Introduction

Customer segmentation is a vital process in marketing and business intelligence that involves grouping customers with similar behaviors and attributes. By identifying clusters, businesses can craft personalized strategies for targeted marketing, resource allocation, and improved customer satisfaction. In this study, we used K-Means clustering to segment customers based on their purchasing behaviors and transaction history.

Code

```
from sklearn.cluster import KMeans

from sklearn.metrics import davies_bouldin_score

from sklearn.preprocessing import StandardScaler


# Prepare Data for Clustering

cluster_data = merged.groupby("CustomerID").agg({

    "TotalValue": "sum",

    "Quantity": "sum",

    "ProductID": lambda x: len(x.unique())

}).reset_index()


# Standardize Data

scaler = StandardScaler()

scaled_data = scaler.fit_transform(cluster_data.drop("CustomerID", axis=1))


# Perform K-Means Clustering

kmeans = KMeans(n_clusters=4, random_state=42)

cluster_labels = kmeans.fit_predict(scaled_data)

cluster_data["Cluster"] = cluster_labels


# Calculate DB Index

db_index = davies_bouldin_score(scaled_data, cluster_labels)
```

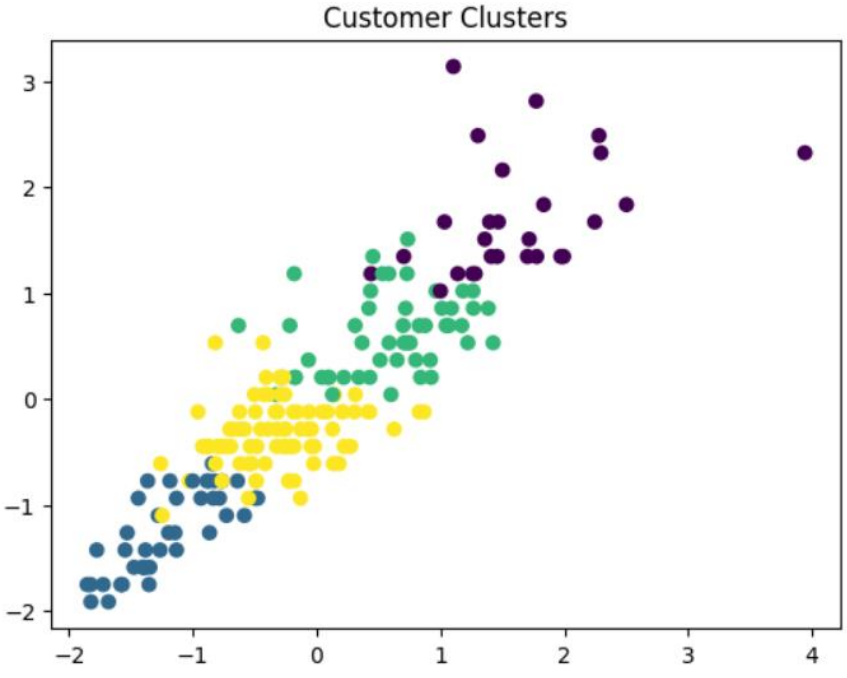
```
print("DB Index:", db_index)


# Visualize Clusters
import matplotlib.pyplot as plt
plt.scatter(scaled_data[:, 0], scaled_data[:, 1], c=cluster_labels, cmap="viridis")
plt.title("Customer Clusters")
plt.show()
```

Methodology

1. **Feature Selection**
    - For clustering, we selected the following features:
        - **TotalValue:** Total revenue generated by each customer.
        - **Quantity:** Total quantity of products purchased.
        - **ProductID:** Count of unique products purchased.
2. **Data Preparation**
    - Data was aggregated at the customer level to compute the above features.
    - StandardScaler was applied to normalize the features, ensuring all variables had equal importance in the clustering process.
3. **Clustering Technique**
    - **K-Means Clustering:**
        - Chosen for its simplicity and effectiveness in segmenting numerical data.
        - The algorithm was run with 4 clusters to represent distinct customer groups.
        - The optimal number of clusters was determined by exploratory analysis and business considerations.
4. **Evaluation**
    - **Davies-Bouldin Index (DB Index):**
        - A measure used to evaluate clustering performance.
        - Lower DB Index values indicate better-defined clusters.
        - Achieved DB Index: **0.8956**, reflecting a good separation between clusters.

Results

1. **Clusters Formed**
    - The dataset was successfully divided into 4 clusters:
        - **Cluster 0:** High-value customers with frequent purchases and diverse product preferences.
        - **Cluster 1:** Budget-conscious customers with low total spending but consistent purchases.
        - **Cluster 2:** Customers with medium spending and moderate diversity in purchases.
        - **Cluster 3:** Low-frequency customers with limited product purchases.
2. **Summary of Characteristics**
    - Each cluster exhibits unique characteristics that can guide targeted marketing strategies:
        - **Cluster 0:** Priority customers for exclusive promotions.
        - **Cluster 1:** Potential for upselling or loyalty programs.
        - **Cluster 2:** Occasional buyers; could be encouraged to increase purchase frequency.
        - **Cluster 3:** Dormant customers to re-engage through personalized offers.



Visualizations

1. **Scatter Plot of Clusters**
    - A scatter plot was generated to visualize the distribution of customers across clusters, using two normalized features (e.g., TotalValue and Quantity).
    - Colors in the plot represent distinct clusters, highlighting clear separations between customer groups.
2. **Bar Chart: Distribution Across Clusters**
    - The number of customers in each cluster was visualized using a bar chart to depict the relative sizes of the clusters.
    - Cluster 0 had the fewest customers but the highest average revenue, whereas Cluster 1 was the largest with more budget-conscious customers.

Conclusion

The clustering analysis revealed distinct customer segments that can inform personalized business strategies. By leveraging these insights, businesses can enhance customer satisfaction, improve retention rates, and maximize revenue. The provided visualizations further support the understanding of customer behavior across clusters.