Problem 1

In this experiment, we used the Auto MPG dataset and selected several continuous features (such as mpg, displacement, horsepower, weight, and acceleration) for analysis. After standardizing these features, we applied hierarchical clustering (Agglomerative Clustering) to divide the data into 3 clusters. Then, we compared the clustering results with the origin label (which represents the car’s origin: 1 for the USA, 2 for Europe, and 3 for Japan).

From the clustering results, cluster 1 is mainly composed of cars from origin=1 (USA), meaning that American cars have distinct characteristics compared to cars from other countries, which allows the clustering to differentiate them well. Cluster 0 and cluster 2 mainly contain cars from origin=2 (Europe) and origin=3 (Japan). These two origins have somewhat similar features, which makes it difficult for the clustering to clearly separate them.

Through the cross-tabulation and heatmap, we can also see the relationship between the clusters and the original class labels. Overall, there is some correlation between the clustering results and the origin labels, especially for cars from the USA, where the feature differences are more prominent, allowing the clustering to distinguish them well. Although there is some overlap between cars from Europe and Japan, we can still see some common traits between them.

Problem 2

Based on the K-Means clustering analysis results, the most suitable number of clusters is 2 because the Silhouette score is highest at k = 2, reaching 0.36, which indicates a good clustering result. For the optimal clustering, the average values of each feature in both clusters were calculated, and there are noticeable differences between the first and second clusters. These means reflect the overall characteristics of the data within each cluster. The centroid coordinates, however, are the cluster centers computed by the K-Means algorithm, and they are positioned in the standardized data space, so they differ from the actual feature means. The values of the centroids are typically smaller because the data has been standardized, and their directions may also differ from the clustering means. This is because the centroids are determined by distance calculations, while the clustering means are simply averages.

Problem3

Given the actual class labels, homogeneity and completeness at the optimal K value are used to evaluate the clustering performance. Homogeneity checks whether most of the samples in each cluster belong to the same actual class; the higher the value, the more similar the samples within the cluster are. Completeness checks whether all samples of a given actual class are assigned to the same cluster; the higher the value, the more the clustering result reflects the true classes. By trying different K values and calculating homogeneity and completeness, we can find the most suitable K value and thus achieve the best clustering results.