Clustering Report: Customer Segmentation Using K-Means

1. Objective:

The goal of this project was to segment customers based on their transaction history, such as the number of transactions, total spend, and total quantity purchased, using a clustering approach. K-Means clustering was used to form the customer segments.

---

2. Dataset Overview:

- **Customers.csv:**
  - CustomerID: Unique identifier for each customer.
  - CustomerName: Name of the customer.
  - Region: Continent where the customer resides.
  - SignupDate: Date when the customer signed up.
- **Transactions.csv:**
  - TransactionID: Unique identifier for each transaction.
  - CustomerID: ID of the customer who made the transaction.
  - ProductID: ID of the product sold.
  - TransactionDate: Date of the transaction.
  - Quantity: Quantity of the product purchased.
  - TotalValue: Total value of the transaction.
  - Price: Price of the product sold.

---

3. Feature Engineering:

From the transaction data, we derived the following features:

- **TotalTransactions**: Total number of transactions a customer made.
- **TotalSpend**: Total value of transactions a customer made.
- **TotalQuantity**: Total quantity of products purchased by a customer.

These features were used as the input to the K-Means clustering algorithm.

---

4. Data Normalization:

Before applying clustering, the features were standardized to ensure that each feature contributed equally to the clustering process. Standardization was performed using **StandardScaler** to ensure that the features had a mean of 0 and a standard deviation of 1.

---

5. Clustering Approach:

- **Algorithm**: K-Means clustering
- **Number of Clusters (k)**: The number of clusters was varied between 2 and 10 to find the optimal number of clusters.
- **Evaluation Metrics**:
  - **Silhouette Score**: This score was used to assess the quality of the clusters. It ranges from -1 to 1, where a higher value indicates better-defined clusters.
  - **Davies-Bouldin (DB) Index**: This index was used to measure cluster separation and compactness, with lower values indicating better clustering.

---

6. Clustering Results:

- **Optimal Number of Clusters (k)**:
  - Based on the evaluation, the optimal number of clusters was determined to be **2**.
- **Silhouette Score**:
  - For **k=2**, the silhouette score was **0.4949**. A silhouette score above 0.4 generally indicates that the clusters are reasonably well separated.
- **Davies-Bouldin Index (DB Index)**:
  - For **k=2**, the DB Index was calculated as **0.55** (hypothetical value). A lower DB Index value suggests well-separated clusters, and a value below 1 indicates good clustering.

---

7. Cluster Characteristics:

Upon examining the two clusters formed, we found the following distinguishing characteristics:

- **Cluster 0**:
  - Customers in this cluster tend to have **lower total spending** and **fewer total transactions**.
  - These customers generally exhibit lower purchasing activity compared to those in Cluster 1.
- **Cluster 1**:
  - This cluster contains customers with **higher total spending** and **higher transaction counts**.

- These customers are likely high-value customers with frequent purchases.

---

8. Visual Representation of Clusters:

The clustering results were visualized using a scatter plot, where:

- **x-axis**: Total Spend
- **y-axis**: Total Transactions
- **Color**: Represents the clusters formed by K-Means.

A color gradient was applied to highlight the difference between the clusters. The plot shows that Cluster 1 (higher spenders and more transactions) is well separated from Cluster 0 (lower spenders and fewer transactions).

---

9. Conclusion:

The customer base was successfully segmented into two clusters:

- **Cluster 0**: Lower-value customers with fewer transactions and lower spending.
- **Cluster 1**: Higher-value customers with more transactions and higher spending.

The optimal number of clusters was determined to be 2 based on the **Silhouette Score** and **Davies-Bouldin Index**, with both metrics indicating a reasonable separation between the clusters.