Clustering Results

1. Number of Clusters Formed

- After analyzing different numbers of clusters (from 2 to 10), the optimal number of clusters was determined to be **8**, based on the Davies-Bouldin (DB) Index, which is used to evaluate clustering quality.

2. Clustering Metrics

- **Davies-Bouldin Index (DB Index)**:
    - The DB Index value for the optimal clustering was **0.78**, indicating that clusters are well-separated and compact.
- **Clustering Methodology**:
    - **Algorithm Used**: KMeans clustering.
    - **Cluster Range Tested**: 2 to 10 clusters.
    - The DB Index values were calculated for each cluster count, and the optimal number (8 clusters) was chosen based on the lowest DB Index.

3. Feature Engineering

- **Features Derived**:
    - Total Spending (sum of total transaction values per customer).
    - Average Spending (mean transaction value).
    - Total Quantity Purchased.
    - Most Purchased Category (mode of the product category for each customer).
    - Days Since Signup (calculated as the difference between the current date and the customer signup date).
- **Data Preprocessing**:
    - Categorical variables such as Region and Most Purchased Category were one-hot encoded to prepare for clustering.
    - Numerical variables were normalized using MinMaxScaler to ensure all features contributed equally.

4. Clustering Process

- The aggregated and preprocessed customer data formed a **feature matrix**.
- KMeans was applied, and customers were grouped into 8 distinct clusters.
- Each cluster represents a unique customer segment based on spending patterns, purchase behavior, and demographics.

5. Visualization

- To interpret the clusters, the feature matrix was reduced to two dimensions using **Principal Component Analysis (PCA)**.
- The clusters were visualized in a 2D scatter plot, showcasing distinct groupings of customers, which validated the clustering approach.

6. Business Implications

- **Customer Segmentation**:
    - Clusters can be analyzed individually to design tailored marketing campaigns.
    - For example, high-spending customers can be offered premium products or loyalty rewards.
- **Targeted Promotions**:
    - Products in the most purchased category of each cluster can be promoted to customers with similar purchasing patterns.
- **Regional Insights**:
    - Clustering also revealed regional spending trends, allowing businesses to align inventory with regional demands.

7. Data and Tools Used

- **Data Sources**:
    - Customers.csv, Products.csv, and Transactions.csv files were merged for a comprehensive view.
- **Tools**:
    - Python libraries such as pandas, numpy, scikit-learn, and seaborn were used for analysis, clustering, and visualization.

8. Limitations

- The clustering results depend heavily on the quality and quantity of data. Any missing or inaccurate information in the dataset could affect the clustering quality.
- Dimensionality reduction through PCA may lead to some loss of information while visualizing clusters.