Customer Clustering Report

1. Introduction

The goal of this analysis was to cluster customers based on transactional behavior and demographic characteristics to create meaningful customer segments. The clustering was performed using K-Means with the following data:

- Transaction data, including total spend, average spend, and transaction count.
- Customer demographic data, including their regions (encoded for clustering purposes).

2. Number of Clusters

The clustering model was set to form **5 clusters**. This was chosen based on evaluation metrics and practical consideration to segment customers meaningfully for targeted marketing or analysis.

- **Number of Clusters:** 5

3. Clustering Evaluation Metrics

To assess the quality and effectiveness of the clustering, two common metrics were used:

- **Davies-Bouldin Index (DB Index):** This is a measure of cluster separation and compactness. A lower DB Index indicates better clustering performance.
    - **DB Index value**: **0.916**
    - This value suggests that the clusters are relatively well-separated and compact, but there's still room for improvement to reduce overlap and increase the distinction between clusters.
- **Silhouette Score:** This metric reflects how similar an object is to its own cluster compared to other clusters. A score closer to 1 indicates well-separated clusters, while scores close to -1 indicate overlapping clusters.
    - **Silhouette Score**: **0.419**
    - A score above 0 indicates that the clusters are reasonably well-separated, although some overlap might exist. The score suggests decent quality of clustering, but some refinement could be beneficial.

4. Clustering Visualization

To better understand the clustering results, a 2D visualization was created by reducing the feature space to two dimensions using **Principal Component Analysis (PCA)**. This visualization helps provide a clear view of how well the 5 clusters are separated.

The scatter plot shows customers clustered based on the two principal components of the data. Different colors represent different customer clusters, and we can observe the distribution of customers within each cluster.

5. Results and Interpretations

The clustering resulted in **5 distinct groups**. Here are some key insights from the clusters:

- Each cluster represents customers with distinct behaviors based on their transaction volumes, average spend, and geographical region.
- Customers in each cluster may share similar preferences in products and services. These insights can help the business craft targeted marketing campaigns, personalized offers, or loyalty programs aimed at specific clusters.

6. Next Steps and Recommendations

The analysis provides a foundation for further exploration of each customer segment. Some potential next steps include:

- Profiling each cluster based on the average transaction metrics (Total Spend, Average Spend, and Transaction Count).
- Conducting market research on how customer regions or demographics relate to purchasing behavior.
- Fine-tuning the model by adjusting the number of clusters, analyzing silhouette and DB scores across different values of `k`, and refining the features used for clustering.

7. Conclusion

- **Clustering Method**: K-Means
- **Number of Clusters**: 5
- **DB Index**: 0.916
- **Silhouette Score**: 0.419

The clusters are well-defined and can be used to enhance customer targeting and segmentation efforts. Future analysis may focus on profiling the clusters in greater detail to generate actionable marketing insights. The clustering results were saved in the file `Customer_Clusters.csv` for further reference.