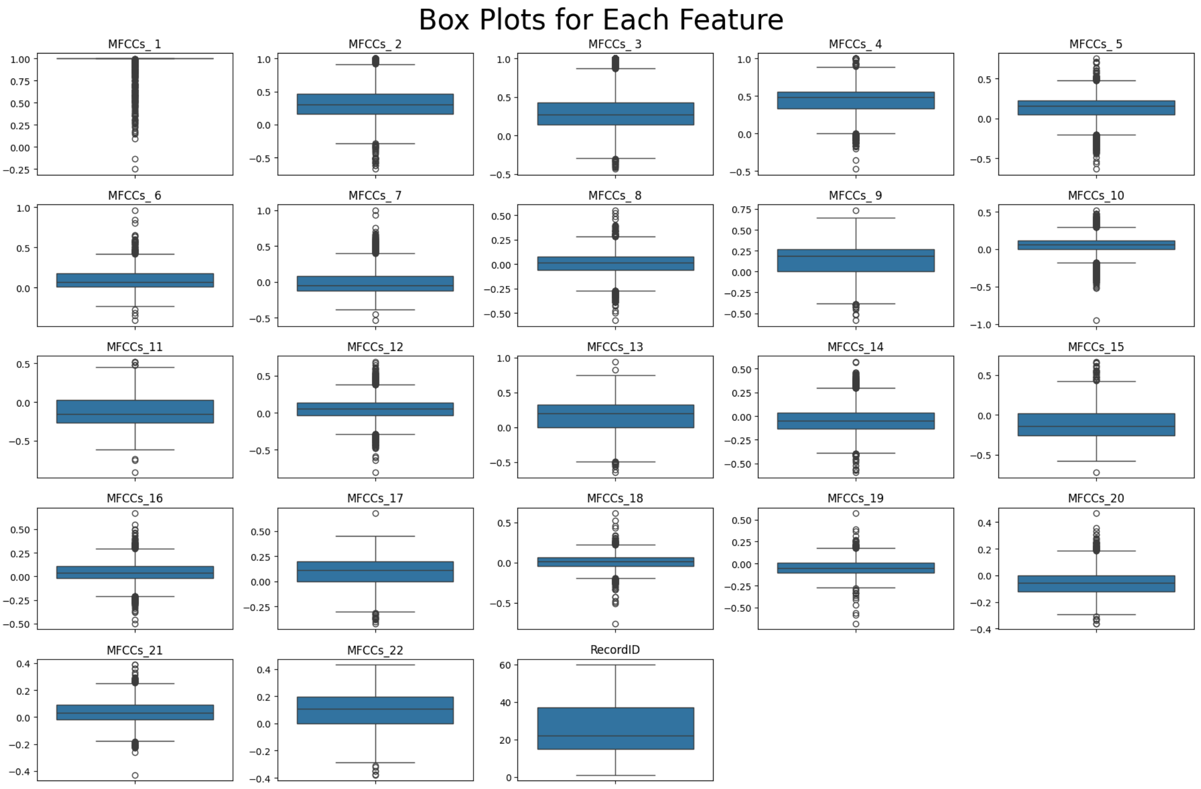
G.M.M.Prabhash
22CS30027

REPORT : ASSIGNMENT 3
PART 2 : K MEANS

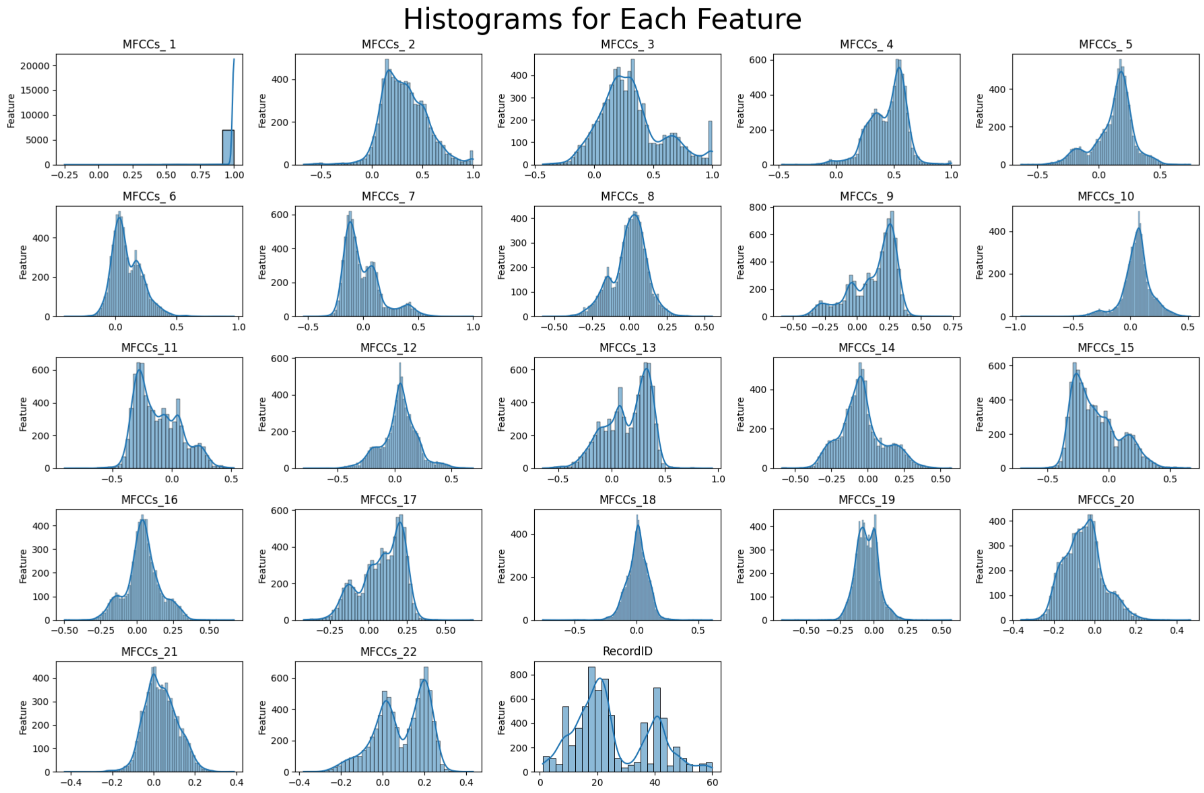
Box Plot For Each :

Box plots serve as a fundamental tool for visualising the distribution of individual features within the dataset. For the Frogs_MFCCs dataset, which comprises 22 MFCC coefficients, box plots reveal critical insights into the central tendency, variability, and potential outliers for each coefficient. The presence of outliers in the box plots may indicate variations in frog calls or anomalies in the data collection process. Analysing these distributions is crucial, as features with significant outliers or skewed distributions may adversely affect clustering performance, leading to less reliable cluster assignments.



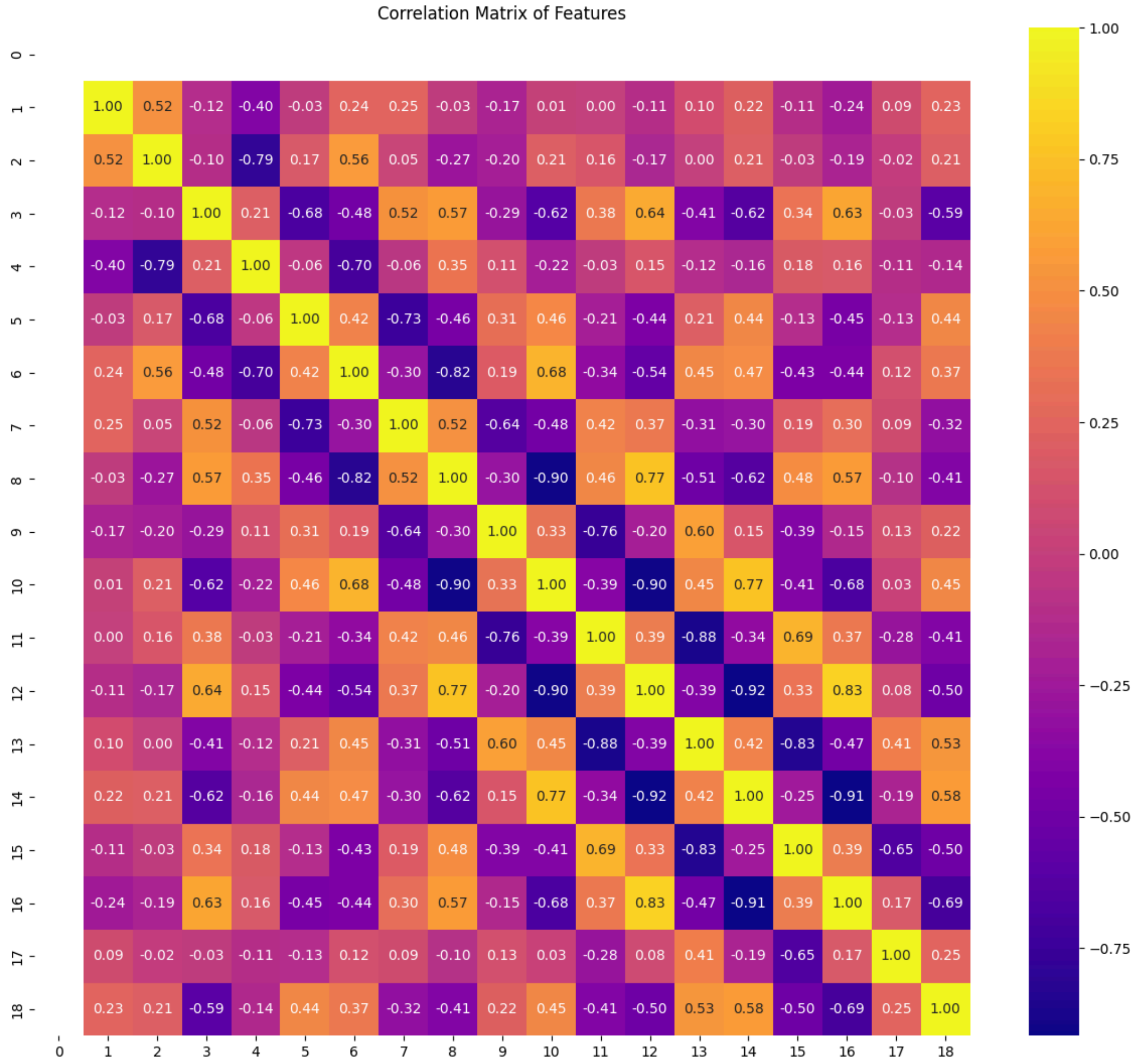
<u>**Histogram Plot For Each :**</u>

Histograms provide a detailed view of the frequency distribution of each MFCC feature. By plotting the frequency of values for each feature, histograms can identify patterns such as normality, skewness, and multimodal distributions. Understanding the shape of these distributions is essential for clustering, as certain algorithms may assume that data follows a specific distribution. For **i**nstance, features with normal distributions may allow for more effective application of clustering algorithms, while heavily skewed features may require transformation or normalization prior to analysis.



Correlation Matrix Between Features :

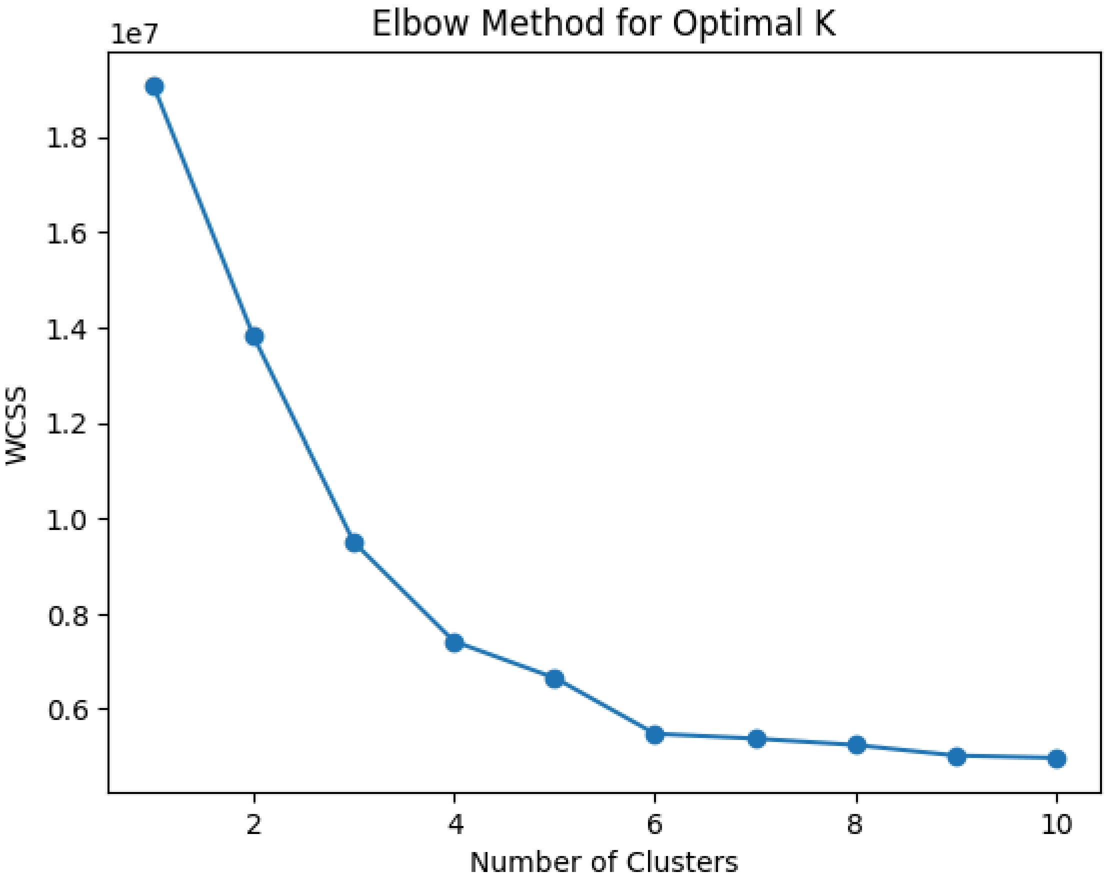
A correlation matrix is instrumental in evaluating the relationships between the MFCC coefficients. By calculating and visualising the correlation coefficients, this matrix highlights pairs of features that exhibit strong correlations, either positive or negative. Understanding these relationships is vital, as high correlation between features can lead to multicollinearity, which may impair the performance of clustering algorithms. Identifying correlated features allows for feature selection or dimensionality reduction strategies, such as PCA, to improve clustering outcomes by focusing on more independent features.

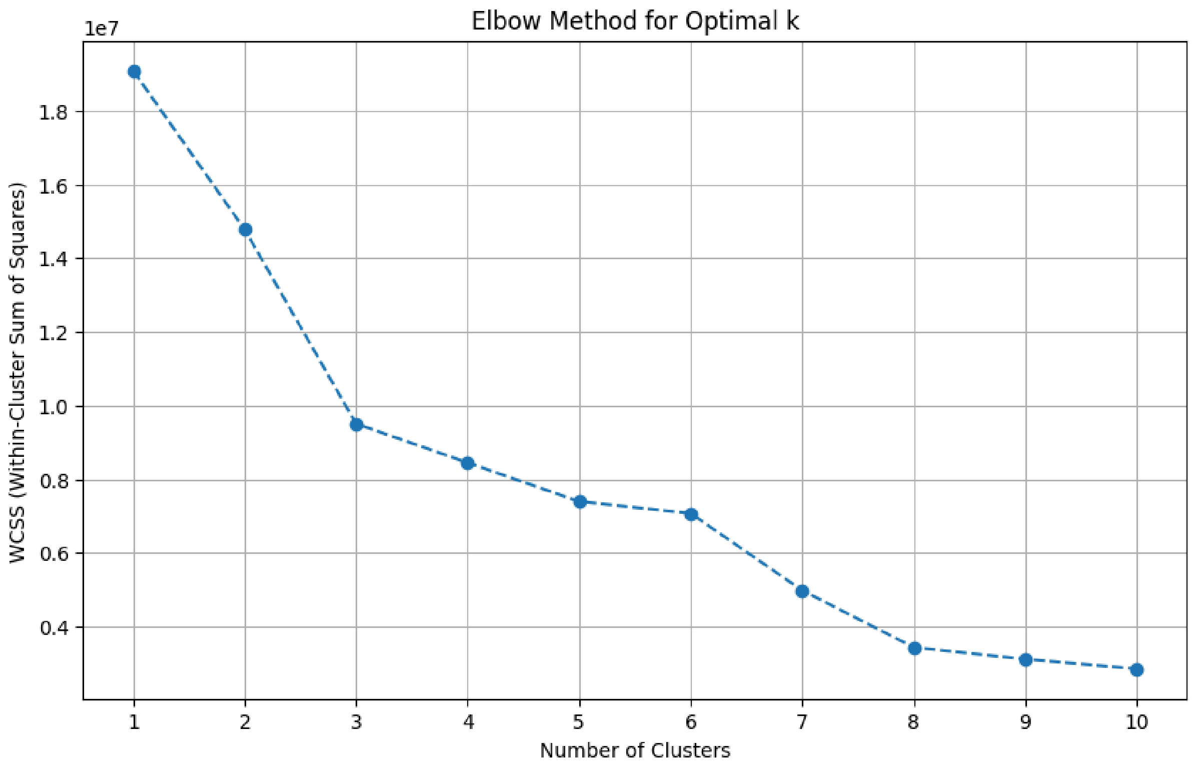


	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1	1.00	0.52	-0.12	-0.40	-0.03	0.24	0.25	-0.03	-0.17	0.01	0.00	-0.11	0.10	0.22	-0.11	-0.24	0.09	0.23
2	0.52	1.00	-0.10	-0.79	0.17	0.56	0.05	-0.27	-0.20	0.21	0.16	-0.17	0.00	0.21	-0.03	-0.19	-0.02	0.21
3	-0.12	-0.10	1.00	0.21	-0.68	-0.48	0.52	0.57	-0.29	-0.62	0.38	0.64	-0.41	-0.62	0.34	0.63	-0.03	-0.59
4	-0.40	-0.79	0.21	1.00	-0.06	-0.70	-0.06	0.35	0.11	-0.22	-0.03	0.15	-0.12	-0.16	0.18	0.16	-0.11	-0.14
5	-0.03	0.17	-0.68	-0.06	1.00	0.42	-0.73	-0.46	0.31	0.46	-0.21	-0.44	0.21	0.44	-0.13	-0.45	-0.13	0.44
6	0.24	0.56	-0.48	-0.70	0.42	1.00	-0.30	-0.82	0.19	0.68	-0.34	-0.54	0.45	0.47	-0.43	-0.44	0.12	0.37
7	0.25	0.05	0.52	-0.06	-0.73	-0.30	1.00	0.52	-0.64	-0.48	0.42	0.37	-0.31	-0.30	0.19	0.30	0.09	-0.32
8	-0.03	-0.27	0.57	0.35	-0.46	-0.82	0.52	1.00	-0.30	-0.90	0.46	0.77	-0.51	-0.62	0.48	0.57	-0.10	-0.41
9	-0.17	-0.20	-0.29	0.11	0.31	0.19	-0.64	-0.30	1.00	0.33	-0.76	-0.20	0.60	0.15	-0.39	-0.15	0.13	0.22
10	0.01	0.21	-0.62	-0.22	0.46	0.68	-0.48	-0.90	0.33	1.00	-0.39	-0.90	0.45	0.77	-0.41	-0.68	0.03	0.45
11	0.00	0.16	0.38	-0.03	-0.21	-0.34	0.42	0.46	-0.76	-0.39	1.00	0.39	-0.88	-0.34	0.69	0.37	-0.28	-0.41
12	-0.11	-0.17	0.64	0.15	-0.44	-0.54	0.37	0.77	-0.20	-0.90	0.39	1.00	-0.39	-0.92	0.33	0.83	0.08	-0.50
13	0.10	0.00	-0.41	-0.12	0.21	0.45	-0.31	-0.51	0.60	0.45	-0.88	-0.39	1.00	0.42	-0.83	-0.47	0.41	0.53
14	0.22	0.21	-0.62	-0.16	0.44	0.47	-0.30	-0.62	0.15	0.77	-0.34	-0.92	0.42	1.00	-0.25	-0.91	-0.19	0.58
15	-0.11	-0.03	0.34	0.18	-0.13	-0.43	0.19	0.48	-0.39	-0.41	0.69	0.33	-0.83	-0.25	1.00	0.39	-0.65	-0.50
16	-0.24	-0.19	0.63	0.16	-0.45	-0.44	0.30	0.57	-0.15	-0.68	0.37	0.83	-0.47	-0.91	0.39	1.00	0.17	-0.69
17	0.09	-0.02	-0.03	-0.11	-0.13	0.12	0.09	-0.10	0.13	0.03	-0.28	0.08	0.41	-0.19	-0.65	0.17	1.00	0.25
18	0.23	0.21	-0.59	-0.14	0.44	0.37	-0.32	-0.41	0.22	0.45	-0.41	-0.50	0.53	0.58	-0.50	-0.69	0.25	1.00

For Finding Best K ⇒ Elbow Method :

The Elbow Method is a widely used heuristic for determining the optimal number of clusters in K-Means clustering. By plotting the Within-Cluster Sum of Squares (WCSS) against a range of cluster numbers, this method reveals how WCSS decreases as more clusters are added. The point at which the reduction in WCSS begins to slow significantly indicates the ideal number of clusters, known as the "elbow." For the Frogs_MFCCs dataset, the Elbow Method indicated an optimal cluster count of 4, suggesting that the dataset is best represented by four distinct groups.





Optimal number of clusters from Elbow Method: 4

Silhouette scores quantify the quality of clusters by measuring how similar an object is to its own cluster compared to other clusters. Higher silhouette scores indicate better-defined clusters. For the K-Means clustering with random initialization, the silhouette score was **0.7176**, while the score for K-Means with K-Means++ initialization was **0.6087**. This result suggests that random initialization produced better-defined clusters compared to K-Means++, highlighting the sensitivity of K-Means to initial centroid placement.

Silhouette Score:

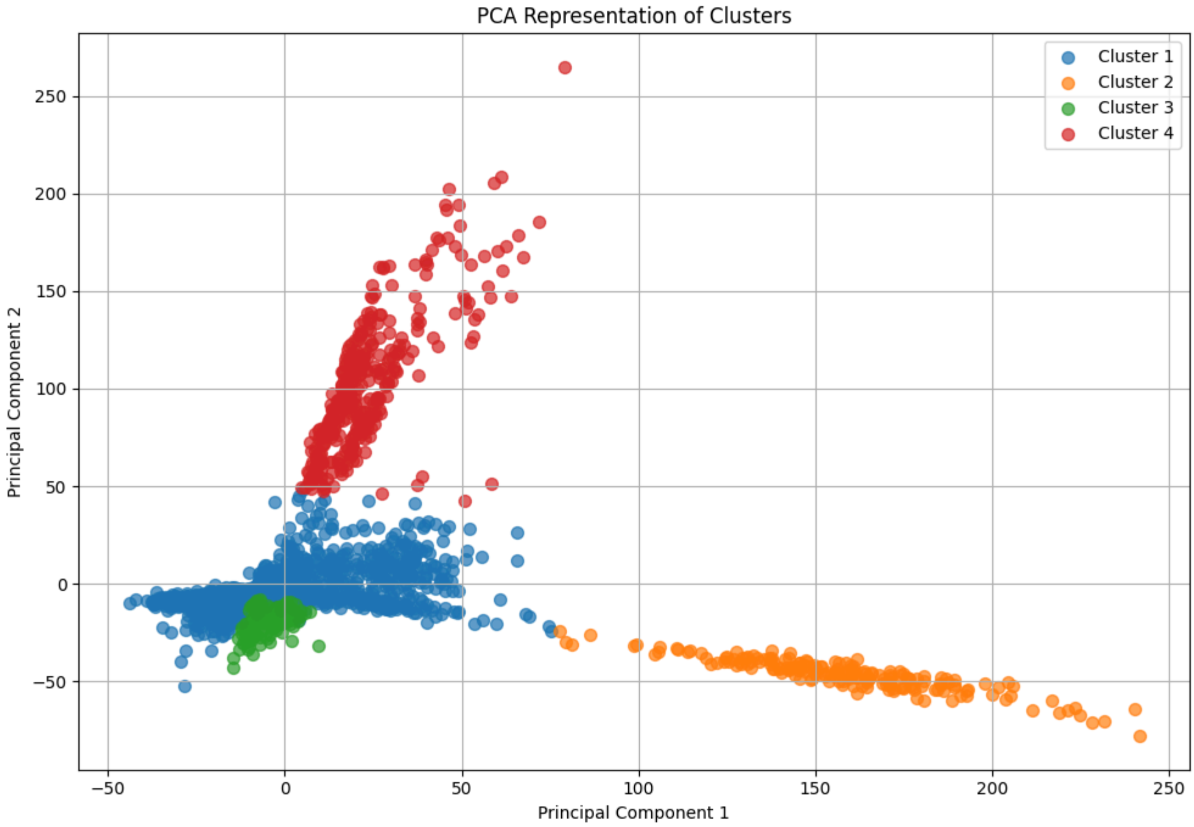
- For Random Initialization: **0.7176**
- For K-Means++ Initialization: **0.6087**

Conclusion:

Random initialization is more effective than K-Means++ initialization for this dataset, likely due to the inherent clustering structure that aligns more favourably with the initial random centroids.

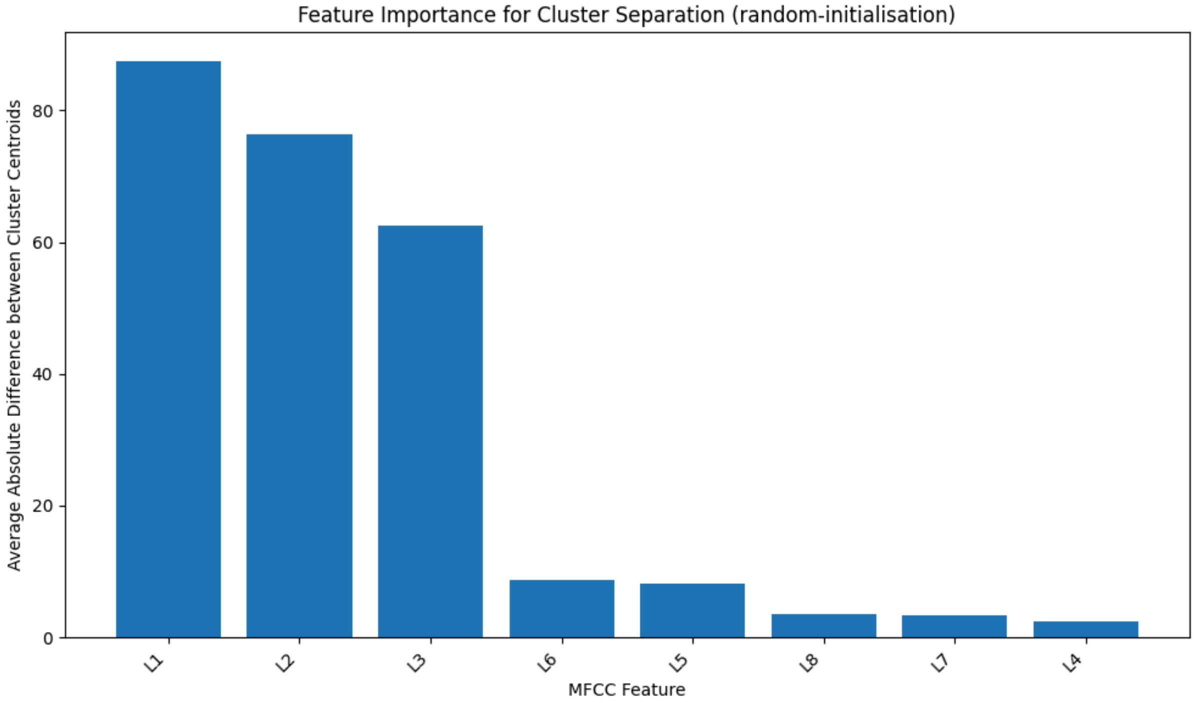
PCA Representation of Clusters :

Principal Component Analysis (PCA) provides a powerful means to visualise high-dimensional data in lower dimensions, facilitating the understanding of clustering results. By transforming the 22 MFCC features into two or three principal components, PCA enables clear visualisation of the clusters formed through K-Means. This representation aids in evaluating the effectiveness of the clustering algorithm and identifying the spatial distribution of clusters.



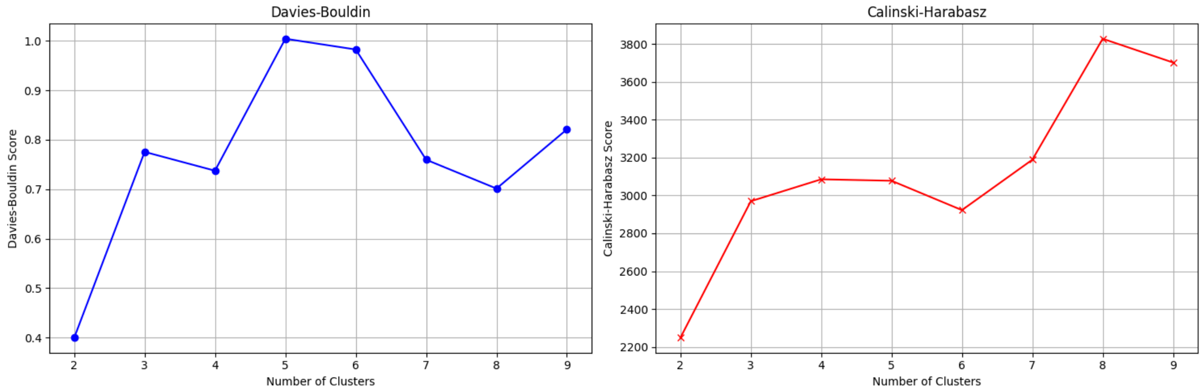
Feature Importance for Cluster Separation (random-initialisation) :

Feature importance in clustering can be assessed by evaluating the average absolute differences between cluster centroids. This analysis highlights which features contribute most significantly to the separation of clusters. In the case of the Frogs_MFCCs dataset, identifying these important features can inform subsequent analyses, helping to understand the characteristics that define each cluster of frog calls.



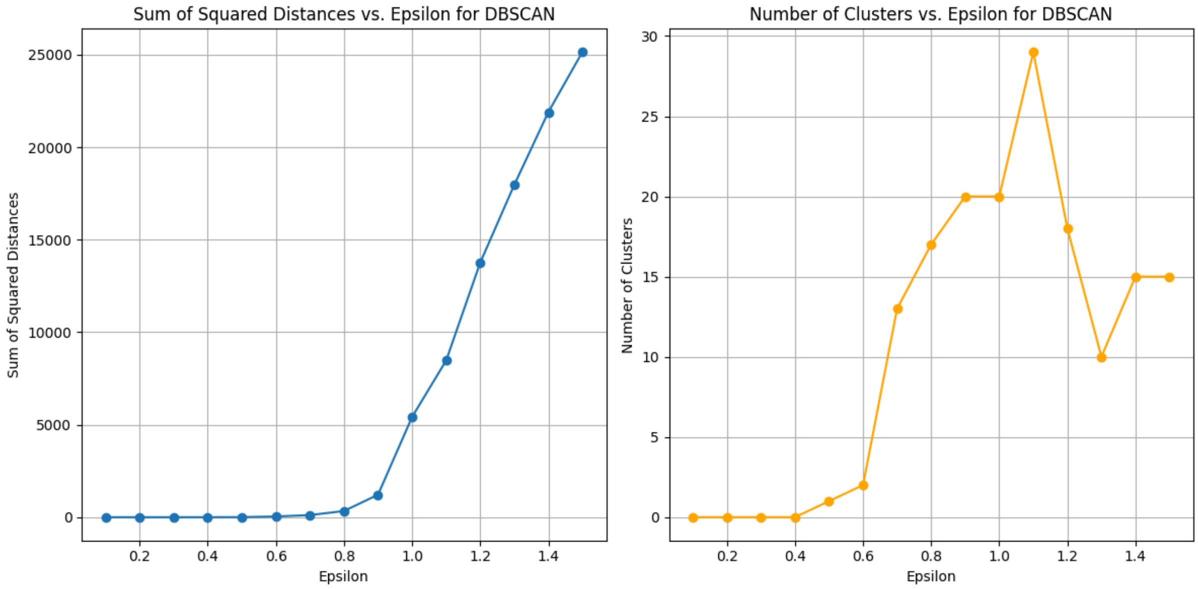
Davies-Bouldin Index & Calinski-Harabasz Index :

The Davies-Bouldin Index and Calinski-Harabasz Index provide additional metrics for evaluating clustering quality. A lower Davies-Bouldin Index indicates better clustering, as it suggests that clusters are farther apart relative to their size. Conversely, a higher Calinski-Harabasz Index reflects a well-structured clustering solution. Both indices reinforce findings from silhouette scores, enabling a comprehensive evaluation of clustering effectiveness.



DBSCAN :

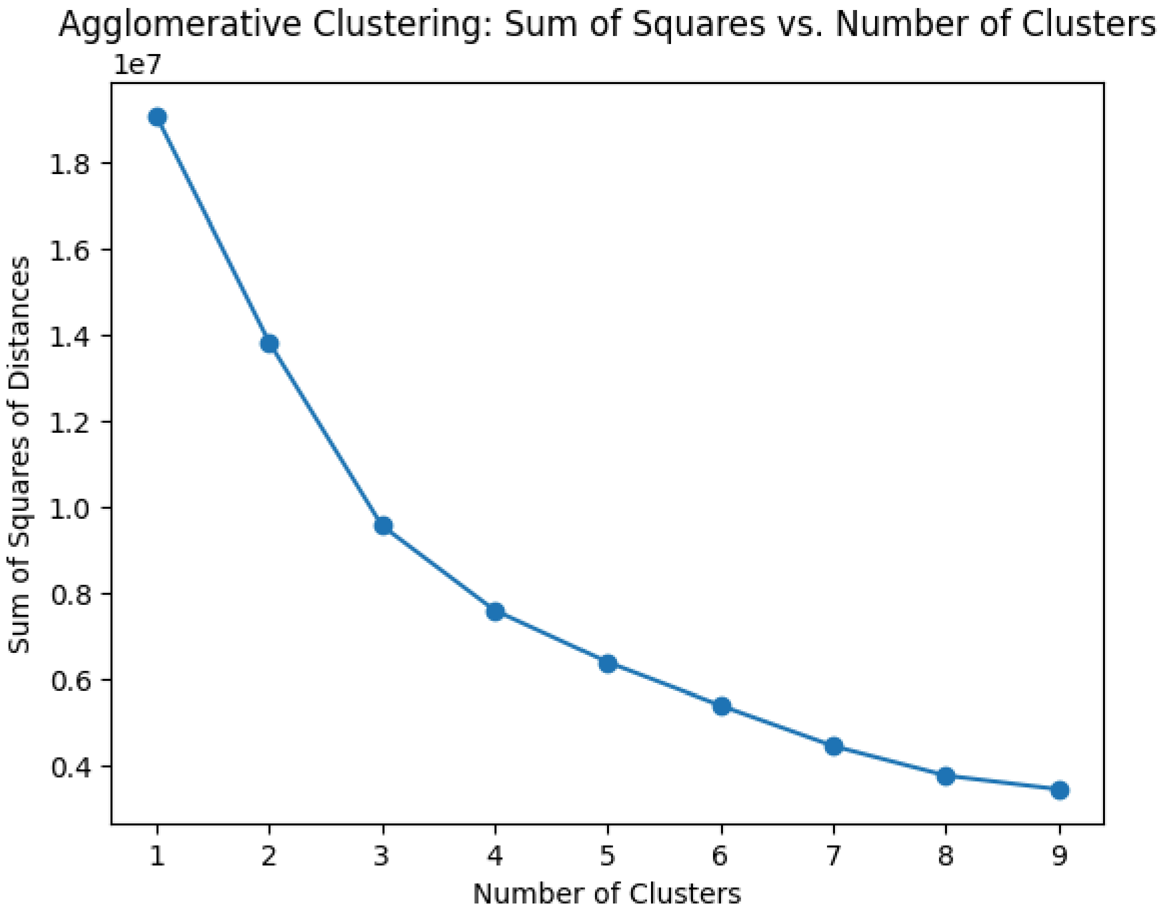
DBSCAN (Density-Based Spatial Clustering of Applications with Noise) excels in identifying clusters of varying shapes and sizes. Its ability to handle noise and outliers makes it particularly suitable for the Frogs_MFCCs dataset, where certain call types may be less frequent or consist of irregular patterns. The silhouette score for DBSCAN was **0.9559**, indicating excellent cluster separation and confirming its effectiveness in this context.



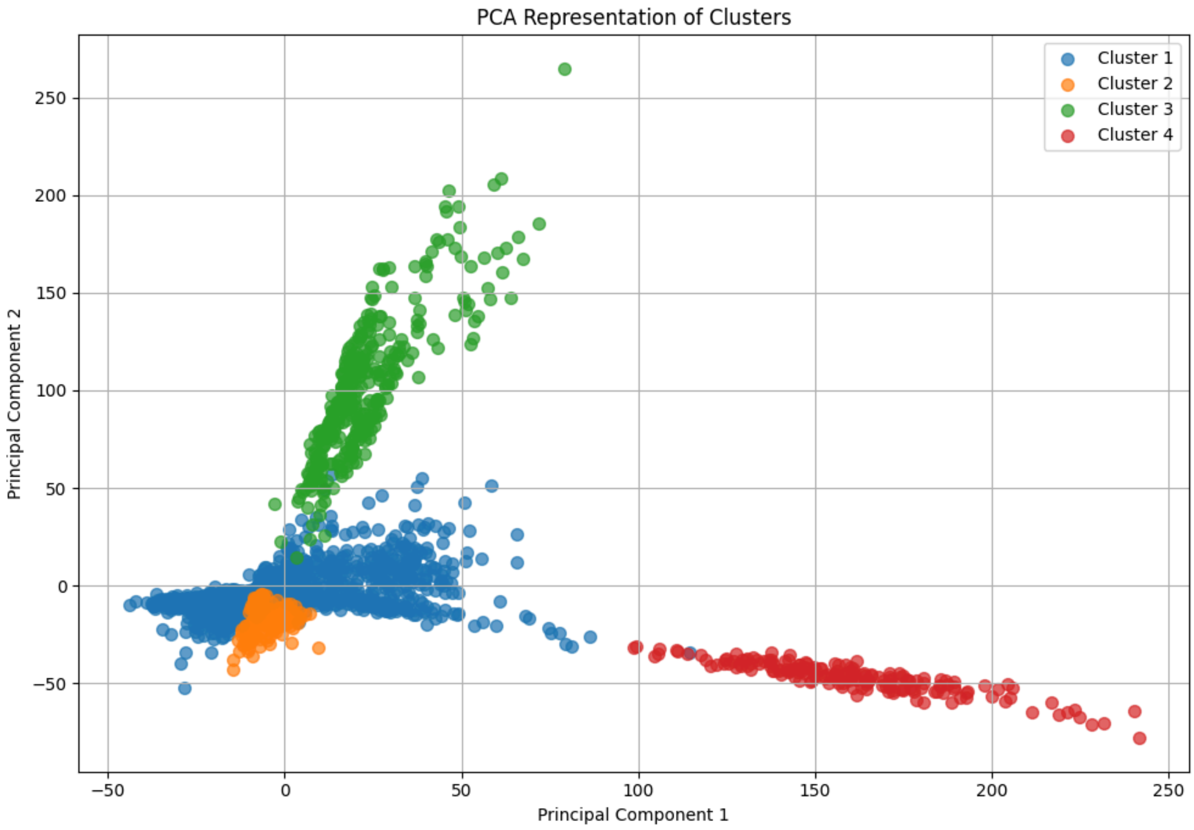
Silhouette score for DBSCAN: .0.9559224783745093

Agglomerative Clustering :

Agglomerative clustering offers a hierarchical approach to clustering, revealing nested group structures. The silhouette score for Agglomerative Clustering was 0.6802, suggesting decent clustering quality but indicating that it may not perform as well as DBSCAN or K-Means with random initialization in this dataset.



Silhouette score for Agglomerative Clustering: 0.6801959845351576



Performance Comparison (Based on Silhouette Score):

- **DBSCAN > K-Means (Random Initialization) > Agglomerative Clustering > K-Means++**

Based on silhouette scores, DBSCAN emerged as the best clustering algorithm for the Frogs_MFCCs dataset, demonstrating its superior capability to identify meaningful clusters amidst the complex patterns inherent in the data.

Algorithm Insights:

1. **K-Means Clustering**
   - **Strengths**: Efficient and suitable for large datasets; good performance on high-dimensional data with spherical clusters.
   - **Weaknesses**: Assumes spherical shapes; sensitive to outliers and requires predefined cluster numbers.
2. **Agglomerative Clustering**
   - **Strengths**: Flexible with cluster shapes; does not require a predefined number of clusters; robust to outliers.

   - **Weaknesses**: Computationally intensive for large datasets; sensitive to the choice of linkage criteria.
3. **DBSCAN Clustering**
   - **Strengths**: Handles arbitrary-shaped clusters and identifies noise; no need for predefined cluster count.
   - **Weaknesses**: Sensitive to parameter choices; may struggle with clusters of uneven density.

Clustering Analysis Summary for Frogs_MFCCs Dataset

In summary, this analysis has demonstrated the effectiveness of various clustering algorithms applied to the Frogs_MFCCs dataset. The exploration and evaluation of K-Means, Agglomerative Clustering, and DBSCAN reveal distinct strengths and weaknesses, with DBSCAN standing out as the most effective method for this dataset, especially considering its ability to manage noise and capture the irregularities in the data.