ID: A20563448
Name: Zibo Nie **Data Mining HW_4.0** **Date: 5/9/2025**

1 Problem 1: Hierarchical Clustering on Auto-MPG Dataset

Dataset: The Auto-MPG dataset from the UCI Machine Learning Repository contains 398 samples with features including miles per gallon (mpg), displacement, horsepower, weight, acceleration, and origin (1, 2, 3). The continuous features (mpg, displacement, horsepower, weight, acceleration) are used for clustering.

Task:

- Load the dataset into a Pandas DataFrame.
- Impute missing values with the mean of each feature.
- Perform **hierarchical clustering** using `sklearn.cluster.AgglomerativeClustering` with:
  - `n_clusters = 3`
  - `linkage = average`
  - `affinity = euclidean`
- Compute mean and variance for each cluster and compare with origin-based class statistics.
- Analyze the relationship between cluster assignments and origin labels using a crosstab.

Analysis:

- Is there a clear relationship between cluster assignments and origin labels?
- How do cluster statistics compare to origin class statistics?

Data Preprocessing: Missing values (e.g., '?' in horsepower) were replaced with the mean of the respective feature. The continuous features were standardized using:

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

where μ is the mean and σ is the standard deviation.

Results: Table 1 shows the mean and variance of standardized features for each cluster. **Table 2** presents the crosstab of cluster assignments versus origin labels.

Table 1: Cluster Mean and Variance (Standardized Data)

Cluster	mpg	displacement	horsepower	weight	acceleration
Mean					
0	0.3411	-0.4717	-0.4677	-0.4398	0.3113
1	-1.1511	1.4845	1.4915	1.3875	-1.0627
2	2.5858	-0.9764	-1.4429	-0.9892	2.6530
Variance					
0	0.6778	0.3238	0.1998	0.4182	0.6427
1	0.0783	0.1927	0.4561	0.2710	0.4205
2	0.0049	0.0011	0.0027	0.0303	0.3044

Table 2: Cluster vs. Origin Crosstab

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

Cluster 1 is predominantly composed of origin 1 samples, indicating a strong relationship. Cluster 0 contains a mix of all origins, suggesting no clear dominance, while Cluster 2 has few samples, primarily from origin 2.

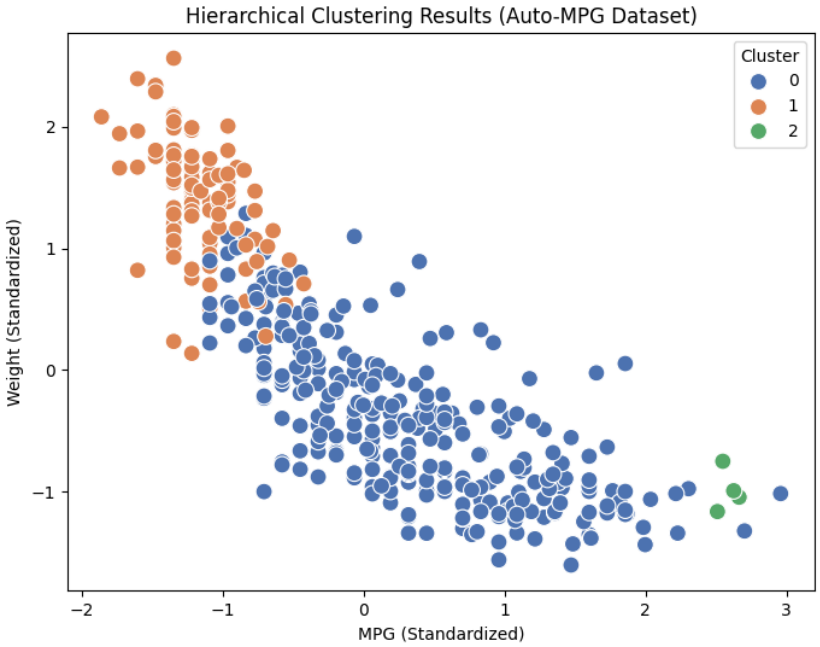

Figure 1: Hierarchical Clustering Results (mpg vs. weight)

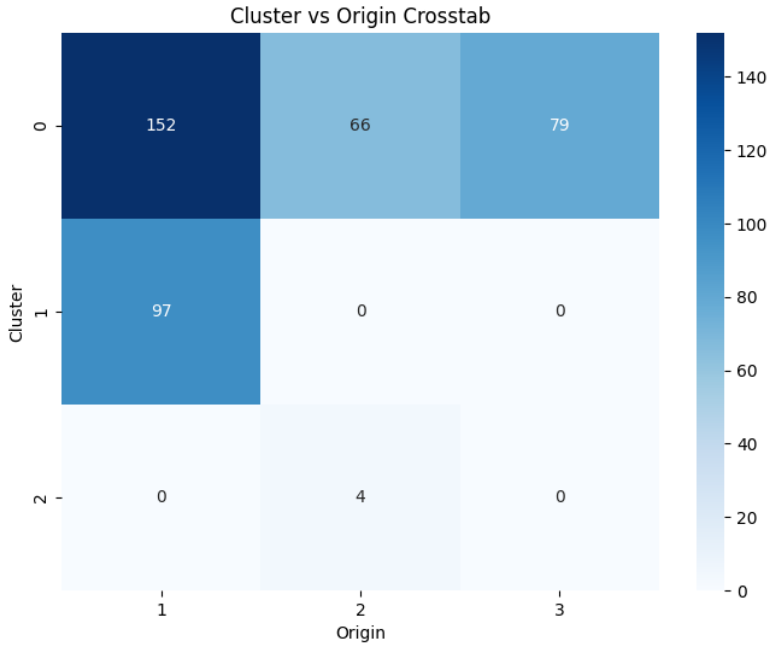

Figure 2: Cluster vs. Origin Crosstab Heatmap

Figure 1 shows the clustering results using standardized mpg and weight, with distinct separation for Cluster 1. **Figure 2** visualizes the crosstab, highlighting the strong association of Cluster 1 with origin 1.

Discussion: The hierarchical clustering partially aligns with origin labels. Cluster 1's exclusivity to origin 1

suggests that features like lower mpg and higher displacement are characteristic of origin 1 vehicles. However, the mixed composition of Cluster 0 indicates that the clustering does not fully capture the origin-based structure.

2 Problem 2: K-Means Clustering on Boston Dataset

Dataset: The Boston Housing dataset contains 506 samples with 13 features, including crime rate (CRIM), average number of rooms (RM), and lower status population (LSTAT).

Task:

- Load the dataset into a Pandas DataFrame.
- Perform **K-Means clustering** on standardized data for `n_clusters` from 2 to 6.
- Compute **silhouette scores** to identify the optimal number of clusters.
- Calculate mean feature values for each cluster and compare with centroids for the optimal k.

Analysis:

- Which k yields the highest silhouette score, and why?
- How do cluster means differ from centroid coordinates?

Data Preprocessing: The data were standardized to ensure scale invariance for K-Means clustering.

Results: Table 3 shows the silhouette scores for $k = 2$ to 6, with $k = 2$ achieving the highest score (0.3601). Table 4 presents the cluster means for $k = 2$.

Table 3: Silhouette Scores for Different k

k	**Silhouette Score**
2	0.3601
3	0.2575
4	0.2658
5	0.2878
6	0.2625

Table 4: Cluster Means for $k = 2$

Feature	**Cluster 0**	**Cluster 1**
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 differences between cluster means and centroids are negligible (on the order of 10^{-16}), indicating accurate centroid representation.

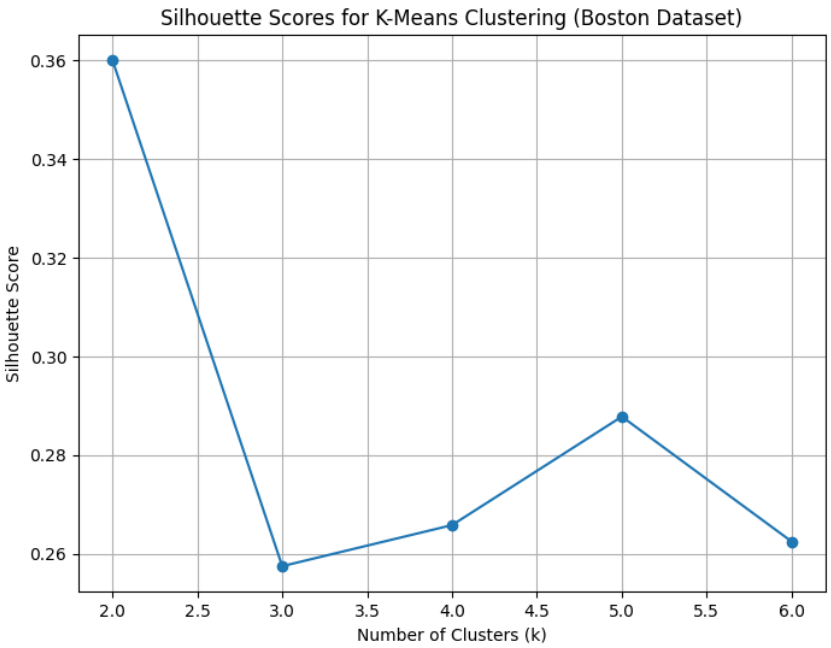

Figure 3: Silhouette Scores for Different k

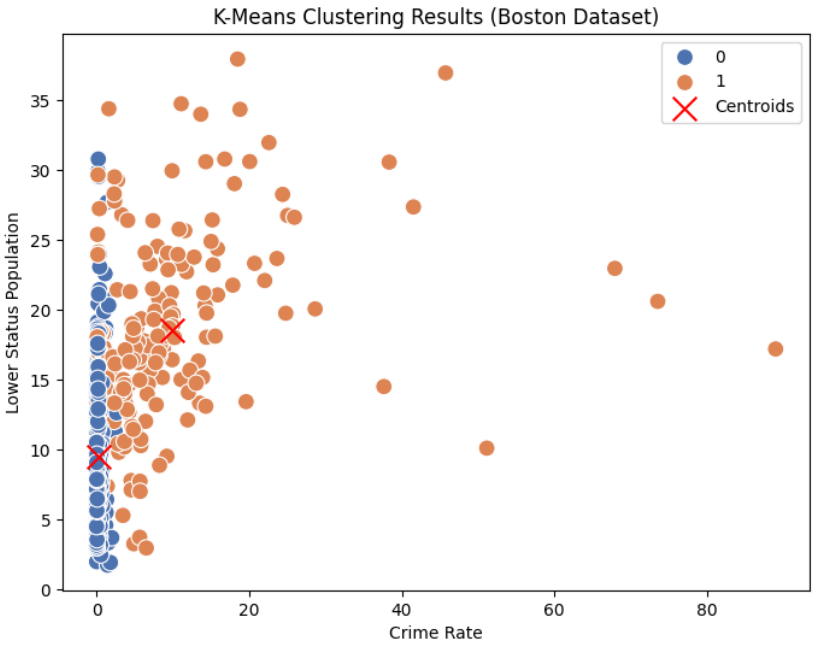

Figure 4: K-Means Clustering Results (CRIM vs. LSTAT)

Figure 3 illustrates the silhouette scores, confirming $k = 2$ as optimal. **Figure 4** shows the clusters with centroids, indicating clear separation based on crime rate and LSTAT.

Discussion: The optimal $k = 2$ suggests two distinct groups in the Boston dataset, likely reflecting socioeconomic differences. Cluster 0 has lower crime rates and younger houses, while Cluster 1 has higher crime rates and older houses. The high silhouette score validates the clustering quality.

3 Problem 3: K-Means Clustering on Wine Dataset

Dataset: The Wine dataset contains 178 samples with 13 chemical features (e.g., alcohol, flavonoids) and three true class labels representing wine types.

Task:

- Load the dataset into a Pandas DataFrame.
- Perform **K-Means clustering** with `n_clusters = 3` on standardized data.
- Compute **homogeneity** and **completeness** scores using true class labels.

Analysis:

- What do homogeneity and completeness reveal about clustering quality?
- How well do the clusters align with true class labels?

Data Preprocessing: The data were standardized to ensure equal feature contributions.

Results: The clustering achieved a homogeneity score of 0.8788 and a completeness score of 0.8730. **Table 5** shows the mean values of key features for each cluster.

Table 5: Cluster Means for $k = 3$

Feature	**Cluster 0**	**Cluster 1**	**Cluster 2**
Alcohol	12.2509	13.1341	13.6768
Malic Acid	1.8974	3.3073	1.9979
Ash	2.2312	2.4176	2.4663
Flavanoids	2.0500	0.8188	3.0032
Proline	510.1692	619.0588	1100.2258

- **Homogeneity**: Measures whether each cluster contains samples from a single class.

$$\text{Homogeneity} = 0.8788$$

- **Completeness**: Measures whether all samples of a class are assigned to the same cluster.

$$\text{Completeness} = 0.8730$$

Figure 5: K-Means Clustering Results (PCA)

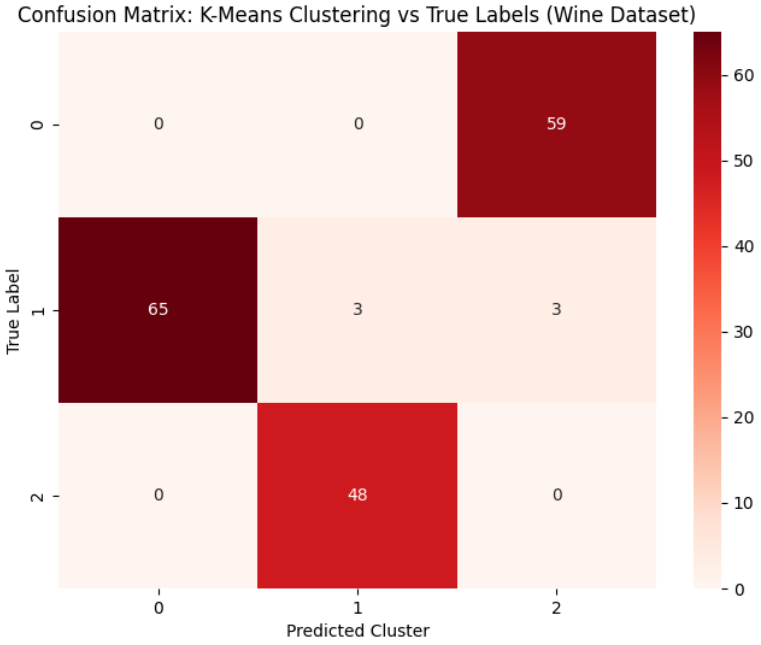

Figure 6: Confusion Matrix for K-Means Clustering

Figure 5 shows the clusters in a PCA-reduced 2D space, with centroids marked. **Figure 6** presents the confusion matrix, indicating strong alignment with true labels.

Discussion: The high homogeneity and completeness scores suggest that the clusters closely correspond to the true wine types. Cluster 2 is distinguished by high proline and flavanoids, aligning with one wine type. The confusion matrix confirms minimal misclassifications, supporting the clustering quality.