Clustering Analysis Report

1. Overview of Clustering Approaches

We applied multiple clustering algorithms and evaluated them using key metrics such as Silhouette Score and Davies-Bouldin Index (DB Index). The clustering techniques used were:

- K-Means Clustering
- Agglomerative Clustering
- Gaussian Mixture Model (GMM)
- DBSCAN

Our goal was to identify meaningful clusters while ensuring well-separated and compact groups.

2. Clustering Results and Evaluation Metrics

K-Means Clustering
- Number of Clusters: 8
- Silhouette Score: 0.4224
- DB Index: 0.7268

Agglomerative Clustering
- Number of Clusters: 4
- Silhouette Score: 0.3064
- DB Index: 0.8792

Gaussian Mixture Model (GMM) Clustering
- Number of Clusters: 4
- Silhouette Score: 0.3719
- DB Index: 0.7885

DBSCAN Clustering
- Number of Clusters: 1 (+ Noise points)
- Silhouette Score: 0.4985
- DB Index: 0.5352 (Best Performance)

3. Insights and Observations

- DBSCAN had the best DB Index (0.5352), meaning it formed the most compact clusters.
- K-Means provided a balanced structure with 8 clusters, making it suitable for multi-cluster analysis.

- Agglomerative performed the worst with a high DB Index (0.8792), indicating poor cluster separation.
- GMM performed better than Agglomerative but was weaker than K-Means.

4. Task 3: Clustering Optimization

- Tuned Clustering Parameters (eps=1.6, min_samples=10 for DBSCAN).
- Applied PCA for dimensionality reduction before clustering.
- Evaluated clusters with Silhouette Score and DB Index.
- Visualized clusters using scatter plots with PCA-transformed data.

5. Conclusion and Recommendations

- If multiple clusters are needed: K-Means (8 clusters) is the best choice.
- If identifying one main cluster + outliers: DBSCAN is optimal.
- Future improvements: Experiment with distance metrics, hierarchical DBSCAN (HDBSCAN), and additional feature engineering for better clustering results.