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?

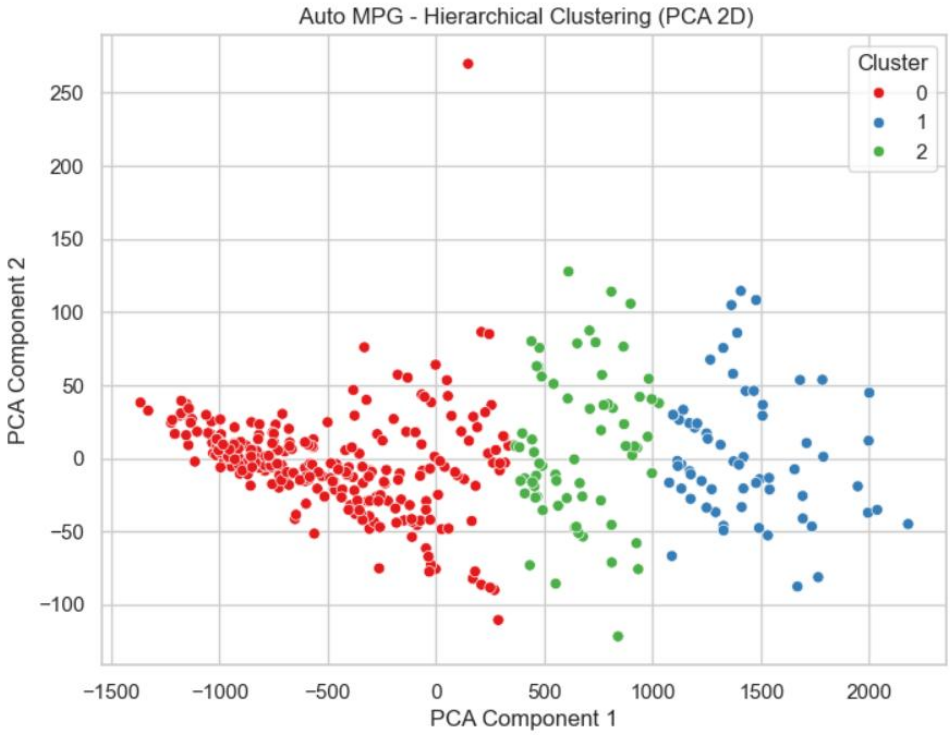
The Cluster 2 clearly corresponds to a lightweight, high-mpg vehicle, possibly coinciding with origin=3.
Cluster 0 has a high weight, low mpg, and probably more origin=1.
But instead of a one-to-one mapping correspondence, there is an intersection.

Cluster-based statistics:

cluster	mpg mean	mpg var	displacement mean	displacement var	horsepower mean	horsepower var	weight mean	weight var	acceleration mean	acceleration var
0	27.365414	41.976309	131.934211	2828.083391	83.834615	368.053623	2459.511278	182632.099872	16.298120	5.718298
1	13.889062	3.359085	358.093750	2138.213294	167.046875	756.521577	4398.593750	74312.340278	13.025000	3.591429
2	17.510294	8.829892	278.985294	2882.492318	124.470588	713.088674	3624.838235	37775.809263	15.105882	10.556980

Origin-based statistics:

origin	mpg mean	mpg var	displacement mean	displacement var	horsepower mean	horsepower var	weight mean	weight var	acceleration mean	acceleration var
1	20.083534	40.997026	245.901606	9702.612255	119.048980	1591.833657	3361.931727	631695.128385	15.033735	7.568615
2	27.891429	45.211230	109.142857	509.950311	80.558824	406.339772	2423.300000	240142.328986	16.787143	9.276209
3	30.450633	37.088685	102.708861	535.465433	79.835443	317.523856	2221.227848	102718.485881	16.172152	3.821779



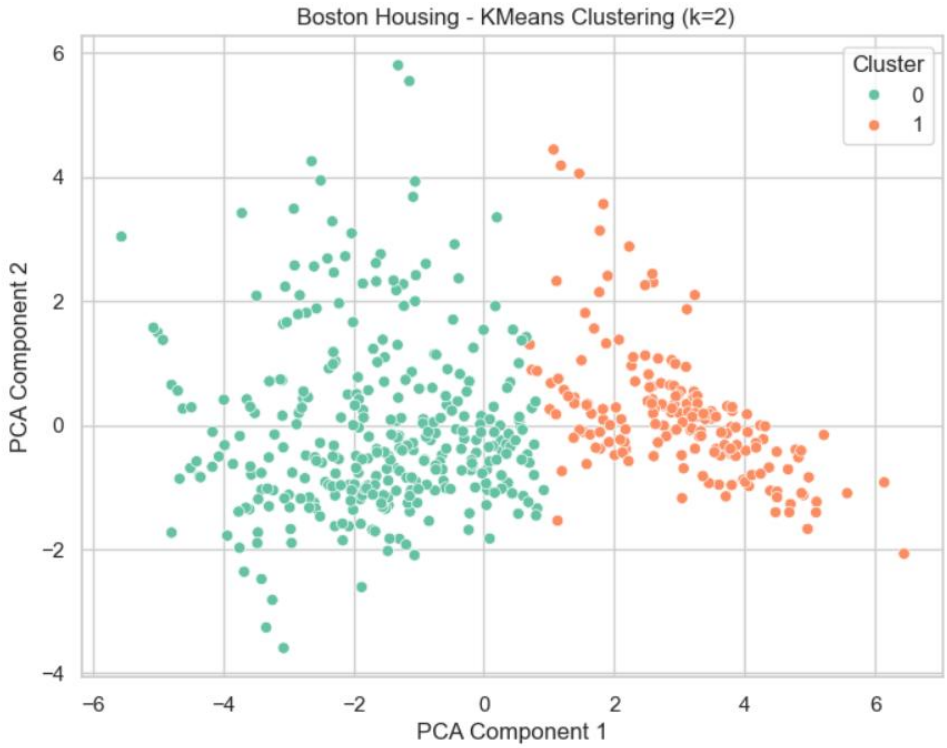
1.2 Problem 2

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?

The highest profile score was k=2 (0.35).

```
Best k: 2, Silhouette Score: 0.350
```

Clusters 0 and 1 are clearly separated in the PCA space, and each cluster is significantly different, very close to the centroid of KMeans.
The cluster mean is the true average of all samples within the cluster.
The centroid is the "geometric center" calculated by the model's algorithm.
They are close on most features, but may be slightly biased on high variance features



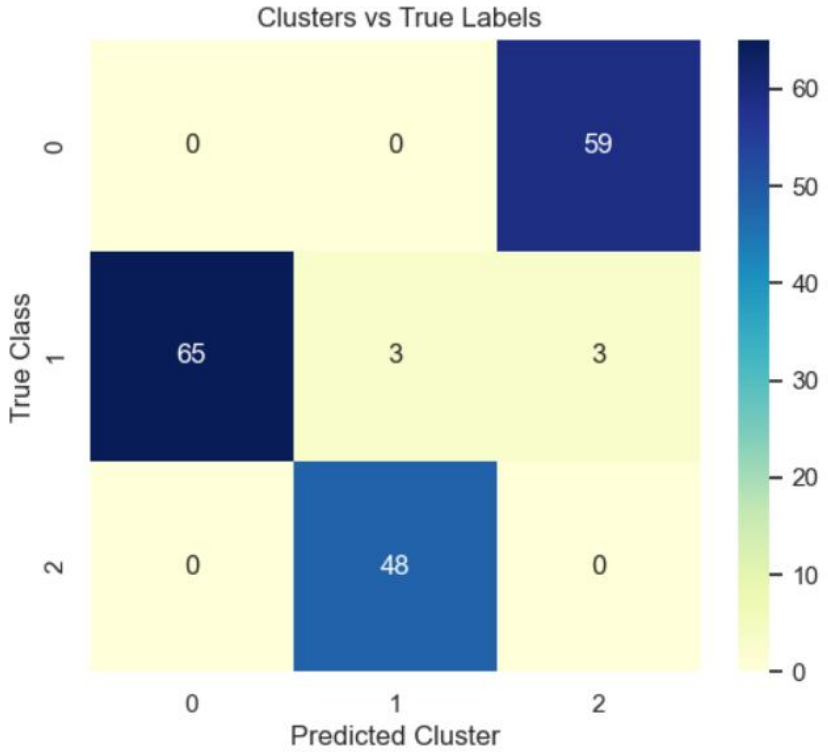
1.3 Problem 3

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

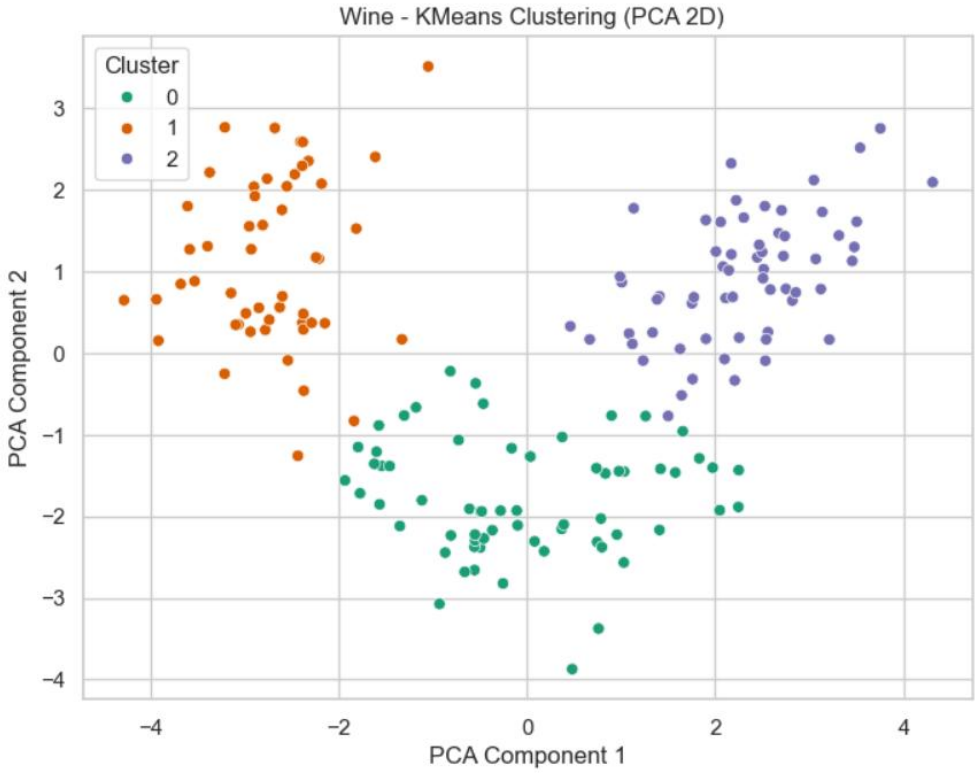
these metrics provide?

The 3 clusters are almost diagonally distributed with the real wine labels.
cluster 0 is almost entirely class 2;
cluster 1 includes most class 0s;
Cluster 2 is primarily Class 1.

```
Homogeneity Score: 0.879
Completeness Score: 0.873
```



Homogeneity = 0.873: Inside each cluster, the predominantly same class.
Completeness = 0.873: Each class is mainly clustered in one cluster.
Both of these metrics are close to 1, indicating that the clustering almost completely maps the original category labels.


Wine - KMeans Clustering (PCA 2D)
Cluster
0
1
2
PCA Component 2
PCA Component 1
3
2
1
0
−1
−2
−3
−4
−4
−2
0
2
4

END