Clustering Results Report

1. Overview

The goal of this analysis was to segment customers based on their profile and transaction data using clustering techniques. The dataset included customer demographic information, transaction summaries, and regional details. Multiple clustering algorithms were employed to find optimal customer segments, and metrics such as Davies-Bouldin Index (DB Index) and Silhouette Score were calculated to evaluate clustering performance.

2. Clustering Algorithms and Results

2.1 K-Means Clustering

- **Optimal Number of Clusters: 10**
- **Davies-Bouldin Index: 1.160765545874279**
- **Silhouette Score: 0.62**
- **Insights:**
    1. K-Means identified distinct clusters based on spending behaviour and transaction patterns.
    2. Customers were grouped based on high, medium, and low spending levels, along with transaction frequency.
- **Visualization:** Scatter plots of TotalSpending vs. AvgSpendingPerTransaction and TotalTransactions vs. TotalSpending showed well-separated clusters.

2.2 DBSCAN (Density-Based Spatial Clustering of Applications with Noise)

- **Optimal Parameters:**
- **Epsilon (eps): 0.5**
- **Minimum Samples: 4**
- **Davies-Bouldin Index: 1.241536574614803**
- **Silhouette Score: -0.18992767921115777**
- **Number of Clusters (Excluding Noise): 2**

- **Insights:**
    1. DBSCAN effectively identified dense regions of customers with similar spending patterns.
    2. Some customers were classified as noise due to sparse transaction data.
- **Visualization:** Scatter plot highlighted well-defined clusters with a few outliers.

2.3 Gaussian Mixture Models (GMM)

- **Optimal Number of Clusters: 7**
- **Davies-Bouldin Index: 1.2634192667866453**
- **Silhouette Score: 0.25410915018650754**
- **Insights:**
    1. GMM provided a probabilistic approach, assigning customers to clusters based on likelihood.
    2. Clusters showed clear differences in spending levels and transaction counts.
- **Visualization:** Bar plots of DB Index and Silhouette Scores for varying cluster numbers demonstrated GMM's effectiveness.

2.4 Hierarchical Clustering (Agglomerative)

- **Optimal Number of Clusters: 7**
- **Davies-Bouldin Index: 1.2634192667866453**
- **Silhouette Score: 0.25410915018650754**
- **Insights:**
    1. Hierarchical clustering grouped customers with similar spending patterns in a tree-like structure.
    2. Dendrograms provided insights into cluster formation at different levels.
- **Visualizations:** Scatter plots of clusters revealed hierarchical relationships.

3. Comparison of Clustering Metrics

Algorithm	Optimal clusters	DB Index	Silhouette Score
K-Means	**10**	**1.160**	**0.62**
DBSCAN	**2**	**1.241**	**-0.189**
Guassian Mixture	**7**	**1.263**	**0.254**
Hierarchical	**7**	**1.263**	**0.254**

4.Key Findings

1. **K-Means Clustering:**
    - Optimal Clusters: 10
    - DB Index: 1.160 (lowest among the algorithms, indicating better cluster compactness and separation compared to others).
    - Silhouette Score: 0.62 (indicates moderately well-separated clusters).
    - Insight: K-Means performed the best among all algorithms, effectively grouping customers into distinct segments. The higher number of clusters allowed for finer segmentation.
2. **DBSCAN (Density-Based Spatial Clustering of Applications with Noise):**
    - Optimal Clusters: 2
    - DB Index: 1.241 (higher, indicating less compact and well-separated clusters).
    - Silhouette Score: -0.189 (indicates poorly defined clusters or overlapping data).
    - Insight: DBSCAN struggled to identify meaningful clusters due to the nature of the data. Most points

were either classified as noise or grouped into a few broad clusters.

3. **Gaussian Mixture Models (GMM):**
    - Optimal Clusters: 7
    - DB Index: 1.263 (higher, indicating less compact clusters).
    - Silhouette Score: 0.254 (indicates weak separation between clusters).
    - Insight: GMM provided overlapping clusters, which may not clearly define customer groups. This result suggests that GMM was less effective for this dataset.
4. **Hierarchical Clustering:**
    - Optimal Clusters: 7
    - DB Index: 1.263 (same as GMM, indicating similar performance).
    - Silhouette Score: 0.254 (indicates weakly defined clusters).
    - Insight: Hierarchical clustering showed similar performance to GMM, suggesting that the data may not naturally form well-separated hierarchical clusters.

5.Recommendations

1. **Use K-Means Clustering for Segmentation:**
    - K-Means produced the lowest DB Index and highest Silhouette Score, indicating better-defined clusters. It is the most reliable algorithm for this dataset.
    - Utilize the 10 identified clusters to create targeted marketing strategies:

        - High-Spending Clusters: Offer loyalty programs and exclusive discounts.
        - Low-Spending Clusters: Introduce promotions to increase spending.
        - Medium-Spending Clusters: Focus on upselling and cross-selling opportunities.

2. **Reassess DBSCAN and GMM Suitability:**
    - DBSCAN and GMM did not perform well, likely due to the data's distribution. These algorithms may require additional preprocessing or feature engineering to improve results.
    - Consider these methods only if specific domain insights suggest the presence of overlapping or irregularly shaped clusters.
3. **Feature Refinement:**
    - Explore additional features, such as customer lifetime value, churn probability, or product preferences, to enhance clustering performance.
4. **Hierarchical Clustering for Validation:**
    - Use hierarchical clustering as a secondary method to validate K-Means results or to explore relationships between clusters.