ZEOTAP INTERNSHIP ASSIGNMENT

TASK 3: Customer Segmentation

INTRODUCTION

- Clustering is a vital unsupervised machine learning technique used to group similar data points together based on their inherent characteristics.
- This report presents an analysis of clustering results obtained using the K-Means algorithm.
- Various metrics, including the number of clusters formed, the Davis-Bouldin (DB) Index, and the Silhouette Score, are evaluated to identify the optimal clustering configuration.

METHODOLOGY

The clustering analysis was performed using the KMeans algorithm, which partitions the dataset into a predefined number of clusters by minimizing the within-cluster sum of squares. The algorithm iteratively refines cluster centroids to improve compactness and separation.

EVALUATION METRICS

To assess the clustering quality, the following metrics were utilized:

- **Davis-Bouldin (DB) Index**: Measures the compactness and separation of clusters. Lower DB Index values indicate better clustering performance, with clusters being compact and well-separated.
- **Silhouette Score**: Quantifies how similar an object is to its cluster compared to other clusters. The score ranges from -1 to 1, with higher values indicating well-defined clusters.

OBSERVATIONS

The results are summarized for 2 to 10 clusters, as shown in the table below:

Clusters	DB Index	Silhouette Score
2	1.831694	0.194528
3	1.428838	0.311165
4	1.195446	0.337076
5	1.105905	0.354516
6	1.091585	0.314243
7	1.029169	0.331628
8	0.982020	0.352756
9	1.042394	0.337134
10	1.008522	0.343505

RESULTS

1. **Optimal Number of Clusters:**
    - The DB Index decreases as the number of clusters increases, reflecting an improvement in compactness and separation. However, after 8 clusters, the reduction becomes less significant.
    - The Silhouette Score peaks at 5 clusters (0.354516), suggesting this is the optimal number of clusters, where the balance between intra-cluster similarity and inter-cluster separation is maximized.
2. **Cluster Compactness and Separation:**
    - For fewer clusters (e.g., 2 or 3), the DB Index is relatively high, indicating less compact and poorly separated clusters.
    - As the cluster count increases, the DB Index improves, and the clusters become more compact and distinct. However, overly increasing the number of clusters can lead to overfitting, where clusters become too specific.

3. **Silhouette Score Trends:**
    - The Silhouette Score is relatively low for 2 clusters, reflecting poorly defined cluster boundaries.
    - The score improves consistently up to 5 clusters, where the optimal balance is achieved, before slightly fluctuating for higher cluster counts.

CONCLUSION

The clustering analysis highlights the importance of selecting an appropriate number of clusters for optimal performance. Using the DB Index and Silhouette Score, it is concluded that 5 clusters provide the best balance between compactness and separation. These findings are critical for further data exploration and insights extraction, ensuring that the data is partitioned in a meaningful and interpretable manner.

Future work could involve:

- Testing alternative clustering algorithms, such as Hierarchical Clustering or DBSCAN, to compare results.
- Using domain knowledge to validate cluster relevance.
- Applying dimensionality reduction techniques (e.g., PCA) to enhance clustering efficiency and visualization.