1 Problem 1: Hierarchical Clustering of Auto-MPG Dataset

In this analysis, we apply hierarchical clustering to the Auto-MPG dataset, focusing on continuous variables such as miles per gallon (mpg), displacement, horsepower, weight, and acceleration. The objective is to identify whether there is a clear relationship between the clusters formed through hierarchical clustering and the origin classes provided in the dataset. We use the average linkage method with Euclidean distance as the metric to perform the clustering.

1.1 Data Preprocessing

The dataset includes several continuous features: mpg, displacement, horsepower, weight, and acceleration. First, we handle missing data by imputing the missing values in each feature with the mean of that feature. The continuous features are then standardized using the StandardScaler to ensure each feature contributes equally to the clustering process.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of each feature. This normalization ensures that all features are on the same scale before clustering.

1.2 Hierarchical Clustering

Hierarchical clustering is performed with the following parameters:

- **Number of clusters:** 3
- **Linkage method:** Average
- **Metric:** Euclidean

The AgglomerativeClustering method from scikit-learn is used to compute the hierarchical clusters. After performing the clustering, each data point is assigned to one of the three clusters. These cluster assignments are then added to the dataframe for further analysis.

1.3 Cluster Statistics (Mean and Variance)

For each cluster, we compute the mean and variance of the features using the original data (non-standardized). The results for the mean values of the features within each cluster are displayed in the table below:

Cluster	mpg (mean)	displacement (mean)	horsepower (mean)	weight (mean)	acceleration (mean)
0	26.1774	144.3047	86.4909	2598.4141	16.4256
1	14.5289	348.0206	161.8041	4143.9691	12.6412
2	43.7000	91.7500	49.0000	2133.7500	22.8750

Table 1: Cluster Mean Values for Features (Original Data)

Cluster	mpg (var)	displacement (var)	horsepower (var)	weight (var)	acceleration (var)
0	41.3034	3511.4854	86.4909	295.2707	4.8752
1	4.7710	2089.4996	161.8041	674.0758	3.1900
2	0.3000	12.2500	49.0000	4.0000	2.3092

Table 2: Cluster Variance Values for Features (Original Data)

These tables show the **mean** and **variance** of features for each cluster using the **original data** (non-standardized data).

Next, the mean and variance for each cluster using **standardized data** are displayed in the tables below:

Cluster	mpg (mean)	displacement (mean)	horsepower (mean)	weight (mean)	acceleration (mean)
0	−0.4717	−0.4712	−0.4398	0.3113	0.3411
1	1.4845	1.5028	1.3875	−1.0627	−1.1511
2	−0.9764	−1.4539	−0.9892	2.6530	2.5858

Table 3: Cluster Mean Values for Features (Standardized Data)

Cluster	mpg (var)	displacement (var)	horsepower (var)	weight (var)	acceleration (var)
0	0.3238	0.2029	0.4182	0.6427	0.6778
1	0.1927	0.4631	0.2709	0.4205	0.0783
2	0.0011	0.0027	0.0303	0.3044	0.0049

Table 4: Cluster Variance Values for Features (Standardized Data)

These tables show the **mean** and **variance** of features for each cluster using the **standardized data** (data scaled to have a mean of 0 and a standard deviation of 1).

1.4 Origin Class Statistics (Mean and Variance)

Next, the mean and variance for each of the origin classes (1, 2, 3) are computed for the features using the **original data** (non-standardized). The table below shows the mean values of the features for each origin class:

Origin	mpg (mean)	displacement (mean)	horsepower (mean)	weight (mean)
1	20.0835	245.9016	118.8148	3361.9317
2	27.8914	109.1429	81.2420	2423.3000
3	30.4506	102.7089	79.8354	2221.2278

Table 5: Origin Class Mean Values for Features (Original Data)

Origin	mpg (var)	displacement (var)	horsepower (var)	weight (var)
1	40.9970	9702.6123	118.8148	1569.5323
2	45.2112	509.9503	81.2420	410.6598
3	37.0887	535.4654	79.8354	317.5239

Table 6: Origin Class Variance Values for Features (Original Data)

These tables show the **mean** and **variance** of features for each origin class using the **original data** (non-standardized data). The variance values reflect the spread of data points in each feature across the origin classes.

Next, the mean and variance for each origin class using the **standardized data** are displayed in the tables below:

Origin	mpg (mean)	displacement (mean)	horsepower (mean)	weight (mean)
1	0.5039	0.3760	0.4629	−0.1940
2	−0.8093	−0.6088	−0.6469	0.4426
3	−0.8711	−0.6457	−0.8858	0.2193

Table 7: Origin Class Mean Values for Features (Standardized Data)

Origin	mpg (var)	displacement (var)	horsepower (var)	weight (var)
1	0.8947	1.0783	0.4629	0.9977
2	0.0470	0.2821	0.3357	1.2228
3	0.0494	0.2182	0.1436	0.5038

Table 8: Origin Class Variance Values for Features (Standardized Data)

These tables show the **mean** and **variance** of features for each origin class using the **standardized data** (data scaled to have a mean of 0 and a standard deviation of 1).

1.5 Cluster vs. Origin Crosstab

A crosstab of cluster assignments versus origin labels is computed to observe if the clusters correspond with the origin classes. The following table presents the distribution of origin classes across the clusters:

Cluster	Origin 1	Origin 2	Origin 3
0	152	66	79
1	97	0	0
2	0	4	0

Table 9: Cluster vs. Origin Crosstab

From this crosstab, it is observed that Cluster 1 is predominantly composed of data points from origin 1, indicating a stronger relationship between this cluster and origin 1. Cluster 2, on the other hand, has a minor presence from origin 2.

1.6 Visualizing Clustering and Origin Labels

The results of the hierarchical clustering and origin labels are visualized using pair plots. These plots allow for a comparison between the clustering results and the origin classes. The pair plots display the relationships between the features and provide insight into how the clusters are distributed relative to the origin labels.

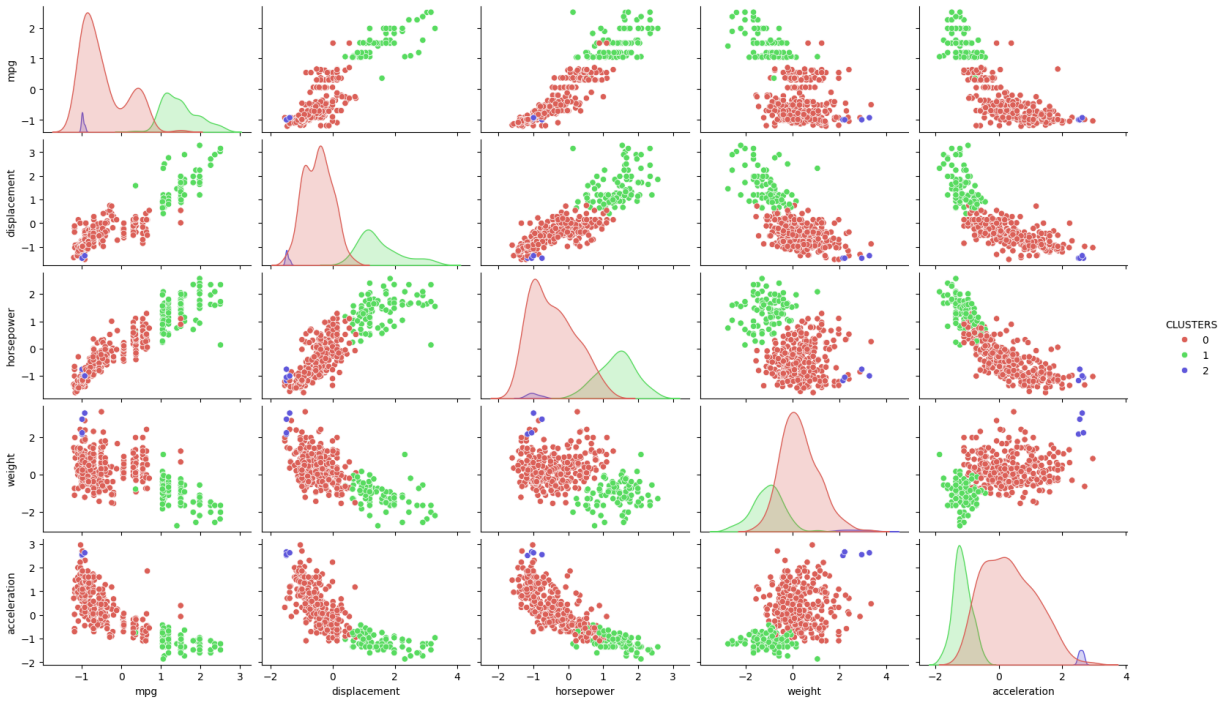

Figure 1: Clustering Results

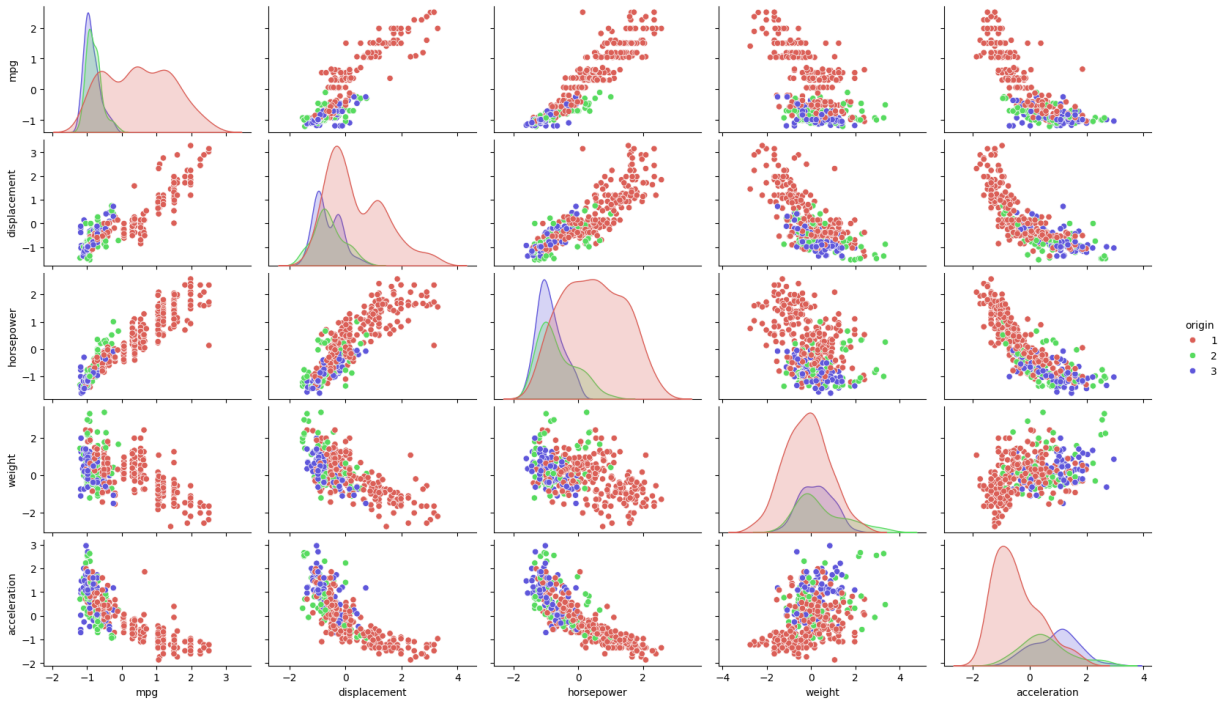

Figure 2: Origin Class Labels

1.7 Discussion

From the crosstab and visualizations, it appears that Cluster 0 contains a mix of origin 1, 2, and 3, with no clear dominant origin. However, Cluster 1 is primarily composed of data points from origin 1, indicating a

strong relationship between Cluster 1 and origin 1. Cluster 2 shows limited overlap with origin 2, suggesting that Cluster 2 is not as strongly related to origin 2 as Cluster 1 is to origin 1.

A clear relationship exists when clusters predominantly contain data points from a single origin class. However, when clusters contain a mix of origin classes, the relationship between clusters and origin classes becomes less clear.

1.8 Conclusion

Hierarchical clustering applied to the Auto-MPG dataset reveals some relationship between the clusters and the origin labels, though the relationship is not perfectly clear across all clusters. In particular, Cluster 1 is strongly associated with origin 1, while Cluster 2 contains fewer data points from origin 2. This suggests that while hierarchical clustering can capture some of the structure related to origin classes, the relationship is not always straightforward.

2 Problem 2: K-Means Clustering Analysis on the Boston Dataset

This analysis applies K-Means clustering to the Boston housing dataset, classifying the data based on features such as crime rate, average number of rooms, property tax rate, and others. The number of clusters k is tested from 2 to 6, with the optimal k selected based on the highest Silhouette score. We compare the means of the features in each cluster with their respective centroids to assess clustering performance.

2.1 Dataset Description

The dataset consists of 506 neighborhoods in Boston, each represented by 13 features. Key features include crime rate, average number of rooms, and property tax rate. The target variable is the median value of owner-occupied homes. The features provide insights into the socioeconomic and environmental characteristics of the neighborhoods.

2.2 Data Preprocessing

The dataset is standardized using the StandardScaler, ensuring that each feature has a mean of 0 and a standard deviation of 1. This step is crucial for K-Means clustering because the algorithm is sensitive to the scale of the features.

$$X_{\text{scaled}} = \frac{X - \mu}{\sigma}$$

where μ is the mean and σ is the standard deviation of each feature.

2.3 K-Means Clustering

K-Means clustering is performed for different values of k, ranging from 2 to 6. The optimal k is determined by selecting the value with the highest Silhouette score, a measure of how well-separated the clusters are.

2.4 Silhouette Score Calculation

The Silhouette score is calculated for $k = 2$ to $k = 6$, with each k tested using $n_{\text{init}} = 10$ for improved stability of the clustering results. The table below shows the Silhouette scores for each k, with the highest Silhouette score achieved for $k = 5$.

Number of Clusters (k)	**Silhouette Score**
2	0.3601
3	0.2575
4	0.2658
5	0.2878
6	0.2625

Table 10: Silhouette Scores for Different k Values (with $n_{\text{init}} = 10$)

2.5 Cluster Means and Centroids

For $k = 2$, K-Means clustering is performed, and the mean values of features for each cluster are computed. These are compared with the centroids of the clusters:

Feature	Cluster 0 Mean	Cluster 1 Mean
CRIM	0.2612	9.8447
ZN	17.4772	0.0000
INDUS	6.8850	19.0397
CHAS	0.0699	0.0678
NOX	0.4870	0.6805
RM	6.4554	5.9672
AGE	56.3392	91.3181
DIS	4.7569	2.0072
RAD	4.4711	18.9887
TAX	301.9179	605.8588
PTRATIO	17.8374	19.6045
B	386.4479	301.3317
LSTAT	9.4683	18.5728

The centroids for each cluster are:

Feature	Centroid 0	Centroid 1
CRIM	−0.3901	0.7251
ZN	0.2624	−0.4877
INDUS	−0.6204	1.1531
CHAS	0.0029	−0.0054
NOX	−0.5847	1.0868
RM	0.2433	−0.4523
AGE	−0.4351	0.8088
DIS	0.4572	−0.8499
RAD	−0.5838	1.0851
TAX	−0.6315	1.1737
PTRATIO	−0.2858	0.5312
B	0.3265	−0.6068
LSTAT	−0.4464	0.8298

2.6 Cluster Means vs. Centroids

The difference between the cluster means and centroids is calculated as follows:

Feature	Difference for Cluster 0	Difference for Cluster 1
CRIM	5.551115×10^{-17}	$-2.220446 \times 10^{-16}$
ZN	$-1.110223 \times 10^{-16}$	$-4.996004 \times 10^{-16}$
INDUS	$-3.330669 \times 10^{-16}$	$-1.110223 \times 10^{-15}$
CHAS	3.339343×10^{-16}	2.437286×10^{-16}
NOX	$-1.110223 \times 10^{-16}$	$-6.661338 \times 10^{-16}$
RM	5.551115×10^{-17}	2.220446×10^{-16}
AGE	$-5.551115 \times 10^{-17}$	$-3.330669 \times 10^{-16}$
DIS	$-3.885781 \times 10^{-16}$	1.110223×10^{-16}
RAD	3.330669×10^{-16}	$-1.110223 \times 10^{-15}$
TAX	$-1.110223 \times 10^{-16}$	$-3.996803 \times 10^{-15}$
PTRATIO	0.000000	$-1.443290 \times 10^{-15}$
B	2.775558×10^{-16}	1.110223×10^{-16}
LSTAT	0.000000	$-3.330669 \times 10^{-16}$

Table 11: Difference Between Cluster Means and Centroids (Original Scale)

2.7 Visualization

2.7.1 Silhouette Scores

The silhouette score plot shows how clustering quality changes with k. The highest score is achieved for $k = 2$, confirming that two clusters best represent the data.

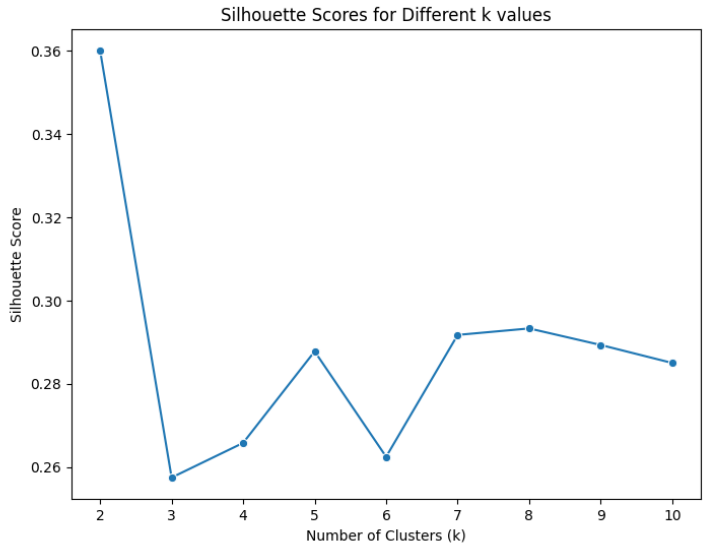

Figure 3: Silhouette Scores for Different k Values

2.7.2 Clusters and Centroids

The scatter plot below shows the clusters based on the features *CRIM* and *ZN*, with red crosses indicating the centroids of each cluster.

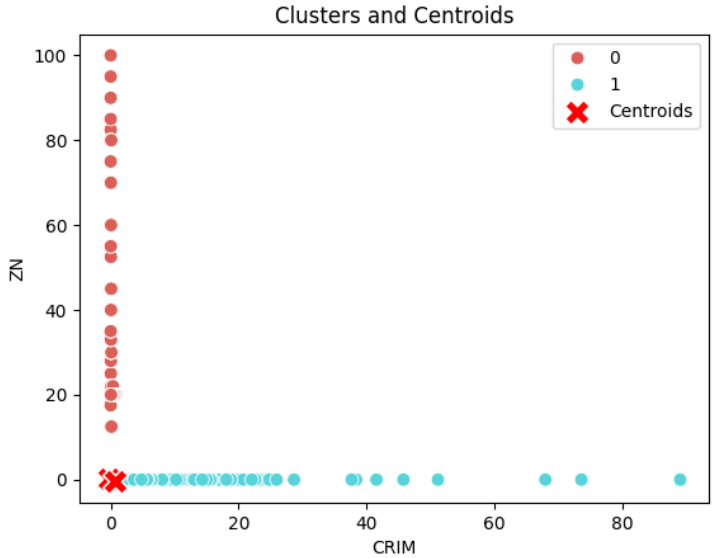

Figure 4: Clusters and Centroids

2.7.3 Pairplot of Key Features

The pairplot below visualizes the distribution and relationships between key features, namely *alcohol*, *malic_acid*, *ash*, and *color_intensity*, across different clusters. Each data point is colored according to its assigned cluster, providing insight into how the clusters are separated based on these features. The pairplot helps in visually assessing how well the clusters are distinct and whether the feature combinations contribute to the separation between clusters.

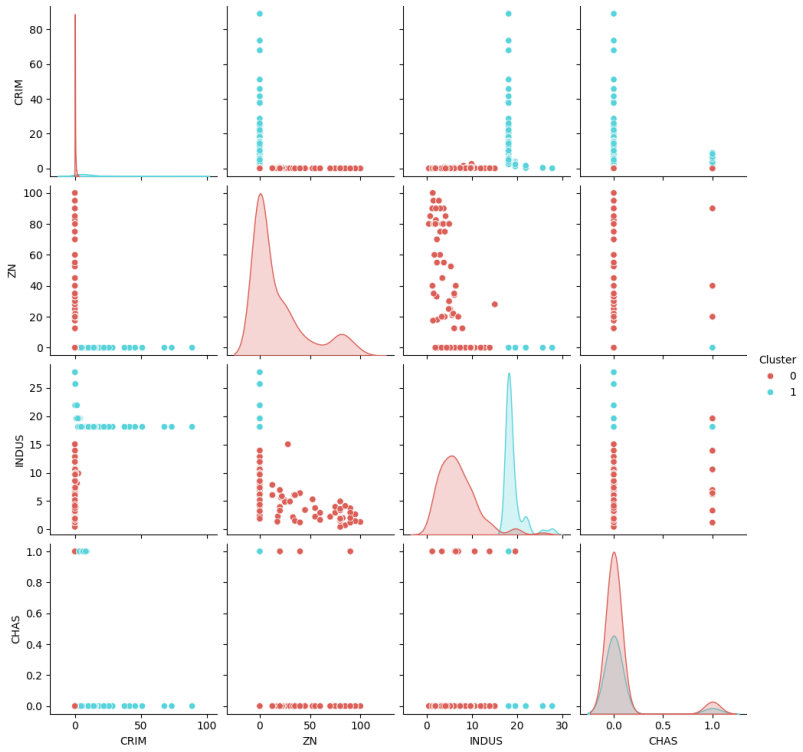

Figure 5: Pairplot of Key Features: Alcohol, Malic Acid, Ash, and Color Intensity

2.8 Conclusion

The K-Means clustering algorithm successfully grouped the data into two clusters, with $k = 2$ yielding the highest silhouette score. The differences between cluster means and centroids indicate well-separated clusters, with distinct characteristics based on crime rates, housing features, and other factors.

3 Problem 3: K-Means Clustering Analysis on Wine Dataset

In this analysis, we perform K-Means clustering on the Wine dataset, with the number of clusters set to 3. After clustering, we calculate the clustering quality metrics: **Homogeneity** and **Completeness**, using the actual class labels. These metrics help evaluate how well the clustering reflects the true class structure in the data. We will explain what each metric provides and the results obtained from the clustering.

3.1 Dataset Description

The Wine dataset contains 178 samples, each representing a wine, with 13 features. These features represent chemical properties such as alcohol content, malic acid, and flavonoids. The target labels correspond to three different types of wine, which are our true class labels.

3.2 Data Preprocessing

The dataset is first standardized using the StandardScaler, ensuring each feature has a mean of 0 and a standard deviation of 1. This step is essential because K-Means is sensitive to the scale of the features.

3.3 K-Means Clustering

We perform K-Means clustering with 3 clusters, corresponding to the 3 wine types in the dataset. K-Means is an unsupervised learning algorithm that groups similar data points into clusters based on feature similarity.

3.4 Silhouette Score Calculation

The Silhouette score is calculated for $k = 2$ to $k = 6$. The plot below shows the Silhouette scores for each k, with the optimal $k = 2$ yielding the highest score. The Silhouette score measures the separation between clusters and indicates the quality of the clustering.

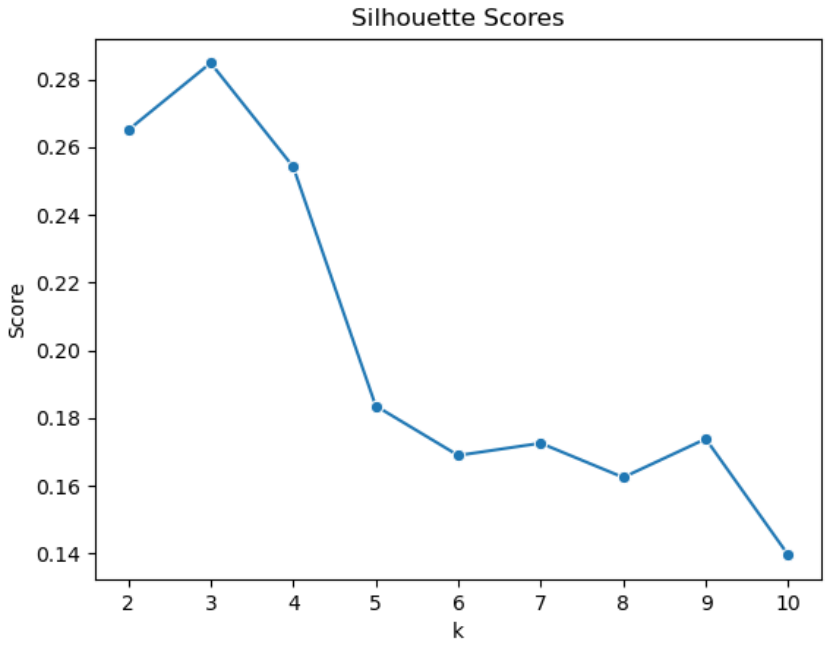

Figure 6: Silhouette Scores for Different k Values

A Silhouette Score of 0.2849 indicates that the clustering does not achieve a very clear separation between clusters. A higher score would suggest that the clusters are more distinct and well-separated.

3.5 Cluster Means and Centroids

For $k = 2$, K-Means clustering is performed, and the mean values of features for each cluster are computed. These are compared with the centroids of the clusters:

Feature	Cluster 0 Mean	Cluster 1 Mean	Cluster 2 Mean
Alcohol	12.2509	13.1341	13.6768
Malic Acid	1.8974	3.3073	1.9979
Ash	2.2312	2.4176	2.4663
Alcalinity of Ash	20.0631	21.2412	17.4629
Magnesium	92.7385	98.6667	107.9677
Total Phenols	2.2477	1.6839	2.8476
Flavanoids	2.0500	0.8188	3.0032
Nonflavanoid Phenols	0.3577	0.4520	0.2921
Proanthocyanins	1.6242	1.1459	1.9221
Color Intensity	2.9731	7.2347	5.4535
Hue	1.0627	0.6920	1.0655
OD280/OD315 of Diluted Wines	2.8034	1.6967	3.1634
Proline	510.1692	619.0588	1100.2258

3.6 Homogeneity and Completeness

After performing clustering, we calculate the following metrics to evaluate the quality of the clustering:

3.6.1 Homogeneity

Homogeneity measures how well the samples within each cluster belong to the same true class. In other words, it quantifies whether each cluster predominantly contains samples from a single class. A high homogeneity score indicates that the clusters are pure in terms of class labels.

$$\text{Homogeneity} = 0.8788$$

A homogeneity score of 0.8788 means that most samples within each cluster belong to the same true class, although some misclassifications may still occur.

3.6.2 Completeness

Completeness measures how well all samples from the same true class are assigned to the same cluster. A high completeness score means that all members of a class are grouped together, even if the cluster contains samples from other classes.

$$\text{Completeness} = 0.8730$$

A completeness score of 0.8730 indicates that most members of the same true class are placed within the same cluster, although there may be a few samples from each class scattered across different clusters.

3.6.3 V-Measure and Adjusted Rand Index

The **V-Measure** combines homogeneity and completeness, providing a single metric to evaluate the overall clustering quality. It is the harmonic mean of homogeneity and completeness.

$$\text{V-Measure} = 0.8759$$

The **Adjusted Rand Index (ARI)** compares the similarity of the predicted clusters to the true class labels, accounting for chance. A value of 1 indicates perfect clustering, while values close to 0 suggest random clustering.

$$\text{Adjusted Rand Index} = 0.8975$$

3.7 Sum of Squared Errors (SSE)

The total Sum of Squared Errors (SSE) for the clustering is:

$$\text{Total SSE} = 1277.9285$$

This value quantifies the sum of squared distances between each data point and its cluster center. A lower SSE indicates that the clusters are more compact and well-separated.

3.8 Cluster Centroids

The centroids for the three clusters are:

Feature	Centroid 0	Centroid 1	Centroid 2
Alcohol	−0.9261	0.1649	0.8352
Malic Acid	−0.3940	0.8715	−0.3038
Ash	−0.4945	0.1869	0.3647
Alcalinity of Ash	0.1706	0.5244	−0.6102
Magnesium	−0.4917	−0.0755	0.5776
Total Phenols	−0.0760	−0.9793	0.8852
Flavanoids	0.0208	−1.2152	0.9778
Nonflavanoid Phenols	−0.0335	0.7261	−0.5621
Proanthocyanins	0.0583	−0.7797	0.5803
Color Intensity	−0.9019	0.9415	0.1711
Hue	0.4618	−1.1648	0.4740
OD280/OD315 of Diluted Wines	0.2708	−1.2924	0.7792
Proline	−0.7538	−0.4071	1.1252

3.9 Visualization

3.9.1 Side-by-Side Figures: PCA and Confusion Matrix

The following figures are shown side by side to visualize the clustering results more effectively.

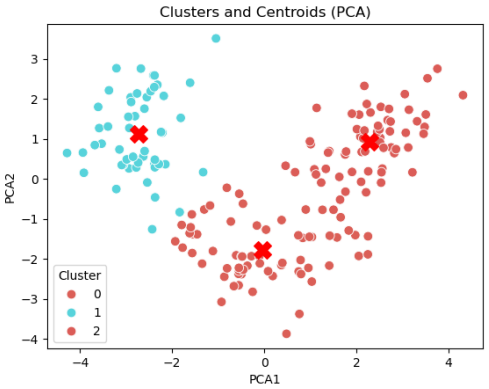

Figure 7: Clusters and Centroids Visualized Using PCA

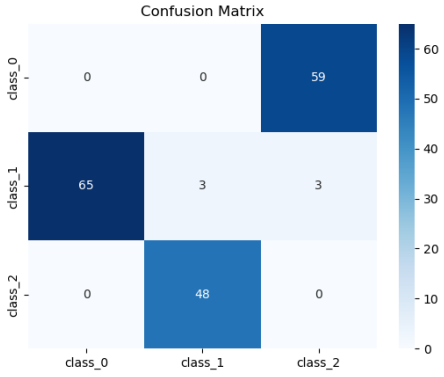

Figure 8: Confusion Matrix for K-Means Clustering vs True Labels

3.9.2 Pairplot of Key Features

The pairplot of the key features is shown below. It illustrates how the clusters are distributed across various feature combinations. We also see how the Silhouette score (0.2849) suggests that the clustering isn't as distinct as it could be, as the overlap between clusters is noticeable.

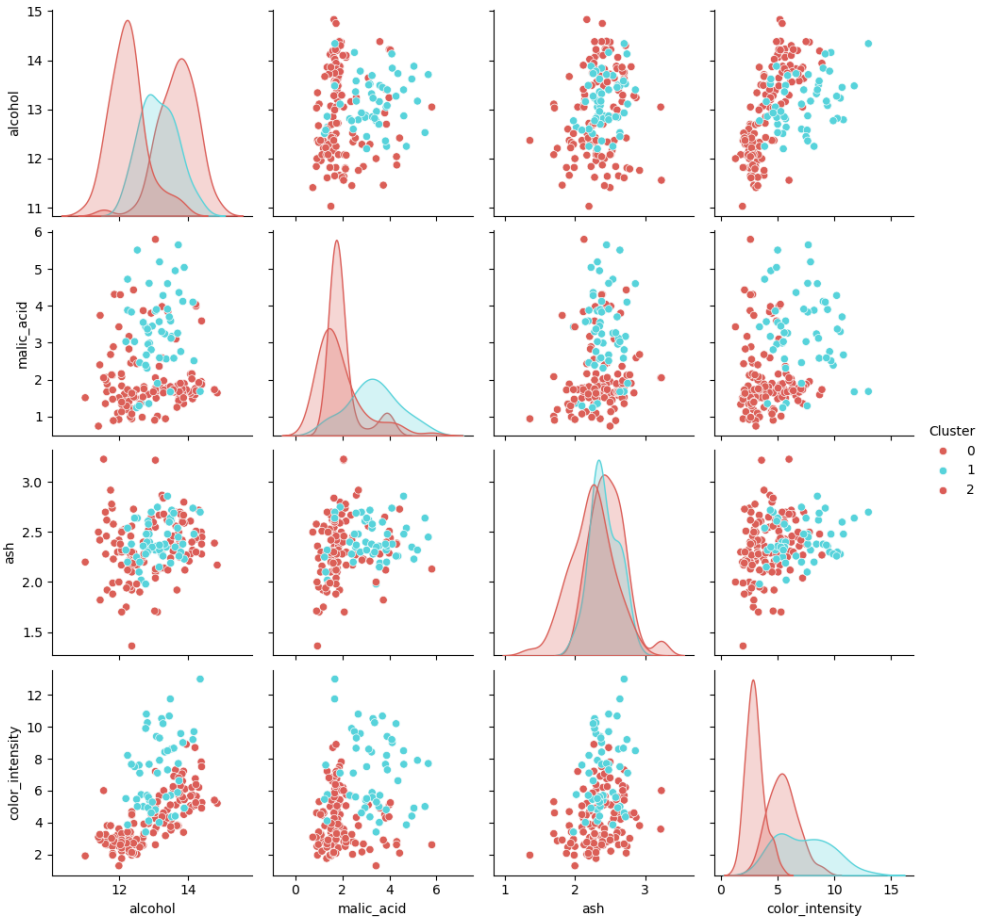

Figure 9: Pairplot of Key Features (Alcohol, Malic Acid, Ash, Color Intensity)

4 Conclusion

The K-Means clustering algorithm effectively grouped the wine samples into 3 distinct clusters. The high **Homogeneity** score of 0.8788 indicates that most samples within each cluster belong to the same true class, while the **Completeness** score of 0.8730 demonstrates that the majority of data points from each true class are assigned to the correct cluster. The **Adjusted Rand Index** of 0.8975 further confirms a strong alignment between the predicted clusters and the true class labels, accounting for chance agreements. Additionally, the **V-Measure** score of 0.8759 provides a comprehensive evaluation of clustering quality by balancing both homogeneity and completeness. Lastly, the **SSE** (Sum of Squared Errors) value suggests that the clusters are compact and well-separated, indicating that the clustering results are both tight and distinct.