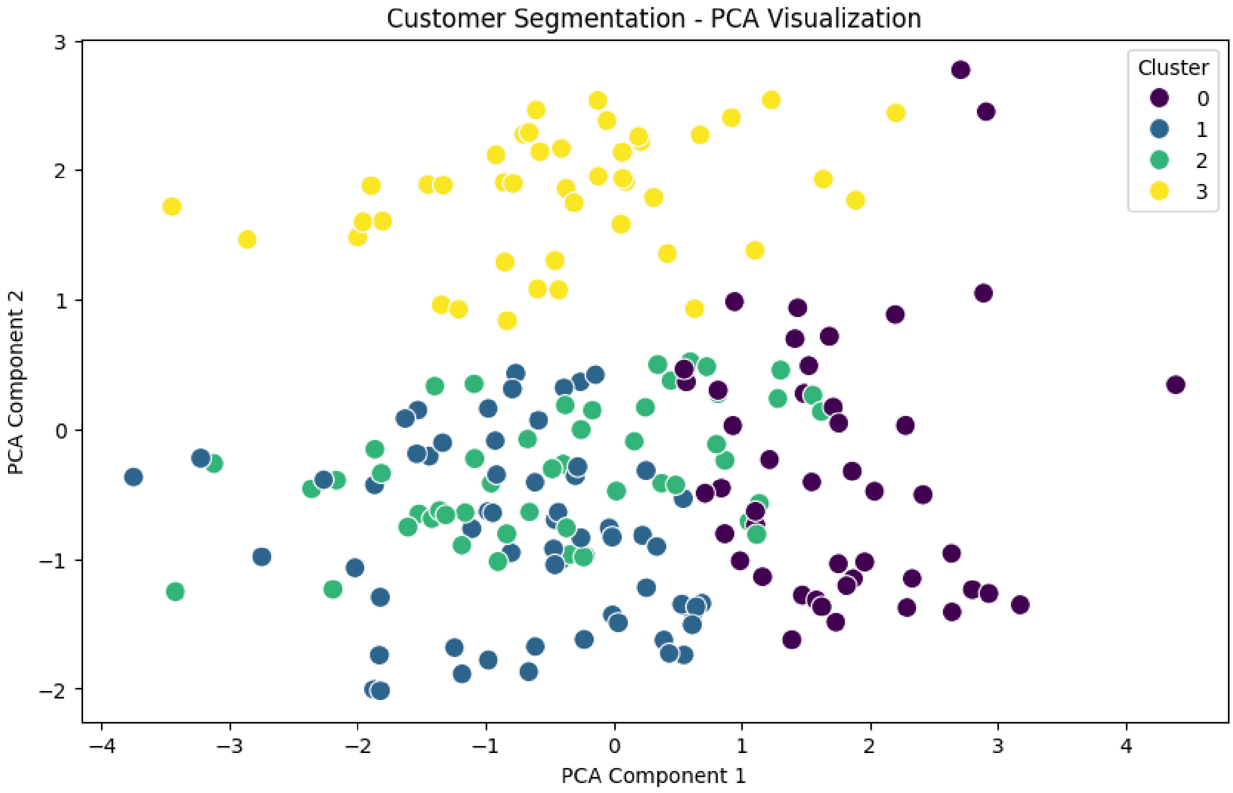
Data Science Intern Assignment | Zeotap| Ipshita Das

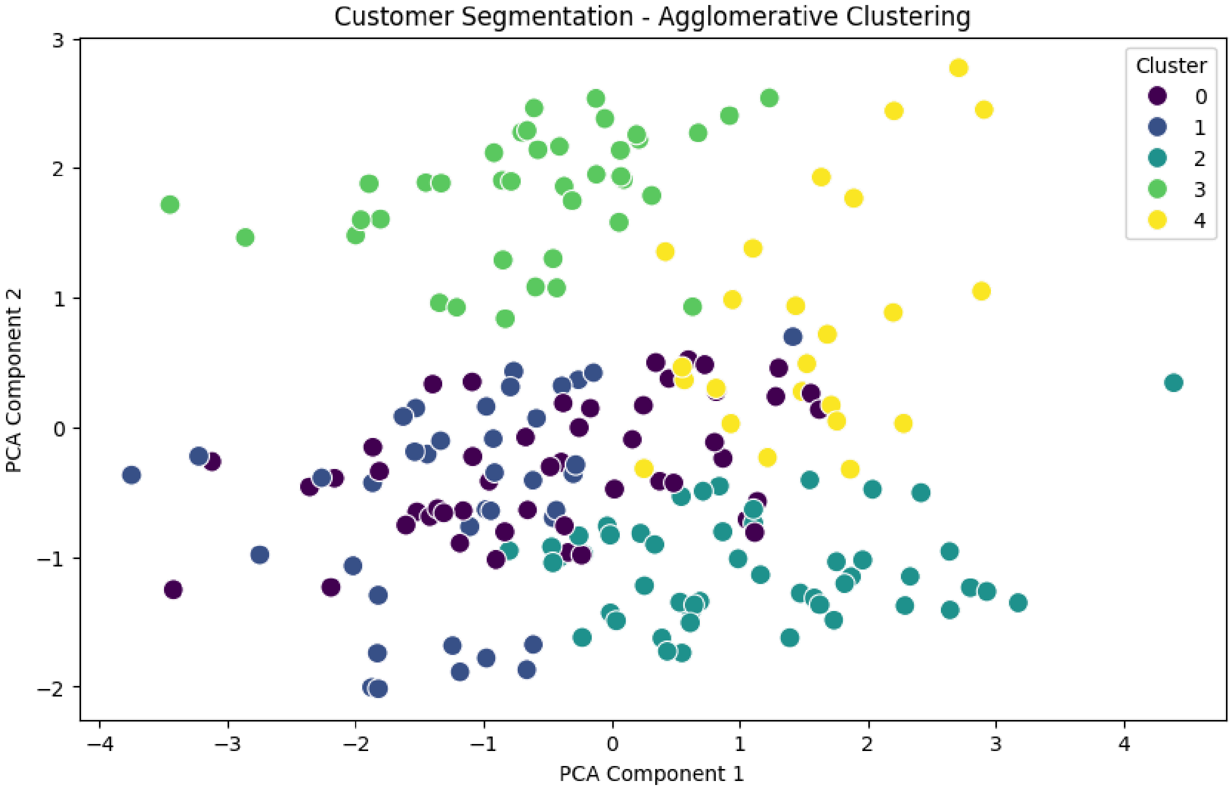
Task 3

Business Insights from EDA

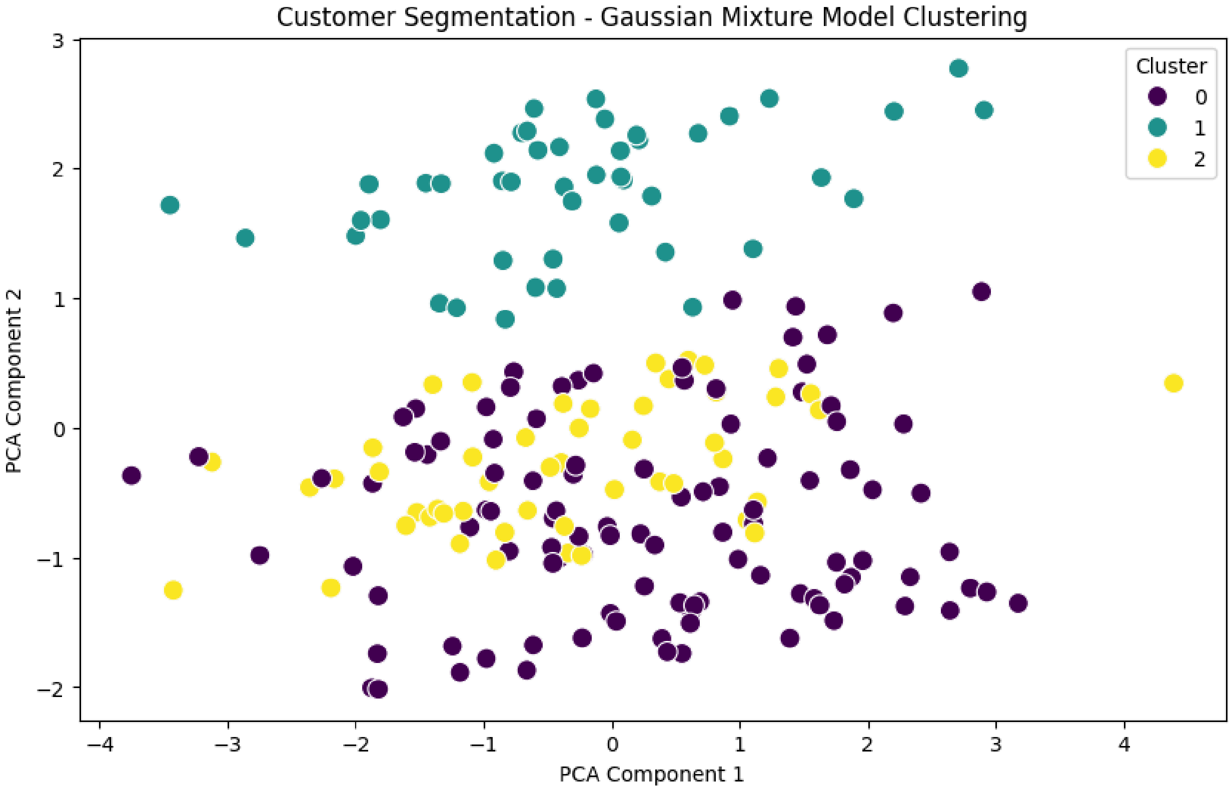
Visualization of KMeans



Visualization of AgglomerativeClustering



Visualization of GMM



We performed customer segmentation using KMeans, Aggregative Clustering, and the Gaussian Mixture Model (GMM) and evaluated them using the Davies-Bouldin Index (DBI) and Silhouette Score. There's no single "best" model. The optimal model depends on the specific priorities and interpretation of the evaluation metrics.

Summary of Model Performance:
We evaluated each clustering algorithm across a range of cluster numbers (2-10), plotting DBI and Silhouette scores to identify potential optimal cluster numbers. Then we applied each clustering method with a specific number of clusters (decided by the metrics or arbitrarily set). Finally, we calculated DBI and Silhouette scores for each algorithm's selected number of clusters.

Model	**Cluster Centroids Count**
K-means	4
Agglomerative Clustering	5
GMM	3

The visualization helps us understand the separation of clusters in 2D space. Evaluated its DBI and Silhouette scores to judge the quality of the clusters against the other models.

Choosing the Best Model:

Lower DBI is better: A lower DBI indicates better-separated clusters. Compare the DBI scores of the three models after they have been run with their respective "optimal" cluster counts.

Higher Silhouette Score is better: A higher Silhouette Score signifies that data points are well-matched to their clusters and poorly matched to neighbouring clusters. Compare the Silhouette Scores of the three models in the same way as above.

Visual Inspection: The PCA visualizations provide a 2D representation of the clusters. Observe if the clusters appear well-separated and distinct in each model's visualization. A model with visually distinct and well-separated clusters is generally preferred.

Domain Knowledge: The "best" model might also depend on business context or domain expertise. For example, if a certain number of customer segments are already expected, or if prior knowledge indicates certain customer groups are likely to exist, the results should be reviewed in this light.

Best Model: Agglomerative Clustering
Reasoning: Although GMM has the highest silhouette score (0.28), Agglomerative Clustering has the lowest Davies-Bouldin Index (1.36), indicating better cluster separation compared to GMM (1.50) and KMeans (1.40).Trade-off: While the silhouette

score is slightly lower for Agglomerative Clustering, its superior DBI suggests better cluster cohesion and separation overall.

Best Model: Agglomerative Clustering

Reasoning: Although GMM has the highest silhouette score (0.28), Agglomerative Clustering has the lowest Davies-Bouldin Index (1.36), indicating better cluster separation compared to GMM (1.50) and KMeans (1.40).

Trade-off: While the silhouette score is slightly lower for Agglomerative Clustering, its superior DBI suggests better cluster cohesion and separation overall.