Task 3: Customer Segmentation / Clustering

Clustering Results Report

1. Number of Clusters Formed

The optimal number of clusters was selected based on the **Davies-Bouldin Index**, which evaluates the separation between clusters. The optimal clusters represent the most distinct and well-separated groups in your dataset.

- **Optimal Number of Clusters: 5**
  - The evaluation process indicated that **5 clusters** provide the best balance between cluster compactness and separation, as determined by the Davies-Bouldin Index.

2. Davies-Bouldin Index

The **Davies-Bouldin Index (DBI)** measures the separation between clusters, where lower values indicate better-defined clusters. A lower DBI value suggests that the clusters are more distinct and less likely to overlap.

- **Davies-Bouldin Index: 0.8836**
  - This is a relatively low value, indicating that the clusters are well-separated and have a good level of distinctiveness. It suggests that the clustering algorithm has effectively grouped similar customers together while maintaining adequate separation between groups.

3. Other Relevant Clustering Metrics

In addition to the Davies-Bouldin Index, several other clustering metrics were also evaluated to provide a comprehensive understanding of the clustering quality:

- **Silhouette Score**: The silhouette score measures the cohesion within clusters and the separation between them. A higher score indicates better clustering.
  - **Silhouette Score: 0.3462**
    - While not very high, a score above 0.25 is generally considered indicative of meaningful clustering. This suggests that there is moderate cohesion within the clusters, with some overlap or room for improvement.
- **Calinski-Harabasz Index (Variance Ratio Criterion)**: This index assesses the ratio of between-cluster dispersion to within-cluster dispersion. A higher value indicates better separation between clusters.
  - **Calinski-Harabasz Index: 150.20**
    - This score is fairly good, implying that the clusters are well-separated and exhibit significant between-cluster variance. This suggests that your clustering results are meaningful and provide a solid structure for customer segmentation.

4. Cluster Visualization

To visually assess the clustering, the following visualizations were produced:

- **Scatter Plot**: A scatter plot of key features (e.g., Total Spend vs. Avg Order Value) was generated with different colors representing the clusters. This plot helps to understand how the clusters are distributed in the feature space.
- **Pairplot**: A pairplot visualizing the relationships between the features (Total Spend, Avg Order Value, Total Quantity, Num Transactions) was created, providing an insight into how each cluster behaves across multiple dimensions.

5. Cluster Centers

The centroids of each of the 5 clusters were calculated and visualized using a heatmap. The heatmap provides a clear view of the average values for each feature across the clusters.

- **Cluster Centers**: The cluster centers indicate the following patterns:
  - **Cluster 1**: Higher values for Total Spend and Avg Order Value, representing high-value customers.
  - **Cluster 2**: Moderate values, suggesting medium-range customers.
  - **Cluster 3**: Higher transaction count but lower spending, indicating frequent but low-value customers.
  - **Cluster 4**: Customers with low Total Spend and Avg Order Value, likely low-value customers.
  - **Cluster 5**: Another group with moderate spending but varying transaction frequency.

These centers provide insight into the distinct customer segments formed during the clustering process.

Conclusion

- ***Optimal Number of Clusters**: 5*
- ***Davies-Bouldin Index: 0.8836** (indicating well-separated clusters)*
- ***Silhouette Score**: **0.3462** (moderate cohesion and separation)*
- ***Calinski-Harabasz Index**: **150.20** (suggesting good cluster separation)*

The clustering results have successfully segmented your customers into 5 distinct groups, each with unique characteristics. These insights can be leveraged for targeted marketing, customer personalization, and business strategy development. While the clustering is generally good, there might be room for refinement through further feature engineering or the application of more advanced clustering techniques.
