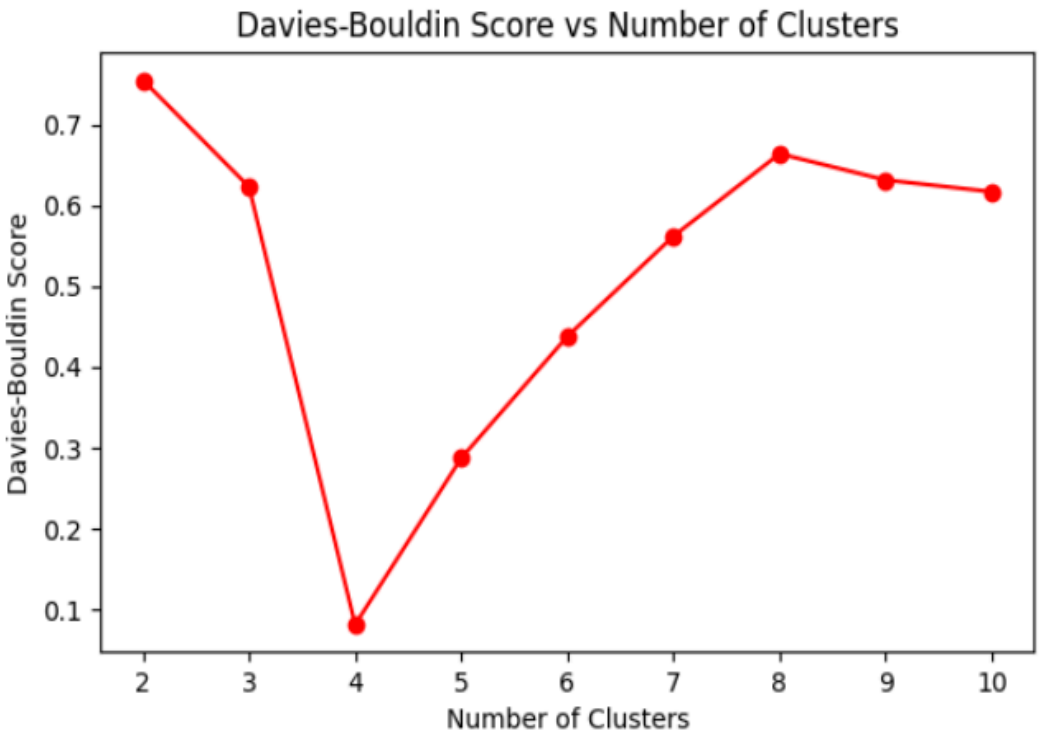
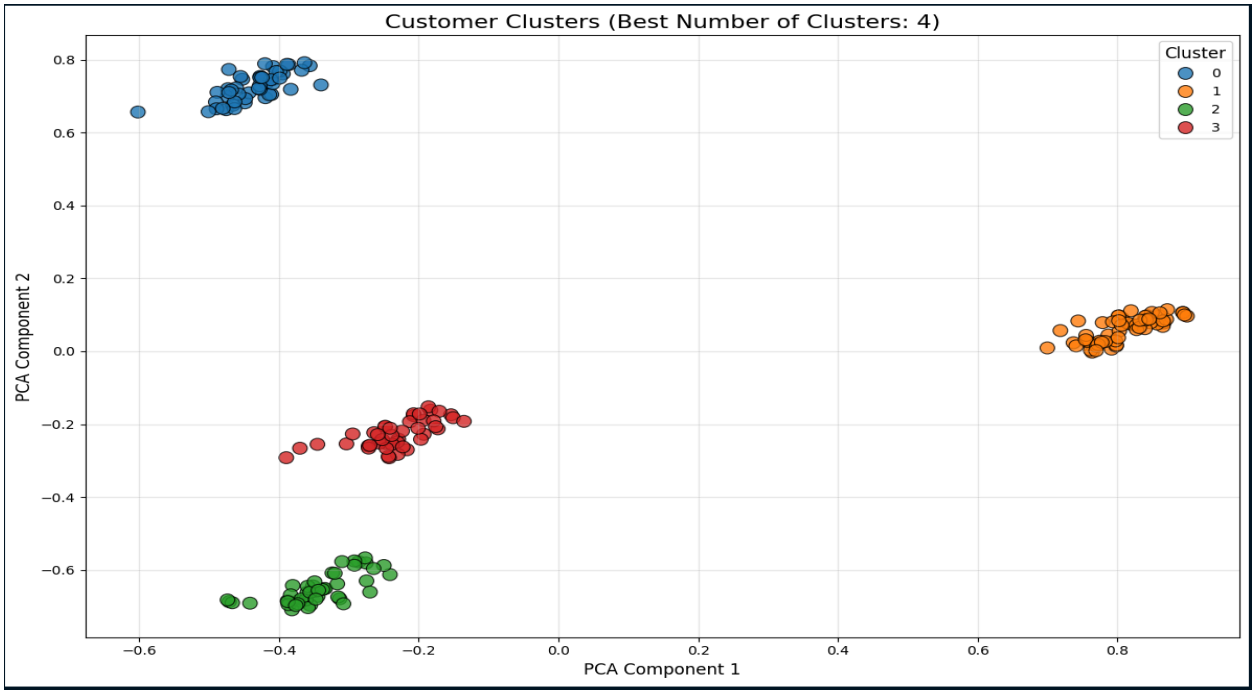
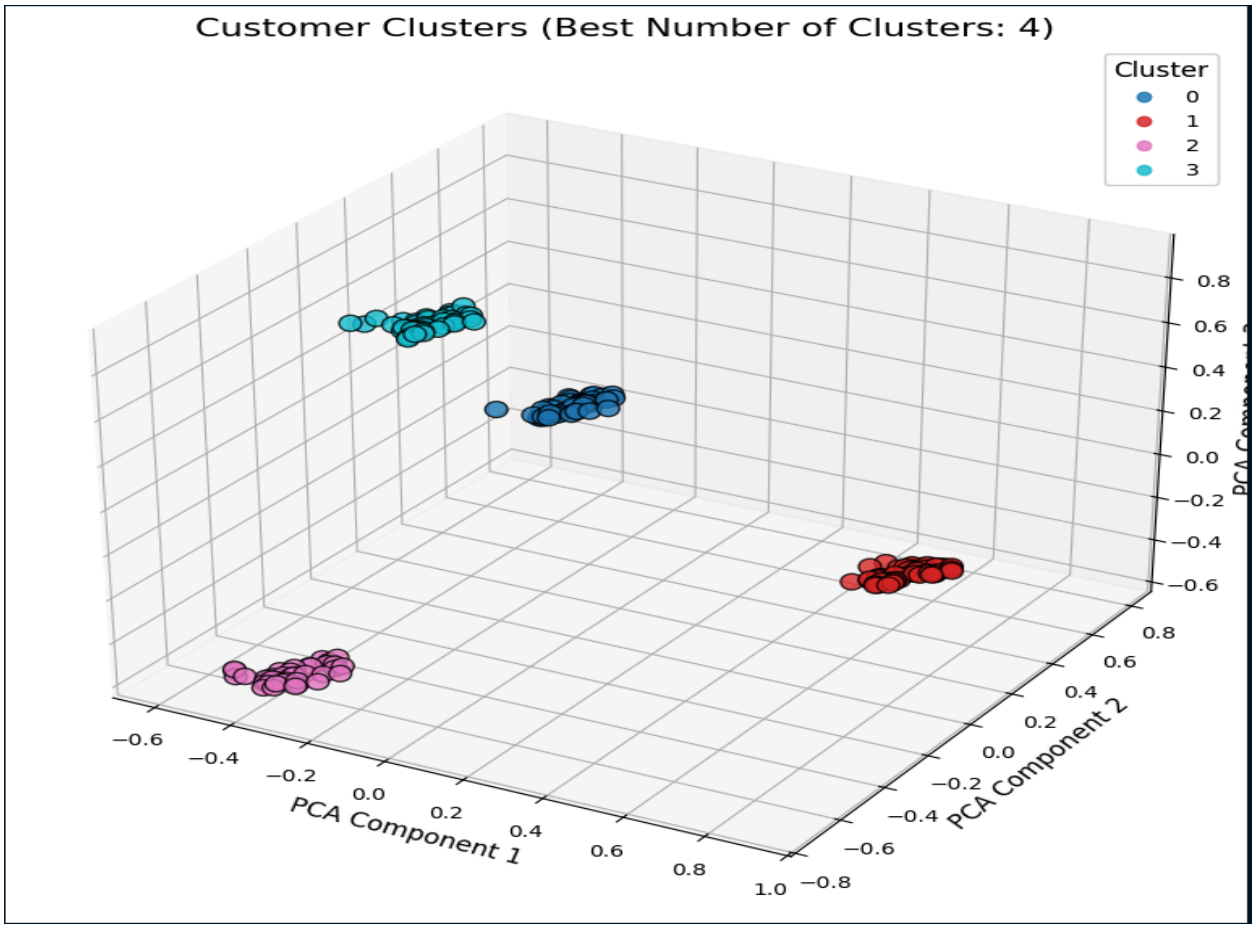
Clustering part by Atharva Shetty

Clustering Metrics:

- **Number of Clusters Formed**: 4 (based on the PCA results and cluster validation)
- **Davies-Bouldin Index**: 0.0814
  A lower value indicates that the clusters are well-separated and compact.
- **WCSS**: 0.83
- **Silhouette Score**: 0.9433
  The silhouette score measures how similar each point is to its own cluster compared to other clusters. A higher score indicates better-defined clusters.
- **Calinski-Harabasz Score**: 11640.67
  This score assesses the dispersion of clusters. A higher score generally suggests well-separated and well-formed clusters.




Customer Clusters (Best Number of Clusters: 4)
Cluster
0
1
2
3
PCA Component 2
PCA Component 1
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8


Customer Clusters (Best Number of Clusters: 4)
Cluster
0
1
2
3
PCA Component 1
PCA Component 2
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
−0.8

Further Analysis:
As the main intention of the problem was to reduced the DB score, hence I used 4 clusters which are foreshadowing the importance of Region column in the dataset. These means that customer in same region are more similar than different region. Moreover if we want to remove the foreshadowing of the region column, we can try mean target encoding instead of OHE to reduce region column influence. The cluster may be different if we follow the MTE approach but other features will get more importance.