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?

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?

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?

1.1 Problem 1

There is a clear relationship between cluster assignments and origin labels:
Cluster 0 captures almost all European and Japanese cars, which are generally small, light, and fuel-efficient.
Cluster 1 almost exclusively includes powerful American cars, aligning well with origin = 1.
Cluster 2 represents mid-range American cars.
Thus, hierarchical clustering successfully reflects the underlying class structure, especially in identifying American muscle cars versus compact foreign cars.

1.2 Problem 2

We loaded the Boston dataset from the official website and performed K-Means clustering, finding that k=2 gave the highest silhouette score, indicating the best clustering. By comparing the feature mean values and centroid coordinates, we found that the feature means represent the actual values, while the centroids are in the standardized space, with the difference mainly due to the standardization.

1.3 Problem 3

Homogeneity: 0.8788: This score indicates that the clustering is relatively good in terms of purity. About 87.88% of the samples in each cluster belong to the same true class. However, there's still some mix between classes within the clusters.

Completeness: 0.8730: This score indicates that the clustering has done a decent job of grouping all samples of the same true class into one cluster. 87.30% of the samples of each class are assigned to the same cluster. There's some room for improvement, as not all samples of a class are perfectly grouped.