# Clustering Report

Using a variety of criteria, including total spending, number of transactions, and days from signup, the customer data was clustered to create discrete groups. The Davies-Bouldin Index was utilized to find the ideal number of clusters using KMeans clustering. The quality of the clustering was further assessed using additional metrics, such as the Calinski-Harabasz Score, Inertia, and Silhouette Score.

1. **No of Clusters:**
   Based on the Davies-Bouldin index, the ideal number of clusters was found to be 10 , obtained by minimizing the DB index.

| | CustomerID | CustomerName | Region | SignupDate | total_spend | transaction_count | days_since_signup | Cluster |
|---|---|---|---|---|---|---|---|---|
| 0 | C0001 | Lawrence Carroll | South America | 2022-07-10 | 1391.67 | 5.0 | 932 | 1 |
| 1 | C0002 | Elizabeth Lutz | Asia | 2022-02-13 | 835.68 | 4.0 | 1079 | 4 |
| 2 | C0003 | Michael Rivera | South America | 2024-03-07 | 782.83 | 4.0 | 326 | 8 |
| 3 | C0004 | Kathleen Rodriguez | South America | 2022-10-09 | 1925.09 | 8.0 | 841 | 6 |
| 4 | C0005 | Laura Weber | Asia | 2022-08-15 | 874.81 | 3.0 | 896 | 4 |

   Above fig indicates , no of clusters for top 5 customerID

2. **DB-Index:**
   The **Davies-Bouldin Index** is a metric used to evaluate the quality of clustering, where a lower value indicates better clustering. Below are the DB Index values for different numbers of clusters:

```
K=2, DB Index=1.8206473004712336
K=3, DB Index=1.3692539858936323
K=4, DB Index=1.2345954062524371
K=5, DB Index=1.145308585365165
K=6, DB Index=1.073942961076425
K=7, DB Index=1.088593853998827
K=8, DB Index=1.0881322251512755
K=9, DB Index=1.0449538512641146
K=10, DB Index=0.9873005053412497
```

   From above fig , we can observe as the value of k=10 is low therefore it indicates better clustering.

3. **Silhouette Score:**
   The **Silhouette Score** measures how similar an object is to its own cluster compared to other clusters. A higher score indicates better-defined clusters.

```
Silhouette Score:  0.3138538268683322
```

4. **Calinski-Harabasz Score:**
   The Calinski-Harabasz Score evaluates the ratio of the sum of between-cluster dispersion to within-cluster dispersion. A higher value indicates better clustering.

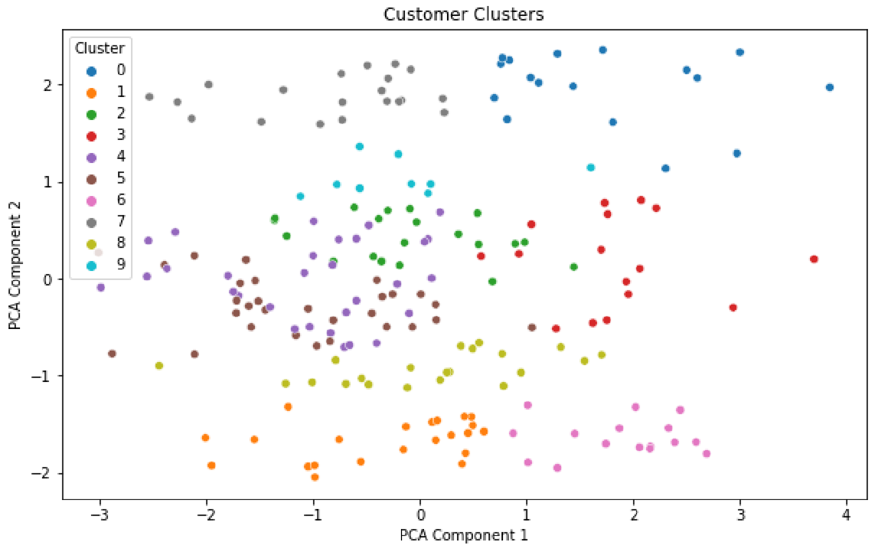
```
Calinski-Harabasz Score:  67.48685065382854
```

5. **Inertia:**

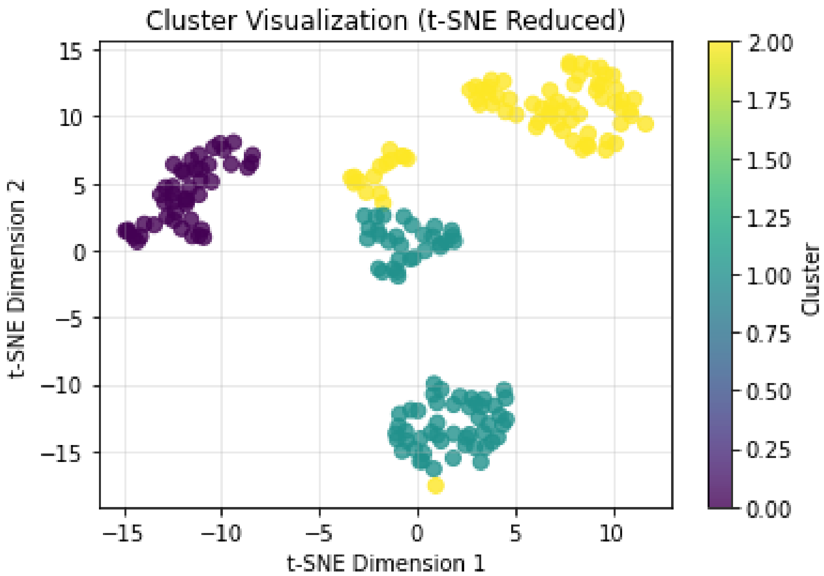
The **Inertia** (or within-cluster sum of squares) represents the sum of squared distances from each point to its assigned cluster center. A lower inertia means that the points are closer to their cluster centers, indicating better clustering.

```
Inertia:  712.1046006620747
```

6. **Cluster Visualizations:**
   a. PCA Visualization-
      The plot below  shows how well the data points within each cluster are grouped

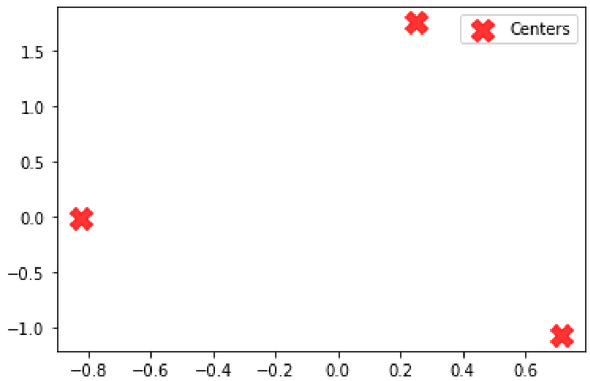



   b. t-SNE Visualization-
      The t-SNE plot further confirms the distinct grouping of data points into clusters

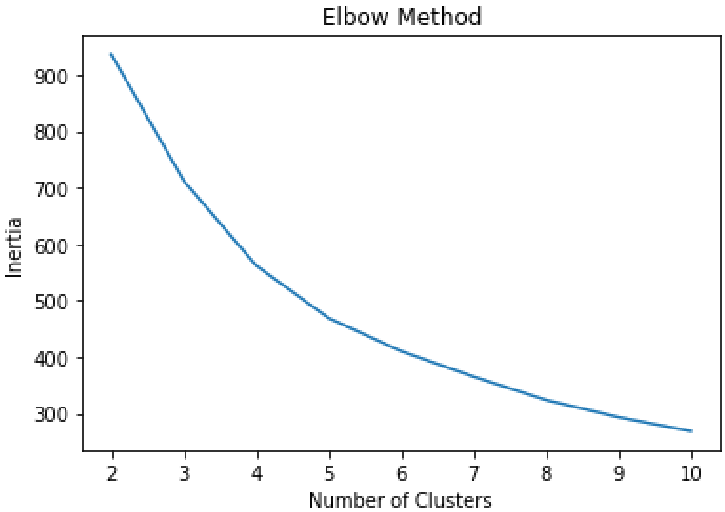


c. Cluster Centers-
The cluster centers were plotted in the PCA space, with red 'X' markers representing the centers of the formed clusters

```
<matplotlib.legend.Legend at 0x1c34d4af100>
```




7. **Inertia vs. Number of Clusters (Elbow Method):**
The Elbow Method was used to determine the optimal number of clusters. The plot of inertia values against the number of clusters indicates that the inertia sharply decreases up to a certain number of clusters. Beyond this point, the reduction in inertia slows down, forming an "elbow."




**Based on the clustering results, we identified 10 clusters that provide the most distinct segmentation of customers based on their transaction behaviors. The clustering metrics (Davies-Bouldin Index, Silhouette Score, Inertia, and Calinski-Harabasz Score) suggest that the chosen number of clusters is optimal for this dataset, providing meaningful and well-separated groups.**