REPORT ON CUSTOMER SEGMENTATION USING KMEANS CLUSTERING

Objective:

The goal of this analysis is to segment customers into distinct groups based on their purchasing behavior using transaction data. The KMeans clustering algorithm is applied to group customers, and clustering performance is evaluated using the Davies-Bouldin Index (DB Index) and Silhouette Score. The results are visually represented using PCA (Principal Component Analysis) to reduce dimensionality and plot the customer clusters.

Data Overview:

- **Customers.csv**: Contains demographic information about customers, including CustomerID, Region, and SignupDate.
- **Transactions.csv**: Contains transaction data with CustomerID, ProductID, Category, and TotalValue (spending per product).
- **Products.csv** (used during merge): Contains ProductID and Category.

Preprocessing Steps:

1. **Data Cleaning**:
    - Stripped extra spaces from column names in the Transactions dataset.
    - Renamed columns for clarity (Category_x to Category).
2. **Merging Data**:
    - Merged product category information into the Transactions dataset based on ProductID.
    - Aggregated transaction data by CustomerID and Category using TotalValue as the aggregation metric (i.e., total spending per category).
3. **Feature Engineering**:
    - Created a new feature SignupDays, representing the number of days since each customer signed up, calculated as the difference between the current date and the SignupDate.
4. **Handling Missing Data**:
    - Missing values in the customer transaction data were filled with zeros to ensure completeness of the dataset.
5. **Normalization**:
    - Normalized the transaction values for each product category using StandardScaler to ensure that all features have the same scale and are comparable during clustering.

Clustering:

- **KMeans Algorithm**: KMeans clustering was applied with 5 clusters. The fit_predict method was used to assign each customer to one of the clusters based on their transaction history.

Evaluation Metrics:

1. **Davies-Bouldin Index (DB Index)**:
    - The DB Index evaluates the average similarity ratio of each cluster with the one that is most similar to it. A lower value indicates better clustering.
    - **DB Index = 1.354**: This score indicates moderate clustering performance. Ideally, a lower value would be preferred for more distinct clusters.
2. **Silhouette Score**:
    - The Silhouette Score measures how similar each point is to its own cluster compared to other clusters. A higher score (close to 1) indicates well-separated and well-formed clusters, while a lower score indicates poorly defined clusters.
    - **Silhouette Score = 0.211**: This score suggests that the clusters are not very well-separated, indicating that the clustering might not be optimal.

Dimensionality Reduction & Visualization:

- **PCA (Principal Component Analysis)** was applied to reduce the dimensionality of the customer data to 2 principal components for easy visualization.
- The scatter plot below represents the customer clusters in the reduced 2D space, with different colors indicating different clusters.

Clustering Visualization:

The following scatter plot shows the result of customer segmentation using PCA:

- **X-Axis**: PCA1 (First Principal Component)
- **Y-Axis**: PCA2 (Second Principal Component)
- **Cluster Colors**: Each cluster is represented by a different color, allowing for visual differentiation of the customer segments.

Conclusion:

- **Clustering Performance**: The clustering results suggest that the customer segments may not be distinctly separated, as indicated by the moderate DB Index and low Silhouette Score.
- **Potential Improvements**: The clustering results could be improved by experimenting with a different number of clusters or trying other clustering algorithms (e.g., DBSCAN, Agglomerative Clustering). Fine-tuning the features used for clustering or incorporating additional data might also enhance the segmentation.

Recommendations:

- **Cluster Analysis**: Further investigation of the characteristics of each cluster is recommended. Profiling customers based on their transaction behavior (e.g., high spenders, frequent shoppers) could help in developing targeted marketing strategies.

- **Model Tuning**: Experiment with different numbers of clusters and clustering techniques to achieve better-defined customer segments.