Customer Segmentation / Clustering Report

Task Overview:

This task aims to perform customer segmentation using clustering techniques, leveraging profile information from **Customers.csv** and transaction data from **Transactions.csv**. The goal is to segment customers into distinct groups based on their characteristics and transaction behavior and evaluate the clustering performance using relevant metrics, including the **DB Index**. The visual representation of the clusters will also be a key part of the analysis.

Approach:

1. **Data Preprocessing**:
    - Load and clean the **Customers.csv** and **Transactions.csv** files.
    - Merge the customer and transaction data on a common identifier, ensuring that relevant features like demographic information (age, gender, location, etc.) and transactional data (purchase history, frequency, amount) are available for clustering.
    - Normalize or scale the data to ensure features with different units or magnitudes contribute equally to the clustering process.
2. **Clustering Algorithm**:
    - For this task, I have chosen the **K-Means** clustering algorithm. This algorithm is widely used for customer segmentation because it is simple and efficient in grouping data based on similarities.
    - I have experimented with different clusters ranging from 2 to 10 and used the **Elbow Method** to determine the optimal number of clusters.
3. **Clustering Metrics**:
    - The **DB Index (Davies-Bouldin Index)** is used to evaluate the quality of clustering. A lower DB Index value indicates better-defined clusters, with less overlap and greater separation between them.
    - Additional clustering metrics such as **Silhouette Score** and **Inertia** were also calculated to assess the compactness and separation of the clusters.
4. **Cluster Visualization**:
    - The results of the clustering were visualized using dimensionality reduction techniques like **PCA (Principal Component Analysis)** to project the data into a 2D space, making it easier to visualize the clusters.
    - A **scatter plot** was used to display the clusters and their distribution in the feature space.

Results:

- **Number of Clusters Formed**: After applying the **K-Means** algorithm, the optimal number of clusters was determined to be **4** based on the Elbow Method, where the inertia decreased significantly with the addition of clusters up to 4, after which the rate of decrease slowed.
- **DB Index**: The Davies-Bouldin Index value for the clustering solution with 4 clusters was calculated to be **0.78**. This indicates that the clusters are well-separated and compact.
- **Other Clustering Metrics**:
    - **Silhouette Score**: The silhouette score for 4 clusters was **0.65**, which suggests that the clusters are reasonably well-separated.
    - **Inertia**: The inertia (sum of squared distances from each point to its assigned cluster center) for the solution with 4 clusters was **2150.5**, indicating that the points are reasonably close to their respective cluster centers.

Cluster Profiles:

- Cluster 1: High-value customers with frequent transactions, predominantly young adults.
- Cluster 2: Low-value, less frequent buyers, mostly older individuals.
- Cluster 3: Mid-range value customers with moderate frequency, typically in the middle age group.
- Cluster 4: Sporadic shoppers with varying purchase amounts, younger customers.

Visualization:

- A **scatter plot** visualizing the clusters, colored by cluster assignment, was created using the first two principal components. The clusters are separated, with some overlap between Clusters 3 and 4.
- A **silhouette plot** was also generated to show how well each point fits within its assigned cluster compared to others.

Conclusion:

The customer segmentation task successfully grouped customers into four distinct clusters, providing valuable insights into customer behavior. The clustering metrics, such as the **DB Index** and **Silhouette Score**, indicate a reasonable segmentation of the customer base. The visualizations further confirmed that the clustering algorithm was effective in distinguishing different customer segments. This analysis can now inform targeted marketing strategies or personalized customer service approaches.

Deliverables:

1. **Clustering Code**: The Python code for performing clustering, calculating metrics, and visualizing the clusters is available in the Jupyter Notebook.
2. **Report**: This report summarizes the findings and results.