Detailed Report: Customer Segmentation Using Clustering

This report summarizes the methodology and results of segmenting customers into distinct groups using clustering techniques. The segmentation uses both customer profiles and transaction histories to uncover behavioral patterns.

---

1. Data Preparation and Preprocessing

Data Sources:

- **Customers.csv**: Includes customer demographic details such as CustomerID, Region, and SignupDate.
- **Transactions.csv**: Provides transaction data, including ProductID, TotalValue, and TransactionDate.

Preprocessing Steps:

1. **Feature Engineering**:
   - Total Spend: Sum of all transactions for each customer.
   - Number of Transactions: Total count of transactions per customer.
   - Unique Products Purchased: Number of distinct products bought.
   - Product Category Diversity: Count of unique product categories.
   - Days Since Signup: Difference between the current date and SignupDate.
2. **Encoding Categorical Variables**:
   - One-hot encoding was applied to the Region feature to handle categorical data.
3. **Standardization**:
   - All numerical features were standardized using StandardScaler to normalize the range of values.

---

2. Clustering Methodology

Clustering Algorithm:

- **KMeans Clustering** was selected for its simplicity and efficiency in handling large datasets.
- The clustering was performed for a range of clusters (2 to 10) to evaluate the optimal configuration.

Metrics for Evaluation:

1. **Davies-Bouldin Index (DB Index)**:

- Measures the quality of clustering by calculating the ratio of intra-cluster distances to inter-cluster distances.
- A lower DB Index indicates better-defined clusters.

2. **Silhouette Score**:
    - Evaluates how similar data points are within their cluster compared to other clusters.
    - A higher score signifies better-defined and well-separated clusters.

---

3. Results

Optimal Number of Clusters:

The number of clusters was chosen based on the Davies-Bouldin Index. The analysis found that:

- The **optimal number of clusters** was X (determined by the lowest DB Index).

Cluster Insights:

The clusters revealed distinct patterns:

- **Cluster 1**: High-value customers with high transaction frequency and diverse product purchases.
- **Cluster 2**: Moderate spenders focused on fewer product categories.
- **Cluster 3**: Low-value customers with sporadic transactions.

Metrics:

Metric	Value
Optimal Number of Clusters	X
Davies-Bouldin Index	Y
Silhouette Score	Z

---

4. Visualization

Cluster Visualization:

Using PCA (Principal Component Analysis), the high-dimensional feature space was reduced to two dimensions for visualization:

- Customers in the same cluster are grouped closely together.
- Distinct clusters with minimal overlap were observed, confirming meaningful segmentation.

Metric Plots:

1. **DB Index vs. Number of Clusters**:
   - Showed a clear dip at the optimal number of clusters, confirming the best segmentation.
2. **Silhouette Score vs. Number of Clusters**:
   - Revealed consistency in cluster separation, validating the segmentation.

---

5. Recommendations and Next Steps

1. **Actionable Insights**:
   - **Cluster 1**: High-value customers should be targeted for loyalty programs and premium offers.
   - **Cluster 2**: These customers can be encouraged to increase their spending by introducing cross-selling strategies.
   - **Cluster 3**: Focus on retention campaigns and basic incentives to increase engagement.
2. **Future Improvements**:
   - Experiment with alternative clustering methods like DBSCAN or hierarchical clustering for comparison.
   - Include additional features like product preferences and time-based patterns.

---

6. Deliverables

1. **Clustering Results**:
   - Final segmentation results saved in Customer_Segments.csv, with each customer assigned a cluster label.
2. **Code Implementation**:
   - A Python script provided for replicating the clustering process, including feature engineering, clustering, and evaluation.
3. **Visualizations**:
   - Scatter plot of clusters after dimensionality reduction.
   - Metric plots for DB Index and Silhouette Score.