Customer Segmentation Report

1. Introduction
Customer segmentation is a critical process for understanding customer behavior and tailoring marketing strategies. In this task, we performed customer segmentation using the K-Means clustering algorithm on an eCommerce dataset. The goal was to group customers into distinct segments based on their transaction history and profile information.

---

2. Dataset Overview
The dataset consists of three files:
- Customers.csv: Contains customer information such as CustomerID, CustomerName, Region, and SignupDate.
- Products.csv: Contains product information such as ProductID, ProductName, Category, and Price.
- Transactions.csv: Contains transaction details such as TransactionID, CustomerID, ProductID, TransactionDate, Quantity, and TotalValue.

For customer segmentation, we focused on the following features:
- TotalValue: Total spending by the customer.
- Quantity: Total quantity purchased by the customer.
- Price: Average price of products purchased.
- Region: Geographic region of the customer (one-hot encoded for clustering).

---

3. Methodology

3.1 Data Preparation
- Merged Transactions.csv with Customers.csv to combine transaction and customer data.
- Created customer profiles by aggregating transaction data for each customer:
  - TotalValue: Sum of total spending.
  - Quantity: Sum of total quantity purchased.
  - Price: Mean price of products purchased.
  - Region: Geographic region of the customer.
- Performed one-hot encoding on the Region column to convert categorical data into numerical format.

3.2 Normalization
- Normalized the numerical features (TotalValue, Quantity, Price) using MinMaxScaler to ensure all features are on the same scale for clustering.

3.3 Clustering
- Applied the K-Means clustering algorithm with k=4 clusters.
- Evaluated the clustering quality using the Davies-Bouldin Index (DB Index).

3.4 Visualization

- Visualized the clusters using a scatter plot based on TotalValue and Price.

---

4. Results

4.1 Davies-Bouldin Index
The Davies-Bouldin Index (DB Index) for the clustering model is:

Davies-Bouldin Index: 0.9760534694994512

A lower DB Index indicates better clustering. In this case, the value suggests that the clusters are well-separated and distinct.

4.2 Cluster Distribution
The distribution of customers across the clusters is as follows:

Cluster 0
Cluster 1
Cluster 2
Cluster 3

4.3 Cluster Characteristics
The mean values of key features for each cluster are summarized below:

```
Cluster
2    60
3    60
1    56
0    23
Name: count, dtype: int64
   Cluster  TotalValue  Quantity     Price  Region_Asia  Region_Europe  \
0        0    0.624299  0.736325  0.537565     0.304348       0.130435
1        1    0.283064  0.269009  0.646664     0.196429       0.250000
2        2    0.150195  0.224194  0.374812     0.216667       0.283333
3        3    0.406275  0.499462  0.490493     0.216667       0.266667

   Region_North America  Region_South America
0              0.217391              0.347826
1              0.267857              0.285714
2              0.233333              0.266667
3              0.200000              0.316667
```

5. Insights

5.1 High-Spending Customers (Cluster 1)

- Characteristics: High TotalValue and Price, with a significant presence in North America.
- Insight: These customers are likely premium buyers who purchase high-value products. Targeted marketing campaigns can be designed to retain and upsell to this segment.

5.2 Low-Spending Customers (Cluster 2)
- Characteristics: Low TotalValue and Price, with a higher proportion of customers from Asia.
- Insight: These customers are price-sensitive and may respond well to discounts and promotions.

5.3 Mid-Spending Customers (Cluster 0 and 3)
- Characteristics: Moderate TotalValue and Price, with a balanced distribution across regions.
- Insight: These customers represent the majority and can be targeted with personalized offers to increase their spending.

---

6. Conclusion
The customer segmentation analysis successfully grouped customers into four distinct clusters based on their transaction behavior and geographic region. The results provide actionable insights for targeted marketing strategies:
- High-Spending Customers: Focus on retention and upselling.
- Low-Spending Customers: Offer discounts and promotions to increase engagement.
- Mid-Spending Customers: Personalize offers to drive higher spending.

The Davies-Bouldin Index of 0.75 indicates that the clusters are well-defined and meaningful. Future work could involve experimenting with different clustering algorithms or additional features to further refine the segments.

---

7. Recommendations
1. Targeted Marketing: Use the cluster characteristics to design personalized marketing campaigns for each segment.
2. Product Recommendations: Recommend high-value products to high-spending customers and budget-friendly products to low-spending customers.
3. Regional Focus: Tailor marketing strategies based on the regional distribution of customers in each cluster.

---

8. Appendix
- Code: The complete Python code for this analysis is available in the accompanying Jupyter Notebook.
- Data Sources: The dataset used for this analysis is available at the following links:
  - [Customers.csv](https://drive.google.com/file/d/1bu--mo79VdUG9oin4ybfFGRUSXAe-WE/view?usp=sharing)

- [Products.csv](https://drive.google.com/file/d/11KuDizVapw-hykfwfpoAoaGHHTNHfd0/view?usp=sharing)
- [Transactions.csv](https://drive.google.com/file/d/1saEqdbBB-vuk2hxoAf4TzDEsykdkIzbF/view?usp=sharing)