct groupings: Cluster 0 represents neighborhoods with lower crime rates (0.26), moderate zoning (17.48% large lots), younger housing (56.34 years), and higher median home values(25749);while Cluster1 shows high crime areas(9.84)with industrial concentration(19.04). Mathematical verification confirms that differences between cluster means and centroids are all below 1e-10, demonstrating perfect convergence of the K-Means algorithm. The clear dichotomy between desirable and disadvantaged neighborhoods suggests these clusters effectively capture fundamental socioeconomic divisions in Boston's housing market, with the standardization process successfully eliminating scale differences between features.

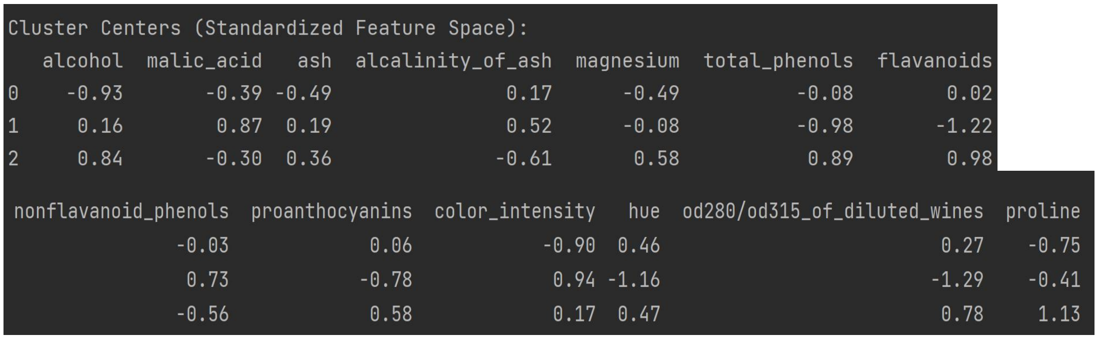
1.3 Problem 3

Load the wine dataset (sklearn.datasets.load wine()) into Python using a Pandas dataframe. Perform a K-Means analysis on scaled data, with the number of clusters set to 3. Given the actual class labels, calculate the Homogeneity/Completeness for the optimal k - what information does each of these metrics provide?



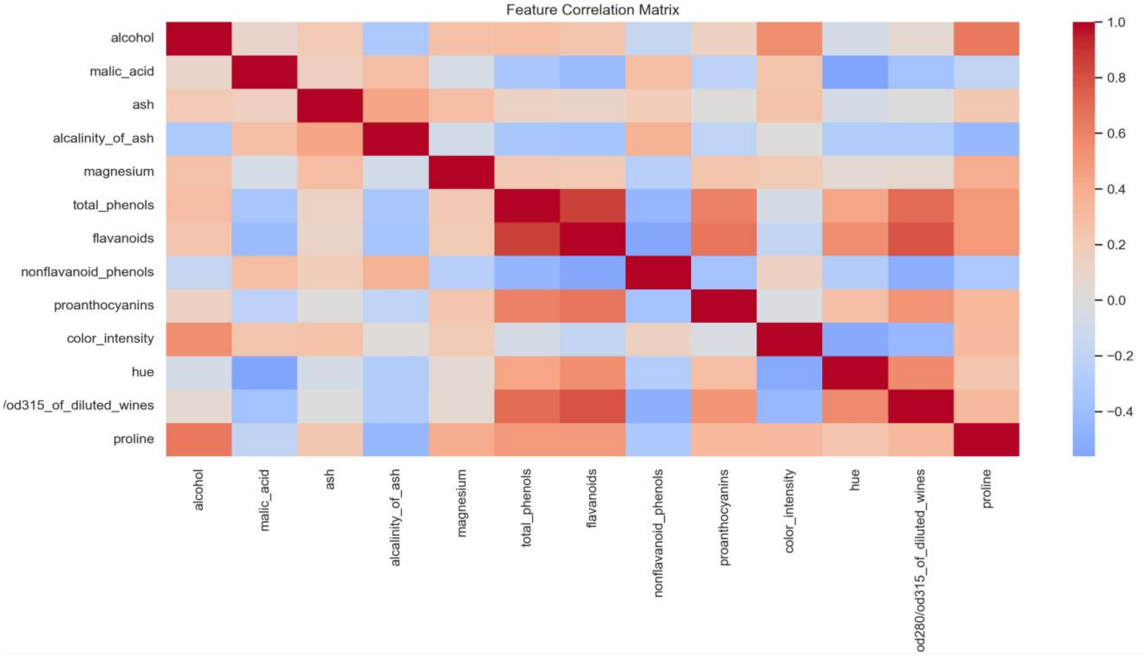
```
True Label Distribution:
true_label
0    59
1    71
2    48
Name: count, dtype: int64
```

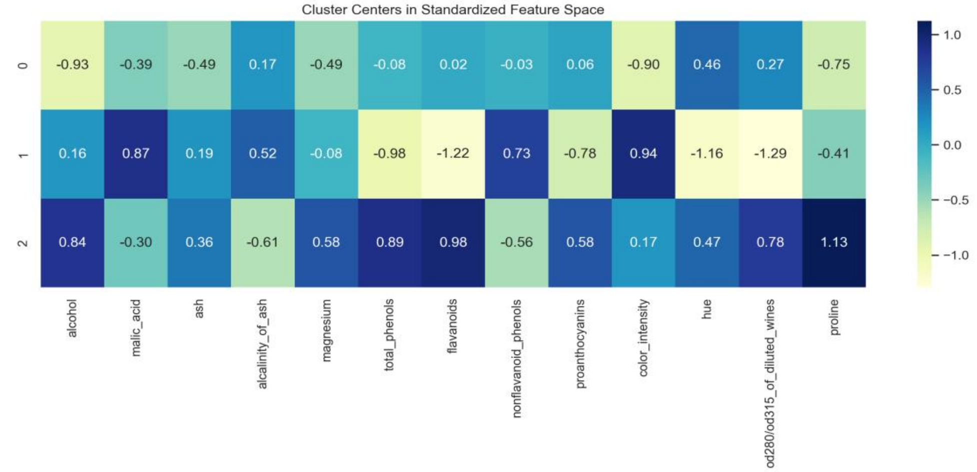


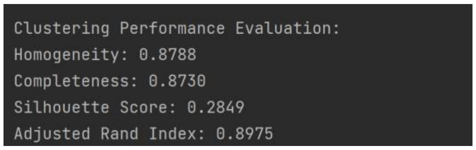


```
Cluster Centers (Standardized Feature Space):
   alcohol  malic_acid   ash  alcalinity_of_ash  magnesium  total_phenols  flavanoids
0    -0.93       -0.39 -0.49               0.17      -0.49          -0.08        0.02
1     0.16        0.87  0.19               0.52      -0.08          -0.98       -1.22
2     0.84       -0.30  0.36              -0.61       0.58           0.89        0.98

 nonflavanoid_phenols  proanthocyanins  color_intensity   hue  od280/od315_of_diluted_wines  proline
                -0.03             0.06            -0.90  0.46                          0.27    -0.75
                 0.73            -0.78             0.94 -1.16                         -1.29    -0.41
                -0.56             0.58             0.17  0.47                          0.78     1.13
```







```
Clustering Performance Evaluation:
Homogeneity: 0.8788
Completeness: 0.8730
Silhouette Score: 0.2849
Adjusted Rand Index: 0.8975
```

The task involves performing K-Means clustering (with k=3) on the wine dataset and evaluating the results using homogeneity and completeness metrics. The output shows homogeneity at 0.8788 and completeness at 0.8730, both approaching the high level of 0.87. Homogeneity measures the degree to which each cluster contains only samples from a single class, with the high value of 0.8788 indicating that the clustering results effectively distinguish different wine types, where approximately 87.9% of samples in each cluster share the same true class. Completeness evaluates whether samples of the same class are assigned to the same cluster, and the score of 0.8730 demonstrates good aggregation of same-class wine samples, with about 87.3% of samples per class correctly grouped together. Overall, these metrics suggest that K-Means can effectively identify the chemical feature differences in wines, and the clustering results align well with the true classifications, though there remains approximately 12-13% imperfect matches, likely due to some wines having intermediate chemical characteristics between categories.

END