Customer Segmentation

Data Preparation

The data was pre-processed and aggregated at the customer level to create a comprehensive view of their spending behavior. We used the following columns:

- **TotalValue**: The total amount spent by a customer.
- **Quantity**: The number of items purchased.
- **Region**: The most frequent region for each customer.

We then performed one-hot encoding on the **Region** column to convert it into numerical features for clustering and scaled the features to standardize the data.

Clustering Logic and Deliverables

The customer data was grouped into 5 clusters using **KMeans** clustering. The logic behind the formation of these clusters is based on the combination of customer spending behavior (total value and quantity) and the region where the customer is located.

Number of Clusters Formed

The optimal number of clusters was determined through the **Elbow Method**. After evaluating the within-cluster sum of squares (WCSS) for different values of **k**, we observed that **5 clusters** offered the best balance between compactness and separation, making it the most appropriate choice for segmentation.

The Logic Behind the 5 Clusters

These 5 clusters represent distinct groups of customers with varying behaviors:

1. **High-Value, Frequent Shoppers (Cluster 1)**
   - **Behavior**: These customers spend large amounts and purchase frequently, indicating that they are likely to be loyal, high-value customers.
   - **Characteristics**: High total spending with a relatively high quantity of items purchased.
   - **Marketing Insight**: Target these customers with loyalty programs or premium product offers.
2. **Moderate-Value, Occasional Shoppers (Cluster 2)**

   - **Behavior**: Customers in this group spend moderately and make purchases less frequently.
   - **Characteristics**: Moderate total value and purchase quantity, with occasional large purchases.
   - **Marketing Insight**: Ideal candidates for targeted campaigns aimed at increasing shopping frequency.

3. **Low-Value, Frequent Small Purchasers (Cluster 3)**
   - **Behavior**: These customers make small, frequent purchases, resulting in lower total value but higher purchase frequency.
   - **Characteristics**: Low total spending with a relatively high frequency of smaller purchases.
   - **Marketing Insight**: Encourage these customers with promotions for small-ticket items or frequent-buyer incentives.
4. **High-Value, Low-Quantity Shoppers (Cluster 4)**
   - **Behavior**: This group includes customers who make fewer but high-value purchases, possibly buying premium or luxury products.
   - **Characteristics**: High total spending but low purchase quantity, suggesting a preference for high-ticket items.
   - **Marketing Insight**: Market high-end, luxury products to these customers or offer personalized services.
5. **Low-Value, Infrequent Shoppers (Cluster 5)**
   - **Behavior**: These customers make infrequent purchases with low total value.
   - **Characteristics**: Low spending with low frequency of purchases, possibly new or occasional customers.
   - **Marketing Insight**: Use re-engagement strategies or seasonal promotions to convert these customers into more frequent buyers.

Clustering Techniques

Several clustering algorithms are used to segment customers:

1. K-Means Clustering

K-Means is a popular partitioning method that divides data into a predefined number of clusters. It works by assigning each data point to the nearest cluster centroid and then recalculating the centroids iteratively.

To determine the optimal number of clusters (K), the **Elbow Method** is used. The point where the sum of squared distances to centroids starts to level off is considered the ideal number of clusters.

Deliverables for K-Means:

- **Number of Clusters Formed**: The Elbow Method is applied to find the optimal K value. In this case, K=5 was chosen as the ideal number of clusters after observing that the Elbow point occurred at this value. This means that the variance explained by the clusters begins to plateau after 5 clusters.
- **Logic Behind the Number of Clusters**: The Elbow Method reveals the point at which the addition of more clusters results in minimal improvement in the clustering performance. Choosing 5 clusters allows for a balance between model simplicity and performance.

Evaluation Metrics for K-Means:

- **Silhouette Score**: Measures how similar an object is to its own cluster compared to other clusters. A higher score indicates better-defined clusters.
- **Davies-Bouldin Index**: Measures the average similarity ratio of each cluster with the cluster that is most similar to it. A lower score indicates better clustering.

2. Hierarchical Clustering

Hierarchical clustering builds a tree of clusters, which is visualized using a **Dendrogram**. It allows the user to decide the number of clusters by cutting the tree at a desired level.

- **Number of Clusters Formed**: The number of clusters is determined from the dendrogram. Based on the observation, 5 clusters were selected as optimal by cutting the tree at an appropriate level.
- **Logic Behind the Number of Clusters**: The dendrogram shows the hierarchical relationships between customers. By cutting the tree at a level where the clusters have well-defined separations, we ensure that the clusters are meaningful.

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

DBSCAN is a density-based algorithm that does not require the number of clusters to be specified. It identifies clusters of varying shapes and can also detect outliers.

- **Number of Clusters Formed**: The number of clusters is not predefined. DBSCAN groups data points based on density and can label some points as outliers (assigned a label of -1).
- **Logic Behind the Number of Clusters**: The algorithm uses two parameters: **eps** (maximum distance between two samples for them to be considered as in the same neighborhood) and **min_samples** (the number of samples in a neighborhood for a point to be considered a core point). By adjusting these parameters, DBSCAN forms clusters based on density rather than a fixed number of clusters.

4. Gaussian Mixture Model (GMM)

GMM is a probabilistic model that assumes the data is generated from a mixture of several Gaussian distributions. It provides a soft assignment of data points to clusters, meaning that each point can belong to multiple clusters with different probabilities.

- **Number of Clusters Formed**: 5 clusters were selected after evaluating the **AIC (Akaike Information Criterion)** and **BIC (Bayesian Information Criterion)** scores. These criteria penalize the likelihood for having more components, ensuring a balance between fit and complexity.
- **Logic Behind the Number of Clusters**: The optimal number of components (clusters) is selected based on the lowest AIC and BIC scores. In this case, 5 components were identified as the optimal choice.

Cluster Comparison

The performance of each clustering method is evaluated using the following metrics:

- **Silhouette Score**: This measures how well-defined the clusters are. Higher values indicate better-defined clusters.
- **Davies-Bouldin Index**: A lower value indicates better separation between clusters.

The following clustering techniques were compared:

- **K-Means**: Moderate silhouette score and Davies-Bouldin index.
- **DBSCAN**: Lower silhouette score, indicating weaker clustering quality compared to other methods.

- **GMM**: Similar silhouette score to K-Means but with a higher Davies-Bouldin index, indicating less clear separation.
- **Hierarchical**: Comparable performance to K-Means, with slightly better cluster separation in some cases.

Conclusion

By identifying and analyzing these 5 customer segments, we can better understand customer behaviors and tailor marketing strategies accordingly. Each cluster has unique characteristics that suggest how different types of customers interact with the business, allowing for more targeted and effective engagement strategies.