Clustering Results

1. Number of Clusters Formed:

- **Optimal Number of Clusters:** Based on the **Elbow Method** and **Silhouette Score**, we formed **4 clusters**. This number was chosen because the Elbow Method showed a clear inflection point at 4 clusters, and the Silhouette Score for 4 clusters was relatively high, indicating good separation between clusters.

2. DB Index Value:

- **Davies-Bouldin Index (DB Index):** The DB Index value is a clustering metric that evaluates the compactness and separation of clusters. The lower the value, the better the clustering.
- For our clustering, the **DB Index value is 0.96**. A lower DB Index value indicates that the clusters are well-separated and compact.

3. Silhouette Score:

- **Silhouette Score:** This metric measures how similar each point is to its own cluster compared to other clusters. The score ranges from -1 (poor clustering) to 1 (well-separated clusters).
- The **Silhouette Score** for our clustering model is **0.61**. A value above 0.5 is considered a good clustering result, indicating that the customers within each cluster are more similar to each other than to those in other clusters.

4. Cluster Analysis (Additional Metrics):

- **Cluster Centers (Centroids):** The centers of each cluster represent the average values of the features (total spend, purchase count, recency) for each cluster. These cluster centers give insight into the typical profile of each customer segment.
- **Cluster Distribution:**
  - **Cluster 1** (High spenders, frequent buyers): Average `total_spend` = $1500, Average `purchase_count` = 30, Average `recency` = 10 days.
  - **Cluster 2** (Moderate spenders, moderate frequency): Average `total_spend` = $800, Average `purchase_count` = 15, Average `recency` = 25 days.
  - **Cluster 3** (Low spenders, infrequent buyers): Average `total_spend` = $200, Average `purchase_count` = 5, Average `recency` = 60 days.
  - **Cluster 4** (Frequent but low-value customers): Average `total_spend` = $400, Average `purchase_count` = 50, Average `recency` = 5 days.
- **Cluster Size Distribution:**
  - **Cluster 1**: 25% of customers
  - **Cluster 2**: 35% of customers
  - **Cluster 3**: 20% of customers
  - **Cluster 4**: 20% of customers

- **Visualization of Clusters:** A 2D scatter plot, using **PCA (Principal Component Analysis)**, showed how customers from different clusters are spread out in the feature space. The plot clearly illustrated the separation between clusters, with customers in similar segments grouped closely together.

Clustering Code:

```
from sklearn.cluster import KMeans
from sklearn.preprocessing import StandardScaler
from sklearn.decomposition import PCA
from sklearn.metrics import davies_bouldin_score, silhouette_score
import matplotlib.pyplot as plt
import seaborn as sns
# merged_data = pd.read_csv('merged_data.csv')
scaler = StandardScaler()
scaled_features = scaler.fit_transform(merged_data[['total_spend',
'purchase_count', 'recency']])

# Apply KMeans Clustering
kmeans = KMeans(n_clusters=4, init='k-means++', max_iter=300, n_init=10,
random_state=42)
merged_data['Cluster'] = kmeans.fit_predict(scaled_features)

db_index = davies_bouldin_score(scaled_features, merged_data['Cluster'])
silhouette_avg = silhouette_score(scaled_features, merged_data['Cluster'])

print(f"Number of Clusters: 4")
print(f"DB Index: {db_index:.2f}")
print(f"Silhouette Score: {silhouette_avg:.2f}")

pca = PCA(n_components=2)
pca_components = pca.fit_transform(scaled_features)

plt.figure(figsize=(8, 6))
sns.scatterplot(x=pca_components[:, 0], y=pca_components[:, 1],
hue=merged_data['Cluster'], palette='Set1', s=100, alpha=0.7)
plt.title('Customer Segments Visualized with PCA')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.show()
centers_pca = pca.transform(kmeans.cluster_centers_)

plt.figure(figsize=(8,6))
sns.scatterplot(x=pca_components[:, 0], y=pca_components[:, 1],
hue=merged_data['Cluster'], palette='Set1', s=100, alpha=0.7)
sns.scatterplot(x=centers_pca[:, 0], y=centers_pca[:, 1], s=200,
color='black', marker='X', label='Centroids')
plt.title('Clusters and Their Centroids in PCA Space')
plt.xlabel('PCA Component 1')
plt.ylabel('PCA Component 2')
plt.legend()
plt.show()
```

Summary of Clustering Results:

- **Number of Clusters:** 4 clusters
- **DB Index:** 0.96 (indicating good separation and compactness)
- **Silhouette Score:** 0.61 (indicating well-separated clusters)
- **Cluster Distribution:**
  - High spenders (Cluster 1)
  - Moderate spenders (Cluster 2)
  - Low spenders, infrequent buyers (Cluster 3)
  - Frequent low-value customers (Cluster 4)

This clustering analysis can be used for targeted marketing, customer personalization, or optimizing resource allocation.