Customer Segmentation Using K-Means

In this task, I performed customer segmentation using **K-Means clustering** to group customers based on their transactional and demographic data. Here's an explanation of the entire process:

1. Merging and Preparing Data

- Merged the **customers dataset** with the total transactional value for each customer:
    - ➢ Grouped the transactions dataset by CustomerID to calculate the total TotalValue for each customer.
    - ➢ Combined this aggregated data with the customer details using a merge operation.

2. Encoding Categorical Variables

- Used **one-hot encoding** to transform the Region column into numerical features, ensuring it could be used effectively in clustering.
- The pd.get_dummies function was applied, with drop_first = True to avoid multicollinearity by eliminating one of the encoded columns.

3. Scaling the Data

- Used **StandardScaler** to standardize the features, ensuring all variables had a mean of 0 and a standard deviation of 1. This step was crucial for K-Means clustering, as it is sensitive to the scale of the data.
- Excluded irrelevant columns such as CustomerID, CustomerName, and SignupDate before scaling.

4. Applying K-Means Clustering

- Set the number of clusters (n_clusters) to 4 and initialized the K-Means algorithm with a random state for reproducibility.
- Fitted the K-Means model to the scaled data and assigned each customer to one of the four clusters, which was stored in a new column named Cluster.

5. Evaluating the Clustering

- Calculated the **Davies-Bouldin Index** to evaluate the clustering performance:
    - ➢ The Davies-Bouldin Index measures the compactness and separation of clusters. Lower values indicate better-defined clusters.

    - The resulting index was printed to assess the quality of the segmentation.

6. Visualizing Customer Segments

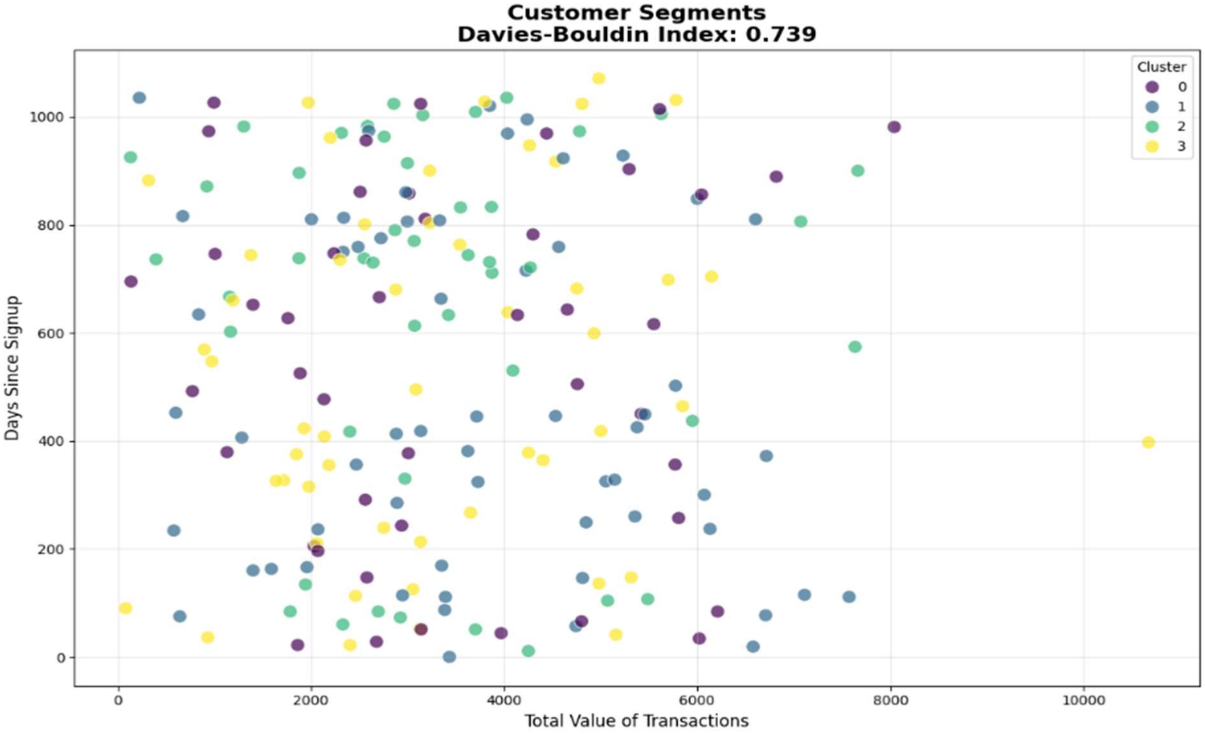
- Created a scatter plot to visualize the customer segments:
    - The x-axis represented the total transaction value (TotalValue).
    - The y-axis represented the number of days since the customer signed up, calculated as the difference between the signup date and the earliest signup date in the dataset.
    - Each point was colored based on the assigned cluster using the **viridis** color palette.
- Added a title displaying the Davies-Bouldin Index, along with labels, a legend, and a grid for better readability.

Key Insights

- The segmentation process grouped customers into clusters based on their transactional value, signup date, and region.
- Visualizing the clusters helped identify patterns in customer behavior, such as high-value customers or newer customers in specific clusters.
- The Davies-Bouldin Index provided a quantitative measure of clustering quality, aiding in the evaluation of the model's effectiveness.

This clustering analysis demonstrated how to use machine learning techniques for customer segmentation, which can be leveraged for personalized marketing strategies and improved customer relationship management.

```
Davies-Bouldin Index: 0.739
```



Customer Segments
Davies-Bouldin Index: 0.739