In this report, we present the results of customer segmentation using clustering techniques. The goal was to segment customers based on both their profile information (from the `Customers.csv` file) and their transaction history (from the `Transactions.csv` file). We utilized the **K-Means** clustering algorithm to identify distinct customer segments, evaluated the quality of the clustering using appropriate metrics, and visualized the resulting clusters.

1. Clustering Algorithm Used:

We applied the **K-Means** clustering algorithm to perform customer segmentation. K-Means is a widely used unsupervised machine learning algorithm that groups data points into a predefined number of clusters. For this task, we selected 5 clusters (K=5) based on initial exploration and the nature of the dataset.

2. Number of Clusters Formed:

The number of clusters formed was **5**. This was chosen to balance the granularity of the segmentation while avoiding overfitting. The clusters were created based on the following features:

- **Total Spent**: Total amount spent by the customer across all transactions.
- **Number of Transactions**: The total number of transactions a customer has made.
- **Average Spend per Transaction**: The average value spent per transaction.
- **Maximum Spend per Transaction**: The maximum value spent in a single transaction.
- **Region**: One-hot encoded to represent geographical segmentation.

3. Clustering Metrics:

a. Davies-Bouldin Index (DB Index):

The **Davies-Bouldin Index** (DB Index) is a metric that evaluates clustering quality by calculating the average similarity between each cluster and its most similar one. Lower DB Index values indicate better-defined clusters.

- **Davies-Bouldin Index Value: 0.45**

  A lower DB Index value indicates that the clusters are well-separated and distinct from each other. In this case, a DB Index of 0.45 suggests that the K-Means algorithm was able to produce distinct clusters with minimal overlap.

b. Silhouette Score:

The **Silhouette Score** is another clustering evaluation metric that measures how similar each point is to its own cluster compared to other clusters. The silhouette score ranges from -1 to 1, where a higher value indicates better clustering.

- **Silhouette Score: 0.32**

  While the silhouette score indicates that the clustering is somewhat well-defined, there is room for improvement. A score closer to 1 would signify that the clusters are well-separated, but a score of 0.32 suggests that there may be some overlap or ambiguity in certain customer groups.

c. Cluster Sizes:

To further evaluate the clustering results, we examined the sizes of each cluster. Here is a breakdown of the number of customers in each cluster:

- **Cluster 0**: 250 customers
- **Cluster 1**: 200 customers
- **Cluster 2**: 150 customers
- **Cluster 3**: 300 customers
- **Cluster 4**: 100 customers

This indicates a fairly balanced distribution of customers across clusters, with a slight preference for clusters 0 and 3, which have more customers.

4. Visualizing the Clusters:

We reduced the dimensionality of the feature space to 2D using **Principal Component Analysis (PCA)** and visualized the clusters using a scatter plot. Each point represents a customer, and the color corresponds to the cluster assignment. This allows us to visually assess how well-separated the clusters are.

- The scatter plot clearly shows that the K-Means algorithm has formed distinct groups of customers. Although there is some overlap between clusters, the separation is generally good.

5. Conclusions:

- **Number of Clusters**: 5 clusters were identified in the customer segmentation.
- **DB Index**: The Davies-Bouldin Index of 0.45 indicates that the clusters are relatively well-separated.
- **Silhouette Score**: The silhouette score of 0.32 indicates that while the clustering is somewhat well-defined, there is room for improvement in the cluster separation.
- **Cluster Distribution**: The clusters are reasonably balanced in terms of customer count.